Is Anisotropy Truly Harmful? A Case Study on Text Clustering

Mira Ait-Saada[§†] and **Mohamed Nadif**[§]
[§]Centre Borelli UMR9010, Université Paris Cité, 75006, Paris
[†]Caisse des Dépôts et Consignations, 75013, Paris
{mira.ait-saada,mohamed.nadif}@u-paris.fr

Abstract

In the last few years, several studies have been devoted to dissecting dense text representations in order to understand their effectiveness and further improve their quality. Particularly, the anisotropy of such representations has been observed, which means that the directions of the word vectors are not evenly distributed across the space but rather concentrated in a narrow cone. This has led to several attempts to counteract this phenomenon both on static and contextualized text representations. However, despite this effort, there is no established relationship between anisotropy and performance. In this paper, we aim to bridge this gap by investigating the impact of different transformations on both the isotropy and the performance in order to assess the true impact of anisotropy. To this end, we rely on the clustering task as a means of evaluating the ability of text representations to produce meaningful groups. Thereby, we empirically show a limited impact of anisotropy on the expressiveness of sentence representations both in terms of directions and L_2 closeness.

1 Introduction

Contextualized pre-trained representations are now widely used as input to various tasks such as information retrieval (Lin et al., 2021), anomaly detection (Ait-Saada and Nadif, 2023) and document clustering (Boutalbi et al., 2022). In parallel, several studies have investigated the intrinsic properties of Transformers (Peters et al., 2018; Ait Saada et al., 2021; Ethayarajh, 2019; Kovaleva et al., 2019) in order to demystify these black-box models and the reasons behind their impressive performance levels. Particularly, it has been observed that language models in general (Gao et al., 2019) and Transformer word embedding models in particular (Ethayarajh, 2019; Wang et al., 2020) produce an anisotropic embedding space. This concretely means that the directions of trained dense word representations do not occupy uniformly the embedding space, which is suspected to limit their expressiveness and thus their expected performance on downstream tasks. The main question addressed in this paper is how harmful this anisotropy really is regarding the quality of text representations.

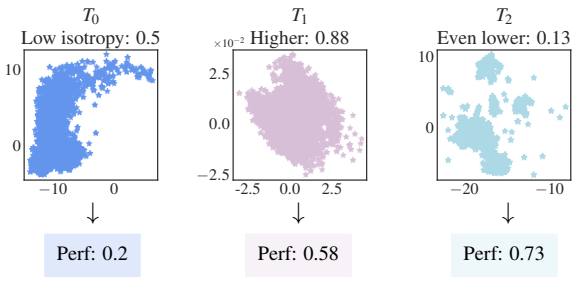

Figure 1: Isotropy versus performance with different transformations (T_0 = no transformation).

Several approaches have been proposed to increase the isotropy of dense representations, based on different strategies. In the context of static word embeddings like GloVe and word2vec, both Raunak et al. (2019) and Mu and Viswanath (2018) propose a post-processing method that consists in removing the first principal components before reconstructing the word vectors as opposed to the traditional approach of removing the weakest components. This approach improves the quality of word vectors on several downstream tasks while reducing their anisotropy (Mu and Viswanath, 2018).

As to contextualized representations provided by Transformer models, several approaches have been proposed in order to alleviate the anisotropy problem. For instance, based on the idea that anisotropic representations tend to have high expected pairwise cosine similarity, Wang et al. (2020) propose to apply a cosine similarity regularization term to the embedding matrix. In the same vein, Gao et al. (2019) propose a method named "spectrum control" that allows for increasing the isotropy of Transformer representations and improving the performance of the machine translation task. To this purpose, they propose regularization

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1194–1203

terms that hamper the singular value decay of the embedding matrix. However, despite the success of these *optimization* tricks in lowering the anisotropy of Transformer representations, Ding et al. (2022) have recently shown that they do not bring any improvement, relying on several tasks like summarization and sentence similarity (STS). They even observed a certain deterioration of the performance brought by anisotropy mitigation techniques.

