Investigating Language Model Cross-lingual Transfer for NLP Regression Tasks Through Contrastive Learning With LLM Augmentations

Stanford CS224N Custom Project

Raghav Ganesh
Department of Computer Science
Stanford University
graghav@stanford.edu

Raj Palleti
Department of Computer Science
Stanford University
rpalleti@stanford.edu

Abstract

Many language models are not as performant when evaluated on languages other than English, due to limited high-quality data. In this work, we experiment with contrastive learning, a technique used for pretraining to generate learned data representations that improve performance on downstream NLP tasks in a low-data regime, Tamil movie reviews. Specifically, we employ LLaMa 3 to create augmentations through sentence paraphrasings and use Tamil-LLaMa to generate English to Tamil translations. This produces a large unlabeled dataset in Tamil on which we can apply contrastive learning, specifically training on a SimCSE objective. We first train BERT on this contrastive learning objective, and then fine-tune it for our regression objective, which is rating movie reviews. We see that incorporating LLM paraphrasings and translations through contrastive learning increases model performance on our downstream Tamil task but decreases performance on our downstream English task. In the Tamil case, our results show a strong potential in using contrastive learning paired with our LLM driven augmentations and translations to improve downstream NLP regression task performance in the context of low-resource languages. We believe that our English downstream task performs worse with contrastive learning since the large size of the IMDB dataset is better suited for directly training on the final objective. Lexical differences between the corpus used for contrastive learning and IMDB may also worsen learned feature representations through contrastive learning.

1 Key Information to include

- Mentor: Soumya
- Team Contributions: Raghav was primarily responsible for implementing SimCSE as well as using Tamil-LlaMa to obtain translations. Raj took the lead for writing the report as well as using LlaMa 3 to obtain data augmentations by prompting Llama to paraphrase data.
- External Collaborators (if you have any): None
- Sharing project: None

2 Introduction

Labeling data in NLP is often time-consuming and expensive (Whitehouse et al., 2023). Self-supervised learning, which involves learning representations from an unlabeled dataset, can often help when the size of the labeled dataset is small in comparison to the size of the unlabeled dataset (Chen et al., 2020b). One approach to self-supervised learning is contrastive learning. Unlike traditional pretraining techniques, contrastive learning focuses on learning strong representations by maximizing similarity between semantically similar data and minimizing similarity between semantically farther data (Chen et al., 2020a). This technique has proven to be effective in learning discriminative representations to increase performance on downstream supervised tasks (Gao et al., 2022; Abaskohi et al., 2023). In this work, we focus on contrastive learning for cross-lingual transfer for low-data regimes, particularly languages with less data available. We aim to show that contrastive learning helps for these non-English languages when data is limited and also propose a novel framework using LLMs for data augmentations in these low-data regimes.

3 Related Work

Contrastive learning has been highly successful in computer vision, in which unlabeled images are used to improve performance on a downstream image classification task (Chen et al., 2020a). The SIMCLR approach proposed by Chen et. al introduced a contrastive learning framework, in which every unlabeled image is augmented twice and fed into a pretrained encoder. As shown in Figure 1, the pretrained encoder is taught to distinguish between augmentations from the same image versus different images. This process improves its representations, which it can leverage for a downstream task. This paper inspired the work of this project. In NLP, contrastive learning has been less explored but has still shown promising potential as an effective approach for self-supervised learning in low-data regimes (Gao et al., 2022). We begin by discussing current approaches for contrastive learning in NLP and then discussing data augmentation as a method cross-lingual transfer. Our project merges both approaches by introducing a contrastive learning pipeline for cross-lingual transfer.

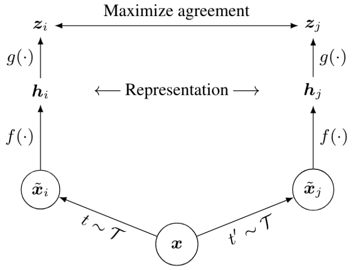


Figure 1: SimCLR framework for contrastive learning. Two data augmentations are generated for each unlabeled example (Chen et al., 2020a)

.

