How many words does it take to understand a low-resource language?

Emily Chang
Department of Computer Science
University of Virginia
ec5ug@virginia.edu

Dr. Nada Basit
Department of Computer Science
University of Virginia
basit@virginia.edu

Abstract

When developing language technology, researchers have routinely turned to transfer learning to solve the data scarcity conundrum presented in low-resource languages. To our knowledge, this study is the first to evaluate the amount of documentation needed for transfer learning, specifically the smallest vocabulary size needed to create a sentence embedding space. In adopting widely spoken languages as a proxy for low-resource languages, our experiments show that the relationship between a sentence embedding's vocabulary size and performance is logarithmic with performance leveling at a vocabulary size of 25,000. It should be noted that this relationship cannot be replicated across all languages, and this level of documentation does not exist for many low-resource languages. We do observe, however, that performance accelerates at a vocabulary size of $\leq$ 1000, a quantity that is present in most low-resource language documentation. These results can aid researchers in understanding whether a low-resource language has enough documentation necessary to support the creation of a sentence embedding and language model.

1 Introduction

More than 43% of the languages spoken in the world are endangered (Zhang et al., 2022). Due to globalization and neocolonialism, language loss occurs at an accelerated rate (Zhang et al., 2022). Saving and revitalizing endangered languages has become very important for maintaining cultural diversity (Zhang et al., 2022). In times of crisis, these language technologies allow first responders to save lives. For example, the Low Resource Languages for Emergent Incidents (LORELEI) provides situational awareness based on information from any language and supports humanitarian assistance/disaster relief, peacekeeping, and infectious disease response (Strassel and Tracey, 2016).

Working with minimal data—as would be the case with endangered languages—makes it difficult to train natural language models from scratch. For these reasons, transfer learning is a potential method for language models to adapt to endangered languages (Alnajjar et al., 2023; Chen et al., 2019; Lee et al., 2021; Tran, 2020). We focus our research questions on cross-lingual transfer learning for low-resource languages to:

- **RQ 1**: What is the lower bound of documentation needed?
- **RQ 2**: When the target low-resource language is linguistically distant from the source high-resource language, does this lower bound of documentation change?

By establishing this lower-bound, we can better assess whether a low-resource language has enough documentation to support the creation of a sentence embedding space and language model.

2 Methodology

We analyze sentence embeddings as they are highly important in the creation of language models (Mao et al., 2024). In a survey of cross-lingual transfer learning learning methodologies, we found that Alnajjar et al. (2023)'s methodology to be the simplest. Alnajjar et al. (2023) draws on Finnish word embeddings to create embedding spaces and sentiment classifiers for endangered Uralic languages. The choice of Finnish as the source language is ideal as Finnish is part of the same language family as the endangered Uralic languages. We proceeded to modify the cross-lingual transfer methodology described in the paper.

When performing cross-lingual transfer learning, we select Dutch as the "high-resource" source language and English to train a Dutch sentiment classifier. To evaluate whether vocabulary size varies by proximity to the high-resource source language

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 207–224

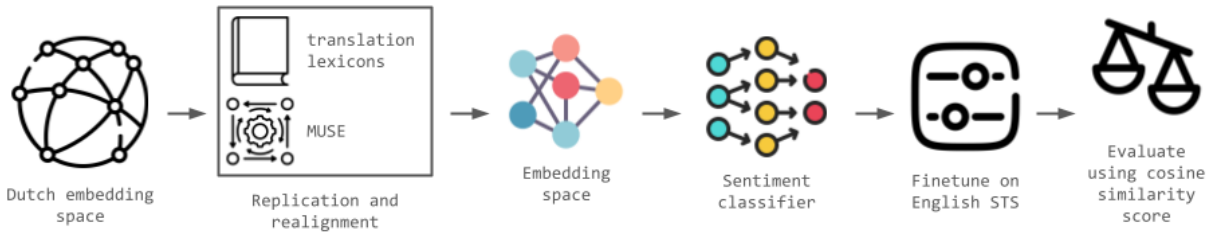


Figure 1: Overview of methodology. Using translation lexicons, the Dutch embedding space is replicated for each proxy language and aligned using MUSE. Sentiment classifiers are then built from the embedding spaces and finetuned on STS English examples. These classifiers are evaluated on their respective language in MTEB.

Dutch, we select four widely spoken languages as proxies for low-resource target languages: German, Turkish, Arabic, and Mandarin. We test Arabic and Mandarin separately to determine how replicable cross-lingual transfer is across different languages. Adopting high-resource languages as proxies allow us to experiment with varying degrees of language documentation, from the very small to the very large.

Our methodology is illustrated in Figure 1. We select a classic tokenizer that splits on whitespace and punctuation as an acknowledgment of the reality faced by many low-resource languages: a lack of data to train a more sophisticated tokenizer. With the help of translation lexicons, we replicate Dutch word embeddings for each proxy language before aligning all word embeddings. We then create sentence embeddings, each finetuned on English data as done in Alnajjar et al. (2023). We then evaluate these sentence embeddings by injecting sentence pairs into the sentence embedding space and comparing the model's cosine similarity score with the actual similarity score using the Spearman correlation (Spearman, 1904).

Language	Text
Dutch	Hij stierf dinsdag in Osaka.
German	Er verstarb am Dienstag in Osaka.
Turkish	Salı günü Osaka'da vefat etti.
Arabic	مات في أوساكا يوم الثلاثاء.
Mandarin	周二，他在大阪去世

Table 1: Languages analyzed in the study. Translations are provided for the phrase: "He died in Osaka on Tuesday" NLLB Team et al. (2024). Turkish uses a similar script similar to Dutch.

2.1 Evaluating the impact of genetic proximity using proxies

To account for genetic proximity, we adopt four high-resource languages as proxies for low-resource languages: German for its proximity to Dutch, Turkish because its typology is similar to Dutch but is in a different language family, and Arabic and Mandarin as their typologies are dissimilar to Dutch and are in a different language family (see Table 1 and Appendix A). Transfer learning is performed between two groups: (1) transfer of Dutch word embeddings to German, Turkish, and Arabic, and (2) transferring Dutch word embeddings to German, Turkish, and Mandarin, to see the relative performance of the languages most dissimilar to Dutch.

2.2 Tokenizing text

Word tokenizers facilitate the creation of organized representations of language, which is useful for language modeling (Dagan et al., 2024). The development of these tokenizers requires data (Dagan et al., 2024). For example, byte-pair encoding (BPE) tokenizers require training on text corpora to learn how to split words into frequently occurring subword units. While such tokenizers have proven successful for certain languages and have been used in state-of-the-art language models, their applicability to *low-resource* languages remains debated. Arnett and Bergen (2024) writes that differences in tokenizer performance can be attributed to disparities in dataset size. If a BPE tokenizer is exposed to limited data and does not segment words along morphological boundaries —a common occurrence in morphologically-rich languages —it may be difficult for the language model to efficiently learn the language (Arnett and Bergen, 2024). While less robust when compared to a BPE tokenizer, a classic tokenizer that splits on whitespace and punctuation is a nod to the reality of low-resource languages:

there may not exist sufficient data to train a well-performing tokenizer.

2.3 Using translation lexicons to generate word embeddings

To simulate our proxy languages under low-resource conditions, we adopt translation lexicons—dictionaries that translate from one language to another—provided by Facebook's Multilingual Unsupervised and Supervised Embeddings (MUSE) (Conneau et al., 2017) as the most common types of resource available for low-resource and endangered languages are translation lexicons and universal dependencies (Alnajjar et al., 2023). We chained together lexicons that translated from our proxy languages to English and English to Dutch. These translation lexicons allowed us to replicate the Dutch word embedding space and vocabulary as the proxy's. We forwent additional fine-tuning as performance remained unchanged (see Appendix B).

