Efficient Zero-Shot Cross-lingual Inference via Retrieval

Genta Indra Winata, Lingjue Xie, Karthik Radhakrishnan,
Yifan Gao, Daniel Preoţiuc-Pietro
Bloomberg
gwinata@bloomberg.net

Abstract

Resources for building NLP applications, such as data and models, are usually only created and curated for a limited set of high resource languages. Thus, the ability to transfer knowledge to a new language is a key way in which to enable access to NLP technology for a wider population. This paper presents a framework to perform zero-shot inference in a target language by using cross-lingual retrieval from another language where limited annotated data for a comparable domain is available. Results on two large-scale multilingual datasets show that, in this setup, this framework improves over fine-tuning multilingual models or translating annotated data, and achieves results relatively close to fine-tuning the model on the target language directly. These results show that models can be transferred efficiently across languages for a given task and domain, even for languages not covered by multilingual model training approaches.

1 Introduction

Multilingual pre-trained language models (LMs) allow for sharing and transfer of knowledge across languages (Conneau and Lample, 2019; Pires et al., 2019; Wu and Dredze, 2019; Goyal et al., 2021; Lin et al., 2022; Muennighoff et al., 2023; Scao et al., 2022; Shliazhko et al., 2022). This limits the need of gathering annotated data for a specific task and/or domain and language pair to obtain good performance by bootstrapping the model using higher resource source language(s) (Siddhant et al., 2020). This is beneficial to enabling access to NLP technology across the globe, and especially in low-resource or regional languages and dialects, because collecting new datasets is costly and requires effort in finding or training annotators for a given language and task (Adelani et al., 2022a,b; Mahendra et al., 2021; Aji et al., 2022; Ebrahimi et al., 2022; Winata et al., 2023). Recent research has shown that few-shot learning abilities are able to carry over to some extent even to languages unseen in the pre-training data of the multilingual model (Scao et al., 2022; Srivastava et al., 2023; Winata et al., 2022; Yong et al., 2023).

A common approach to zero-shot cross-lingual inference involves fine-tuning a model on the source language, then applying it to the target language (Artetxe and Schwenk, 2019; Liu et al., 2019; Lauscher et al., 2020; Phang et al., 2020; Nooralahzadeh et al., 2020; Bari et al., 2021; Kanakagiri and Radhakrishnan, 2021; Nozza, 2021), with the assumption that the underlying learned representations are aligned and will transfer to the task in another language. This approach also requires a full fine-tuning for each language and domain which makes scaling across multiple languages cumbersome.

Separately, multilingual sentence representations are trained to obtain a joint representation of utterances across multiple languages and can be directly used as inputs to train classifiers that can be applied across languages (**?**). Further, fine-tuning encoder models for sentence representations, for example using the natural language inference task, shows an ability to generalize for both monolingual (Yin et al., 2019) and multilingual classification tasks (Winata et al., 2021). However, these approaches are less robust and do not perform as well as full fine-tuning on downstream tasks (Ma et al., 2021).

In this paper, we present a simple, yet effective framework for zero-shot inference in a target language via cross-lingual retrieval. Effectively, for each utterance in the target language, we use a multilingual sentence representation model to retrieve similar examples from a pool of labeled data in the source language and project their labels onto the target by combining label distributions and averaging across multiple samples. This framework is efficient for zero-shot cross-lingual inference, as it does not require any training or parameter updates,

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pages 93–104
November 1–4, 2023.