In contrast, Rajaee and Pilehvar (2022, 2021) show that *post-processing* methods made for increasing isotropy are also responsible for a performance increase in the STS task in both monolingual and cross-lingual settings. Similarly, the whitening operation, which consists in using the principal components normalized by their inertia, has shown an increase in isotropy as well as enhanced performance in STS (Su et al., 2021; Huang et al., 2021) and document clustering (Ait-Saada et al., 2021). However, there is no evidence that the decrease of anisotropy brought by such transformations is directly responsible for the gain of performance, as shown in Figure 1, which gives an initial idea of the question addressed in this paper.

Indeed, despite the great energy devoted to studying and mitigating the anisotropy of dense text representations, there is no clear connection between isotropy and performance, which seems to depend, inter alia, on the sought task. In order to contribute to settling this question, we consider using a task that has never been used for this purpose: document clustering. The rationale behind this choice is to evaluate, under different degrees of isotropy, the capability of text representations to facilitate the clear separation and identification of meaningful groups of documents through clustering algorithms.

The main contributions of this paper are:

- We extend the isotropy study of word embeddings to document representations.
- We investigate the correlation between different isotropy measures.
- We assess the connection between isotropy and quality of representation.

2 Background

2.1 Isotropy measures

Let $\mathcal{X} = \{\mathbf{x}_i\}$ be a set of n vector representations, characterizing n words or documents by d features. In Mu and Viswanath (2018), the isotropy is assessed using the partition function ψ as follows:

$$\frac{\min_{\|\mathbf{c}\|=1} \psi(\mathbf{c})}{\max_{\|\mathbf{c}\|=1} \psi(\mathbf{c})}; \quad \text{where } \psi(\mathbf{c}) = \sum_{i=1}^{n} e^{\langle \mathbf{x}_i, \mathbf{c} \rangle}$$

This approach is inspired by the theoretical findings issued by Arora et al. (2016) who prove that, for isotropic representations $\mathcal{X}$, the partition function ψ can be approximated by a constant for any unit vector $\mathbf{c}$, thus leading to a min/max ratio score of 1. As there is no analytic solution $\mathbf{c}$ that maximizes or minimizes $\psi(\mathbf{c})$, Mu and Viswanath propose to use the eigenvectors of the covariance matrix as the set of unit vectors, which leads to:

$$\mathcal{I}_{pf}(\mathcal{X}) = \frac{\min_{\mathbf{w}_j} \psi(\mathbf{w}_j)}{\max_{\mathbf{w}_j} \psi(\mathbf{w}_j)}$$

where *pf* stands for *partition function*, $\mathbf{w}_j$ is the jth eigenvector of $\mathbf{X}^\top \mathbf{X}$ ($\mathbf{X}$ being the representation matrix). In our experiments, $\mathcal{X}$ contains representations of either words or sentences/documents. In addition to this measure, Wang et al. (2020) quantify the anisotropy by the standard deviation of the partition function normalized by the mean:

$$\mathcal{A}(\mathcal{X}) = \sqrt{\frac{\sum_{j=1}^{d} (\psi(\mathbf{w}_j) - \bar{\psi}))^2}{d\,\bar{\psi}^2}}$$

where $\bar{\psi}$ is the average value of the partition function. Perfectly isotropic representations lead to $\mathcal{A}(\mathcal{X}) = 0$ and greater values denote a higher anisotropy. For our purpose, we derive the isotropy score as the square root of the precision score $\tau = 1/\sigma$, which leads to:

$$\mathcal{I}_{pf_2}(\mathcal{X}) = \frac{1}{\sqrt{\sigma}} = \frac{1}{\mathcal{A}(\mathcal{X})}$$

σ being the variance normalized by $d\bar{\psi}^2$.

