Balanced Topic Aware Sampling for Effective Dense Retriever: A Reproducibility Study

Shuai Wang
The University of Queensland
Brisbane, Australia
shuai.wang5@uq.net.au

Guido Zuccon
The University of Queensland
Brisbane, Australia
g.zuccon@uq.edu.au

ABSTRACT

Knowledge distillation plays a key role in boosting the effectiveness of rankers based on pre-trained language models (PLMs); this is achieved using an effective but inefficient large model to teach a more efficient student model. In the context of knowledge distillation for a student dense passage retriever, the balanced topic-aware sampling method has been shown to provide state-of-the-art effectiveness. This method intervenes in the creation of the training batches by creating batches that contain positive-negative pairs of passages from the same topic, and balancing the pairwise margins of the positive and negative passages.

In this paper, we reproduce the balanced topic-aware sampling method; we do so for both the dataset used for evaluation in the original work (MS MARCO) and for a dataset in a different domain, that of product search (Amazon shopping queries dataset) to study whether the original results generalize to a different context. We show that while we could not replicate the exact results from the original paper, we do confirm the original findings in terms of trends: balanced topic-aware sampling indeed leads to highly effective dense retrievers. These results partially generalize to the other search task we investigate, product search: although we observe the improvements are less significant compared to MS MARCO.

In addition to reproducing the original results and studying how the method generalizes to a different dataset, we also investigate a key aspect that influences the effectiveness of the method: the use of a hard margin threshold for negative sampling. This aspect was not studied in the original paper. With respect to hard margins, we find that while setting different hard margin values significantly influences the effectiveness of the student model, this impact is dataset-dependent – and indeed, it does depend on the score distributions exhibited by retrieval models on the dataset at hand. Our reproducibility code is available at https://github.com/ielab/TAS-B-Reproduction.

CCS CONCEPTS

• **Information systems → Language models**.

KEYWORDS

BERT, Topic clustering, Knowledge Distillation, Dense Retriever

ACM Reference Format:
Shuai Wang and Guido Zuccon. 2023. Balanced Topic Aware Sampling for Effective Dense Retriever: A Reproducibility Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591915

1 INTRODUCTION

A dense retriever is an information retrieval approach based on pre-trained language models (PLMs) and where each passage and query is encoded separately using the backbone PLM to obtain embedding representations (dense vectors) [13, 20, 24]. The encoding of passages is performed offline, and embeddings are stored in a vector index. At query time, the query is encoded in an embedding and an approximate nearest neighbour search is performed on the vector index to identify the k passages that are closest to the query. Because passages are not encoded at query time, dense retrievers are characterized by lower query latency time than alternative PLM-based architectures like cross-encoders (e.g., monoBERT [17]), but this better query efficiency often comes at the expense of lower search effectiveness.

Knowledge distillation has been used to address the lower effectiveness of dense retrievers while still retaining their efficiency compared to cross-encoder architectures [6, 10, 13, 14, 19]. In knowledge distillation approaches, highly effective cross-encoders are employed as teachers to transfer knowledge to the less complex and more efficient student models (the dense retrievers). One such approach is the Balanced Topic Aware Sampling by Hofstätter et al. [8], which is the focus of this paper: this is the state-of-the-art knowledge distillation method for dense retrievers at the time of writing. The Balanced Topic Aware Sampling approach leverages a simple but powerful intuition: within each training batch, queries in a batch should pertain to the same topic. This intuition, implemented in the TAS method, is further complemented by controlling the pairwise margin between positive and negative documents in each training batch so that the margin of positive-negative document pairs is uniformly distributed (or 'balanced') in the margin range (TAS-B), and setting a maximum margin value (hard margins) above which pairs are not selected (TAS-B-HM). The pairwise margin between two passages is the difference in their retrieval score. The results reported in the original study showed these methods to be effective in training competitive dense retrievers via knowledge distillation. The original study, however, left a number of directions unexplored, including whether the results obtained on the popular

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591915

MS MARCO dataset [16] generalise to other datasets, and a thorough study of the hard margins conditions. The objective of this reproducibility paper is to study these directions; we do so through the investigation of the following research questions:

RQ1: Do we obtain the same results when replicating the original study within the same experimental setting (MS MARCO dataset)?

RQ2: To what extent do findings obtained on MS MARCO generalise to a different dataset? For this, we evaluate the effectiveness of the method on the Amazon shopping queries dataset[18].

RQ3: What is the impact of varying the maximum hard margin value on the effectiveness of models trained using the TAS-B-HM method?

2 REPRODUCING TAS, TAS-B AND TAS-B-HM

In this section, we describe the training architecture used in the original study and technical details about TAS, TAS-B and TAS-B-HM [8]. Figure 1 provides a visual summary of the key aspects of the training architecture.

2.1 Dual Supervision

In the original study, all experiments rely on a dual supervision training architecture, that is to utilize (1) a Pairwise cross-encoder teacher [7, 9]; and (2) an In Batch ColBERT teacher [13].

Pairwise cross-encoder teacher is based on the finding that ensemble multiple strong cross-encoder models can create a stronger teacher, which benefits the teaching of a more effective student dense retriever. Three cross-encoder teachers are used in the original study: BERT-base-uncased, AlBERT, and BERT-Large. To ensemble the three models, relevance scores are computed using the average of the scores generated from all three models. For example, for a query q and a passage p, the relevant score of the Pairwise teacher model is computed using the following formula:

$$RS_t(q,p) = (RS_{BERT-base-uncased}(q,p) + RS_{AlBERT}(q,p) + RS_{BERT-Large}(q,p))/3 \quad (1)$$

To teach a student dense retriever model score RS_s using teacher RS_t for a training pair p^+, p^- with respect to a query q, the loss function can be defined as follows:

$$Loss_{pair}(q,p^+,p^-) = Loss(RS_s(q,p^+) - RS_s(q,p^-), RS_t(q,p^+) - RS_t(q,p^-)) \quad (2)$$