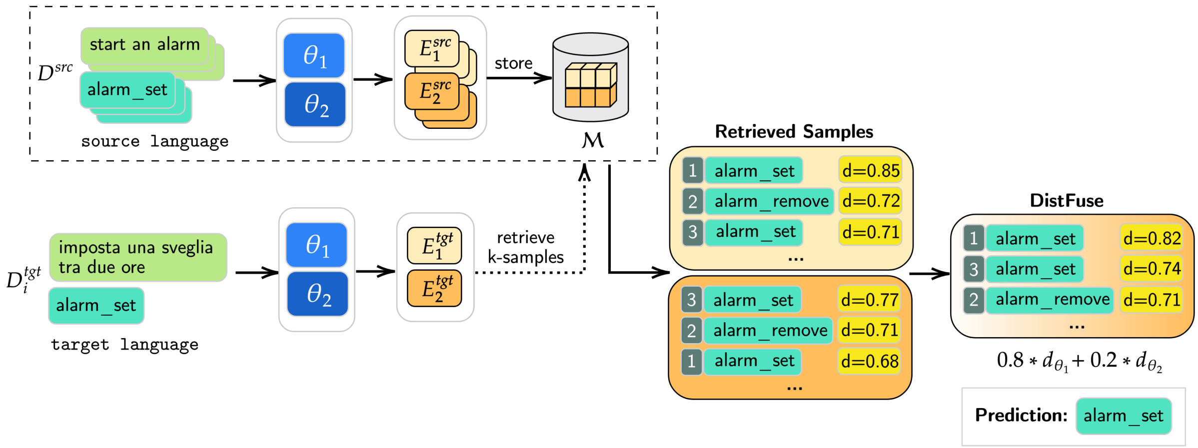

Figure 1: Inference using our proposed zero-shot framework. In this example, we use two different models θ_1 and θ_2, and the model weights are $w_1 = 0.8$ and $w_2 = 0.2$.

allowing it to scale effectively to multiple target languages. It is also lightweight, as it only requires the availability of a multilingual sentence representation model. Different from Bari et al. (2021), this framework does not require any few-shot samples or prior adaptation using a source language.

We evaluate this method for classification across two large-scale multilingual datasets, NusaX (Winata et al., 2023) and MASSIVE (FitzGerald et al., 2023), where annotated data in multiple languages from the same domain is available. Results show that this method outperforms cross-lingual fine-tuning on the source language and fine-tuning on the translated training data. Further, based on the findings in Winata et al. (2022), we evaluate the ability of this framework when the target language is not seen in pre-training of the LMs or training of the sentence representations. Results on these unseen languages show that our framework is more robust and obtains a greater relative improvement over the fine-tuning on source language data approach, albeit with a wider gap to the upper bound performance of fine-tuning with data from the target language.

Our contributions are as follows:

- We propose a lightweight and efficient inference framework for cross-lingual zero-shot text classification without any gradient updates.
- We benchmark cross-lingual zero-shot learning approaches on two large-scale multilingual datasets, and study the robustness of our framework on languages that are unseen in the training on the LMs.
- We show the effectiveness of merging output distribution from multiple models, showing the ability to capture complementary information.

2 Methods

2.1 Problem Definition

Our goal is zero-shot cross-lingual text classification, where no labeled data from the target language is seen in training, and labeled data for the same task and domain is available in a different source language.

2.2 Proposed Framework

Our framework for zero-shot inference is based on the intuition that similar documents across languages should have the same label. We use multilingual sentence representation models to find similar samples to the target language utterance. Figure 1 presents an illustration of the framework. We formalize this as follows:

Models We define θ_j as multilingual pre-trained encoder LM to which we can pass samples from the source and target languages to generate embeddings E_j^{src} and E_j^{tgt}.

Data D^{src} is the labeled dataset from the source language and D_i^{tgt} is the labeled dataset from the target language i, where each dataset has input-label pairs.

Memory We store embeddings E_j^{src} and the corresponding labels to a memory $\mathcal{M}$ that will be used as a source for retrieval.

Sample Retrieval We pass the test sample to the models θ_j to get test sample embeddings E_j^{tgt}. We then retrieve the k most similar embeddings from $\mathcal{M}$ from each model by calculating their cosine similarity $d_{\theta_j} = sim(E_j^{src}, E_j^{tgt})$.

Model	seen			unseen								
	ind	jav	sun	ace	ban	bbc	bjn	bug	mad	min	nij	**avg.**
Baselines												
Random	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33
Majority	38.25	38.25	38.25	38.25	38.25	38.25	38.25	38.25	38.25	38.25	38.25	38.25
Zero-shot XLM-R_{XNLI}	59.28	55.11	53.50	44.74	44.20	37.67	53.97	40.43	51.24	53.36	47.52	49.18
Fine-tune (src lang)	87.16	71.66	52.25	40.62	51.90	29.99	63.84	28.55	46.30	57.31	43.25	52.08
Fine-tune (translate train)	85.18	77.11	46.48	-	-	-	-	-	-	-	-	-
Fine-tune (translate test)	79.10	62.35	44.42	-	-	-	-	-	-	-	-	-
Our Zero-Shot Framework												
XLM-R_{BASE}	71.29	56.32	52.64	33.52	41.59	31.71	54.91	35.24	34.30	48.94	37.86	45.30
CMLM	73.60	72.21	74.29	64.92	68.41	55.09	72.31	51.56	63.32	69.47	61.49	66.06
LaBSE	74.10	74.50	76.08	65.52	66.76	64.38	70.99	58.55	64.11	71.80	67.09	68.54
DistFuse	78.75	78.50	78.75	65.50	70.50	65.25	75.25	58.00	67.25	73.50	70.25	**71.05**
Target Language Data (Upper Bounds)												
Fine-tune (tgt lang)†	88.40	78.90	80.10	73.90	72.80	62.30	76.60	66.60	69.70	79.10	75.00	74.85
Fine-tune (src + tgt lang)	90.50	82.60	81.33	76.90	81.41	72.47	81.41	70.19	74.62	80.54	74.77	78.79

Table 1: Results on the NusaX dataset in the zero-shot cross-lingual setting. †The results are taken from Winata et al. (2023), showing the upper bound model performance when the training data on the target language is available.

