Exploring Anisotropy and Outliers in Multilingual Language Models for Cross-Lingual Semantic Sentence Similarity

Katharina Hämmerl[1,2] and **Alina Fastowski**[1]

Jindřich Libovický[3] and **Alexander Fraser**[1,2]

[1]Center for Information and Language Processing, LMU Munich, Germany
{haemmerl,fraser}@cis.lmu.de
[2]Munich Centre for Machine Learning (MCML), Germany
[3]Faculty of Mathematics and Physics, Charles University, Czech Republic

Abstract

Previous work has shown that the representations output by contextual language models are more anisotropic than static type embeddings, and typically display outlier dimensions. This seems to be true for both monolingual and multilingual models, although much less work has been done on the multilingual context. Why these outliers occur and how they affect the representations is still an active area of research. We investigate outlier dimensions and their relationship to anisotropy in multiple pre-trained multilingual language models. We focus on cross-lingual semantic similarity tasks, as these are natural tasks for evaluating multilingual representations. Specifically, we examine sentence representations. Sentence transformers which are fine-tuned on parallel resources (that are not always available) perform better on this task, and we show that their representations are more isotropic. However, we aim to improve multilingual representations in general. We investigate how much of the performance difference can be made up by only transforming the embedding space without fine-tuning, and visualise the resulting spaces. We test different operations: Removing individual outlier dimensions, cluster-based isotropy enhancement, and ZCA whitening. We publish our code for reproducibility.[1]

1 Introduction

Since BERT-like (Devlin et al., 2019) language models rose to popularity, much has been made of the study of their hidden states and parameters (cf. Rogers et al., 2020). Thanks to their ability to incorporate context, they have been a major improvement for most tasks over static input embeddings. However, a certain issue has been shown in a number of works to affect contextual language models to a greater degree: outlier dimensions in the weights and hidden states (Kovaleva et al., 2021) and correspondingly, high anisotropy (Gao et al., 2019; Ethayarajh, 2019, inter alia). At the same time, the raw pre-trained embeddings work surprisingly badly for semantic similarity tasks, prompting efforts to train better sentence embeddings such as done by Reimers and Gurevych (2019).

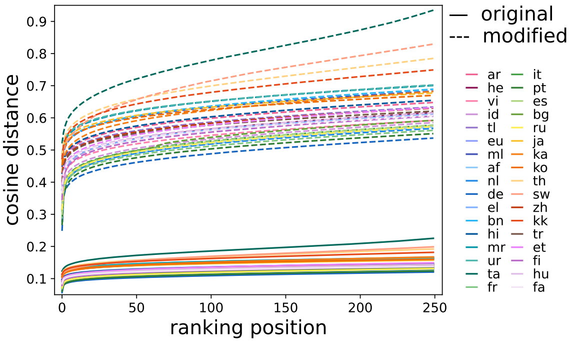

Figure 1: Effect of removing dimension 588 from layer 8 on Tatoeba cosine similarities. The x-axis is the ranking position of candidate sentences; the y-axis is their cosine distance from the query sentence, which should be relatively large for all but the correct translation. After removing 588 from the sentence representations, the highest ranking candidate sentence is much more clearly differentiated from the lower-ranking candidates.

In this paper, we are interested in multilingual sentence embedding quality. We discuss both outliers and anisotropy as two related aspects of embedding quality. Outlier dimensions are typically defined as dimensions that consistently produce values of a magnitude more than three or five times the standard deviation of all dimensions (Kovaleva et al., 2021). If a model has outlier dimensions in its hidden states, it will necessarily have higher anisotropy, since these dimensions create a consistent shift towards a certain direction in the embedding space. On the other hand, high anisotropy can also occur without individual dimensions meet-

[1] https://github.com/kathyhaem/outliers

Findings of the Association for Computational Linguistics: ACL 2023, pages 7023–7037
July 9-14, 2023 ©2023 Association for Computational Linguistics

ing the outlier definition, namely if some principal components composed of multiple dimensions are much larger than others. Therefore, as we understand it, anisotropy is the wider phenomenon of which outliers are a subset.

From a theoretical perspective, high anisotropy is considered a problem because it means that the model is not using the full representation space available, and because it translates to high average cosine similarity even between unrelated words or sentences. Figure 1 illustrates this problem clearly. This can increase the odds of picking a wrong candidate on word and sentence similarity tests, and makes representations produced by the model less expressive and less interpretable.

Outlier dimensions, since they contribute to anisotropy, entail similar challenges. On the other hand, they are easy to spot, easy to manipulate, and a straightforward entry point to the anisotropy issue. Previous work has sometimes found that models rely strongly on outlier weights for certain tasks, and are overly vulnerable to pruning a select few weights, e.g. (Kovaleva et al., 2021). Further, outliers have been found to present a challenge in model quantisation (Bondarenko et al., 2021).

Because they are aspects of the output representations, studies of anisotropy and outliers often use semantic similarity tests that rely directly on these representations, without fine-tuning the model. We follow this approach as well. In this work, we specifically consider sentence representations.

Only a small amount of work has been done on outliers and isotropy in multilingual models, which we focus on. Rajaee and Pilehvar (2022) found that mBERT does not contain outlier dimensions, while XLM-R does. However, both models nevertheless exhibit high anisotropy.