In Batch ColBERT teacher uses an efficient ColBERT architecture proposed by Lin et al. [13], namely TCT-ColBERT. Similar to a dense retriever, this architecture allows the ColBERT model to generate two separate vector representations for the query and passages, and compute relevance scores through dot product. The loss when teaching a student model using an In Batch teacher can be then defined as:

$$Loss_{inbatch}(Q,P^+,P^-) = \frac{1}{2\times|P^+|}(\sum_i^{P^+}\sum_j^{P^-} Loss_{pair}(q,p_i^+,p_j^-) + \sum_i^{P^+}\sum_{j\neq i}^{P^+} Loss_{pair}(q,p_i^+,p_j^+)) \quad (3)$$

For dual supervision, the loss function is computed using a weighted fusion from the Pairwise teacher loss $Loss_{pair}$ and the In Batch ColBERT teacher loss $Loss_{inbatch}$, by setting a weighting parameter α. The calculation of dual supervision loss can be defined as follows:

$$Loss_{dual}(Q,P^+,P^-) = (\sum_i^{P^+}\sum_j^{P^-} Loss_{pair}(q,p_i^+,p_j^-) + Loss_{inbatch}(Q,P^+,P^-)\times\alpha \quad (4)$$

2.2 Topic-aware Sampling

In the original study by Hofstätter et al. [8], the authors propose topic-aware sampling as a way to group similar training pairs in the same batch, which can improve the effectiveness of knowledge distillation. To implement this approach, the authors add an additional step to the training process, which involves clustering the training pairs. However, since the original study uses only queries to represent the training pairs, clustering is limited to the queries in the training files.

To cluster the queries, the authors use a dual-encoder BERT model trained using a Pairwise teacher to generate query representations [7]. These representations are then subjected to k-means clustering [15] on the query set in the training file using the following formula, which results in K clusters:

$$\underset{C}{argmin}\sum_{i=1}^{K}\sum_{q\in\mathbb{C}_i} ||q-v_i||^2 \quad (5)$$

2.3 Balanced Topic-aware Sampling

Another contribution from the original paper is the implementation of balanced topic-aware sampling, which suggests that under each training batch, the margin of training pairs should be balanced. To calculate a margin of a training pair $\{p^+, p^-\}$, the difference between the relevance scores of the positive and negative passages to a query q is defined, and it is calculated as follows:

$$M(p^+,p^-) = RS_t(q,p^+) - RS_t(q,p^-) \quad (6)$$

Then for a minimum margin M_{min} and a maximum margin M_{max}, H margin bins is defined, each margin bin h ($h \in \mathbb{H}$) contains training pairs with a margin size s ($s = (M_{max} - M_{min})/|H|$) that falls within the range of:

$$M_{min} + i\times s \leq M(p^+,p^-) < M_{min} + (i+1)\times s \quad (7)$$

2.4 Hard margin balanced topic-aware sampling

In addition to the original paper, the published code from the study introduced the concept of Hard margin balanced topic-aware sampling (TAS-B-HM). This method involves setting a maximum margin, M_{max}, during the balanced topic-aware sampling process. The motivation for using a hard margin is that a large margin of a training pair may indicate that the negative sample is easier to distinguish, and including these samples in the training process could potentially harm the effectiveness of the student model.

After conducting a preliminary study, we found that using a hard margin can lead to improved effectiveness in certain cases. To further investigate the impact of hard margin on the effectiveness

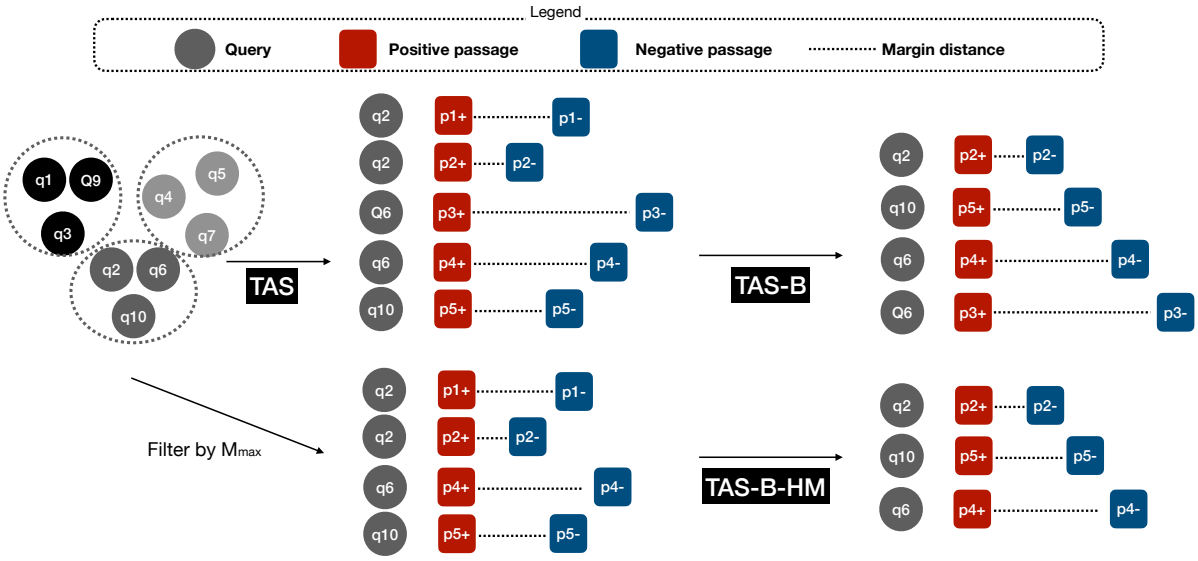

Figure 1: Process of extracting samples for one training batch using topic-aware sampling (TAS), balanced topic-aware sampling (TAS-B) and hard margin balanced topic-aware sampling (TAS-B-HM)

of knowledge distillation, additional experiments are conducted to discuss the effectiveness of the trained student model when different maximum margin values are used.