DistFuse If there is more than one model, we take the distance of the label distributions from the models θ and merge them using a linear combination: $d_{\text{FUSE}} = \sum_{j=1}^{M} w_j d_{\theta_j}$, where w_j is the weight for model θ_j.

Aggregate We aggregate the nearest k samples by taking the majority label.

Note that this framework does not involve any model training or parameter updates.

3 Experimental Setup

3.1 Datasets

We use two multilingual datasets. NusaX (Winata et al., 2023) is a multilingual sentiment analysis dataset comprising 12 languages, including 10 Indonesian regional languages. MASSIVE (FitzGerald et al., 2023) is a multilingual natural language understanding dataset with 51 languages for which we use the intent detection data.

In all our experiments, we use English as the source language for cross-lingual transfer to maintain the uniformity and tractability of experiments. Identifying the best language to transfer from is an orthogonal direction of exploration (Lin et al., 2019; Eronen et al., 2023) we consider beyond our scope and thus leave it for future work.

3.2 Models

Our framework uses XLM-R_{BASE} (Conneau et al., 2020) as the base LM, and LaBSE (Feng et al., 2022) and CMLM (Yang et al., 2021) as the multilingual sentence representation models. The parameter count for the models are: XLM-R_{BASE} – 270M parameters, LaBSE and CMLM – 471M parameters, and M2M100 – 1.2B parameters. We define *seen* languages those included in pre-training or training of the models; otherwise, we classify languages as *unseen*. For the translation methods, we use the M2M100 `1.2B` model (Fan et al., 2021) to obtain the translated text. We pick this over commercial systems as it is a high performing system that is both open-source and transparent, which makes our results easily reproducible and can help isolate the effects of training data and languages covered by this model.

3.3 Baselines

We use the following models as baselines for comparison:

- **Random:** Assigns each sample with a random label uniformly chosen from possible labels.
- **Majority:** Assigns each sample with the majority label from the training set.
- **Zero-shot:** Zero-shot prediction using an existing cross-lingual fine-tuned model on XNLI data (Conneau et al., 2018).[1]
- **Fine-tune (src lang):** The base LM fine-tuned on data from the source language only.
- **Fine-tune (translate train):** The base LM fine-tuned on the training set translated from the source language to the target language.
- **Fine-tune (translate test):** The base LM fine-tuned with the training set and evaluated with the

[1]The model can be accessed at `https://huggingface.co/joeddav/xlm-roberta-large-xnli`.

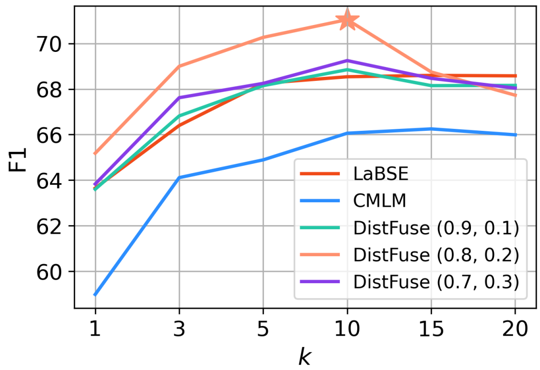

Figure 2: Performance with different k and DistFuse weights on NusaX dataset. The star marker $\star$ shows the optimal performance.

test data translated from the target to the source language.

In addition, for comparison purposes, we include the two following methods which use data from the target language and that should be considered an upper bound for classification performance, as these are trained :

- **Fine-tune (tgt):** The base LM fine-tuned with the target language data.
- **Fine-tune (src + tgt):** The base LM fine-tuned with data from both the source and target languages.

3.4 Hyper-parameters

We run the fine-tuning baselines with five seeds and report the average F1 scores for NusaX and average Accuracy scores for MASSIVE. We train for a maximum of 20 epochs with a batch size of 32 on a V100 32GB GPU. We do early stopping after three consecutive epochs without performance improvement, and use a learning rate of 1e-5 for NusaX and 5e-5 for MASSIVE. We explore different retrieval samples size $k \in \{1, 3, 5, 10, 15, 20\}$ and DistFuse weights $(0.9, 0.1), (0.8, 0.2), (0.7, 0.3)(0.6, 0.4), (0.5, 0.5)$. We report results with the best parameters $k = 10$ and DistFuse weights of $w_1 = 0.8$ (LaBSE) and $w_2 = 0.2$ (CMLM).