Another important aspect to consider in the multilingual case is that even if representations are more or less isotropically distributed, the subspaces for different languages can still be misaligned, which further affects cross-lingual performance. Training with parallel data, as done in Reimers and Gurevych (2020), is one way to radically improve cross-lingual alignment. However, we are interested in pushing models to perform well without parallel data. The present work therefore attempts to separate the effect of anisotropy from other factors that could account for the performance gap, such as the use of parallel data objectives and internal misalignment of languages.

Our contributions. This work provides an in-depth exploration of outlier dimensions and anisotropy in XLM-R and other pre-trained multilingual language models, using the Tatoeba (Artetxe and Schwenk, 2019), multilingual STS (Cer et al., 2017), and BUCC 2018 (Zweigenbaum et al., 2018) semantic similarity tasks and looking directly at the relevant hidden state representations.

We confirm that certain outlier dimensions have a negative effect on similarity search in the cross-lingual setting (§ 5). We find that outlier dimensions can differ between languages, although the largest outliers occur in all or most tested languages (§ 5). Anisotropy also varies across languages, and we observe a possible relationship to pre-training data size (§ 4). In our experiments, mBERT does exhibit outlier dimensions (§ 4).

Looking at semantic similarity task performance, we show that zeroing outliers and isotropy-enhancing transformations are quick ways to improve model performance on such tasks (§ 5, 6). However, a multilingual sentence-transformer performs much better out-of-the-box, and benefits little to not at all from further increasing isotropy. As we show in § 4, this model is already much more isotropic than XLM-R, its pre-trained equivalent.

Finally, we give a clearer intuition of the phenomena in question by using tSNE (van der Maaten and Hinton, 2008) to visualise embedding spaces (§ 7). This allows us to grasp more intuitively how anisotropy is one aspect of misalignment between languages in multilingual models.

2 Related Work

BERT-like models have dominated NLP research in recent years. Multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) are two popular models whose variants are used for many different ends. Accordingly, some amount of research has focused on analysing properties of the models, sometimes called “BERTology” (Rogers et al., 2020). The phenomena we discuss in this paper—outlier dimensions and anisotropy—are just two aspects of model analysis.

2.1 Describing the phenomena

First, we discuss outlier dimensions specifically. Kovaleva et al. (2021) focus specifically on outlier dimensions in the LayerNorm weights of English BERT. Around the same time that the LayerNorm outliers arise, training loss and evaluation perplexity start to fall off sharply. The exact cause is unknown but this suggests the outliers help the model, which they corroborate by showing that task performance decreases significantly when zeroing out outlier weights after fine-tuning. If zeroing the weights is done before fine-tuning, the model recovers most of the performance, but a slight disadvantage is still observed.

Timkey and van Schijndel (2021) take a different view of outliers in that they analyse hidden representations instead of weights. They also focus on similarity measures and find that in this context, the outlier dimensions "obscure representational quality". Rajaee and Pilehvar (2022) are one of few to focus on outliers in multilingual models: They find no outliers in mBERT, but do find them in XLM-R. The paper also looks at the embeddings of different languages separately, an approach we follow for the majority of our experiments.

As we mention above, outliers are one way to look at anisotropy in hidden representations. Ethayarajh (2019) is one of the first to present evidence for unusually high anisotropy in contextual embedding models, including BERT and GPT-2. Gao et al. (2019) describe the *representation degeneration problem* and suggest using cosine regularisation to mitigate it. We discuss mitigation approaches in more detail below (§ 2.3).

There are multiple ways to measure (an)isotropy, including but not limited to:

- average cosine similarity (cf. Ethayarajh, 2019; Timkey and van Schijndel, 2021)
- based on principal components (Mu and Viswanath, 2018)
- IsoScore (Rudman et al., 2022)

These are continuous measures, with value ranges depending on the method. While lower anisotropy is theoretically desirable, it can be hard to decide at what point a space is "isotropic enough". In the present work, we stick to the first measure, that is, average cosine similarity between random pairs (see § 4).

2.2 Searching for causes

It has been shown that word frequency plays a significant role in how representations are distributed in contextual models: For instance, rare words tend to be pushed further from the origin during pretraining, leading to a separation of tokens by frequency. Yu et al. (2022) show that rare token embeddings are the first to become anisotropic during pre-training, and seem to "take down with them" the rest of the space. Puccetti et al. (2022) similarly find that outliers are "driven by token frequency".

On the other hand, Luo et al. (2021) argue that outliers are caused by positional embeddings which display outliers, and this propagates forward through the model. They demonstrate this by training RoBERTa models with and without positional embeddings. The model without positional embeddings has much worse perplexity, but no outliers. This idea has not been confirmed by other works, and Rajaee and Pilehvar (2022) find that multilingual BERT, despite having positional embeddings, does not display outliers. We use a different mBERT checkpoint in our experiments which does exhibit outliers, but we draw no conclusions about positional embeddings.

2.3 Attempts at mitigation

Various methods have been suggested to increase isotropy in the contextual embedding space.