3 EXPERIMENTAL SETTINGS

3.1 Dataset

We use four datasets in our experiment, MS MARCO passage ranking dataset [16], TREC DL 2019 [3], TREC DL 2020 [2] and Amazon shopping queries dataset [18].

- **MS MARCO Passage Ranking task** contains a collection of queries and passages that is created for the purpose of evaluating the performance of information retrieval (IR) systems. The dataset contains over 8.8 million passages and contains over 500K queries from the training subset, sampled from real user search queries from the Bing search engine. MS MARCO dev dataset is a validation dataset containing 6980 sparsely judged (1-2 positive judgement per query) queries, and is used to evaluate the effectiveness of any trained model. In our experiment, we use MS MARCO Dev set to test the effectiveness of the models trained.
- **TREC DL 2019 & 2020** is an annual event focusing on using deep learning methods for IR tasks. Unlike the MS MARCO dev set, which is sparsely judged, TREC DL provides densely judged query sets on graded relevance; In our experiment, we use TREC DL 2019 and 2020 datasets, containing 43 queries and 54 queries, respectively.
- **Amazon Shopping Queries** is a product search dataset containing over 100K queries and 2 million relevance judgments. The dataset provides a list of up to 40 potentially relevant results with their relevance assessment for each query. In this experiment, the reduced version of the dataset is used, which filters out easy queries, and only English queries were used. A total of 20k training queries and 10k validation queries were obtained, and a validation set was created by sampling 3k queries from the training set. The product titles are used to represent the products, and only products that obtained a relevance assessment of 'exact' were treated as relevant products, while all other types were treated as irrelevant (Substitute, Complement and Irrelevant). To create training triples, all possible positive-negative product pairs were generated from each query's potentially relevant results set. The Amazon Shopping Queries dataset was originally introduced as a re-ranking set, so all products were combined to create a collection set that could be used for retrieval. The effectiveness of the trained student model was evaluated for both retrieval and re-ranking. The set of potentially relevant results originally included in the dataset is used as a starting point for re-ranking.

3.2 Model Parameters

We use the original study's model parameters whenever possible; otherwise, we default to the parameters provided in the original authors' published code.

We use the original study's Pairwise teacher result for the MS MARCO dataset, available at https://zenodo.org/record/4068216. For the Amazon collection, we create a Pairwise teacher by fine-tuning BERT-base-uncased, BERT-large, and ALBERT cross-encoder models using ranknet [1] as the loss function. During the training of the student dense retriever, we use the mean ensemble scores from the three fine-tuned teacher models as the Pairwise teacher. We use the same model, sebastian-hofstaetter/colbert-distilbert-margin_mse-T2-msmarco', for both In Batch ColBERT and In Batch teacher models in both collections. We believe the same In Batch teacher is used in the original study. We employ a dual-teacher combination hyperparameter α of 0.75 for dual supervision, and

use margin-mse' [7] with a batch size of 32 to teach all student models. We also use an Adam optimizer with a learning rate of 7×10^{-6}, which we believe were also used in the original study.

For topic-aware sampling, we sample 2k query clusters using a pairwise trained dual-encoder model ("sebastian-hofstaetter/distilbert-dot-margin_mse-T2-msmarco" on Hugging Face [7]) and select queries only from the same cluster for each batch.

For balanced-topic-aware sampling, we use the same settings as topic-aware sampling and set the number of margin bins $|H|$ to 10, as it is the default in the published code base.

For hard margin balanced topic aware sampling, we set the maximum margin value to the default value of $M_{max} = 6$ for the MS MARCO dataset and a smaller value of $M_{max} = 2$ for the Amazon shopping queries dataset. Our preliminary study showed that the largest possible margin value in the training pairs for Amazon shopping queries is 4.681.

To investigate the effectiveness of student dense retrievers for different M_{max} values, we used our preliminary study on the distribution of margin values in the MS MARCO and Amazon Shopping Queries datasets. The density plot of these distributions is shown in Figure 2. The plot revealed that the training pairs in the MS MARCO dataset had significantly higher margin values than those in the Amazon Shopping Queries dataset. This observation motivated us to investigate whether the user information needs in these two datasets was distinct from each other. As a result, we set the maximum margin values for the MS MARCO dataset in the range of 1-20 (each number from 1-10, then 15, 20) and limited our investigation to the range of -1 to 2 for the Amazon Shopping Queries dataset. This ensured that a similar proportion of data was covered for the maximum margin between the two datasets.

For the BM25 baseline, we use the Pyserini implementation with the default values of $k1 = 0.82$ and $b = 0.68$ as described in [12]. For the dense retriever baseline, we trained a model on the MS MARCO dataset using the original training pairs associated with the dataset [16] without Pairwise teacher scores and used ranknet as the loss function. For the Amazon shopping queries dataset, we generate training triples by pairing all positive products with all negative products for each query, and then train a baseline dense retriever using the same method as for MS MARCO.

3.3 Early Exit for Model Training

The original study employs an early exit evaluation approach, which involves evaluating the model's performance on a validation set every n steps (where $n = 4000$). The early stopping patience is set to p (where $p = 30$), which means that the model would exit training if, for $n \times p$ steps, there is no improvement in validation effectiveness, as measured by ndcg@10. To ensure that the validation set is efficiently evaluated, only a small subset of queries from the original collection are uniformly sampled. We follow the same approach as the original study and sample 3200 queries from a large validation set for MS MARCO. For the Amazon shopping queries dataset, we uniformly sample 3200 queries from the training set to use as the validation set.