4 Results and Discussion

Tables 1 and 2 show results on the NusaX and MASSIVE datasets, respectively. We observe that the proposed zero-shot framework (`DistFuse`) significantly outperforms the Fine-tune (src lang) by $\sim$19% F1 on NusaX. Our proposed model achieves similar performance as Fine-tune (src lang) on MASSIVE with a minor improvement. Moreover, the XLM-R$_{\text{XNLI}}$ results are also lagging behind the sentence transformers models, LaBSE and CMLM, as the model does not use any labeled data from the same domain. We also see that LaBSE obtains better results than CMLM on both datasets. We hypothesise this is because LaBSE is optimized for bitext mining.

We also calculate the average performance of each model from Table 1 on seen and unseen languages on the NusaX dataset and summarize the results in Table 7. The breakdown performance analysis again suggests our proposed methods are effective on both seen and unseen languages, and often surpass the baselines by a large margin.

Fine-grained results for each language on the MASSIVE dataset are available in Table 5 and Table 6.

4.1 Generalization to Unseen Languages

As shown in Table 1, `DistFuse` is able to handle unseen languages significantly better than `Fine-tune (src lang)` baseline on the NusaX dataset ($\sim$21% F1 on the average of nine languages), showing the strong generalization ability on languages that are not supported by the encoder LMs. This is feasible because the unseen languages share subword tokens with the LM vocabulary (Winata et al., 2022).

4.2 Retrieved Samples

Figure 2 shows the zero-shot cross-lingual performance when varying k. LaBSE's performance increases with larger k but tapers off after $k = 10$, while CMLM's performance drops after $k = 15$. Thus, we fix $k = 10$ in all our experiments for optimal results.

4.3 DistFuse Weights

The optimal fusion is obtained when $k = 10$ with the weight proportion $w_1 = 0.8$ (LaBSE) and $w_2 = 0.2$ (CMLM), showing the need to give a higher weight on a more robust model LaBSE when combining the two distributions. In general, combining LaBSE and CMLM by fusing distributions is shown to boosts performance (+2.51 on NusaX, +0.58 on MASSIVE), showing the two methods can capture complementary information. An analysis of performance on the validation set for different DistFuse weights is presented in Table 8.

4.4 Qualitative Analysis

Table 3 shows the top 10 retrieved sentences with the LaBSE (top) and XLM-R (bottom) models. We

Model	low	mid	high	avg. all langs
Baselines				
Random	1.67	1.67	1.67	1.67
Majority	7.03	7.03	7.03	7.03
Zero-shot XLM-R$_{\text{XNLI}}$	29.26	33.74	34.98	32.53
Fine-tune (src lang)	63.31	75.95	69.43	70.96
Fine-tune (translate train)#	39.49	62.59	53.09	57.22
Fine-tune (translate test)#	51.72	73.81	59.87	68.73
Our Zero-Shot Framework				
XLM-R$_{\text{BASE}}$	25.10	13.25	27.27	23.62
CMLM	66.83	70.81	71.05	69.60
LaBSE	68.73	71.84	72.01	70.89
DistFuse	69.31	72.46	72.48	**71.47**
Target Language Data (Upper Bounds)				
Fine-tune (tgt lang)	76.09	81.85	71.88	78.48
Fine-tune (src + tgt lang)	75.03	80.91	69.20	77.23

Table 2: Results on the MASSIVE dataset in the zero-shot cross-lingual setting. The languages are grouped into three vitality classes based on Joshi et al. (2020): 1-2→low, 3-4→mid, 5→high. The full mapping is in Table 4 from the Appendix. #The M2M100 model does not support `te-IN` and `zh-TW`.