On the other hand, the study of anisotropy provided in (Ethayarajh, 2019) has been applied to word representations and the empirical results have been obtained using a high number of words picked randomly. The authors rely on the assumption that the expected similarity of two words uniformly randomly sampled from an isotropic embedding space is zero and that high similarities characterize an anisotropic embedding space. They hence use the expected pairwise cosine similarity in order to assess the anisotropy level of word representations. The *isotropy* is thus obtained by:

$$\mathcal{I}_{cos} := \mathbb{E}_{i \neq i'}(1 - \cos(\mathbf{x}_i, \mathbf{x}_{i'}))$$

where the score is computed over m random pairs $(\mathbf{x}_i, \mathbf{x}_{i'})$ of vector representations.

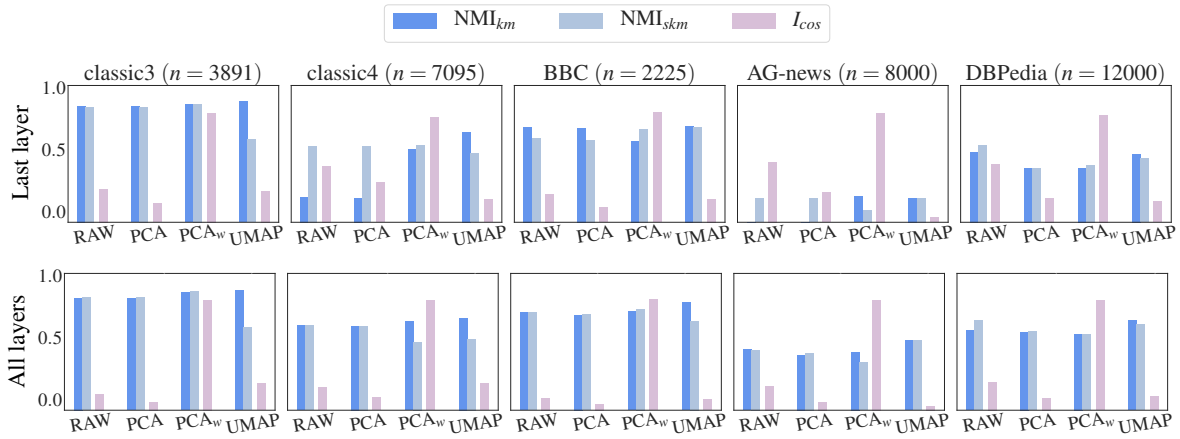

Figure 2: Isotropy against clustering performance. The first row is obtained using the last layer of BERT while the second row uses all the layers averaged together. NMI$_{km}$ and NMI$_{skm}$ correspond to the NMI score obtained by k-means and spherical k-means respectively. $\mathcal{I}_{cos}$ is the cosine isotropy score computed using Equation 2.1.

2.2 Quality measures

In order to assess the quality of text representations $\mathcal{X}$ of size n, we rely on the document clustering task, with the aim of estimating the ability of a clustering algorithm to accurately distinguish groups of documents in a corpus represented by $\mathcal{X}$. As the accuracy measure is not reliable when the classes are dramatically unbalanced, this is achieved using two well-known measures: Normalized Mutual Information (NMI, Strehl and Ghosh 2002), and the Adjusted Rand Index (ARI, Hubert and Arabie 1985; Steinley 2004).

Thereby, to compare two partitions A and B into g clusters, the NMI metric takes the following form: $\text{NMI}(A,B) = \frac{\text{MI}(A,B)}{\sqrt{\mathcal{H}(A)\,\mathcal{H}(B)}}$ where MI(A,B) denotes the mutual information while $\mathcal{H}(.)$ denotes the entropy; NMI(A,B) is hence given by:

$$\frac{\sum_{k,\ell} \frac{n_{k\ell}}{n} \log \frac{n n_{k\ell}}{n_k \hat{n}_\ell}}{\sqrt{(\sum_k n_k \log \frac{n_k}{n})(\sum_\ell \hat{n}_\ell \log \frac{\hat{n}_\ell}{n})}}$$

where n_k represents the number of samples contained in the class $A_k (1 \leq k \leq g)$, $\hat{n}_\ell$ the number of samples belonging to the class $B_\ell (1 \leq \ell \leq g)$, and $n_{k\ell}$ the number of samples that are at the intersection between the class A_k and the class B_ℓ.