It is worth noting that the method used to sample the validation set in the published code-base differs from the approach described in the original study. In the code base, the validation set is sampled based on the per-query effectiveness using the baseline model. To accomplish this, h bins were created, with each bin having a size of s (where $s = ((S_{max} - S_{min})/h)$). Queries are assigned an evaluation metrics score based on their bin range using the baseline model, and those with scores falling within the bin are included. The final validation metrics are stratified uniformly across h bins. In our reproduction, we use the uniform sampling technique described in the original study to sample the validation set. However, in our preliminary study, we find that the differences in effectiveness between the trained student dense retrievers are negligible when using these two different validation sampling techniques.

3.4 Evaluation Measures

We utilized common rank-based measures to represent the effectiveness of our trained model, including $recip_rank$@10, $ndcg$@10, and $recall$@1000. These same measures were used in the original study to report the effectiveness of the model's retrieval results.

4 MAIN RESULTS

4.1 Replication of the Original Study

We start by testing whether we are able to replicate the results reported in the original study, thus addressing our RQ1.

In Table 1, we report the results we obtained using TAS, TAS-B and TAS-B-HM on the dataset considered in the original study, MS MARCO, including the three query sets MS MARCO dev (sparse labels), and TREC 2019 & 2020 (dense labels). These results are completed by those obtained when sampling is performed randomly (which represents the standard baseline practice), and by those obtained by a baseline keyword matching method (BM25) and a baseline dense retriever (DR). To help facilitating a comparison between our results and those reported in the original study, we copy the original in Table 2. Note that for the topic-aware sampling strategy, the original results were only provided for TAS and TAS-B; the TAS-B-HM method, in fact, was not mentioned in the original paper, despite it being in the code base that was released along the publication. As we shall discuss in our analysis, it may be that the original paper actually reported the results for TAS-B-HM in place of TAS-B but without mentioning the imposition of a threshold for the hard margins.

We first focus on comparing the effectiveness of different sampling strategies against that of the baseline DR, which does not use any sampling strategy. We also compare our findings with the effectiveness of the keyword-matching model, BM25. Our results indicate that the methods that use TAS and TAS-B sampling strategies outperform the baselines (DR and BM25) across all evaluation metrics on the MS MARCO dev set, with the exception of recall@1000 when using Pairwise teacher with the TAS sampling strategy. We find similar results on TREC DL, despite not finding any statistically significant differences between the methods, unlike for the MS MARCO dev set. These findings match those reported in the original paper (see our Table 2), at least in terms of relative ordering in effectiveness, i.e. BM25 < DR < Random < TAS <TAS-B.

Our comparison between the sampling strategies of each supervision technique reveals that TAS and TAS-B consistently outperform random sampling on the MS MARCO dev set. However, on the TREC DL queries, the effectiveness of the proposed sampling

		MS MARCO dev			TREC DL 2019			TREC DL 2020		
Method	Sampling	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000
BM25	none	0.234	0.187	0.857	0.497	0.682	0.745	0.488	0.655	0.803
DR	none	0.334	0.280	0.930	0.607	0.815	0.702	0.609	0.765	0.759
Pairwise	Random	0.382*	0.325*	0.954*	0.678	0.846	0.775	0.655	0.809	0.818
	TAS	0.381*	0.324*	0.950*	**0.680**	**0.856**	0.768	0.663	0.833	0.813
	TAS-B	0.387*	0.329*	0.952*	0.674	0.848	0.786	0.663	**0.847**	0.825
	TAS-B-HM	0.382*	0.326*	0.933†	0.639	0.843	0.708	0.628	0.826	0.748
In Batch	Random	0.337	0.279	0.944*	0.646	0.848	0.760	0.581	0.744	0.799
	TAS	0.361*†	0.299*†	0.958*†	0.656	0.829	0.785	0.624	0.803	0.829
	TAS-B	0.362*†	0.299*†	**0.959***†	0.655	0.816	**0.794**	0.615	0.776	0.828
	TAS-B-HM	0.370*†	0.308*†	0.957*†	0.680	0.872	0.807*	0.630	0.804	0.829
Pairwise + In Batch	Random	0.330	0.279	0.901*	0.614	0.784	0.663	0.595	0.812	0.745
	TAS	0.348*†	0.294*†	0.926†	0.639	0.850	0.731	0.615	0.823	0.776
	TAS-B	0.363*†	0.308*†	0.934†	0.661	0.845	0.747	0.617	0.765	0.786
	TAS-B-HM	**0.403***†	**0.343***†	0.958*†	0.640	0.831	0.782	**0.665**	0.813	**0.834**

Table 1: Reproduction results of TAS, TAS-B and TAS-B-HM on MS MARCO dev, TREC DL 2019 and TREC DL 2020, for TAS-B-HM, the maximum hard margin value is set as default $M_{max} = 6$. Statistical significance with paired t-test($p < 0.05$) between baseline DR and all other methods is shown in *, between random sampling method under each teacher training pipeline with all other sampling methods is shown in †.

		MS MARCO dev			TREC DL 2019			TREC DL 2020		
Method	Sampling	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000
BM25	none	0.241	0.194	0.868	0.501	0.689	0.739	0.475	0.649	0.806
DR	none	0.353	0.298	0.935	0.602	0.781	0.714	0.602	0.782	0.757
Pairwise	Random	0.385	0.326	0.958	0.687	0.851	0.767	0.654	0.812	0.801
	TAS	0.385	0.325	0.957	0.677	0.851	0.769	0.650	0.820	0.819
	TAS-B	0.393*†	0.334*†	0.963*	0.686	0.866	0.0.783	0.665	0.823	0.825
In Batch	Random	0.372	0.315	0.951	0.680	0.857	0.745	0.631	0.773	0.792
	TAS	0.396*	0.336*	0.968*	0.706	0.886	0.799	0.667*	0.821	0.826*
	TAS-B	0.397*	0.338*	0.968*	0.716	0.910	0.800	0.677*	0.810	0.820*
Pairwise + In Batch	Random	0.391	0.331	0.968	0.695	0.891	0.787	0.673	0.812	0.839
	TAS	0.401*	0.338*	0.973*	0.713	0.878	0.831	0.689	0.815	0.862*
	TAS-B	**0.402***	**0.340***	**0.975***†	0.712	0.892	**0.845**	**0.693**	**0.843**	**0.865***

Table 2: Original results of TAS and TAS-B on MS MARCO dev, TREC DL 2019 and TREC DL 2020. Statistical significance with paired t-test($p < 0.05$) between baseline DR and all other methods is shown in *, between random sampling method under each teacher training pipeline with all other sampling methods is shown in †.

methods is not always higher than that of random sampling, and in some cases, lower average effectiveness is observed. We note these differences are, however, not statistically significant.