2.4 Alignment of word embeddings

We aligned the embedding spaces of English, Dutch, German, Turkish, Arabic, and Mandarin using the state-of-the-art supervised multilingual word embedding alignment technique developed in MUSE, resulting in cross-lingual word embeddings (Conneau et al., 2017). For example, the vector for "dog" in English embeddings points roughly in the same direction as the same word in all other languages. To confirm that realignment improves word translations, see subsection C.1.

2.5 Creating sentence embeddings

The procedure for creating sentence embeddings involves averaging the word embeddings of a given sentence and subsequently feeding them to two fully-connected feed-forward layers, thereby constructing a Deep Averaging Network (DAN) (Iyyer et al., 2015). The sentence embeddings are trained on the English subset of the Massive Text Embedding Benchmark (MTEB) Semantic Textual Similarity (STS) Benchmark (Muennighoff et al., 2023). While training the sentence embedding in its associated language may result in greater improvement in performance, such data may not always be present in a low-resource setting.

The resulting sentence embedding space was evaluated using its corresponding language subset in MTEB. We used the Spearman correlation score (Spearman, 1904) to compare the predicted cosine similarity scores with the actual similarity scores. In evaluating STS systems, researchers recommend using Spearman's rank correlation coefficient (Zesch, 2010). This metric assesses a monotonic relationship by ranking values (Zesch, 2010). Under the Spearman correlation, a model output does not need to match the ground truth; a model output that is *well-correlated* with the ground truth produces a high Spearman correlation, indicating that the sentence embedding can encode meaningful semantic information.

2.6 Creating a sentiment classifier

To assess the robustness of the transfer learning approach introduced by Alnajjar et al. (2023), we replicated Alnajjar et al. (2023)'s sentiment classifier for our proxy languages and compared its performance in our study to the results reported in Alnajjar et al. (2023). The model architecture is depicted in Figure 2.

To train the model, we used English samples from the Stanford Sentiment Treebank (Socher et al., 2013), Amazon Reviews Dataset (McAuley and Leskovec, 2013), and Yelp Dataset (Zhang et al., 2015), and their associated sentiment annotation (positive-negative). To evaluate the model on our target languages, XED (Öhman et al., 2020) —a multilabel sentiment classification dataset —was preprocessed into a binary classification dataset (see Appendix D).

3 Results

Under the methodology described in section 2, the quality of the translations improve as the vocabulary size of the proxy language grows (see subsection C.2). The relationship between vocabulary size and the performance of the sentence embedding is logarithmic. This is evident in the fact that the greatest increases in performance occur at smaller vocabulary sizes. Once the vocabulary size hits 25,000, we begin to see diminishing returns (see Figure 3 and Figure 4). The notable exception is Mandarin as increasing the vocabulary size consistently results in poor performance (see Figure 4). The poor performance in Mandarin can be attributed to its prediction of a constant or near-constant cosine similarity score (see Figure 11).

Interestingly, Turkish and Arabic—two of the languages that are considered linguistically different from the source language Dutch—outperformed German, the language

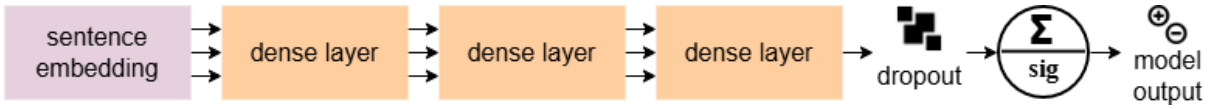


Figure 2: Architecture of sentiment classifier. To determine whether a sentence has a positive or negative connotation, the sentence is processed through a sentence embedding layer, followed by three dense layers, a dropout layer, and a sigmoid activation function.

that was deemed closest to the source language Dutch (see Figure 3). In Figure 4, this trend is replicated only in Turkish. It should be noted that the distributions of the model's predicted similarity scores do not mirror those of the actual similarity scores (see Appendix F).

Using the procedure discussed in subsection 2.6, we compare our results against Alnajjar et al. (2023) in Table 2.

Language	Label	Precision	Recall	F1-Score	Accuracy
Komi-Zyrian	neg	0.57	0.57	0.57	0.56
	pos	0.55	0.55	0.55	
Moksha	neg	0.63	0.65	0.64	0.63
	pos	0.64	0.62	0.63	
Erzya	neg	0.71	0.69	0.70	0.68
	pos	0.67	0.69	0.68	
Udmurt	neg	0.69	0.63	0.66	0.63
	pos	0.58	0.63	0.60	
German	neg	1.00	0.26	0.42	0.47
	pos	1.00	0.73	0.84	
Turkish	neg	1.00	0.46	0.63	0.50
	pos	1.00	0.56	0.72	
Arabic	neg	1.00	0.68	0.81	0.53
	pos	1.00	0.33	0.49	
Mandarin	neg	1.00	0.03	0.06	0.48
	pos	1.00	0.95	0.97	

Table 2: Proxy languages (in red) perform worse compared to the Uralic languages in Alnajjar et al. (2023) study (in black). While the sentiment classifiers in Alnajjar et al. (2023) achieve similar F1 scores for predicting both positive *and* negative labels, the sentiment classifiers for our proxy languages overfit to one of the labels. The classifiers achieve a high F1 score for predicting either positive *or* negative labels, but not both.

4 Discussion

4.1 Minimum tokens

Once a low-resource language's documented vocabulary size reaches 25,000, the performance of its sentence embedding plateaus. Without further finetuning the performance of the model will stagnate as evidenced in Figure 3 and Figure 4. While a vocabulary size of 25,000 exceeds existing documentation in low-resource translation lexicons, the **vocabulary size at which a sentence embedding space most improves ($\leq$ 1000) is accessible in most lexicons** (see Appendix G). This addresses our first research question (**RQ 1**).

4.2 Genetic proximity

Cross-lingual training between typologically-related languages has shown promising results in several NLP tasks especially in low-resource settings (Anastasopoulos and Neubig, 2019; McCarthy et al., 2019). Figure 3 and Figure 4 affirm this finding as German and Turkish—two target languages that share the typology of the source language —Dutch —benefit from cross-lingual transfer learning.

Genetic proximity appears to have little impact on the performance of a proxy language. Interestingly, German STS performance is inferior to that of Turkish's (see Figure 3 and Figure 4). This finding runs counter to Zhao et al. (2020) where researchers chose Lezgian and Tsez as target languages because they belong to the same language family as the source language. Moreover, Arabic—a language that is typologically dissimilar to the source language Dutch—performs the best out of all four languages. However, this trend is not replicated in Mandarin. As shown in Figure 5, naive whitespace tokenization alters the meaning of the sentence and may have negatively contributed to Mandarin's performance. This addresses our second research question (**RQ 2**).

5 Limitations and Future Work

5.1 Proxies

While we are interested in examining how well languages that are typologically dissimilar to the source language perform, the MTEB dataset only includes two such languages: Arabic and Mandarin. Consequently, our analysis was limited by the constraints of this evaluation dataset.