During training. Gao et al. (2019), who described anisotropy early on, proposed a cosine regularisation term to mitigate it. This term simply maximises the angle between any non-identical words. Building on this, Zhang et al. (2020) propose Laplacian regularisation as a way to specifically reduce similarity of word pairs that do not occur in similar contexts. Ferner and Wegenkittl (2022) apply a token-level variational loss to an encoder-decoder Transformer, similar to what is done in Variational Auto-Encoders. All three works add the regularisation terms to a model they train from scratch.

On the other hand, Ding et al. (2022) test several BERT-like models on GLUE tasks before and after "isotropy calibration" (fine-tuning with regularisation terms), and find that task scores do not consistently improve. They reason that this is because the models already benefit from local isotropy, thus further isotropy calibration does not help. We also note that these experiments are all done on tasks that use fine-tuning.

Post-hoc. Rather than training a model from scratch, Li et al. (2020) train normalising flows on STS and similar datasets that they want to test on, starting with a pre-trained BERT model— they call this approach *BERT-flow*. Both Su et al. (2021) and Huang et al. (2021) apply whitening to sentence representations. This operation transforms the mean of the sentence vectors to zero, and the covariance matrix to the identity matrix, as we discuss in more detail in § 6. Su et al. (2021) combine this with a dimensionality reduction strategy.

Timkey and van Schijndel (2021) also test several ways of postprocessing representations, such as standardisation and removing the top few principal components. Liang et al. (2021) and Rajaee and Pilehvar (2021a) remove dominant directions from the embedding space. The former learns a set of parameters for weighted removal (scaling) of principal components, while the latter clusters the data before removing the top principal components from each cluster. Rajaee and Pilehvar (2021b) find that removing the dominant directions after S-BERT training decreases STS performance, while removing them from the vanilla model improves performance. We corroborate these findings for the multilingual case. Jung et al. (2023) apply isotropy-improving methods, namely normalising flows and Whitening, in the context of dense retrieval models, and find score improvements on the target task.

Contrastive fine-tuning. Contrastive learning has become a popular technique in NLP in recent years (Zhang et al., 2022). Among other things, it has been shown to improve sentence embeddings and ensure they are more uniformly distributed. Examples include Gao et al. (2021); Kim et al. (2021); Zhang et al. (2021); Yan et al. (2021), and Reimers and Gurevych (2019). The latter, which we use as a reference in this work, uses in-batch contrastive optimisation in later implementations.

3 Datasets

Because we will show results of each of our experiments as we go along, we start here by introducing the datasets used.

3.1 Tatoeba

This is a cross-lingual sentence retrieval task compiled by Artetxe and Schwenk (2019) and pruned to 36 languages by Hu et al. (2020). We follow the implementation used by the latter. Each language is matched with English, and the objective is to find the correct translation for each query. The subtasks per language contain 1k examples each. The most similar translations are retrieved using the cosine similarity of the mean-pooled hidden representations from layer eight. The metric is accuracy.

3.2 BUCC

This is another similarity search task introduced by Zweigenbaum et al. (2018). However, since it focuses on parallel corpus building, not every query sentence has a match in the target language. Therefore, both precision and recall are important to performance. BUCC has four subtasks: German-English, French-English, Russian-English, and Chinese-English. We again follow the implementation by Hu et al. (2020). The test data contains several hundred thousand examples in each corpus, with between 1900 (Chinese) and 14400 (Russian) matched pairs. The task metric is F1.

3.3 Multilingual STS

Another cross-lingual semantic similarity task is Multilingual STS (Cer et al., 2017) from SemEval 2017. The task here is to score sentence pairs on a scale from 0 to 5 representing their relative similarity. There are four cross-lingual subtasks, namely Arabic-English, two Spanish-English tasks of varying difficulty, and Turkish-English. Each subtask contains 250 examples. The task metric is Pearson correlation with the gold labels.

3.4 Wikipedia

Following Rajaee and Pilehvar (2022), we further use a sample of Wikipedia data in six languages (Arabic, English, Spanish, Sundanese, Swahili, and Turkish) for our analysis. We use these for comparability, as we investigate some of the same multilingual models. The datasets contain between 347 (Sundanese) and 4952 (English) sentences.

4 Outlier and anisotropy analysis

Starting with data from Tatoeba, we derive sentence embeddings for all statements in each dataset. By *deriving sentence embeddings*, we mean encoding each sentence using the model's standard tokeniser, running it through the model in inference mode, then mean-pooling the result while ignoring special tokens. We proceed to calculate anisotropy scores for each language and dataset, as well as the outlier dimensions. We use the 3σ definition of outliers here. Note, however, that by considering sentence embeddings, which are already mean-pooled in one direction, we essentially have a smaller standard deviation and thus a more sensitive measure. For this reason, we also show which outliers are smaller than 5σ by *italicising* them in our tables.

Model	Anisotropy	Outliers	Means	Mean Cosine Contribution
XLM-R	0.92	588	-15.18	0.77
		306	3.08	0.03
		239	*-2.06*	*0.02*
		180	*1.86*	*0.01*
mBERT	0.73	227	-11.64	0.39
		195	-8.01	0.16
		731	*2.70*	*0.02*
Multil. S-BERT	0.35	588	-6.78	0.22
		145	-1.54	0.02
		306	*1.46*	*0.003*
		459	*-1.43*	*0.01*
		741	*1.21*	*0.01*

Table 1: Outliers and anisotropy scores in layer 8 of each model. The numbers in this table are based on Tatoeba data. Outliers are sorted by magnitude. We show all outliers according to the 3σ definition of outlier dimensions. We *italicise* dimensions that do not qualify as outliers under the 5σ definition.