The ARI metric, is a measure of the similarity between two groups of data. From a mathematical point of view, the ARI(A,B) is related to the precision and is given by:

$$\frac{\sum_{k,\ell} \binom{n_{k\ell}}{2} - \left[\sum_k \binom{n_k}{2} \sum_\ell \binom{\hat{n}_\ell}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_k \binom{n_k}{2} + \sum_\ell \binom{\hat{n}_\ell}{2}\right] - \left[\sum_k \binom{n_k}{2} \sum_\ell \binom{\hat{n}_\ell}{2}\right] \Big/ \binom{n}{2}}$$

where the binomial coefficient $\binom{u}{v}$ can be interpreted as the number of ways to choose u elements from a v-elements set.

Intuitively, NMI quantifies how much the estimated clustering is informative about the true clustering, while the ARI measures the degree of agreement between the estimated clustering and the reference partition. Both NMI and ARI are equal to 1 if the resulting clustering partition is identical to the ground truth.

3 Experiments

In this study, we aim to determine to what extent the anisotropy actually affects the quality of the representations and their ability to discriminate data samples through separable clusters. To this end, we use three measures to evaluate the isotropy of the original embedding space before and after post-processing. Then, we compare the changes in isotropy with the corresponding clustering performance in order to establish a potential relationship between the two concepts.

Relying on several isotropy measures allows us to consolidate confidence in our conclusions and, at the same time, verify if the measures agree with each other. In the same spirit, using different clustering methods and performance measures ensures more rigorous assertions.

We make the code and data used publicly available[1].

3.1 Datasets

The datasets used for clustering experiments are described in Table 1, where the balance is the ratio between the smallest and largest cluster sizes. We

[1] https://github.com/miraaitsaada/anisotropy_clustering

used classic3 and classic4 datasets of Cornell University, the BBC news dataset proposed in (Greene and Cunningham, 2006) and random extracts of DBPedia (Lehmann et al., 2015) and AG-news (Zhang et al., 2015) of size 12,000 and 8,000 respectively.

	classic3	classic4	DBPedia	AG-news	BBC
Clusters	3	4	14	4	5
Balance	0.71	0.32	0.92	0.97	0.76
Samples	3 891	7 095	12 000	8 000	2 225

Table 1: Datasets' description.

In addition to the datasets used for clustering, we also make use of an external dataset in order to compute an independent score of isotropy. We make use of the dataset used by Rajaee and Pilehvar (2022) which contains sentences extracted from Wikipedia. We use this dataset to evaluate the isotropy measures like $\mathcal{I}_{cos}$, computed between $m = 5\,000$ pairs of words and sentences. The $10\,000$ resulting representations are also used to compute $\mathcal{I}_{pf}$ and $\mathcal{I}_{pf_2}$.

3.2 Post-processing

In this study, we focus on post-processing operations based on dimension reduction, showing their effectiveness on text clustering and assessing their impact on isotropy. The objective here is to compute a reduced version of $\mathbf{X}_{(n\times d)}$ called $\mathbf{Y}_{(n\times d')}$ that comprises the most useful information present in $\mathbf{X}$ while using only d' dimensions.

As an alternative to removing the dominant principal components (PCs) (Raunak et al., 2019; Mu and Viswanath, 2018), the whitening operation allows to normalize the PCs to unit variance, thus reducing the impact of the first components and producing vectors of better quality. It consists in building a reduced representation $\mathbf{Y}$ whereby each value is computed as:

$$y_{ij} = \mathbf{x}_i \mathbf{w}_j / \sqrt{\delta_j},\ \forall i = 1, \ldots, n;\ j = 1, \ldots, d'$$