The data utilized in this study may not be fully representative of low-resource data. In reality, our proxy languages are high-resource languages and their associated datasets may contain a wider range

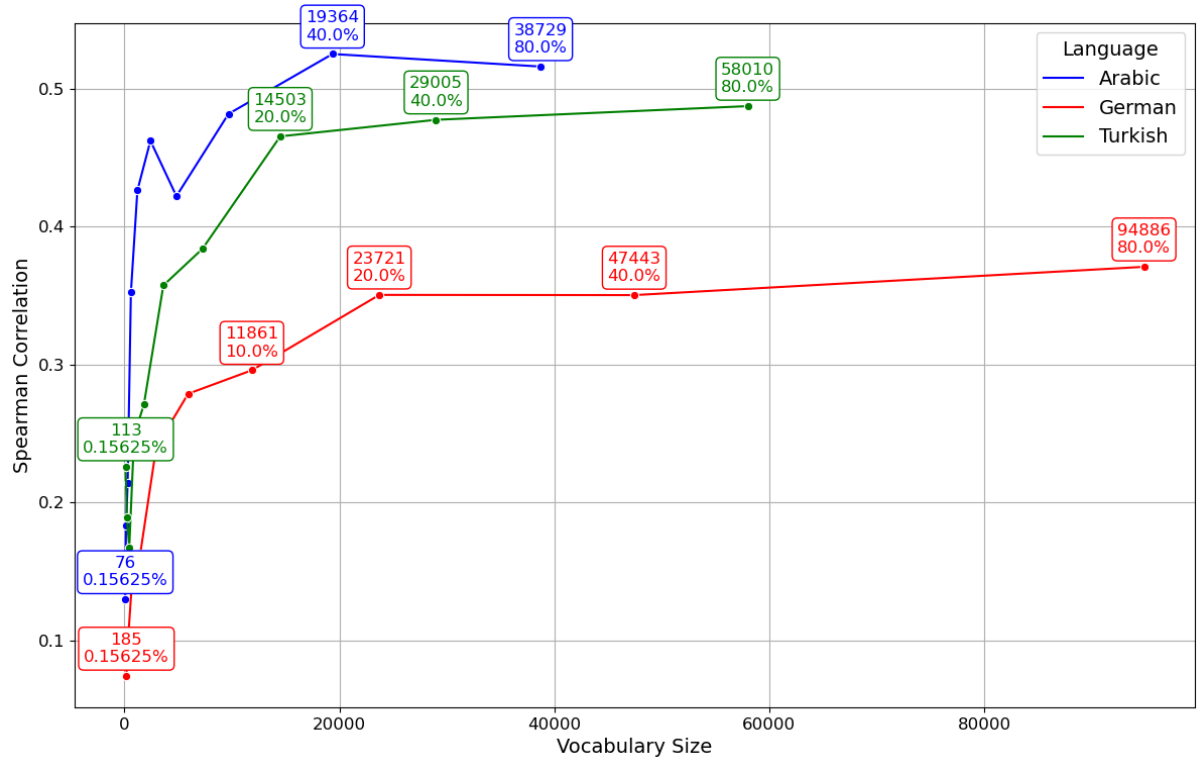


Figure 3: Transfer learning with German, Turkish, and Arabic as target languages. Performance achieves the greatest growth at vocabulary sizes of 371 (German), 906 (Turkish), and 151 (Arabic).

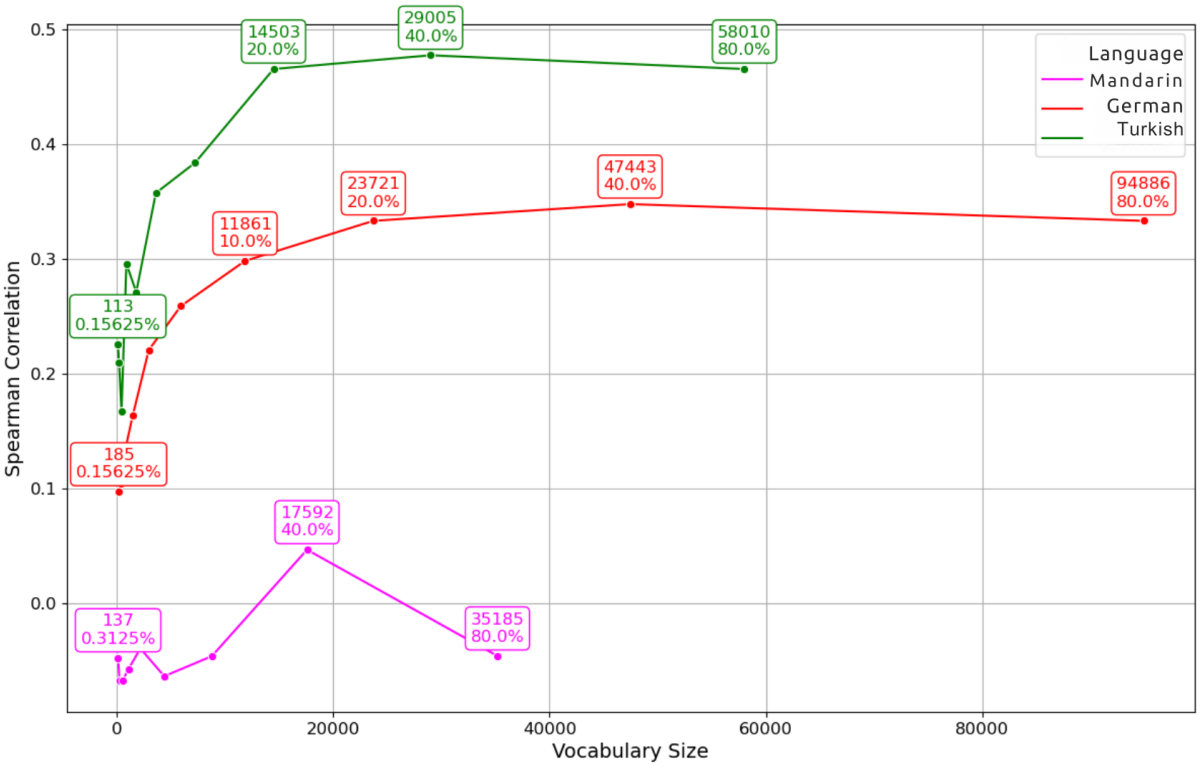


Figure 4: Transfer learning with German, Turkish, and Mandarin as target languages. Performance achieves the greatest growth at vocabulary sizes of 741 (German), 906 (Turkish), and 1100 (Mandarin).

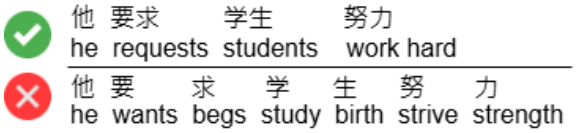


Figure 5: Correct tokenization of Mandarin Chinese (top) versus the study's whitespace tokenization (bottom). The semantic meaning of the sentence changes depending on the tokenization.

of contexts than those for actual low-resource languages (Marashian et al., 2025). Often, the only data available for low-resource languages are small amounts of religious texts (Marashian et al., 2025). Future work could verify findings by replicating the methodology for low-resource languages themselves where sufficient data is available.

5.2 Tokenization

The use of a classic tokenizer and the omission of a more sophisticated tokenizer excludes languages that lack explicit word boundaries. While German, Turkish, and Arabic can be tokenized using whitespace and punctuation, certain languages like Mandarin lack distinct spaces between words. Subword tokenization can better handle languages with non-standard word boundaries. To enhance this work, the study's methodology could be replicated with a subword tokenizer applied to a real-life low-resource language.