Input: Aku nembe ngerti ketemu **kancaku** sek makarya nang **trans tv** **Translation:** I just found out that I met **my friend** while working at **trans tv**	**Label:** neutral	
Retrieved sentence (LaBSE)	**Score**	**Label**
My dad is an employee in **net.tv**	0.5346	neutral
Now I know I've hated that foreign online shop too much	0.4282	negative
My friend works at Gojek	0.4222	neutral
I heard they'll build a **Transmart** there, next to that building	0.4062	neutral
I've been dreaming of travelling abroad for a long time	0.3903	neutral
My friend applies for a position in Tokopedia	0.3762	neutral
Lots of **my friends** also work in Bukalapak	0.3748	neutral
Lots of my family have worked as civil servants.	0.3441	neutral
Last week there was some 4G network in my village for a while.	0.3417	neutral
So bored. I've watched all the films and now I'm drawing a blank	0.3253	negative
Prediction: neutral (k=1); neutral (k=10)		
Retrieved sentence (XLM-R)	**Score**	**Label**
Poor Ungu personnels, can't find a gig after Pasha left	0.9957	negative
Win cool prizes by entering the "Baik untuk Men" photo contest in Alfamart	0.9956	neutral
Pos Indonesia's services are so pathetic nowadays.	0.9953	negative
This person do be blockin' the road like no tomorrow.	0.9951	negative
Rode on the Jayabaya train from Malang to Jakarta, stopped in Gubeng, the ticket costed 35 thousand plus 6k insurance via Traveloka.	0.9949	neutral
How much is the minimal if I may ask, I wanna buy Tiket Kami for Senen - Yogyakarta using the May promo	0.9949	neutral
The PIK Waterboom Jakarta tickets are rising in price.	0.9948	neutral
The employees at Graha Indosat is so rude	0.9946	negative
The denizens found 2.910 KTP-el cards in the bushes.	0.9946	neutral
I wanna help by giving the info connections, but my internet quota is limited	0.9945	neutral
Prediction: negative (k=1); neutral (k=10)		

Table 3: Retrieved English sentences from the NusaX example with LaBSE (top) and XLM-R (bottom). The input is a Javanese sample from the test set.

can observe that 8 out of 10 sentences retrieved using LaBSE have the correct label for the input. Moreover, this sentence transformer model can retrieve English sentences even though the input is in Javanese. It also captures the semantics of the entities in Javanese (e.g., television agencies trans.tv and net.tv) and identifies similar keywords in English, such as *kancaku*, which is the literal translation of *my friend*. The model also retrieves the sentence with an entity *Transmart*, which is an organization that is associated with *trans tv*. This presents the ability of the LaBSE model to not only search for the same keywords during retrieval but also semantically related keywords in another language. However, the XLM-R model performs much worse compared to LaBSE. The retrieved sentences do not have overlapping entities; they generally have different semantic contexts. It shows that the XLM-R representations are not suitable for bitext mining without any additional fine-tuning for cross-lingual alignment. Nevertheless, by taking the majority voting over the 10 sentences, we are still able to predict the correct label.

5 Conclusion

We introduce a simple but effective framework to utilize sentence representation models for text classification without requiring parameter updates. We experiment on two large-scale multilingual datasets and show that our framework outperforms zero-shot cross-lingual fine-tuning. This shows the feasibility of utilizing encoder LMs as zero-shot cross-lingual learners without additional gradient updates. The framework can also be dynamically scaled by updating the memory and combining the output distribution of multiple sentence representation models. Our framework can be further applied to unseen languages that have subword token overlaps with the LM vocabulary.

6 Limitations

This paper only studies text classification tasks with two multilingual datasets; we expect no unseen labels on the test sets. We only experiment using two multilingual sentence transformer models and one variant of the XLM-R model. We only use English as the source language, and we expect better results using the closest language as the source language. We leave the exploration of other models and experiment settings as future work.

7 Ethics Statement

In our experiments, we use publicly available datasets with permissive licenses for research experiments. We do not release new data or annotations as part of this work. There are no potential risks.

Acknowledgments

We would like to thank Rajarshi Bhowmik and the entire Bloomberg AI group for valuable discussions and feedback on the manuscript.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022a. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *arXiv preprint arXiv:2210.12391*.

Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

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.