For the anisotropy score, we adapt Timkey and van Schijndel's (2021) definition to the sentence level. Let S be a sample of n random sentence pairs from a corpus D. The approximate anisotropy $A(f_l)$ of layer l in model f is then:

$$A(f_l) = \frac{1}{n} \cdot \sum_{\{x,y\} \in S} \cos(f_l(x), f_l(y)) \qquad (1)$$

where $\cos(u, v)$ is the cosine similarity.

Further, we calculate the contributions to anisotropy of the largest dimensions. Analogously to the overall anisotropy, if $CC_i(u, v) = \frac{u_i v_i}{\|u\| \; \|v\|}$ is the contribution of dimension i to the total cosine similarity of u and v, then the contribution of dimension i to the overall anisotropy is:

$$CC(f_l^i) = \frac{1}{n} \cdot \sum_{\{x,y\} \in S} CC_i(f_l(x), f_l(y)). \qquad (2)$$

We use hidden representations from layer 8 when applying these techniques on Tatoeba data, since this task is usually done using layer 8. We test XLM-R, mBERT, and a multilingual S-BERT (Reimers and Gurevych, 2020) model which we have found to create good sentence embeddings across many languages.[2]

Results of the analysis are shown in Table 1. XLM-R has an extremely high anisotropy score: Any given random sentence pair is already considered very similar to each other. One of its outlier dimensions (588) contributes far and away the largest part to the expected cosine similarity. This dimension is still present as an outlier, though with a smaller magnitude and cosine contribution, in the multilingual S-BERT which was derived from XLM-R. The S-BERT model also has much lower anisotropy overall.

mBERT shows lower anisotropy than XLM-R but much higher values than the S-BERT. Its two largest dimensions both contribute significantly to anisotropy. Unlike Rajaee and Pilehvar (2022), we do find outlier dimensions in multilingual BERT. It is worth noting that we use a different checkpoint than they do (they use the uncased model, we use the cased version), and we focus on sentence representations rather than individual word embeddings. To verify our findings, we repeat our experiments on the same Wikipedia data they used—this now concerns the final layer of the model. We calculate sentence embeddings in this case as well. These results are listed in Table 2. Note that outlier dimensions can and do differ from layer to layer, which we observe in all three of these models. The multilingual S-BERT has no outliers larger than 5σ in the output layer, but does have larger outlier dimensions in the middle layer 8. It may be that the sentence-transformer tuning affects the later layers first and therefore more thoroughly.

In Table 3, we report anisotropy scores per language for our models. We also use Wikipedia data here, since this includes fewer languages but is of a more natural domain than Tatoeba. XLM-R exhibits such high anisotropy in these sentence embeddings that there is no meaningful difference

[2] The specific model we used is `sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens` and can be found on Huggingface.

Model	Anisotropy	Outliers	Means	Mean Cosine Contribution
XLM-R	0.99	588	17.86	0.89
		741	-5.62	0.09
mBERT (cased)	0.61	423	-1.97	0.03
		731	-1.54	0.02
		373	*-1.22*	*0.01*
		89	*-1.04*	*0.01*
		511	*-0.99*	*0.01*
		761	*-0.92*	*0.01*
		493	*-0.86*	*0.01*
Multil. S-BERT	0.27	*308*	*-0.80*	*0.01*
		281	*0.67*	*0.003*
		176	*0.57*	*0.002*
		152	*-0.57*	*0.002*

Table 2: Outliers and anisotropy scores in the output layer of each model. The numbers in this table are based on the Wikipedia data. Outliers are sorted by magnitude. We use the 3σ definition of outlier dimensions. We *italicise* dimensions that do not qualify as outliers under the 5σ definition.

Model	ar	en	es	su	sw	tr
XLM-R	0.996	0.997	0.996	0.996	0.995	0.996
mBERT (cased)	0.65	0.49	0.56	0.64	0.69	0.6
Multil. S-BERT	0.21	0.17	0.19	0.28	0.59	0.17

Table 3: Anisotropy scores, final layer, per language, on the Wikipedia data.

between the scores across languages. However, the other two models both show an interesting pattern: English and Spanish have the most isotropic spaces, with anisotropy increasing roughly as training data size decreases. This observation fits with the idea that anisotropy is frequency-driven (Yu et al., 2022; Puccetti et al., 2022), i.e., that less frequent tokens tend to be pushed further from the origin. Arabic is more anisotropic than Turkish despite having the same (S-BERT) or double (mBERT) the pre-training data size. Presumably this is due to Arabic using a non-Latin script, since the model has seen more Latin-script data. Sundanese and Swahili are the two languages with the smallest pre-training data of this set. Swahili has the highest anisotropy in both models, and by a large margin in the S-BERT model. This is somewhat surprising, since Sundanese has even smaller pre-training data, but may be down to data quality or tokenisation issues. It may even be that the S-BERT tuning included bad Swahili data—however, this is speculation, since the relevant documentation is lacking.