5.3 Methodology Utilized

Table 2 indicates the methodology adopted for this study overfits to the proxy languages; the study's sentiment classifiers lag well behind those of Alnajjar et al. (2023). Consequently, Alnajjar et al. (2023)'s methodology is unstable and cannot transfer knowledge across *all* languages. Multiple rounds of hyperparameter finetuning did not improve the model's performance (see Appendix E). One possible issue may stem from fine-tuning the sentiment classifier on English STS examples. Even with aligned word embeddings, the model may not possess enough cross-lingual knowledge to map knowledge gained from the English STS examples to the proxy language. The heavily skewed distributions in Figure 10 and Figure 11 suggest that insufficient knowledge is being captured in this step of fine-tuning. It is noted in Stevenson and Merlo (2022) that word embeddings are far from capturing human-like lexical abilities; a more effective vector representation of the language may be necessary to prevent under/over-fitting and pave the way for more efficient learning. Although there may exist other cross-lingual transfer methodologies that are more optimized than Alnajjar et al. (2023), we present one methodology that is simple and intuitive in design. While the languages we evaluated show enough linguistic variation and could generalize to other languages, we feel that such methodologies and results cannot transfer across *all* languages.

While sentiment classification is a foundational task in NLP, additional work could be done to explore how documentation requirements differ for tasks of varying complexity.

6 Conclusion

Genetic proximity between the source and target language may not have an impact on how well the target language performs on the STS task. We note that the performance of the target language plateaus at a vocabulary size of 25,000. This may be dependent on morphology as seen in the case of Mandarin. Based on data from PanLex, low-resource languages lack the level of documentation deemed necessary in this study but embedding spaces experience the greatest level of improvement when vocabularies are relatively small.

While word embeddings are useful in modeling language, they would not exist without a tokenizer. It can be argued that a tokenizer is just as an important area of research as word embeddings, if not more important; without a tokenizer, a model could not extract the relevant semantic features from text. Future research could investigate the minimum amount of data needed to develop this foundational tool in language processing.

Acknowledgments

I wanted to acknowledge the work and help of Caroline Gihlstorf and Jade Gregoire. Their feedback framed the study and laid the groundwork for the initial stages of research.

References

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2023. Sentiment analysis using aligned word embeddings for Uralic languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 19–24, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Catherine Arnett and Benjamin K. Bergen. 2024. Why do language models perform worse for morphologically complex languages? *Preprint*, arXiv:2411.14198.

Vincent Beaufils and Johannes Tomin. 2020. Stochastic approach to worldwide language classification: the signals and the noise towards long-range exploration.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Rozière. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. *Preprint*, arXiv:2402.01035.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, Beijing, China. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chanhee Lee, Kisu Yang, Taesun Whang, Chanjun Park, Andrew Matteson, and Heuiseok Lim. 2021. Exploring the data efficiency of cross-lingual post-training in pretrained language models. *Applied Sciences*, 11(5).

Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. 2024. Ems: Efficient and effective massively multilingual sentence embedding learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2841–2856.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. *Preprint*, arXiv:2412.00966.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Suzanne Stevenson and Paola Merlo. 2022. Beyond the benchmarks: Toward human-like lexical representations. *Frontiers in Artificial Intelligence*, 5.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models. *Preprint*, arXiv:2002.07306.

Torsten Zesch. 2010. *Study of Semantic Relatedness of Words Using Collaboratively Constructed Semantic Resources*. Ph.D. thesis, Technische Universität Darmstadt, Darmstadt, Germany.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Proximity to Dutch

See Table 3 for details on how distant our proxies were from the source language Dutch.

Language	Genetic Proximity to Dutch
German	13.5
Turkish	87.5
Arabic	82.8
Mandarin	83.8

Table 3: A genetic proximity between 1 and 30 indicates two highly related languages while a genetic proximity between 78 and 100 indicates two languages with no recognizable relationship (Beaufils and Tomin, 2020).

B Skipping additional finetuning

While Alnajjar et al. (2023) finetuned the word embeddings with books to expand the embedding's vocabulary, we discovered that this phase was unnecessary for our proxy languages. To evaluate the necessity of this phase, we compared the performance of (1) embedding spaces trained on translation lexicons and finetuned on English STS samples against (2) sentence embedding spaces trained on translation lexicons, *finetuned on Wikipedia articles from their respective languages*, and finetuned on English STS samples. Wikipedia was selected as a data source because its articles cover a wide range of domains. Embedding spaces that underwent this extra phase of finetuning on Wikipedia articles performed only marginally better than embedding spaces that skipped this phase (see Tables 4 and 5). Consequently, this extra phase of finetuning was skipped.

C Qualitative analysis of word embedding alignment

C.1 MUSE

To qualitatively assess how well MUSE alignment worked, we retrieved word embedding vectors that had the highest cosine similarity score with the English word "revolution." Tables 6, 7, and 8 depict how before alignment, the closest words to "revolution" stray from the original definition and take on a positive connotation (e.g. patriot) or negative tone (e.g. riots, uprising). Realignment under MUSE resulted in higher cosine similarity scores as well as words that were denotatively and/or connotatively similar to the word "revolution."

C.2 At varying lexicons sizes

For each proxy language, we examine words that have the highest cosine similarity score with the English word "revolution" across multiple vocabulary sizes. When aligned with small vocabulary sizes, Mandarin embedding spaces output words that are in a different language (see Table 12). At

smaller vocabulary sizes ($\leq$ 200 words), words that are deemed most similar appear to be tangential to the concept of revolution. Certain terms such as “loyalisten" (German: “loyalists") and “japonlar" (Turkish: “Japanese") reflect potential bias (see Table 9 and 10). As vocabulary sizes grow, so do cosine similarity scores (see Tables 9, 10, 11, 12). Even at larger vocabulary sizes, many terms with a high cosine similarity score are ones that reflect a positive and/or negative connotation of revolution, such as “diktaturen" (German: “dictatorships") and “vatansever" (Turkish: “patriotism") (see Tables 9 and 10).

D Cleaning XED

XED is a multilabel classification dataset, annotating samples with labels such as anger, disgust, and anticipation. To convert the dataset into one for binary classification, we labeled samples as positive or negative based on specific rules, resulting in the positive-negative label distribution shown in Table 13.

- **A sample is positive** if it contains only positive labels (i.e. “anticipation", “joy", and “trust"). Samples that combined positive labels with a neutral label (i.e. “surprise") were still considered positive.
- **A sample is negative** if it contains only negative labels (i.e. “anger", “disgust", “fear", “sadness"). Samples that combined negative labels with a neutral label (i.e. “surprise") were also considered negative.

E Impacts of hyperparameter finetuning

Due to resource constraints and the computational load of the sentiment classifier, exhaustively exploring the hyperparameter space was intractable. We focused our efforts on tuning the number of neurons in the hidden layer as the low F1 scores in predicting certain labels indicate that the model was underfitting and potentially lacked sufficient complexity to effectively handle the sentiment analysis of sentences (see Table 2). Setting the dropout rate to 0.2,we fail to identify an optimal hidden layer neuron count as the model consistently predicts positive labels well at the expense of negative labels. This relationship is occasionally reversed: the model consistently predicts negative labels well at the expense of positive labels. Results are shown in Figure 6, Figure 7, Figure 8, and Figure 9.

F Distribution of Semantic Textual Similarity Scores

It is apparent that the distributions of the predicted cosine similarity scores do not mirror that of the actual cosine similarity scores (see Figure 10 and 11). Except for Mandarin, proxy languages show a left-skewed distribution in cosine similarity scores (see Figure 10 and 11). A higher cosine similarity score indicates greater similarity between sentences (Muennighoff et al., 2023), suggesting that the sentence embedding space is more likely to classify a pair of sentences as similar rather than dissimilar.