M Saiful Bari, Batool Haider, and Saab Mansour. 2021. Nearest neighbour few-shot learning for cross-lingual classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1745–1753.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Tushar Kanakagiri and Karthik Radhakrishnan. 2021. Task-oriented dialog systems for dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 85–93.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Zero-shot cross-lingual dialogue systems with transferable latent variables. 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 1297–1303.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. 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 786–796.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. Indonli: A natural language inference dataset for indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. 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 907–914.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8854–8861.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 777–791.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness 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 833–844, Hong Kong, China. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Language	Language Code	Taxonomy (1-5)‡	Vitality*	Seen on Encoders	Seen on M2M100
Afrikaans	af-ZA	3	mid	✓	✓
Albanian	sq-AL	1	low	✓	✓
Amharic	am-ET	2	low	✓	✓
Arabic	ar-SA	5	high	✓	✓
Armenian	hy-AM	1	low	✓	✓
Azerbaijani	az-AZ	1	low	✓	✓
Bengali	bn-BD	3	mid	✓	✓
Burmese	my-MM	1	low	✓	✓
Danish	da-DK	3	mid	✓	✓
Dutch	nl-NL	4	mid	✓	✓
English	en-US	5	high	✓	✓
Finnish	fi-FI	4	mid	✓	✓
French	fr-FR	5	high	✓	✓
Georgian	ka-GE	3	mid	✓	✓
German	de-DE	5	high	✓	✓
Greek	el-GR	3	mid	✓	✓
Hebrew	he-IL	3	mid	✓	✓
Hindi	hi-IN	4	mid	✓	✓
Hungarian	hu-HU	4	mid	✓	✓
Icelandic	is-IS	2	low	✓	✓
Indonesian	id-ID	3	mid	✓	✓
Italian	it-IT	4	mid	✓	✓
Japanese	ja-JP	5	high	✓	✓
Javanese	jv-ID	1	low	✓	✓
Kannada	kn-IN	1	low	✓	✓
Khmer	km-KH	1	low	✓	✓
Korean	ko-KR	4	mid	✓	✓
Latvian	lv-LV	3	mid	✓	✓
Malay	ms-MY	3	mid	✓	✓
Malayalam	ml-IN	1	low	✓	✓
Mandarin (simp)	zh-CN	5	high	✓	✓
Mandarin (trad)‡	zh-TW	5	high	✓	×
Mongolian	mn-MN	1	low	✓	✓
Norwegian	nb-NO	1	low	✓	✓
Persian	fa-IR	4	mid	✓	✓
Polish	pl-PL	4	mid	✓	✓
Portuguese	pt-PT	4	mid	✓	✓
Romanian	ro-RO	3	mid	✓	✓
Russian	ru-RU	4	mid	✓	✓
Slovenian	sl-SI	3	mid	✓	✓
Spanish	es-ES	5	high	✓	✓
Swahili	sw-KE	2	low	✓	✓
Swedish	sv-SE	4	mid	✓	✓
Tagalog	tl-PH	3	mid	✓	✓
Tamil	ta-IN	3	mid	✓	✓
Telugu	te-IN	1	low	✓	×
Thai	th-TH	3	mid	✓	✓
Turkish	tr-TR	4	mid	✓	✓
Urdu	ur-PK	3	low	✓	✓
Vietnamese	vi-VN	4	mid	✓	✓
Welsh	cy-GB	1	low	✓	✓

Table 4: Language category mapping on MASSIVE. * It maps the language taxonomy class to three vitality classes: 1-2→low, 3-4→mid, 5→high. † Mandarin (trad) is considered as Mandarin. ‡ The language taxonomy is taken from Joshi et al. (2020).