where $\mathbf{w}_j$ is the jth eigen vector of $\mathbf{X}^\top \mathbf{X}$ and δ_j its jth eigen value. We also compare the classical and whitened version of PCA with a nonlinear dimension technique called UMAP (McInnes et al., 2018), a faster and more robust manifold technique than t-SNE (van der Maaten and Hinton, 2008) that can be used as a post-processing tool with any d' (while $d' \leq 3$ for t-SNE). UMAP, like t-SNE, is a graph-based method that aims at producing a reduced space that best preserves the (local) connections of a KNN graph. In order to respect the unsupervised context of text clustering, we avoid all kinds of hyperparameter tuning. We thus set d' to 10 for all of the post-processing methods.

Besides, two strategies are used to leverage Transformer models. The first consists simply in taking the last layer as usually performed in the literature (Reimers and Gurevych, 2019). The second strategy used all of the layers by averaging them together (Ait-Saada et al., 2021).

3.3 Euclidean vs. cosine

As a recall, anisotropic vector directions occupy a narrow cone in the geometrical space. Given this definition, we can expect directional techniques based on the angles between vectors to be particularly sensitive to the alleged lack of expressiveness induced by anisotropy. With this in mind, we use, in addition to k-means (MacQueen et al., 1967), Spherical k-means (Dhillon and Modha, 2001) which is made for directional data and based on the cosine distance instead of the L_2 metric. For both algorithms, we use 10 different initializations and keep the partition that yields the best within-cluster inertia. For more details about the datasets used, please refer to Appendix 3.1.

3.4 Correlation estimation

In order to assess the linear correlation between two continuous variables, we use the Pearson correlation coefficient ρ (Pearson, 1896) and test its significance. The ρ coefficient between two random variables X and Y indicates how much does one of the variables increase with the growth of the other. It is computed as:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sqrt{\sigma_X \sigma_Y}}$$

where X et Y are two random variables of variance σ_X et σ_Y respectively and $\mathrm{cov}(X, Y)$ is the covariance between X and Y.

In order to test the significance of ρ we rely on the p-value which is a probability that denotes how likely it is that the observed variables have occurred under the null hypothesis which is that the two variables are perfectly correlated ($\rho_{X,Y} = 0$). Thus, high $\rho_{X,Y}$ values indicate a stronger linear relationship and the closer the p-value gets to zero, the more we consider significant the correlation between X and Y.

		NMI		ARI		Dataset			External (word)			External (sentence)		
		km	skm	km	skm	cos	pf	pf_2	cos	pf	pf_2	cos	pf	pf_2
NMI	km	1.0	0.83	0.94	0.74	-0.05	0.02	-0.05	-0.14	0.02	0.02	-0.07	0.01	0.01
	skm	0.83	1.0	0.74	0.92	-0.09	-0.05	-0.15	-0.05	-0.05	-0.05	-0.08	-0.06	-0.05
ARI	km	0.94	0.74	1.0	0.78	-0.07	0.01	-0.07	-0.12	0.01	0.01	-0.07	-0.0	0.01
	skm	0.74	0.92	0.78	1.0	-0.01	0.04	-0.07	0.07	0.04	0.04	0.02	0.04	0.04
Dataset	cos	-0.05	-0.09	-0.07	-0.01	1.0	0.96	0.9	0.84	0.95	0.95	0.98	0.96	0.95
	pf	0.02	-0.05	0.01	0.04	0.96	1.0	0.95	0.76	1.0	1.0	0.94	1.0	1.0
	pf_2	-0.05	-0.15	-0.07	-0.07	0.9	0.95	1.0	0.73	0.95	0.95	0.89	0.95	0.95
Ext-w	cos	-0.14	-0.05	-0.12	0.07	0.84	0.76	0.73	1.0	0.76	0.76	0.89	0.78	0.75
	pf	0.02	-0.05	0.01	0.04	0.95	1.0	0.95	0.76	1.0	1.0	0.94	1.0	1.0
	pf_2	0.02	-0.05	0.01	0.04	0.95	1.0	0.95	0.76	1.0	1.0	0.94	1.0	1.0
Ext-s	cos	-0.07	-0.08	-0.07	0.02	0.98	0.94	0.89	0.89	0.94	0.94	1.0	0.96	0.93
	pf	0.01	-0.06	-0.0	0.04	0.96	1.0	0.95	0.78	1.0	1.0	0.96	1.0	1.0
	pf_2	0.01	-0.05	0.01	0.04	0.95	1.0	0.95	0.75	1.0	1.0	0.93	1.0	1.0