We normalized the actual similarity scores in Figure 10 and Figure 11 to allow for better comparison.

G Documentation Available in Low-Resource Languages

Table 14 indicates the number of word translation pairs available in PanLex (Kamholz et al., 2014). PanLex is a database that provides over 1.1 billion pairwise translations in about 9,000 language varieties, including 1,603 UNESCO-classified endangered and vulnerable languages. Using the methodology described in the paper, endangered languages in general do not possess 25,000 entries, the amount of data required to see a plateau in embedding performance. Notably, sentence embedding spaces experienced the greatest increase in performance when the vocabulary size was less than 1000, the average number of translations in a translation lexicon for an endangered language (see Table 14 and Figure 12).

H Comparison of proxy language sentence embedding spaces against MTEB models

Except for Mandarin, our best-performing sentence embedding spaces perform as well as the average model on the MTEB leaderboard (see Table 15). How to further improve these sentence embeddings is a matter of future research.

Language	Training and Finetuning Process	Spearman Correlation
German	translation dictionary, MUSE alignment	0.371
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.376
Turkish	translation dictionary, MUSE alignment	0.488
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.517
Arabic	translation dictionary, MUSE alignment	0.516
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.503

Table 4: Fine-tuning the word embedding space on Wikipedia articles resulted in marginal gains in performance for the German, Turkish, Arabic test group.

Language	Training and Finetuning Process	Spearman Correlation
German	translation dictionary, MUSE alignment	0.333
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.363
Turkish	translation dictionary, MUSE alignment	0.466
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.493
Mandarin	translation dictionary, MUSE alignment	-0.046
	translation dictionary, finetuning on Wikipedia articles, MUSE alignment	0.074

Table 5: Fine-tuning the word embedding space on Wikipedia articles resulted in marginal gains in performance for the German, Turkish, Mandarin test group.

Pre-MUSE			Post-MUSE		
Word	**Translation**	**Cosine Similarity**	**Word**	**Translation**	**Cosine Similarity**
muros	*not German*	0.1897	rebellion	rebellion	0.5144
mox	*not German*	0.1897	aufstand	revolt	0.5144
franken	franc	0.1910	radikalisierung	radicalization	0.5150
koadjutor	coadjutor	0.1917	aufstände	riots	0.5311
latein	Latin	0.1918	umwälzungen	upheavals	0.5371
palgrave	*not German*	0.1980	revolutionären	revolutionary	0.6200
neb	*not German*	0.1997	konterrevolution	counterrevolution	0.6209
emeritierung	emeritus	0.2100	revolutionäre	revolutionary	0.6590
emeritierter	emeritus	0.2100	revolutionär	revolutionary	0.6865
avalos	*not German*	0.2181	revolutionen	revolutions	0.6948

Table 6: German translations and cosine similarity scores of “revolution" before and after MUSE alignment. Quality of translation significantly improves following MUSE alignment.

Pre-MUSE			Post-MUSE		
Word	**Translation**	**Cosine Similarity**	**Word**	**Translation**	**Cosine Similarity**
bem	*not Turkish*	0.1850	revolutionibus	*not Turkish*	0.5010
galiçya	galicia	0.1863	diktatörlük	dictatorship	0.5081
gravis	gravis	0.1888	sosyalizm	socialism	0.5123
prism	*not Turkish*	0.1891	isyan	revel	0.5161
lennox	*not Turkish*	0.1905	ayaklanması	uprising	0.5161
gsc	*not Turkish*	0.1906	ayaklanmalar	riots	0.5292
frangı	franc	0.1917	devrimler	revolutions	0.5391
latin	*not Turkish*	0.1970	devrim	revolution	0.6237
palgrave	*not Turkish*	0.1980	devrimciler	revolutionaries	0.6611
neb	*not Turkish*	0.1997	devrimci	revolutionary	0.6611

Table 7: Turkish translations and cosine similarity scores of “revolution" before and after MUSE alignment. Quality of translation significantly improves following MUSE alignment.

Pre-MUSE			Post-MUSE		
Word	**Translation**	**Cosine Similarity**	**Word**	**Translation**	**Cosine Similarity**
回憶	recall	0.1800	推翻	overthrow	0.4902
大主教	archbishop	0.1803	愛國者	patriot	0.4991
退休	retire	0.1841	獨裁	dictatorship	0.5079
稜鏡	*not a phrase*	0.1891	專政	dictatorship	0.5079
pluribus	*not Mondarin*	0.1893	獨裁政權	dictatorship	0.5079
gsc	*not Mandarin*	0.1906	社會主義	socialism	0.5109
mox	*not Mandarin*	0.1910	起義	uprising	0.5115
拉丁文	Latin	0.1970	保皇黨	royalist	0.5148
教育	educate	0.1979	革命性	revolutionary	0.6875
palgrave	*not Mandarin*	0.1970	革命	revolution	0.6928

Table 8: Mandarin translations and cosine similarity scores of “revolution" before and after MUSE alignment. Quality of translation significantly improves following MUSE alignment.

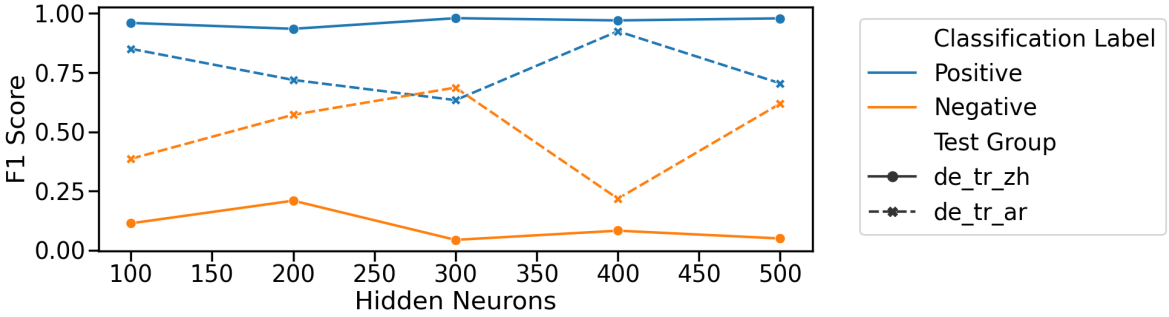


Figure 6: We evaluate the F1 scores of the German sentiment classifier on positive and negative labels across varying amounts of hidden layer neurons. The German classifier from the German, Turkish, and Mandarin test group (abbreviated as `de_tr_zh`) is depicted alongside that from the German, Turkish, and Arabic test group (abbreviated as `de_tr_ar`). Increasing the number of neurons causes a tradeoff in positive and negative label performance as shown in the `de_tr_zh` group. Moreover, increasing the number of neurons does not prevent the model from overfitting to positive labels or underfitting to negative labels as shown in the `de_tr_ar` group.