For TAS-B-HM, we find that imposing a threshold on the maximum margin allowed can improve the effectiveness of the In Batch and Pairwise + In Batch settings (dual supervision) However, when only the Pairwise teacher setting is used, the use of the threshold on the maximum margin does generally harm effectiveness.

While our results found similar trends to those reported in the original study, as discussed above, we also found that we could not obtain the same (or at least similar) absolute values as those reported in the original study. Of course, we did not expect to obtain exactly the same absolute values. There is stochasticity in the training process that most likely leads to small variations in the absolute values. However, we expected these differences to be minor. Notably, we observe considerable differences in absolute values between our reproduced results and the original ones, especially for the In Batch and Pairwise + In Batch settings.

A possible explanation for this considerable mismatch is that we used a ColBERT In Batch teacher model that is different from that used in the original study. Unfortunately, the original study lacked a clear description and associated details of the ColBERT model used: we could not find an indication of the exact parameters used for the training of the teacher models, and we are also uncertain whether

we used the correct model checkpoint for our In Batch teacher model. While we rely on a ColBERT model previously released by the same authors of the original study, we acknowledge that this does not mean that it was the same model they used.

Nevertheless, we found that the TAS-B results in the original study were more similar to our results obtained with TAS-B-HM than with TAS-B. The TAS-B-HM method differs from TAS-B in that it imposes a threshold on the pairwise margin (the difference in score between a positive and a negative passage) above which training pairs are not considered for selection. TAS-B-HM was not described in the original study, although the method is implemented in the code-base released with that work. It may be possible that the original study used the settings we refer to as TAS-B-HM, but called them TAS-B and failed to describe the hard margin filtering because of, for example, space limitations and the believe this is a minor implementation detail. While it is common that minor details cannot find space in a paper, in Section 4.3 we show that in the case of TAS-B-HM, the availability of the hard margin filtering mechanism is quite important in terms of effectiveness on MS MARCO, and importantly is an element of scarce generalisability when studied on a different dataset.

In summary, with respect to RQ1, we found that, despite being unable to replicate the exact results reported in the original study (in terms of absolute value), we could replicate the general findings. Importantly, we confirmed that topic-aware sampling leads to improved effectiveness on MS MARCO (both for dev and TREC queries). Specifically, both TAS and TAS-B sampling methods typically outperform Random sampling across all supervision settings studied; and using TAS-B generally yields higher effectiveness than using TAS. Additionally, we have added to the observations reported in the original study by identifying that applying a hard margin yields increased effectiveness, particularly when using a teacher model in the In Batch or Pairwise + In Batch.

4.2 Generalization to other Dataset

Next, we investigate the generalizability of the topic-aware sampling strategies to a different dataset, thus addressing our RQ2. The dataset we chose for this is the Amazon Shopping Queries dataset. Results for this dataset are reported in Table 3.

We first examine the retrieval task. For both datasets (MS MARCO and Amazon dataset), we find that the sampling strategies lead to far superior dense retrievers than the baseline DR. Among the different sampling methods, we find that TAS or TAS-B significantly improve retrieval effectiveness compared to the random sampling method for the In Batch and In Batch + Pairwise supervision settings. However, in contrast to the findings obtained on MS MARCO, on the Amazon dataset, we find that using a maximum hard margin results in lower effectiveness than TAS-B (though no statistically significant differences).

Next, we examine the re-ranking task. Compared to the baseline effectiveness, the Pairwise and the Pairwise + In Batch supervision settings consistently yield significantly higher effectiveness. However, effectiveness is at par with the baseline when training the student model using solely the In Batch supervision and no significant improvements are recorded. Comparing the proposed sampling techniques with random sampling shows that TAS-based techniques lead to only marginally higher effectiveness across all supervised teaching settings. When we specifically examine TAS-B-HM, we find that using the hard margin has little impact on effectiveness. We provide an explanation of why this may be the case in Section 4.3, where we compare the score distributions observed in MS MARCO and the Amazon datasets, which are displayed in Figure 2.

In summary, with respect to RQ2, we found that on Amazon Shopping Queries Dataset, TAS and TAS-B outperform random sampling for the In Batch and Pairwise + In Batch settings; this result is in line with the results obtained on MS MARCO. However, unlike on MS MARCO, we found that the use of hard margins (TAS-B-HM) does not improve over the other two TAS-based strategies.

4.3 Effectiveness of TAS-B-HM

Next, we examine the impact of different maximum hard margin values on the effectiveness of the trained models to address RQ3. Results on the MS MARCO dev, TREC DL 2019 and TREC DL 2020 are reported in Table 4, while results on the Amazon Shopping Queries dataset are reported in Table 5.

The MS MARCO results and those on TREC DL 2020 indicate that using a hard margin (TAS-B-HM) is more effective than hard the other settings; in particular, improvements are statistically significant for MS MARCO dev queries. While some improvements are also observed in TREC DL 2019 queries, these are inconsistent across measures, and the differences are not statistically significant. The findings on the Amazon dataset are instead in contrast to those previous ones, with no improvements found, although differences are not significant.