3.1 Contrastive Learning in NLP

Developing an algorithm to learn strong sentence embeddings has long been researched in NLP (Kiros et al., 2015; Reimers and Gurevych, 2019). While many existing frameworks often focus on masking out a part of a sentence and training an encoder to reconstruct the original sentence (Yu and Jiang, 2016), contrastive learning has recently been shown to improve sentence embeddings (Kim et al., 2021). SimCSE is a contrastive learning framework for NLP, in which random dropout is applied several times on an unlabeled dataset to produce positive pairs Gao et al. (2022). During pretraining, the encoder in SimCSE is trained to distinguish between postive and negative pairs, and this has been shown to improve performance on downstream NLP classification tasks.

3.2 Data Augmentation for Cross-Lingual Transfer

The performance of NLP models is often contingent upon the abundance of high-quality data. In multilingual NLP, this is not always available since many non-English languages lack abundant high-quality data (Joshi et al., 2021). One method to solve this is multitask training, which involves training a model on a wide range of tasks and languages in order to increase its performance and generalization across languages (Artetxe and Schwenk, 2019). For specific tasks, however, large, general models underperform relative to models trained or fine-tuned on specific tasks. (Lauscher et al., 2020). Recently, data augmentation has been explored as an alternative approach to increasing performance on multilingual NLP tasks. Paraphrasing data through LLMs such as GPT-4 has been shown to be an effective data augmentation approach that yields high-quality text (Whitehouse et al., 2023). This additional synthetic input data can then be used to train multilingual LLMs such as mBERT, which has been shown to increase accuracy on downstream tasks (Whitehouse et al., 2023). However, multilingual NLP is still a relatively recent area of research, and there is not much work involving contrastive learning for multilingual NLP. We aim to integrate contrastive learning through SimCSE as well as data augmentations through LLM-produced paraphrasings and translations to improve performance of NLP models on downstream tasks.

4 Approach

Our end-to-end pipeline is shown in Figure 1. As shown, we first choose a large unlabeled English dataset. We then use LLaMA 3 to create a paraphrasing of each datapoint to serve as data augmentations for contrastive learning. We then use Tamil LLaMA (Balachandran, 2023) to translate this dataset to Tamil. Next, we apply contrastive learning (SimSCE), and we use BERT as our encoder (Devlin et al., 2019). Finally, we finetune our pretrained BERT model on the downstream Tamil regression task and evaluate the finetuned model on our test set. To the best of our knowledge, we are the first to apply contrastive learning for cross-lingual transfer using translations of LLM paraphrasings as our data augmentations.

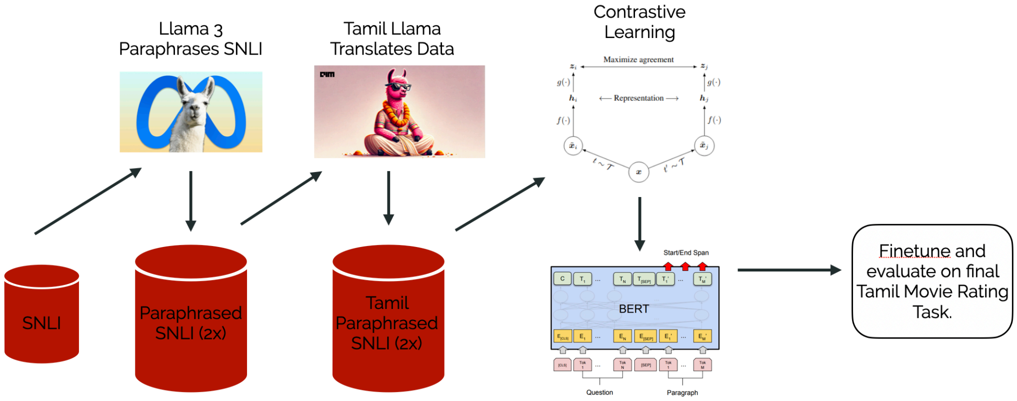
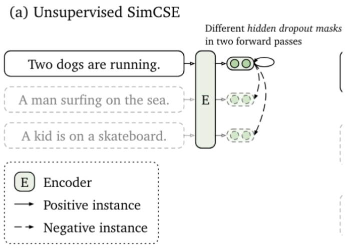


Figure 2: End to end pipeline showing (1) Llama paraphrasings of SNLI, (2) Tamil LLama translations to Tamil, (3) Contrastive learning on the LLama translations (Chen et al., 2020a), and (4) Finetuning and evaluation on the final downstream Tamil Movie Rating task.