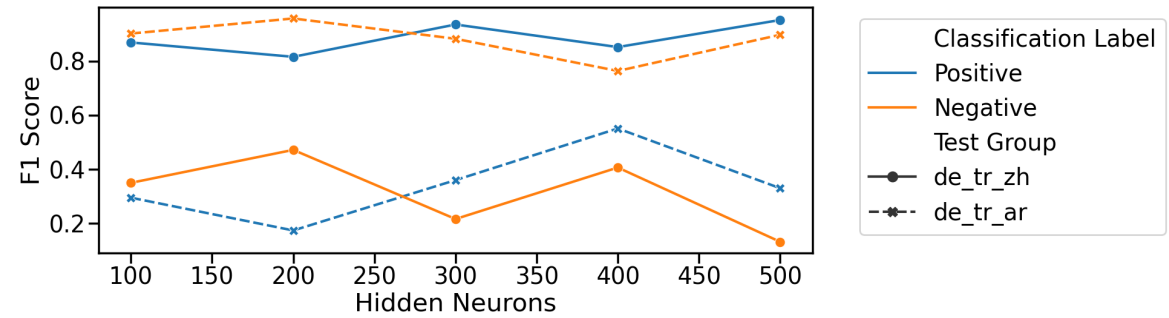


Figure 7: We evaluate the F1 scores of the Turkish sentiment classifier on positive and negative labels across varying amounts of hidden layer neurons. The Turkish classifier from the German, Turkish, and Mandarin test group (abbreviated as `de_tr_zh`) is depicted alongside that from the German, Turkish, and Arabic test group (abbreviated as `de_tr_ar`). In `de_tr_zh`, increasing the number of neurons does not prevent the Turkish sentiment classifier model from overfitting to positive labels and underfitting to negative labels. This is reversed in `de_tr_ar`; the Turkish model overfits to negative labels and underfits to positive labels.

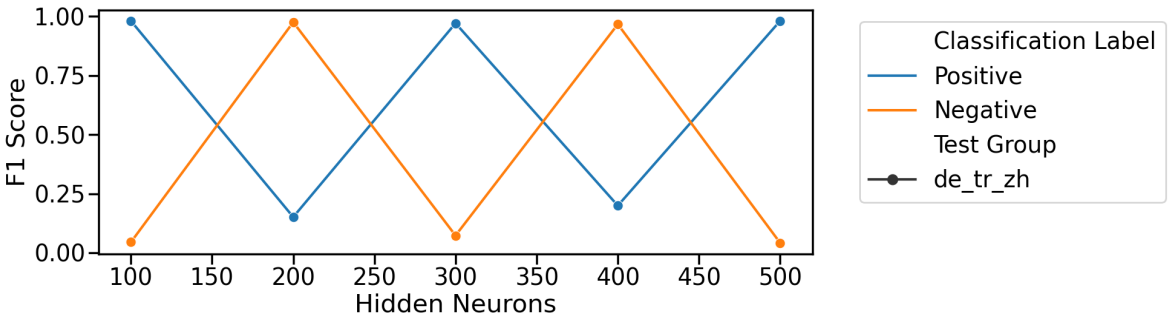


Figure 8: We evaluate the F1 scores of the Mandarin sentiment classifier on positive and negative labels across varying amounts of hidden layer neurons. Increasing the number of hidden neurons causes a tradeoff in positive and negative label performance.

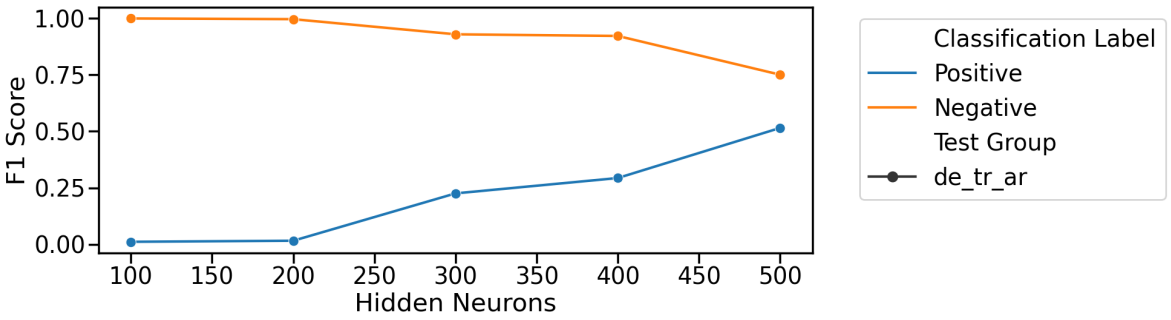
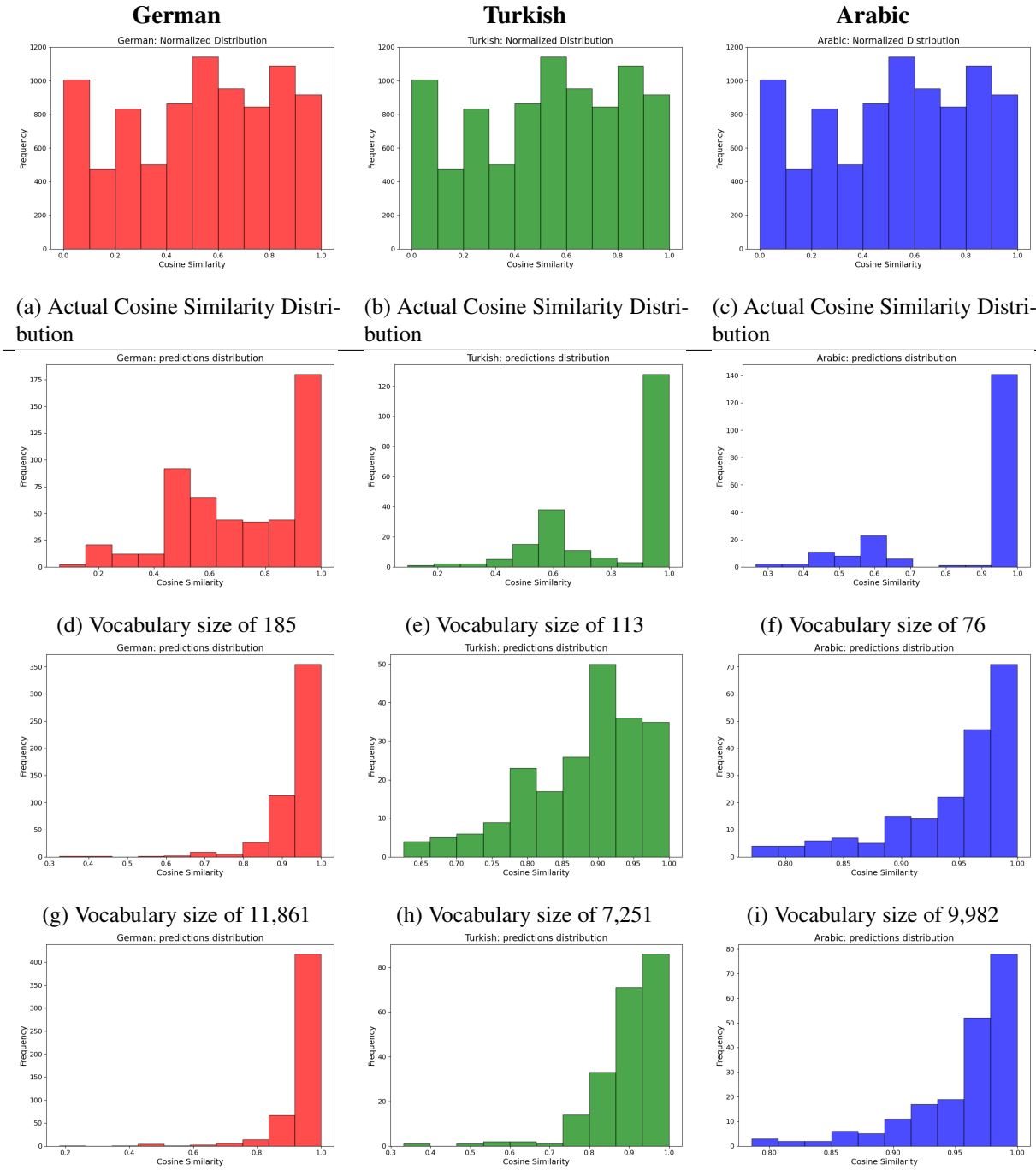


Figure 9: We evaluate the F1 scores of the Arabic sentiment classifier on positive and negative labels across varying amounts of hidden layer neurons. Increasing the number of hidden neurons seemingly causes a convergence in performance but the classifier's ability to correctly positive labels is sacrificed to correctly predict negative labels.