Language	Language Code	Zero-shot NLI	Fine-tune			
			src lang	tgt lang	translate-test	translate-train
Afrikaans	az-AZ	70.9 ± 1.6	86.2 ±1.2	31.04	44.69	45.29
Albanian	sq-AL	67.6 ± 1.7	86.4 ±1.2	31.2	64.51	71.5
Amharic	am-ET	51.9 ± 1.8	81.7 ±1.4	24.92	50.75	20.66
Arabic	ar-SA	62.8 ± 1.7	80.7 ±1.4	30.13	62.13	45.8
Arfikaans	af-ZA	71.7 ± 1.6	85.6 ±1.3	30.87	80.51	57.3
Armenian	hy-AM	71.6 ± 1.6	84.4 ±1.3	29.99	62.06	37.66
Bengali	bn-BD	66 ± 1.7	84.1 ±1.3	32.65	50.07	68.41
Burmese	my-MM	67.6 ± 1.7	83.6 ±1.3	29.52	47.43	30.74
Danish	da-DK	83.1 ± 1.3	86.9 ±1.2	34.26	84.4	78.02
Dutch	nl-NL	82.1 ± 1.4	86.8 ±1.2	35.07	82.64	80.85
English	en-US	88.3 ± 1.2	88.3 ±1.2	36.65	N/A	N/A
Finnish	fi-FI	80.2 ± 1.4	85.5 ±1.3	36.35	82.39	50.03
French	fr-FR	80.8 ± 1.4	86.3 ±1.2	35.74	68.44	79.49
Georgian	ka-GE	61.2 ± 1.8	80.3 ±1.4	26.8	58.63	36.86
German	de-DE	77.6 ± 1.5	85.7 ±1.3	33.86	66.26	76.35
Greek	el-GR	74 ± 1.6	86.2 ±1.2	33.62	64.46	58.85
Hebrew	he-IL	73.2 ± 1.6	85.9 ±1.3	32.92	77.41	54.87
Hindi	hi-IN	74.8 ± 1.6	85.8 ±1.3	32.62	66.12	59.8
Hungarian	hu-HU	77.1 ± 1.5	86.2 ±1.2	32.65	82.38	72.6
Icelandic	is-IS	66.7 ± 1.7	85.3 ±1.3	29.62	48.68	63.11
Indonesian	id-ID	83.1 ± 1.3	87.1 ±1.2	37.53	82.09	67.41
Italian	it-IT	76.4 ± 1.5	86.6 ±1.2	33.25	83.8	75.55
Japanese	ja-JP	44.8 ± 1.8	83.9 ±1.3	37.09	81.73	73.47
Javanese	jv-ID	46.5 ± 1.8	82.9 ±1.4	25.02	47.83	40.24
Kannada	kn-IN	63.5 ± 1.7	84 ±1.3	29.93	31.99	5.62
Khmer	km-KH	61.3 ± 1.8	77.2 ±1.5	26.63	46.52	32.48
Korean	ko-KR	77 ± 1.5	86.5 ±1.2	35.27	57.17	60.67
Latvian	lv-LV	69.2 ± 1.7	86.1 ±1.2	34.26	78.51	64.11
Malay	ms-MY	76.7 ± 1.5	86.1 ±1.2	33.25	63.69	73.82
Malayalam	ml-IN	70.1 ± 1.6	85.1 ±1.3	33.32	71.46	39.59
Mandarin (simp)	zh-CN	61.9 ± 1.7	84.9 ±1.3	36.82	68.43	71.59
Mandarin (trad)	zh-TW	60.4 ± 1.8	83 ±1.3	35.07	63.99	N/A
Mongolian	mn-MN	64.4 ± 1.7	84.3 ±1.3	31.98	50.93	31.22
Norwegian	nb-NO	83.6 ± 1.3	87.3 ±1.2	35.84	85.15	80.03
Persian	fa-IR	81.1 ± 1.4	87 ±1.2	34.77	66.72	67.83
Polish	pl-PL	80.7 ± 1.4	85.8 ±1.3	35.91	83.44	61.76
Portuguese	pt-PT	79.5 ± 1.5	86.7 ±1.2	34.73	83.02	79.09
Romanian	ro-RO	80.8 ± 1.4	86.9 ±1.2	32.08	83.31	76.64
Russian	ru-RU	81.3 ± 1.4	87.2 ±1.2	34.7	83.51	62.08
Slovenian	sl-SI	69.5 ± 1.7	86.3 ±1.2	31.44	74.5	60.11
Spanish	es-ES	78.8 ± 1.5	86.9 ±1.2	34.5	67.96	78.02
Swahili	sw-KE	46.6 ± 1.8	83.1 ±1.3	22.6	63.99	41.39
Swedish	sv-SE	85.2 ± 1.3	87.9 ±1.2	34.57	84.15	65.6
Tagalog	tl-PH	63.7 ± 1.7	84.6 ±1.3	32.08	64.73	45.37
Tamil	ta-IN	68.1 ± 1.7	83.5 ±1.3	31.91	65.31	24.88
Telugu	te-IN	68.2 ± 1.7	84.5 ±1.3	31.1	N/A	N/A
Thai	th-TH	77.4 ± 1.5	84.7 ±1.3	35.61	49.54	66.06
Turkish	tr-TR	78.4 ± 1.5	86.3 ±1.2	35.54	80.59	70.46
Urdu	ur-PK	65.6 ± 1.7	83.2 ±1.3	30.77	71.64	57.54
Vietnamese	vi-VN	79.2 ± 1.5	86.3 ±1.2	36.15	79.74	50.84
Welsh	cy-GB	46.9 ± 1.8	82.6 ±1.4	24.75	39.85	34.8

Table 5: Fine-grained baseline results on MASSIVE. We label “N/A" for English and languages that are not supported by the M2M100 machine translation model.