Table 2: Pearson correlation coefficient values[2]. "Dataset" means that the isotropy is evaluated within the same dataset on which NMI and ARI are computed. "External" means that the isotropy is evaluated using an external dataset either at the "word" or "sentence" level. Values go from the smallest (red) to the largest (green).

4 Discussion

Figure 2 confronts one quality measure (NMI) and one isotropy measure ($\mathcal{I}_{cos}$) using different post-processing techniques. We first observe that PCA$_w$ produces, by far, the most isotropic representations while increasing the performance of the raw vectors. Indeed, an appealing explanation of the success of the whitening operation is that it considerably alleviates the anisotropy of the embedding space (Su et al., 2021). Applying that reasoning, PCA and UMAP should deteriorate the performance since they both exacerbate the anisotropy (in all cases for PCA and in most cases for UMAP). Nonetheless, the performance of PCA is comparable to that of the raw embeddings and UMAP achieves even better performance than PCA$_w$ even though it significantly reduces the isotropy. Overall, averaging the whole set of layer representations achieves better results, even though it clearly decreases the isotropy, compared to using the last layer, as traditionally performed. Also, it is worth noting that even when the *directions* of the vectors are used (skm), the decrease of isotropy has a negligible impact on the performance. All these observations suggest that, although the anisotropy reduces the spectrum of directions taken by sentence vectors, it does not necessarily alter their expressiveness.

In order to confirm this supposition, we directly compare isotropy and quality measures in a wide range of situations. To this end, we compute the correlation (Table 2) between several isotropy measures and performance scores on 2 models (BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019)) with 2 different strategies ("all layers" and "last"), using 5 datasets and 4 transformations, leading to a total of 80 occurrences of each measure. We first observe a high correlation (associated with a near-zero p-value in Table 3) between measures within the same family (e.g. $\mathcal{I}_{cos}$ and $\mathcal{I}_{pf}$). This indicates that the selected measures agree with each other which denotes a certain coherence. However, when looking at the correlation between the two families of measures, it is clear that there is no significant relationship between isotropy and quality measures, since all the values of the correlation coefficient are close to zero, which is corroborated by relatively high p-values, denoting a non-significant correlation. Note that the same observations (not shown in this paper) can be made using the Spearman correlations of ranks (Spearman, 1987).

5 Conclusion

It has been known to happen that transformations that tend to decrease the anisotropy of text representations also improve the performance of downstream tasks. In stark contrast, we observe in the present study that transformations that exacerbate the anisotropy phenomenon may also improve the results, which calls into question the importance of isotropy in text representation. To draw this important conclusion, we relied on the clustering task and several empirical measures to assess the relationship between isotropy and quality of representations, using several datasets. Most importantly, we show that even a directional approach for clustering, which should be primarily affected by anisotropy, does not undergo any performance loss resulting from low-isotropy representations. In addition, we show the advantage of using UMAP as a post-processing step, which provides good-quality representations using only a handful of dimensions, despite a high resulting anisotropy.

[2]Corresponding p-values are given in Table 3 in the Appendix

6 Limitations

In this study, we focused on the clustering task in order to assess the real impact of anisotropy on the quality of representations. The conclusion is clear regarding Euclidean and directional clustering but investigating other tasks like information retrieval and anomaly detection would further strengthen the present findings. Also, the set of post-processing methods is not limited to the ones used in this study, and it would be interesting to conduct a more comprehensive study, including more transformation functions. Finally, an important future direction is to assess the impact of anisotropy on other languages, especially on embedding models trained on a restrained corpus, which can be the case of low-resource languages.