(a) Actual Cosine Similarity Distribution

(b) Actual Cosine Similarity Distribution

(c) Actual Cosine Similarity Distribution

(d) Vocabulary size of 185

(e) Vocabulary size of 113

(f) Vocabulary size of 76

(g) Vocabulary size of 11,861

(h) Vocabulary size of 7,251

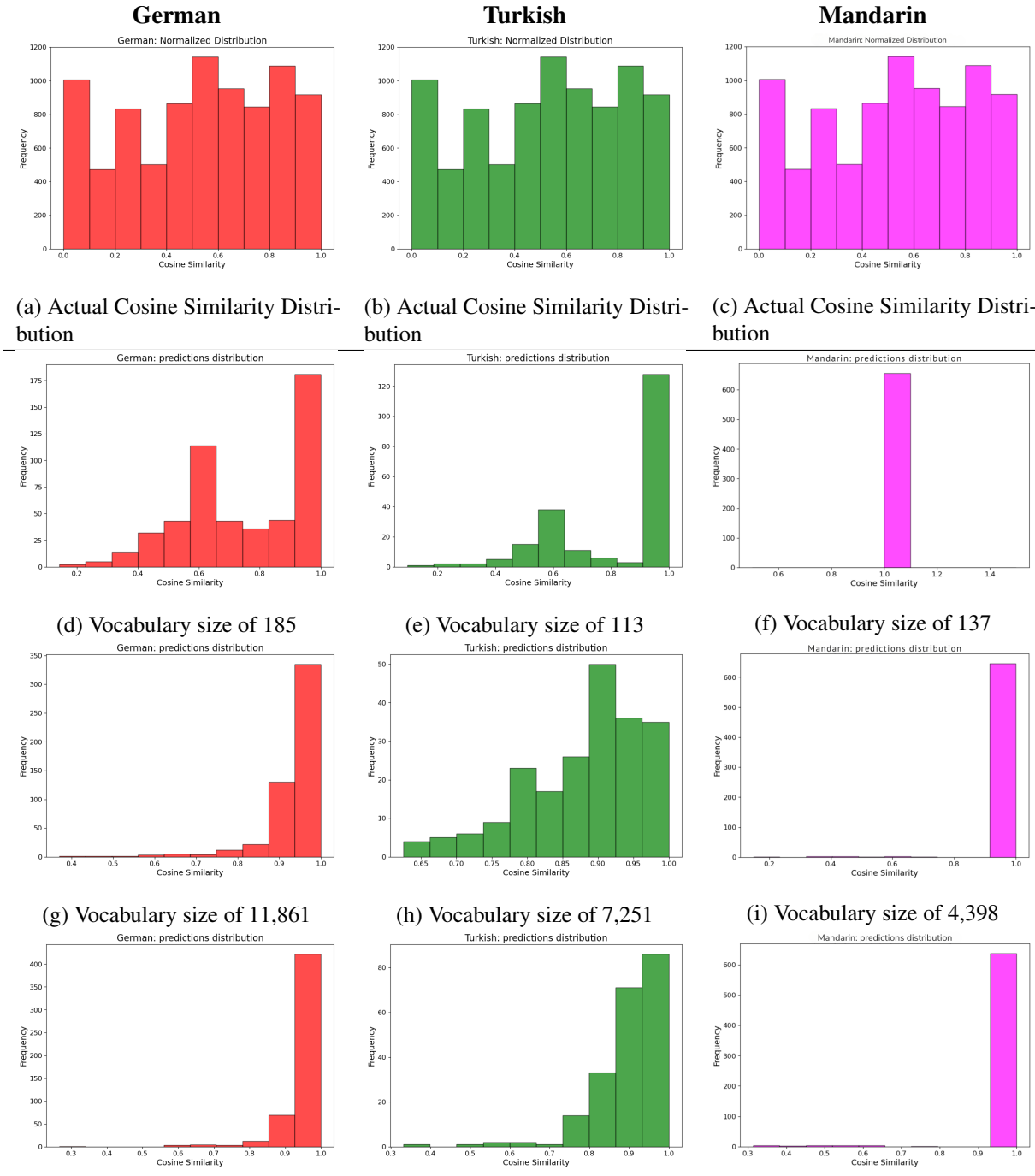
(i) Vocabulary size of 9,982

(j) Vocabulary size of 23,721

(k) Vocabulary size of 14,503

(l) Vocabulary size of 19,364

Figure 10: Distribution of cosine similarity scores across the MTEB evaluation German, Turkish, and Arabic datasets. As the vocabulary size increases, the distribution becomes more left-skewed.



(a) Actual Cosine Similarity Distribution

(b) Actual Cosine Similarity Distribution

(c) Actual Cosine Similarity Distribution

(d) Vocabulary size of 185

(e) Vocabulary size of 113

(f) Vocabulary size of 137

(g) Vocabulary size of 11,861

(h) Vocabulary size of 7,251

(i) Vocabulary size of 4,398

(j) Vocabulary size of 23,721

(k) Vocabulary size of 14,503

(l) Vocabulary size of 17,592

Figure 11: Distribution of cosine similarity scores for the MTEB evaluation German, Turkish, and Mandarin datasets. As the vocabulary size increases, the distribution becomes more left-skewed with the exception of Mandarin.

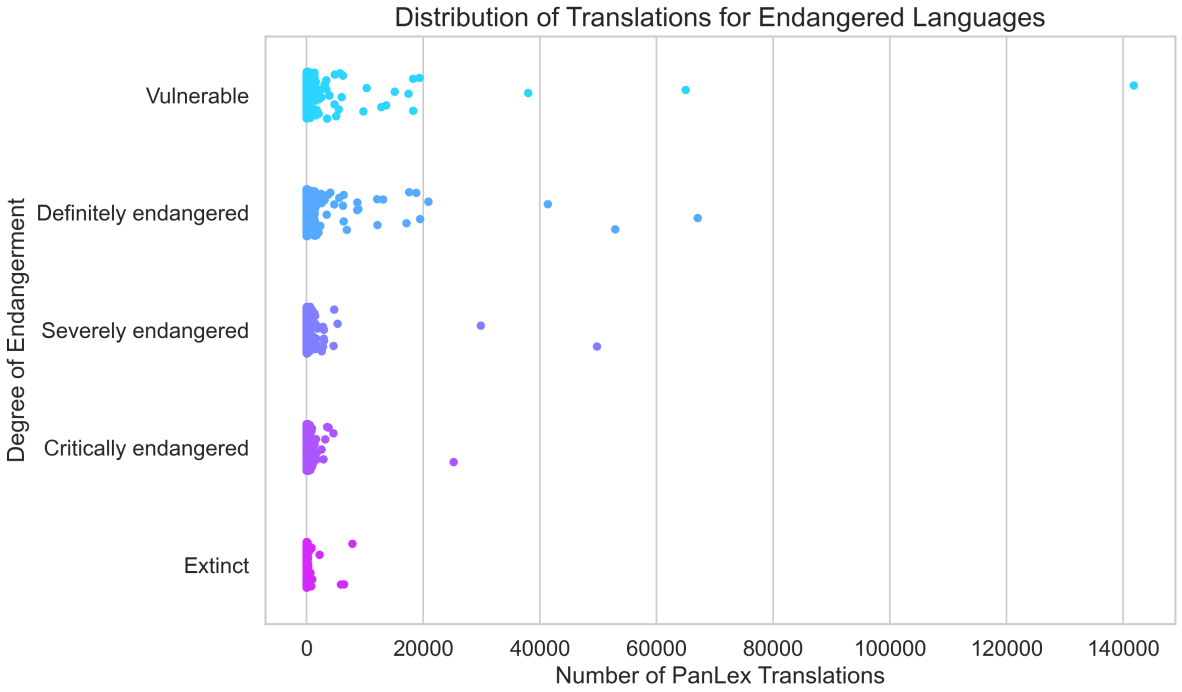


Figure 12: Distribution of PanLex translations for UNESCO-classified vulnerable and endangered languages. Most endangered languages have fewer than 20,000 translations.