4.1 Motivation and Approach behind Contrastive Learning

Our goals with contrastive learning are to allow the model to first learn text representations off of a large unlabeled dataset that then allow the model to perform better when fine-tuned for our downstream tasks. In other words, our hypothesis here is that using the BERT model's weights after training for a contrastive learning objective as a starting point for training, will then lead to better performance on a downstream regression task after fine-tuning, when opposed to using random initial weights or the default pre-trained BERT


(a) Unsupervised SimCSE
Different *hidden dropout masks* in two forward passes
Two dogs are running.
A man surfing on the sea.
A kid is on a skateboard.
E
E Encoder
Positive instance
Negative instance

$$\ell_i = -\log \frac{e^{\mathrm{sim}(\mathbf{h}_i^{z_i},\mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{h}_i^{z_i},\mathbf{h}_j^{z_j'})/\tau}}$$

Figure 3: SimCSE framework and loss function. Two different random dropout masks are applied to every training example. (Gao et al., 2022)

weights and then finetuning for the downstream task. Both the SimCLR and SimCSE papers demonstrated increases in performance on downstream tasks through contrastive learning, which inspired our own approach. We use the SimCSE framework to apply contrastive learning on our unlabeled text datasets. As shown in Figure 3, SimCSE applies two separate random dropout masks z_i and z_i' to every input example x_i and then feeds these into an encoder h. This yields positive pairs, $h_i^{z_i}$ and $h_i^{z_i'}$, as well as negative pairs, which are augmentations created by random dropout applied twice on different input sentences. The SimCSE objective involves distinguishing the positive example among all negative pairs. As shown in the loss function in Figure 3, for any given input example, the model computes the cosine similarity between the positive pair $h_i^z{}_i$ and $h_i^{z_i'}$. The final loss l_i for x_i is computed as the as negative of the cosine similarity between the positive examples divided by the sum of cosine similarity between all negative pairs involving x_i.

4.2 Motivation and Approach behind LLM Paraphrasing

A key ingredient in using contrastive learning is with having a large data corpus. With low-resource languages however, there are often no large human generated datasets available to use for optimizing towards a contrastive learning objective. This is what has motivated our idea of using LLMs to generate paraphrasings of existing sentences to increase the size of a text corpus through as opposed to entirely novel content creation, a task we believe should be more difficult for LLMs to do.

4.3 Motivation and Approach behind LLM Translations

Additionally, we do want to leverage large existing human generated text corpora in higher-resource languages as a resource for training towards our contrastive learning objective in our lower-resource languages. Our approach for this, is to use Tamil-LLaMa, a variant of LLaMa 2 fine-tuned on large translated Tamil corpora, as our tool to translate a large English language corpus to Tamil. We will then use this LLM translated corpus for contrastive learning, while also using a large Tamil human-generated text corpus (Tamil Murasu news articles) to also optimize for the same contrastive learning objective, as a comparsion.

5 Experiments

Our project includes seven major experiments which we have conducted for two major tasks. The first was exploring the use of contrastive learning on English language tasks, and then exploring the use of contrastive learning + LLM paraphrasing & translation on Tamil language tasks.

5.1 Data

We choose the Stanford Natural Language Inference (SNLI) Corpus (Bowman et al., 2015) as our large unlabeled English dataset for contrastive learning and the IMDB movie review dataset with 50K reviews (Maas et al., 2011) as our dataset as our downstream English regression task. SNLI contains 570K sentence pairs amounting to a total of ∼ 1 million sentences, and IMDB contains 50K sentences, each representing a movie review that is labeled as either positive or negative, where each label occurs exactly 50% of the time. Similarly, we choose 127K excerpts from the Tamil Newspaper ”Murasu” as our unlabeled Tamil dataset for contrastive learning and a Tamil Movie Review dataset (∼ 550 web-scraped movie reviews) for our downstream task, which are both smaller than the SNLI and IMDB datasets to simulate the effects of contrastive learning in a lower-data regime.