Our hypothesis to explain these findings is that setting a hard margin works by excluding some easily distinguishable training pairs, allowing the trained model to better separate at inference between positive passages and hard negative ones and thus increasing the score differences between them. This hypothesis would also explain why setting a maximum hard margin does not work on the Amazon Shopping Queries dataset: because the training pairs gathered from the Amazon Shopping Queries dataset are mostly hard-negatives. Thus, setting a maximum margin would only result in less training data being included in the training set.

To test this hypothesis, we evaluate models across different maximum margin values and test if this may impact the margin distance between pairs of relevant-not relevant passages observed during inference. We analyse our hypothesis only on TREC DL 2019 and 2020 datasets, as MS MARCO dev and Amazon shopping queries datasets contain fewer judged documents: it would be misleading to resort to treat un-judged passages as either relevant or non-relevant. To obtain the margin distance at inference, we follow the steps outlined below:

(1) we obtain the top $k = 1,000$ results retrieved by the models we analyze, and normalize their scores using min-max normalization;
(2) for every query q ($q \in Q$), we collect all the negative passages from the relevance assessments and then downsample the collected negative passages to only include those that exist across all retrieval runs (considering the top k only). We indicate the set containing these negative passages as N_q;

Amazon Shopping Queries Dataset		Retrieval Task			Re-ranking Task	
Method	Sampling	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10
DR	none	0.087	0.189	0.375	0.914	0.950
Pairwise	Random	0.164*	0.321*	0.494*	0.922*	0.956*
	TAS	0.167*	0.326*	0.503*	**0.924***	**0.960***
	TAS-B	0.163*	0.324*	0.497*	0.923*	0.958*
	TAS-B-HM	0.161*	0.315*	0.502*	0.923*	0.958*
In Batch	Random	0.226*	0.418*	0.600*	0.911	0.949
	TAS	0.231*	0.422*	0.627*†	0.914	0.951
	TAS-B	0.239*†	0.433*†	0.632*†	0.913	0.951
	TAS-B-HM	0.233*	0.426*	0.630*†	0.914	0.951
Pairwise + In Batch	Random	0.231*	0.424*	0.606*	0.919*	0.955*
	TAS	0.244*†	**0.444***†	0.645*†	0.920*	0.958*
	TAS-B	**0.245***†	0.443*†	**0.645***†	0.920*	0.956*
	TAS-B-HM	0.242*†	0.444*†	0.642*†	0.920*	0.957*

Table 3: Effectiveness of TAS, TAS-B a d TAS-B-HM on Amazon shopping queries dataset, for TAS-B-HM, the maximum hard margin value is set as default $M_{max} = 2$. Statistical significance with paired t-test($p < 0.05$) between baseline DR and all other methods is shown in *, between random sampling method under each teacher training pipeline with all other sampling methods is shown in †.

		MS MARCO dev			TREC DL 2019			TREC DL 2020		
Sampling	M_{max}	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000	ndcg@10	rr@10	recall@1000
TAS-B	none	0.363†	0.308 †	0.934†	0.661	0.845	0.747	0.617	0.765	0.786
TAS-B-HM	1	0.373†	0.312†	0.953*	0.682	0.849	0.799	0.663	0.811	0.821
	2	0.385*†	0.324*†	0.957*	0.682	0.821	**0.813**	0.655	0.841	0.833
	3	0.395*	0.333*	0.958*	0.675	0.826	0.802	0.664	0.833	0.838
	4	0.401*	0.339*	**0.961***	**0.687**	0.860	0.799	0.670	0.839	**0.841**
	5	**0.405***	**0.345***	0.958*	0.669	0.857	0.801	**0.680**	0.846	0.838
	6	0.403*	0.343*	0.958*	0.640	0.831	0.782	0.665	0.813	0.834
	7	0.403*	0.344*	0.956*	0.659	0.863	0.773	0.671	0.842	0.835
	8	0.400*	0.340*	0.956*	0.650	0.844	0.775	0.668	0.847	0.824
	9	0.397*	0.338*	0.953*	0.670	0.880	0.774	0.674	**0.860**	0.823
	10	0.397*	0.337*	0.953*	0.649	0.857	0.770	0.674	0.860	0.820
	15	0.369†	0.313†	0.940†	0.650	**0.884**	0.749	0.634	0.824	0.789
	20	0.369†	0.314†	0.940†	0.650	0.884	0.748	0.634	0.824	0.787

Table 4: Effectiveness of TAS-B-HM with different maximum hard margin (M_{max}) values on MS MARCO, TREC DL 2019 and TREC DL 2020. Statistical significance with paired t-test($p < 0.05$) between TAS-B all TAS-B-HM is shown *, between default TAS-B-HM ($M_{max} = 6$ with all other maximum hard margin values is shown in †.

(3) we extract p_q, the top ranked passage for query q and compute the margin between p_q and each negative passage n_q ($n \in N_q$), i.e. $p_q - n_q$ (which can be reduced to $1 - n_q$ because of the min-max normalization at step 1).
(4) We then calculate the inference margin distance for query q and model m ($Dist(q, m)$) according to:

$$Dist(q, m) = \frac{1}{|N_q|}(\sum_{n}^{N_q}(1 - n) \tag{8}$$

Now, given a set of Q queries, we can compute the cumulative distance for the set:

$$Dist(m) = \frac{1}{|Q|}\sum_{q}^{Q} D(q, m) \tag{9}$$

Figure 3 visualizes the relationship between the inference margin distance for difference hard margin cutoffs used at training (M_{max}), represented by the solid red line, and the effectiveness of the retrieval models (nDCG@10), represented by the solid blue line. The image also reports the inference margin distance and the nDCG@10 scores obtained when using the fixed value we identified in the original work's code-base; these are displayed as dotted lines. From the figures, we observe that the effectiveness of the model

Amazon Shopping Queries Dataset				
Sampling	M_{max}	ndcg@10	rr@10	recall@1000
TAS-B	none	**0.245**	**0.443**	**0.645**
TAS-B-HM	-1	0.205*†	0.388*†	0.593*†
	0	0.218*†	0.405*†	0.615*†
	1	0.237	0.433	0.641
	2	0.242	0.444	0.642

Table 5: Effectiveness of TAS-B-HM with different maximum hard margin (M_{max}) values on Amazon shopping queries dataset. Statistical significance with paired t-test($p < 0.05$) between TAS-B all TAS-B-HM is shown in *, between default TAS-B-HM ($M_{max} = 2$) with all other maximum hard margin values is shown in †.