For XLM-R, we further graph the average hidden representations per layer using Tatoeba data. Layer 8 is shown in Figure 2; all layers in Figure 4 in the Appendix.

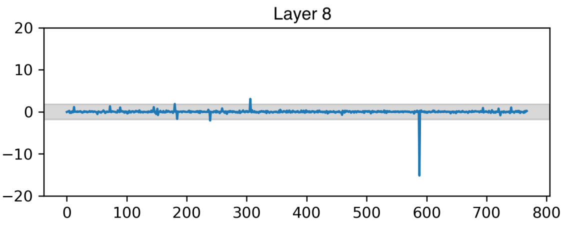

Figure 2: XLM-R mean embedding over all Tatoeba data; layer 8. The grey area denotes 3σ around the mean. The outlier dimensions are clearly visible.

5 Zeroing out dimensions

Based on the outlier analysis, we experiment with zeroing out dimensions from the sentence representations before feeding them to the similarity search functions. The biggest outlier, 588, clearly damages performance by greatly raising the similarity of all sentences. The correct candidate may thus be eclipsed by a false one more easily. Figure 1 illustrates how this occurs. On the x-axis are the ranking positions of candidate sentences, on the y-axis their average cosine distances (inverse to cosine similarity). In the unmodified model, all candidates are highly similar to the query sentence. After removing 588, candidates with lower ranking

Model	Tatoeba	BUCC
XLM-R	50.35	59.1
XLM-R *-588*	52.99	59.6
XLM-R *-306*	50.59	58.0
XLM-R *-239*	51.11	59.2
XLM-R, *18 dims rem.*	**60.09**	**64.4**
Multil. S-BERT	85.17	85.7

Table 4: Average Tatoeba (accuracy) and BUCC (F1) scores for XLM-R and modified versions with large dimensions set to zero. The multilingual S-BERT is included as a reference.

become much more dissimilar, and the difference between the top candidate and the other sentences increases, which is a desirable property (note that the graphic does not show whether and which candidate sentences changed their ranking as a result).

In addition to zeroing the largest outliers, we identified other dimensions of interest by their magnitude. We included the ten largest dimensions in each language of Tatoeba, finding a total of 18 dimensions that are in the top ten for any of the 36 languages.[3] These dimensions include the outliers previously identified, as well as additional large dimensions. We explored removing these dimensions individually and generally found smaller effects, though still a marked effect for some of them. The results are listed in Table 4.

We removed the same dimensions from sentence embeddings of BUCC (Zweigenbaum et al., 2018) data. Interestingly, this sometimes improved precision while also worsening recall. Thus, the overall improvements on this task were small (e.g., 588) or even negated (306). Removing all 18 large dimensions from Tatoeba and BUCC yields +9.7 accuracy and +5.3 F1 over the vanilla XLM-R model, respectively. That said, even with this performance gain, the gap to the sentence-transformer is still very large. In addition, manually zeroing a large number of dimensions depending on the task data cannot be done in a real-world system.

6 Isotropy-enhancing operations

Aside from directly zeroing out individual dimensions, we can apply transformations over the set of embeddings that largely eliminate anisotropy and mean-center the representations. In this work, we test two such transformations:

1. ZCA Whitening (cf. Huang et al., 2021)
2. Cluster-based isotropy enhancement (Rajaee and Pilehvar, 2021a)

6.1 ZCA Whitening

Whitening is an operation originally used in data pre-processing, in order to remove correlations between the input data features to a machine learning system. It is also called a “sphering transformation”, since the resulting data space is a hyperdimensional sphere. However, whitening has recently been used to transform output embeddings of models such as BERT (cf. Huang et al., 2021), before using them for downstream applications.

For a given space X with covariance Σ and mean 0, there are many valid whitening transformations. The resulting matrix $Y = WX$ must have the identity matrix I as its covariance, and the whitening transformation W must satisfy the condition:

$$W^T W = \Sigma^{-1}. \tag{3}$$

Given that Σ can be decomposed into:

$$\Sigma = D\Lambda D^T, \tag{4}$$

a valid W can be found as follows:

$$W = D\Lambda^{-\frac{1}{2}} D^T. \tag{5}$$

6.2 Cluster-based isotropy enhancement

We adopt this method from Rajaee and Pilehvar (2021a). The first step is to separate the provided data into clusters. In their paper, Rajaee and Pilehvar (2021a) use 27 clusters. We make the number of clusters dependent on the number of examples—with too few examples in a single cluster, the concept of “isotropy” becomes meaningless, and it can lead to computation errors. Each cluster is mean-centered, which is necessary for the subsequent steps. Then, PCA is applied to every cluster, and the top k principal components (“dominant directions”) are zeroed out. We follow the original paper in setting $k = 12$.