Language	Language Code	XLM-R	CMLM	LaBSE	DistFuse
Afrikaans	az-AZ	19.22	67.79	68.76	69.29
Albanian	sq-AL	18.2	71.3	71.87	72.5
Amharic	am-ET	6.22	64.5	67.78	68.28
Arabic	ar-SA	15.52	57.96	60.21	60.34
Arfikaans	af-ZA	27.33	71.29	72.78	73.48
Armenian	hy-AM	17.68	68.94	69.48	69.38
Bengali	bn-BD	13.36	70.98	71.82	72.05
Burmese	my-MM	6.68	66.38	66.87	67.59
Danish	da-DK	33.37	71.49	73.12	73.72
Dutch	nl-NL	39.67	73.64	73.77	74.06
English	en-US	60.14	76.7	78.07	78.53
Finnish	fi-FI	32.04	70.65	73.1	72.79
French	fr-FR	34.77	73.82	74.49	74.36
Georgian	ka-GE	14.29	61.45	62.9	63.39
German	de-DE	30.62	69.3	70.35	71.23
Greek	el-GR	24.33	68.48	71.42	72.07
Hebrew	he-IL	20.55	70.9	70.61	71.28
Hindi	hi-IN	20.83	73.51	73.98	74.02
Hungarian	hu-HU	29.96	70.61	72.47	72.91
Icelandic	is-IS	14.1	67.77	68.82	69.21
Indonesian	id-ID	33.69	73.07	74.11	74.26
Italian	it-IT	30.38	70.54	72.95	74.14
Japanese	ja-JP	25.41	74.57	73.79	74.45
Javanese	jv-ID	9.59	63.42	64.39	66.05
Kannada	kn-IN	10.26	71.24	72.1	73.32
Khmer	km-KH	10.82	58.88	61.36	60.01
Korean	ko-KR	24.98	71.32	71.92	73.01
Latvian	lv-LV	19.1	70.66	71.5	71.36
Malay	ms-MY	28.03	69.23	72.04	72.37
Malayalam	ml-IN	13.16	70.89	72.86	73.47
Mandarin (simp)	zh-CN	26.65	72.54	73.2	73.89
Mandarin (trad)	zh-TW	27.83	71.01	71.12	71.5
Mongolian	mn-MN	13.22	68.19	70.58	70.73
Norwegian	nb-NO	30.22	71.11	72.24	74.15
Persian	fa-IR	29.09	72.63	73.65	73.71
Polish	pl-PL	34.42	73.77	72.92	74.42
Portuguese	pt-PT	35.82	73.23	75.45	74.94
Romanian	ro-RO	35.78	71.33	73.14	73.88
Russian	ru-RU	35.56	70.47	71.27	72.38
Slovenian	sl-SI	27.23	73.3	71.85	74.1
Spanish	es-ES	35.34	72.49	74.86	75.51
Swahili	sw-KE	10.35	63.4	65.18	66.06
Swedish	sv-SE	34.18	72.58	73.19	74.66
Tagalog	tl-PH	22.21	68.92	70.63	70.65
Tamil	ta-IN	13.69	69.25	68.85	69.16
Telugu	te-IN	10.16	70	72.77	73.75
Thai	th-TH	22.25	68.46	69.84	69.32
Turkish	tr-TR	23.06	69.98	71.79	72.87
Urdu	ur-PK	14.42	68.4	70.85	71.21
Vietnamese	vi-VN	31.03	70.22	68.67	71.38
Welsh	cy-GB	7.72	57	63.85	64.03

Table 6: Fine-grained zero-shot results on MASSIVE.

Model	seen	unseen
Baselines		
Random	33.33	33.33
Majority	38.25	38.25
Zero-shot XLM-R_{XNLI}	55.96	46.64
Fine-tune (src lang)	70.36	45.22
Fine-tune (translate train)	69.59	-
Fine-tune (translate test)	61.96	-
Our Zero-Shot Framework		
XLM-R_{BASE}	60.75	40.03
CMLM	73.36	63.32
LaBSE	74.89	66.15
DistFuse	**78.66**	**68.19**
Target Language Data (Upper Bounds)		
Fine-tune (tgt lang)†	82.47	72.00
Fine-tune (src + tgt lang)	84.81	76.54

Table 7: Average performance on the seen and unseen languages from the NusaX dataset.

Language	Ratio (LabSE and CMLM)				
	[0.9,0.1]	[0.8,0.2]	[0.7,0.3]	[0.6,0.4]	[0.5,0.5]
acehnese	67.05	61.26	61.42	64.43	66.96
balinese	66.05	65.38	64.35	62.72	64.93
banjarese	62.84	60.54	59.78	65.17	67.33
buginese	58.33	59.68	57.76	57.45	58.41
english	80.24	77.24	79.62	77.36	75.55
indonesian	76.40	77.71	76.44	75.98	77.48
javanese	75.38	76.32	71.86	72.75	72.75
madurese	58.15	59.20	60.12	58.90	59.88
minangkabau	65.67	66.86	69.70	68.83	67.2
ngaju	60.35	61.38	63.15	64.12	63.95
sundanese	76.88	76.02	79.16	77.52	78.56
toba_batak	58.8	59.78	57.05	59.65	59.73
avg.	67.18	66.78	66.70	67.07	67.73

Table 8: Average performance on the validation set from the NusaX dataset.