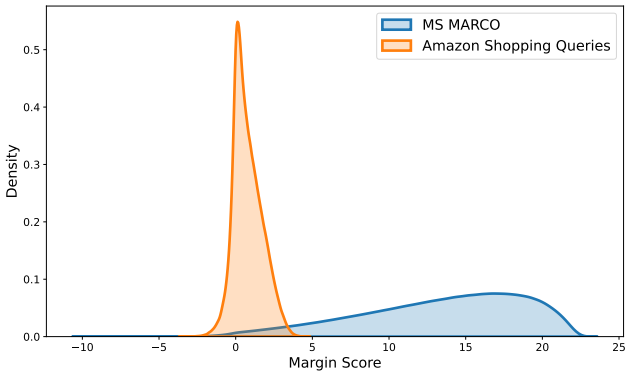

Figure 2: Density plot of MS MARCO and Amazon shopping queries datasets' training triples. The x-axis indicates the margin score of training pairs; the y-axis indicates the density of margin scores.

appears to be correlated with the inference margin distance of the model, especially for TREC DL 2020.

In summary, with respect to RQ3, we found that the impact of hard margin values on model effectiveness is dataset-dependent. In addition, we found a correlation between the inference margin distance of the model and its effectiveness.

5 SUMMARY OF FINDINGS

Below, we summarise the findings of our investigation in the reproduction of TAS-based methods.

RQ1: *Do we obtain the same results when replicating the original study within the same experimental setting (MS MARCO dataset)?*

We were unable to reproduce the exact results reported in the original study, but we could find the same overall trends. Importantly, we found that the proposed TAS and TAS-B methods are effective in improving the effectiveness of the student models in the empirical settings of datasets considered in the original work. We believe that the difference we observed in terms of absolute values between the original experiments and our replication are due to either the stochasticity of the training process or the use of the TAS-B-HB setting we identified in the code-base released with the original paper but not mentioned in the paper itself. Indeed,

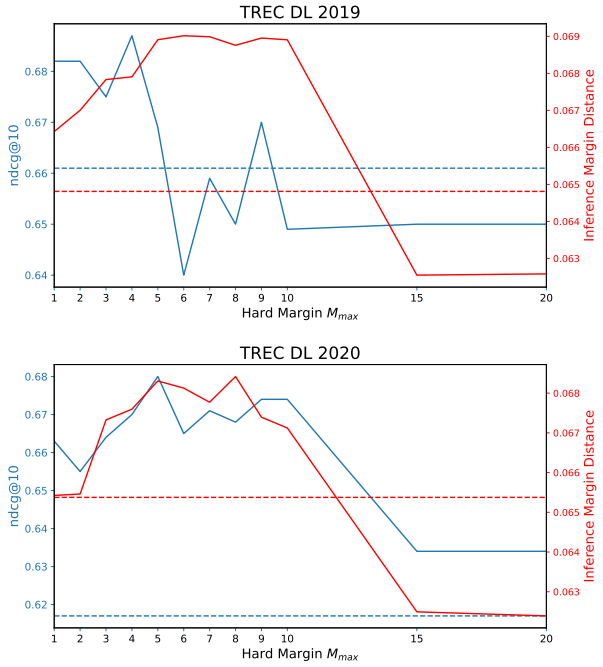

Figure 3: Plot analysing effect of Hard margin value M_{max} to the Margin distance of student models.

the investigation of the use of a maximum margin for TAS-B, i.e. the TAS-B-HM method is an addition to the contributions of the original paper. With respect to TAS-B-HM, we found this method being more effective than the other selection methods across most settings on the MS MARCO dataset.

RQ2: *To what extent do findings obtained on MS MARCO generalise to a different dataset?*

Our result show that the the effectiveness boost from topic-aware sampling techniques we observed on MS MARCO dev and TREC DL datasets does generalise to the Amazon Shopping Queries dataset; although that the performance increase is not as significant as in MS MARCO dev. However, we do find that setting a maximum hard margin value on the Amazon Shopping Queries dataset will rather decrease the performance, this may due to the fact the the training pairs are more difficult than training pairs in Ms MARCO dataset.

RQ3: *What is the impact of varying the maximum hard margin value on the effectiveness of models trained using the TAS-B-HM method?*

Based on the results presented, it is evident that varying the maximum hard margin values has a significant impact on the effectiveness of the trained models, although the impact of the maximum hard margin on the effectiveness of the trained models depends on the dataset used for training and testing. Results from MS MARCO (dev and TREC DL 2020 queries) show that using a hard margin can be more effective than not using one, with a higher average effectiveness observed across all evaluation measures and hard margin values. In contrast, the findings on the Amazon dataset are opposite: setting a maximum hard margin value for sampling the training data results in lower effectiveness in almost all cases. To explain why this may be the case, we put forward and tested the

hypothesis that setting a hard margin works by excluding some easily distinguishable training pairs, allowing at inference the trained model to assign a higher margin for harder-to-distinguish passages. The empirical analysis shows that the effectiveness of the model is somewhat correlated with the inference margin distance score of the model, especially for TREC DL 2020. Overall, the impact of varying the maximum hard margin values on model effectiveness is dataset-dependent, and more research is needed to determine the factors that affect this relationship.

6 RELATED WORK

6.1 Common Architecture of Rank-based PLMs

In ranking tasks that utilize pre-trained language models (PLMs), there are two commonly used architectures: the cross-encoder architecture and the dual-encoder architecture.