6.3 Discussion

The common thread of these methods is that they transform the output representations based on some set of encoded data. This means that either the transformation must be calculated anew for every set of data, or retained from a training set in order to apply it to new data. Though this is not ideal from

[3] [12, 63, 145, 151, 152, 266, 267, 459, 723, 728, 588, 306, 239, 184, 180]

Model	**Anisotropy**	**Tatoeba**	**STS**			
			ar-en	**es-en a)**	**es-en b)**	**tr-en**
XLM-R	0.92	50.35	.114	.04	-.059	.141
XLM-R, *18 dims rem.*	0.47	60.09	—	—	—	—
XLM-R + CBIE	-3.9×10^{-5}	69.01	.316	**.445**	.121	**.37**
XLM-R + Whitening	7.6×10^{-5}	**70.03**	**.355**	.444	**.153**	.36
mBERT	0.73	37.53	.20	.244	.146	.172
mBERT + CBIE	5.7×10^{-5}	**45.79**	**.25**	**.403**	.15	**.217**
mBERT + Whitening	-6.6×10^{-6}	45.14	.208	.395	**.171**	.154
Multil. S-BERT	0.35	85.17	**.772**	**.779**	**.235**	**.762**
S-BERT + CBIE	5.8×10^{-5}	86.36	.722	.742	.233	.724
S-BERT + Whitening	0.0001	**87.35**	.745	.772	.222	.748

Table 5: Anisotropy scores, average Tatoeba (accuracy) scores, and STS cross-lingual subset scores (Pearson correlation) for XLM-R, mBERT, multilingual S-BERT, and modified versions with post-hoc transformations applied to the sentence embeddings.

an application perspective, we follow the approach of calculating the transformation for every new set of encoded data. The tasks in question do not use fine-tuning on any kind of training data, so we transform the embedded test data. An alternative would be to learn and retain a transformation based on some external dataset, then apply this to the task data. Such an approach would be especially helpful when doing inference on only a few queries at a time, or when the overhead of computing the transformation should be avoided at inference time.

6.4 Results

After applying the transformations, we run our anisotropy analysis again. We also test Tatoeba and STS performance before and after the transformations. The results are listed in Table 5. For XLM-R, the transformations lead to a performance boost of almost 20 points on Tatoeba. Recall that removing the top dimensions improved accuracy by only around 10 points. For mBERT, which is more isotropic to begin with, the difference is only eight points. Other factors, such as a more complex misalignment of different languages, seem to be a bigger bottleneck for its performance. The multilingual S-BERT benefits very little from the isotropy-enhancing transformations.

For STS, the multilingual S-BERT in fact performs better without the transformations. mBERT and XLM-R do benefit from the transformations to some degree: In most cases, there is a large improvement, particularly in XLM-R. For mBERT, the **es-en b)** subset only shows a small improvement, and the others benefit more from CBIE than from whitening. Rajaee and Pilehvar (2022) also test on STS, including the monolingual subsets. However, since they report Spearman correlations rather than Pearson, as well as using a different mBERT checkpoint than we do, the numbers are not directly comparable, and we do not show them in our table. The main takeaway here is that using the whitening transformation yields similar results overall to CBIE, and that both work to improve sentence-level representations for semantic similarity. Also, they both have little to no benefit in the S-BERT model, which was tuned with parallel data and is already much more isotropic.

After the transformations, anisotropy scores are very close to zero; that is, the spaces are extremely isotropic. We can also see this in the t-SNE visualisations of these spaces, see § 7. However, applying the outlier definition of three times the standard deviation, we still find outlier dimensions in the transformed spaces. These all have very small magnitude, and are not necessarily related to the dimensions that were outliers before. Since the transformations are not deterministic, these outlier dimensions can also change when recalculating the transformed spaces. Therefore, we do not consider these dimensions true outliers. In an (artificially) highly isotropic space, the traditional outlier definition of larger than three standard deviations may simply not apply.

7 Embedding space visualisation

To visualise the representation space, we use t-SNE (van der Maaten and Hinton, 2008). First, we apply a PCA dimensionality reduction to 50 dimensions. Then, we reduce the dimensionality further using