Dataset	**Sentence**
SNLI	A person on a horse jumps over a broken down airplane.
Paraphrased SNLI	A daring rider leaps over a grounded plane, defying gravity with their trusty steed by their side.
Translated SNLI	ஒரு குதிரையில் ஒரு நபர் விமானத்தின் உடைந்த பாகங்களை கடந்து செல்கிறார்
Paraphrased + Translated SNLI	ஒரு துணிச்சலான சவாரி செய்பவர் தரையில் உள்ள விமானத்தை கடக்க ஒரு தைரியமான சாகசத்தில் ஈடுபடும்போது அவர்களின் நம்பகமான குதிரையுடன் அவர்களுக்கு அருகில் உள்ளது.

Table 1: Example sentence from each of the following datasets: SNLI (Bowman et al., 2015), Llama paraphrasings of SNLI, Tamil Llama translations of SNLI, and Tamil Llama translations of paraphrased SNLI. Raghav understands Tamil and upon random sampling confirmed that the translations and paraphrasings generally make sense, although at times the word choice is very formal and a little convoluted.

5.2 Experimental Design

5.2.1 English Language Experiments

Direct Regression on IMDB.

As IMDB movie review score classification is our downstream task for evaluation, this experiment aims to determine what baseline performance on our English downstream task we can achieve by directly training our BERT model on the final objective.

SimCSE on SNLI and Fine-tuning on IMDB Classification.

This experiment explores the performance impact training on the SimCSE objective with the SNLI corpus prior to fine-tuning on the final IMDB dataset has. Our motivation behind this experiment is that contrastive learning can potentially provide our BERT model with a ”head start” on training on the final IMDB classification task, through learnings derived from training on the SimCSE objective.

Paraphrase SNLI with LLaMa 3, SimCSE on SNLI, and Fine-tuning on IMDB Classification.

In this experiment, we aim to explore what effect on performance first paraphrasing SNLI sentences with LLaMa 3, followed by SimCSE contrastive learning has on pre-training the BERT model on the final objective of IMDB movie reviews. The motivation behind this experiment is that additional data generated by LLMs can better improve the learnings from contrastive learning, in turn leading to better performance on the downstream task.

5.2.2 Tamil Language Experiments

Direct Regression on Tamil Movie Review.

As Tamil movie review score regression is our downstream task for evaluation, this experiment aims to determine what baseline performance on our Tamil downstream task we can achieve by directly training on the final objective, analogous to the same step in the English section above.

SimCSE on Tamil Murasu and Fine-tuning on Tamil Movie Review Regression.

This experiment explores the performance impact training on the SimCSE objective with the Tamil Murasu unlabeled corpus prior to fine-tuning the BERT model on the final Tamil movie review dataset has. Once again, our motivation behind this experiment is that contrastive learning can potentially improve training on the final IMDB classification task, through learnings derived from training on the SimCSE objective.

Translate SNLI with Tamil-LLaMa, SimCSE on Tamil Murasu, and Fine-tuning on Tamil Movie Review Regression.

In this experiment, we aim to explore what effect on performance that translating SNLI sentences with Tamil-LLaMa, followed by SimCSE contrastive learning has on pre-training the BERT model on the final objective of Tamil movie reviews. Although additional data generated by LLMs can better improve the learnings from contrastive learning, in turn leading to better performance on the downstream task, oftentimes the data is simply not there for languages with less resources online. In such cases, such as with Tamil, machine translating text from another language with LLMs could somewhat approximate the human generated text that more resource-abundant languages use for training, which is also something we are exploring here.

Paraphrase SNLI with LLaMA 3, Translate SNLI to Tamil, SimCSE on Tamil Murasu, and Fine-tuning on Tamil Movie Review Regression.