References

Mira Ait-Saada and Mohamed Nadif. 2023. Unsupervised anomaly detection in multi-topic short-text corpora. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1384–1395, Dubrovnik, Croatia. Association for Computational Linguistics.

Mira Ait-Saada, François Role, and Mohamed Nadif. 2021. How to leverage a multi-layered Transformer language model for text clustering: An ensemble approach. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 2837–2841, New York, NY, USA. Association for Computing Machinery.

Mira Ait Saada, François Role, and Mohamed Nadif. 2021. Unsupervised methods for the study of Transformer embeddings. In *Advances in Intelligent Data Analysis XIX*, pages 287–300, Cham. Springer International Publishing.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

Rafika Boutalbi, Mira Ait-Saada, Anastasiia Iurshina, Steffen Staab, and Mohamed Nadif. 2022. Tensor-based graph modularity for text data clustering. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2227–2231, New York, NY, USA. Association for Computing Machinery.

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.

Inderjit S Dhillon and Dharmendra S Modha. 2001. Concept decompositions for large sparse text data using clustering. *Machine learning*, 42(1):143–175.

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.

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*.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 377–384, New York, NY, USA. Association for Computing Machinery.

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.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. 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 4364–4373, Hong Kong, China. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Karl Pearson. 1896. VII. mathematical contributions to the theory of evolution.–III. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. 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.

Vikas Raunak, Vivek Gupta, and Florian Metze. 2019. Effective dimensionality reduction for word embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 235–243, Florence, Italy. 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 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Charles Spearman. 1987. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Douglas Steinley. 2004. Properties of the Hubert-Arable Adjusted Rand Index. *Psychological methods*, 9(3):386.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):583–617.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

A Appendix

		NMI		ARI		Dataset			External (word)			External (sentence)		
		km	skm	km	skm	cos	pf	pf_2	cos	pf	pf_2	cos	pf	pf_2
NMI	km	0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.647	0.858	0.644	0.225	0.88	0.86	0.54	0.964	0.896
	skm	≈ 0.0	0.0	≈ 0.0	≈ 0.0	0.453	0.662	0.195	0.642	0.64	0.658	0.476	0.616	0.635
ARI	km	≈ 0.0	≈ 0.0	0.0	≈ 0.0	0.562	0.921	0.542	0.297	0.946	0.925	0.542	0.98	0.96
	skm	≈ 0.0	≈ 0.0	≈ 0.0	0.0	0.934	0.71	0.517	0.56	0.742	0.722	0.877	0.755	0.741
Dataset	cos	0.647	0.453	0.562	0.934	0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0
	pf	0.858	0.662	0.921	0.71	≈ 0.0	0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0
	pf_2	0.644	0.195	0.542	0.517	≈ 0.0	≈ 0.0	0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0
Ext-w	cos	0.225	0.642	0.297	0.56	≈ 0.0	≈ 0.0	≈ 0.0	0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0
	pf	0.88	0.64	0.946	0.742	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0
	pf_2	0.86	0.658	0.925	0.722	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.0	≈ 0.0	≈ 0.0	≈ 0.0
Ext-s	cos	0.54	0.476	0.542	0.877	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.0	≈ 0.0	≈ 0.0
	pf	0.964	0.616	0.98	0.755	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.0	≈ 0.0
	pf_2	0.896	0.635	0.96	0.741	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	≈ 0.0	0.0

Table 3: p-values of the pearson test of correlation between several isotropy and performance measures. The corresponding correlation coefficients are given in Table 2. Values under 10^{-3} are considered near-zero.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☒ A2. Did you discuss any potential risks of your work?
We did not identify any risk regarding our work.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ 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.
No response.

C ☑ Did you run computational experiments?

3

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

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?
Not applicable. Left blank.

☑ 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?
Left blank.

☐ 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.)?
Not applicable. Left blank.

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.