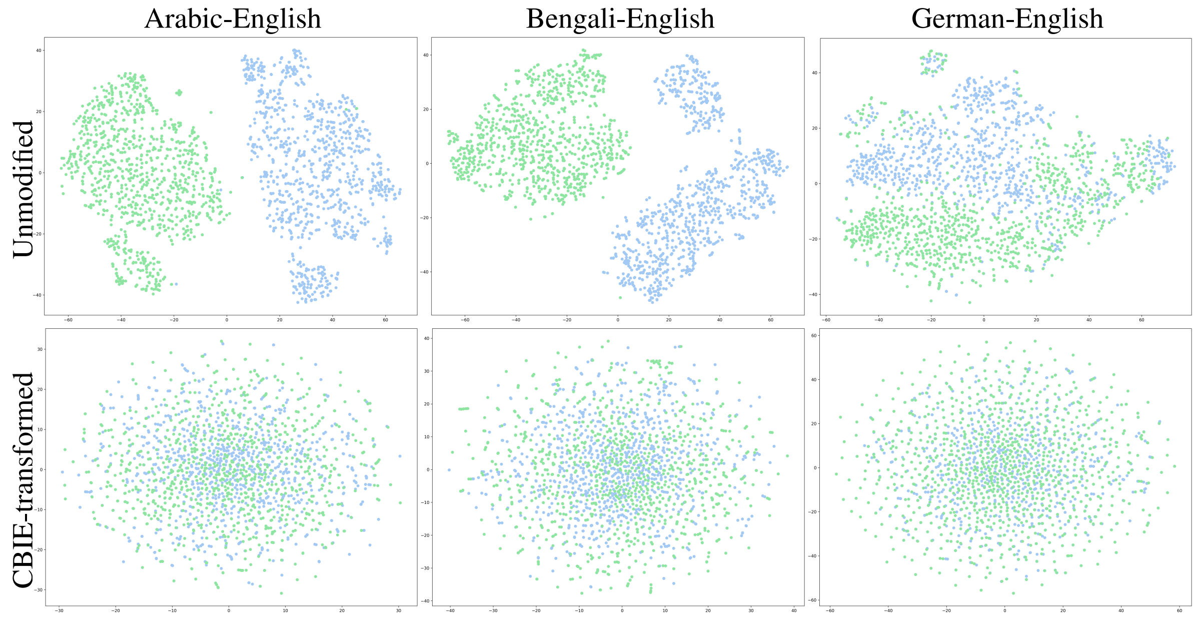

Figure 3: Left to right: Arabic-English, Bengali-English, and German-English Tatoeba sentence embeddings from XLM-R. Top to bottom: Unmodified and CBIE-transformed versions of the embeddings. The source languages are shown in green, the English in blue.

t-SNE and plot the space in two dimensions. In Figure 3, we show examples from Tatoeba data in XLM-R: Arabic, Bengali, and German. For the first two, accuracy increased by more than 20 points after the transformation, while German is already a high-resource language where accuracy only increased by around 5 points. Since CBIE and Whitening produce very similar visualisations, we only show CBIE.

The unmodified spaces very clearly show the problem of internal misalignment between different languages in the model, which disproportionately affects languages with less pre-training data and/or non-Latin scripts. With Arabic-English and Bengali-English, the source and target language spaces are almost disjunct. This issue can be addressed using isotropy-increasing transformations, but they do not solve the problem entirely. For instance, the unmodified sub-spaces of Bengali and English also have markedly different shapes, despite representing a set of parallel sentence pairs. Matching the equivalent sentences to each other starting from such different spaces is more complex than merely applying a linear transformation to increase isotropy.

8 Conclusions

We have analysed how outlier dimensions and anisotropy interact with cross-lingual semantic similarity tasks in pre-trained multilingual language models. In particular, we focused on the sentence representations of multilingual BERT and XLM-R, comparing them to the sentence representations of a multilingual S-BERT model—essentially a modified XLM-R trained with parallel data to optimise for sentence representations. We employed a range of methods on several different tasks to approach the question from multiple angles. The simplest method of increasing isotropy is removing the largest (outlier) dimensions from the sentence embeddings. We compared the results of this with further-reaching isotropy-increasing transformations. Additionally, we examined how changing the representations affected anisotropy measures and outlier dimensions. Finally, we plotted unmodified and transformed sentence representation spaces to illustrate how anisotropy is one aspect that affects sentence similarity, but reducing it does not resolve all issues in the space.

Future Work. Potential future research questions include: Are outliers and anisotropy also relevant when using *fine-tuned* models for cross-lingual transfer? Do larger, particularly generative models, have these issues affecting cross-lingual similarity? Are the pre-training dynamics of anisotropy in multilingual models similar to those of monolingual models? How can we train multilingual models to avoid a degenerating representation space?

Limitations

This paper examines the anisotropy and outlier phenomenon only for a few, relatively similar, models. The isotropy-increasing transformations are non-deterministic and have to be calculated post-hoc based on some set of embedded data, which may not be practical for applications where inference is done on individual or small batches of examples.

Since we specifically consider sentence representations, we first average over word embeddings before calculating the mean and standard deviation for outlier analysis. This in effect reduces the sample size and leads to a smaller standard deviation, making our analysis more sensitive to even slight outlier dimensions. Another reason to work with relatively small datasets is to make computing the transformations simple and fast, but this may limit the ability of these transformations to generalise.

Acknowledgements

This publication was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder; and by the German Research Foundation (DFG; grant FR 2829/4-1). The work at CUNI was supported by Charles University project PRIMUS/23/SCI/023, and by the European Commission via its Horizon research and innovation programme (No. 870930 and 101070350).

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2021. Understanding and overcoming the challenges of efficient transformer quantization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7947–7969, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. 2022. On isotropy calibration of transformer models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Cornelia Ferner and Stefan Wegenkittl. 2022. Benefits from variational regularization in language models. *Machine Learning and Knowledge Extraction*.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the*

Association for Computational Linguistics: EMNLP 2021, pages 238–244, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Euna Jung, Jungwon Park, Jaekoel Choi, Sungyoon Kim, and Wonjong Rhee. 2023. Isotropic representation can improve dense retrieval. In *Advances in Knowledge Discovery and Data Mining*, page 125–137, Cham. Springer Nature Switzerland.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. Association for Computational Linguistics.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT busters: Outlier dimensions that disrupt transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3392–3405, Online. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yuxin Liang, Rui Cao, Jie Zheng, Jie Ren, and Ling Gao. 2021. Learning to remove: Towards isotropic pre-trained BERT embedding. In *Artificial Neural Networks and Machine Learning – ICANN 2021*, page 448–459, Cham. Springer International Publishing.