The rank-based cross-encoder model, such as monoBERT [17], takes two input texts (e.g., a query and a document) and concatenates them with a special token [*SEP*]. The model then encodes the combined text and produces a single output indicating the relevance of the document to the query.

In contrast, the rank-based dual-encoder model, also known as dense retrievers (Example: ANCE [22], RepBERT [23], Condenser [4]), encodes the query and document texts separately, producing representations for each. The relevance of the document to the query is then determined by taking the cosine similarity or dot product of the query and document representations.

One advantage of the dual-encoder model is that it is often more efficient than the cross-encoder model. The encoding of documents can be performed offline, allowing for faster query results during online time and significantly lower latency compared to cross-encoder models. Additionally, studies have demonstrated the effectiveness of dual-encoders when they are combined with sparse models like BM25 [21], or when pseudo relevance feedback is used [11].

6.2 Knowledge Distillation

Knowledge distillation is a technique in machine learning that involves transferring the knowledge learned from one model, called the teacher model, to another model, called the student model. In IR or NLP tasks, knowledge distillation is often used to create smaller and faster models that can perform well on a given task, while reducing the computational resources required [5].

6.2.1 MSE Loss. The mean squared error loss function (MSE) is used extensively in teaching the student model in the original study; this loss function is proposed by Hofstätter et al. [7]. The following formula presents the calculation of MSE loss in a training batch with teacher scores S and student scores T:

$$MSE(S,T) = \frac{1}{|S|} \sum_{s \in S, t \in T} (s-t)^2 \tag{10}$$

7 DISCUSSION AND LIMITATIONS

We investigated the reproducibility of the current state-of-the-art method for knowledge distillation in dense retrieval; this method exploits topic-aware sampling techniques to select subsets of the training data. In our experiments, aside from replicating the original empirical settings, we also studied the effectiveness of these techniques on a different dataset, related to product search, which exhibited the methods have overall limited generalizability. We acknowledge, however, that this result may be restricted to just the additional dataset we chose, and generalizability may be identified in other datasets or tasks. When investigating the impact of hard margin values, we only studied the effectiveness when the dual-supervision architecture is used, and thus the findings observed may not be representative of other settings. While we attempted to follow the methods described in the original study closely, some differences in the implementation or experimental setup could affect the results.

Ultimately, we believe that our study also offers a valuable contribution to the research community in terms of reproducibility. This reproducibility study highlights the critical importance of making research code publicly available to enable the replication of results obtained in original studies. Without access to the underlying code, reproducing results can be challenging, and it may also hinder other researchers from understanding the artifacts used in the research. A good practice for researchers to ensure easy reproducibility is to include a detailed pipeline in published code, as well as publish the original results gathered from the research. By doing so, detailed comparisons can be made between the reproduced and original results to identify any differences in the pipeline. In the context of this study, this information could have been especially crucial for understanding the impact of different maximum hard margin values on the effectiveness of the trained models, where small differences in implementation details can have significant effects on the results obtained. Therefore, we highly recommend researchers follow best practices for publishing research code, such as providing detailed documentation, code, and original results, to facilitate easy replication and reproducibility of research findings.

8 CONCLUSION

In this study, we reproduced a state-of-the-art method for training an effective BERT-based dense retriever that uses topic-aware sampling techniques to select training data to form cohesive training batches. We conducted a reproducibility study on the original datasets these techniques were evaluated upon and an additional dataset in a different retrieval task (the Amazon Shopping Queries dataset for product search). Our study allowed us also to investigate the impact of different maximum hard margin values on the effectiveness of the trained model.

Our empirical results confirm the findings of the original study, showing that TAS and TAS-B are effective methods for dense retrieval. Furthermore, our investigation of maximum hard margins demonstrates that setting a higher value often leads to increased effectiveness during dual-teacher supervision, at least on MS MARCO. However, determining the best value for the maximum hard margin requires further investigation into the behaviour of model training when this parameter is set, especially when datasets other than MS MARCO are used: we showed evidence that the practice related to setting the maximum margin might depend on the score distributions exhibited by the models on the dataset at hand.

ACKNOWLEDGEMENT

Shuai Wang is supported by a UQ Earmarked PhD Scholarship. This research is funded by Australian Research Council Discovery Project DP210104043.

REFERENCES

[1] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[2] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. https://doi.org/10.48550/ARXIV.2102.07662

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the TREC 2019 deep learning track. *arXiv preprint arXiv:2003.07820* (2020).

[4] Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. *arXiv preprint arXiv:2104.08253* (2021).

[5] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision* 129 (2021), 1789–1819.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[7] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666* (2020).

[8] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[9] Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584* (2020).

[10] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351* (2019).

[11] Hang Li, Shuai Wang, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. To interpolate or not to interpolate: Prf, dense and sparse retrievers. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2495–2500.

[12] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. 2356–2362.

[13] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2020. Distilling dense representations for ranking using tightly-coupled teachers. *arXiv preprint arXiv:2010.11386* (2020).

[14] Dugang Liu, Pengxiang Cheng, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2020. A general knowledge distillation framework for counterfactual recommendation via uniform data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 831–840.

[15] J MacQueen. 1965. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob.* 281.

[16] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *choice* 2640 (2016), 660.

[17] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Lin. 2019. Multi-stage document ranking with BERT. *arXiv preprint arXiv:1910.14424* (2019).

[18] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. (2022). arXiv:2206.06588

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[20] Nicola Tonellotto. 2022. Lecture Notes on Neural Information Retrieval. *arXiv preprint arXiv:2207.13443* (2022).

[21] Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval. In *Proceedings of the 2021 ACM SIGIR international conference on theory of information retrieval*. 317–324.

[22] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[23] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2020. Repbert: Contextualized text embeddings for first-stage retrieval. *arXiv preprint arXiv:2006.15498* (2020).

[24] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876* (2022).