In this experiment, we aim to explore what effect on performance first paraphrasing SNLI sentences with LLaMa 3, translating them to Tamil, followed by SimCSE contrastive learning has on pre-training the BERT model on the final objective of Tamil movie reviews. As with the English case, the motivation behind this experiment is that additional data generated by LLMs can better improve the learnings from contrastive learning, in turn leading to better performance on the downstream task.

5.3 Evaluation method

Although the IMDB review classification dataset is formulated as a binary classification problem, we are formulating it as a regression task, to keep it consistent with our Tamil review score regression task. As such, we use Mean Squared Error (MSE) as our evaluation metric since our downstream task for both English and Tamil is a regression task (rating movie reviews). The lower our MSE score, the better our model is at predicting review scores.

5.4 Experimental details

For our English language tasks, we used the pre-trained ”bert-base-uncased” model as our base, from which we trained. For our Non-English tasks we used the pre-trained ”bert-base-multilingual-cased” model as our base. We used the Adam optimizer with a learning rate of 5e-5, while the batch size varied from 8-16 for the movie rating regression tasks, to $\sim$ 500 for training on the contrastive learning objective. For the SimCSE loss, we used a temperature of 0.05. Our training lasted from 10 - 30 epochs, depending on whether there was any non-negligible increase in performance between epochs. Training was conducted on a 52 GB RAM 8 vCPU VM on GCP with a 16GB VRAM NVIDIA V100. Training generally took $\sim$ 1 hour for the regression task and $\sim$ 3 hours for the contrastive learning task. Paraphrase generation for the SNLI dataset with LLaMA 3 took $\sim$ 3 hours. Translation also took an additional $\sim 4 - 5$ hours to run with Tamil-LLaMa.

5.5 Results

As shown in Table 2, SimCSE does not actually help with the IMDB classification task. However, note that both of the MSE values are very small. It is possible that since the IMDB dataset is already quite large, contrastive learning does not really help in this high-data regime. As shown in Table 3, however, SimCSE does help with the Tamil regression task, evidenced through the lower MSE achieved through the integration of our pipeline. This is expected as the Tamil movie dataset is smaller in comparison, allowing the model was able to learn useful represention from the larger Murasu dataset, as well as even better representations from the paraphrased and augmented text corpora. We also saw that using the SNLI translations and paraphrasings for contrastive learning increased the performance of the donwstream task greater than performing contrastive learning on a human generated unlabeled corpus (Tamil Murasu) prior to fine-tuning for the downstream task.

Model	**MSE Regression on IMDB**
No contrastive learning	0.052
SimCSE on SNLI	0.089
SimCSE on Paraphrased SNLI	0.111

Table 2: Impact of Contrastive Learning on English IMDB Classification.

Model	**MSE Regression on Tamil Movie Dataset**
No contrastive learning	0.398
SimCSE on Tamil Murasu	0.392
SimCSE on Translated SNLI	0.342
SimCSE on Paraphrased & Translated SNLI	0.313

Table 3: Impact of SimCSE Contrastive Learning on Tamil Movie Regression.

6 Analysis

With regards to the English regression task, we saw that the incorporation of our contrastive learning methodology and paraphrases actually increased the MSE, indicating that our work was decreasing downstream task performance. One possible explanation for this could be that the paraphrases were overly loquacious, and this in turn shifted the contrastive learning objectives data distribution and vocabulary away from a representation benificial for the downstream task, finally manifesting in lower downstream task performance. The learned representation space from contrastive learning could be unoptimal or counterproductive for the downstream IMDB classification task. We also generally saw that the IMDB movie review task resulted in lower overall loss values when compared with the Tamil movie review task, which makes sense considering that the IMDB movie review task is binary classification, while the Tamil movie review task contains ratings from 0 to 5.