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5312–5327, Online. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. 2022. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1286–1304, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021a. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 575–584, Online. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021b. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3042–3049, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *ArXiv*, abs/2103.15316.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Sangwon Yu, Jongyoon Song, Heeseung Kim, Seongmin Lee, Woo-Jong Ryu, and Sungroh Yoon. 2022. Rare tokens degenerate all tokens: Improving neural text generation via adaptive gradient gating for rare token embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29–45, Dublin, Ireland. Association for Computational Linguistics.

Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise supervised contrastive learning of sentence representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527, Online. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third BUCC shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

A XLM-R Mean Embeddings of Tatoeba in all Layers

See Figure 4.

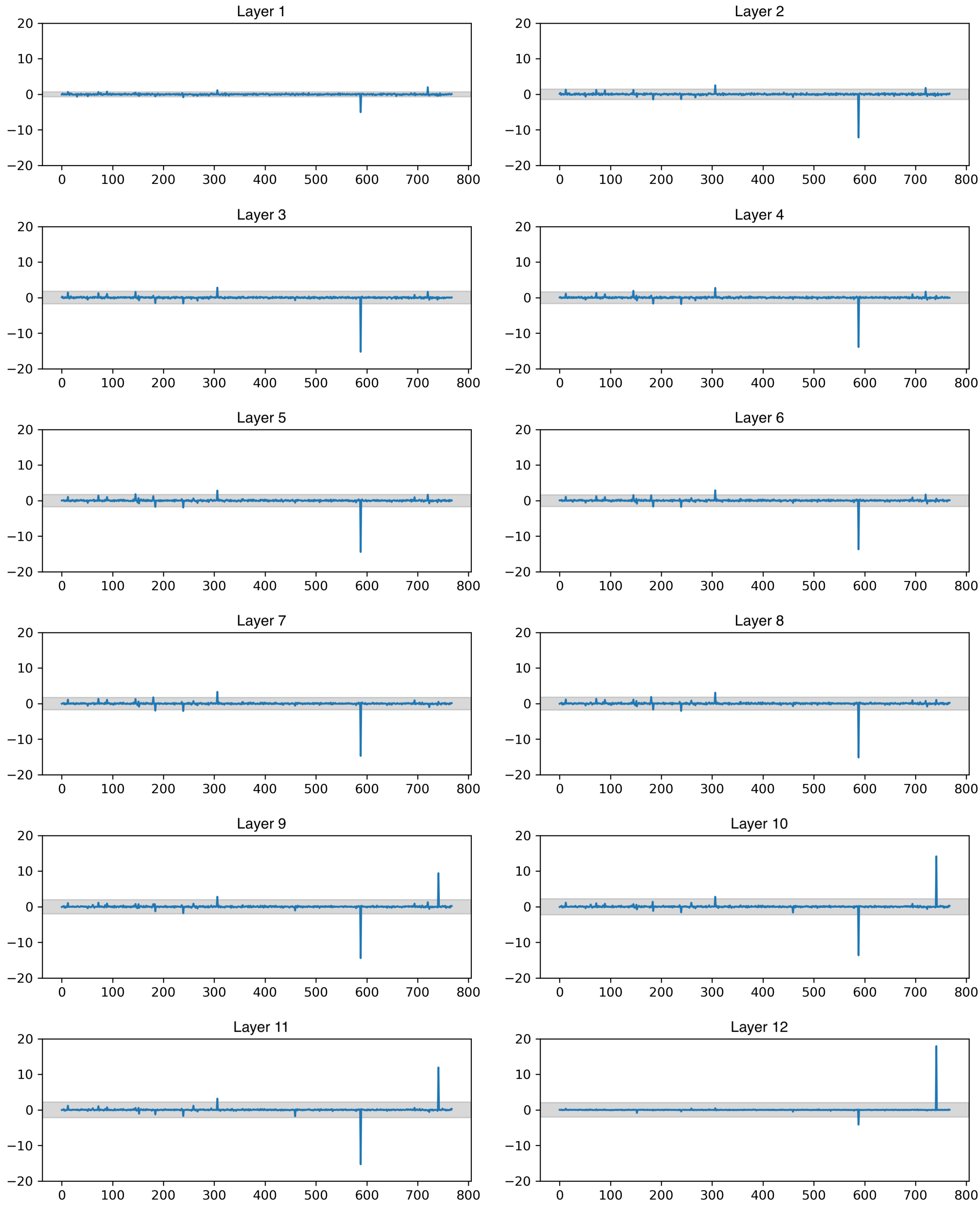

Figure 4: XLM-R mean embeddings on Tatoeba data.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations; Section 6

☒ A2. Did you discuss any potential risks of your work?
We do not see additional risks of this work beyond pre-trained multilingual language models in general.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 for dataset description; in Section 6 we cite implementations from previous work

☑ B1. Did you cite the creators of artifacts you used?
yes, as above

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
we do not release any artifacts at this point

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
we do not release artifacts at this point; existing artifacts were released for further research

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3

C ☑ Did you run computational experiments?

Sections 4-7

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
the majority of the experiments were a) small scripts run on CPU, b) done on many different machines c) done by different authors. we did try to avoid recomputing embeddings if possible

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 4, 5, 6

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
cosine similarity is deterministic, so the results would have been the same

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4-7

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.