With regards to the Tamil tasks, we did see improvements with regards to using contrastive learning and LLM generated paraphrasings. In this case, it very well could have been that the dataset used for contrastive learning and the learned representation from said contrastive learning methods was beneficial for the downstream Tamil movie review regression task. Additionally, we did see that using the translated SNLI dataset and the paraphrases from the SNLI dataset led to greater performance increases for the Tamil movie review downstream task when compared with performing contrastive learning on the Tamil Murasu dataset. This could have been because the language in the Tamil movie review dataset was more similar to the word choice and vocabulary of the Tamil-LLaMa translated and LLaMa 3 generated SNLI paraphrasings. Additionally, given that BERT is primarily trained on English data, the performance increases from pretraining and incorporating a larger Tamil corpus could be greater than replicating the same procedure with an English corpus and English task, solely due to the fact that there is less existing Tamil language knowledge in a BERT model that has not been explicitly fine-tuned yet.

7 Conclusion

Over the course of this project, we explored the use of SimCSE and LLM powered paraphrasing and translation as methods of increasing downstream model performance. We implemented and evaluated these methods on two separate sets of English and Tamil tasks and found that this set of steps increased performance for the Tamil downstream task while decreasing performance for the English downstream task. Additionally, we found that the usage of the LLM translated and paraphrased SNLI dataset for contrastive learning led to better performance for the downstream task when compared with using a human-generated Tamil dataset for contrastive learning.

One limitation of our project is that we only evaluated our model on regression tasks and one language. It would be useful to explore a range of other downstream tasks and explore other low-resource languages as well. One natural extension of our work would be to compare the impact of cross-lingual contrastive learning to languages with more data and try other downsteram tasks, including classification tasks. We could explore how contrastive learning and LLM paraphrasing and translations can affect performance in more resource-deficient languages, such as indigenous languages of South America. It could also be an interesting avenue of exploration to look into additional methods of expanding a text corpus or alternate text augmentations, such as text summarization, text shuffling, synonym replacement. It could also be possible to look into using LLMs to peform several of these augmentations.

8 Ethics Statement

8.1 Biases in Pretraining Dataset

Contrastive learning by pretraining on SNLI increases the chances that biases within this unlabeled dataset are inherited in our model. The SNLI dataset was obtained by crowdsourcing data via Amazon Mechanical Turk (Bowman et al., 2015) and therefore includes biases from each of the crowdsourcing workers, which would likely be exposed in the downstream model. For instance, the model may downplay Tamil movies with people from a certain race, gender, or a cast with a majority of its members from an underrepresented minority. The choice of unlabeled dataset is therefore important. We can mitigate that by choosing datasets that are less likely to have dangerous biases, such as datasets carefully constructed through expert reviewers or those published by governments. Another approach to mitigate that is by preprocessing the unlabeled text corpus before including it in the pretraining dataset. For instance, one can manually examine the data and removing data that perpetuates biases.

8.2 Using the Model for Career Decisions

Furthermore, if the downstream model is used to make impactful decisions, then mistakes in the model's output can cause negative harm. For instance, if the model's predicted movie ratings are used to evaluate a candidate director by averaging predicted ratings for their past movies, making a poor decision can severely harm the candidate's career. In order to mitigate this, we should focus on verifiability and consistency when using the model's predictions to make decisions. There should always be a human in the loop to verify that the accuracy of a predicted score is consistent with at least a small pool of users who watched the given movie. We could also train another model and average predictions across both models to reduce variance.

8.3 Biases in the Translation Model

Furthermore, our translation model, Tamil-LLaMa, carries its own biases which it can potentially carry over when translating SNLI, and therefore affect the final downstream predictions as well. Mitigating the bias caused by the translation model is difficult because many non-English languages like Tamil have a lower abundance of high-quality translation data, so building another, less biased Tamil translation model is difficult and time-consuming. One alternative is to ask a human to verify the translations. While this may still be time-consuming, verification will require less effort than asking a human or training a separate model to produce the translations themselves.

References

Amirhossein Abaskohi, Sascha Rothe, and Yadollah Yaghoobzadeh. 2023. Lm-cppf: Paraphrasing-guided data augmentation for contrastive prompt-based few-shot finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

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.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Taeuk Kim, Kang Min Yoo, and Sang goo Lee. 2021. Self-guided contrastive learning for bert sentence representations.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. 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.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced cross-lingual performance.

Jianfei Yu and Jing Jiang. 2016. Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 236–246, Austin, Texas. Association for Computational Linguistics.