Vocabulary size of 185		
Word	**Translation**	**Cosine Similarity**
verbiete	ban	0.2925
übergelaufen	defected	0.2986
stilllegung	decommissioning	0.3058
ausgleichszahlung	compensation	0.3131
besprechung	meeting	0.3206
überschreitung	exceedance	0.3213
allgemeinen	general	0.3235
passiert	happened	0.3322
loyalisten	loyalists	0.3407
umgestaltung	refactor	0.3612
Vocabulary size of 11,861		
Word	**Translation**	**Cosine Similarity**
demokratisierung	democratization	0.5395
zusammenbrechen	collapse	0.5429
absolutistischen	absolutist	0.5475
radikaler	more radical	0.5542
unterdrückt	suppressed	0.5601
unterdrückten	suppressed	0.5601
bevorstehende	upcoming	0.5727
feindschaft	enmity	0.5752
feindseligkeit	hostility	0.5752
revolutionären	revolutionary	0.6664
Vocabulary size of 23,721		
Word	**Translation**	**Cosine Similarity**
radikaler	more radical	0.5549
diktaturen	dictatorships	0.5596
unterdrückt	suppressed	0.5600
unterdrückten	suppressed	0.5600
feindseligkeit	suppressed	0.5740
feindschaft	enmity	0.5740
bevorstehende	upcoming	0.5748
unterdrückung	suppression	0.5854
verdrängung	displacement	0.5854
revolutionären	revolutionary	0.6645

Table 9: German translations and cosine similarity scores of “revolution" across varying dictionary sizes. Increasing the vocabulary size results in German translations that are semantically closer to “revolution."

Vocabulary size of 227		
Word	**Translation**	**Cosine Similarity**
ihtiyaç	need	0.3368
bazı	some	0.3381
japonlar	japanese	0.3422
saldırılar	attacks	0.3428
izdiham	confluence	0.3490
éluard	eluard	0.3596
hükümdarlık	reign	0.4241
danton	danton	0.4517
başla	start	0.4888
hürriyet	freedom	0.4929
Vocabulary size of 7,251		
Word	**Translation**	**Cosine Similarity**
muhalefet	opposition	0.5261
başarısızlık	failure	0.5275
üstünlüğü	superiority	0.5285
katılım	attendance	0.5317
getirildi	brought	0.5326
çöküş	collapse	0.5415
diriliş	resurrection	0.5435
düşmanlık	hostility	0.5734
vatansever	patriot	0.5926
devrim	revolution	0.6694
Vocabulary size of 14,503		
Word	**Translation**	**Cosine Similarity**
katılım	participation	0.5319
getirildi	brought	0.5321
çöküş	collapse	0.5416
diriliş	resurrection	0.5445
kapitalist	capitalist	0.5531
düşmanlık	hostility	0.5727
vatansever	patriotic	0.5913
vatanseverlik	patriotism	0.5933
devrim	revolution	0.6684
devrimci	revolutionary	0.6794

Table 10: Turkish translations and cosine similarity scores of “revolution" across varying dictionary sizes. Increasing the vocabulary size results in Turkish translations that are semantically closer to “revolution."

Vocabulary size of 76		
Word	**Translation**	**Cosine Similarity**
اخبار	need	0.2582
اجواء	atmosphere	0.2637
اعداء	enemies	0.2711
بصمت	silently	0.2897
السلالة	strain	0.2949
الحرائق	fires	0.3211
بريطاني	British	0.3326
الشرعية	legitimacy	0.3915
الفظائع	atrocities	0.4016
الاعتقاد	belief	0.4858
Vocabulary size of 9,982		
Word	**Translation**	**Cosine Similarity**
انهاء	end	0.5261
زمن	time	0.5275
الزمان	time	0.5285
وقت	time	0.5317
السخط	discontent	0.5326
القلاقل	unrest	0.5415
الديكتاتوريات	dictatorships	0.5435
فشل	failure	0.5734
التدخل	interference	0.5926
الدكتاتورية	dictatorship	0.6694
Vocabulary size of 19,364		
Word	**Translation**	**Cosine Similarity**
الديكتاتوريات	dictatorships	0.5262
فشل	failure	0.5262
سائد	prevalent	0.5269
التدخل	interference	0.5351
الدكتاتورية	dictatorship	0.5614
ناشئة	emerging	0.5748
العداء	hostility	0.5753
الاطاحة	overthrow	0.5833
القمع	suppression	0.5863
الثورات	revolutions	0.6672

Table 11: Arabic translations and cosine similarity scores of “revolution" across varying dictionary sizes. Increasing the vocabulary size results in Arabic translations that are semantically closer to “revolution."

Vocabulary size of 137		
Word	**Translation**	**Cosine Similarity**
北京	Beijing	0.2427
錯位	dislocation	0.2430
動態	dynamic	0.2463
革新	innovation	0.2473
amraam	*not Chinese*	0.2626
税	tax	0.2632
最多	maximum	0.2772
協作	collaboration	0.2779
永久	permanent	0.3895
然後	then	0.3965
Vocabulary size of 4,398		
Word	**Translation**	**Cosine Similarity**
束縛	binding	0.5042
奴隸制	slavery	0.5042
抵制	boycott	0.5070
演示	demo	0.5194
時間	time	0.5197
反對派	opposition	0.5240
混沌	chaos	0.5241
復蘇	recovery	0.5467
動蕩	turmoil	0.5567
敵意	hostility	0.5769
Vocabulary size of 17,592		
Word	**Translation**	**Cosine Similarity**
動亂	unrest	0.5526
獨裁政權	dictatorship	0.5629
受壓迫	oppressed	0.5636
敵意	hostility	0.5765
推翻	overturn	0.5807
愛國主義	patriotism	0.5874
抑制	inhibition	0.5884
政權	regime	0.5952
保皇黨	loyalist	0.6204
革命性	revolutionary	0.7352

Table 12: Mandarin translations and cosine similarity scores of “revolution" across varying dictionary sizes. Increasing the vocabulary size results in Mandarin translations that are semantically closer to “revolution."

Language	Positive Samples	Negative Samples
Arabic	1269	1735
German	2051	2614
Turkish	3570	4552
Mandarin	587	608

Table 13: Number of positive and negative samples in processed XED data. Ratio of positive to negative samples is 1:1.

Degree of endangerment	Average	Standard deviation
Vulnerable	1205.34	8063.48
Definitely endangered	1094.03	5183.69
Severely endangered	541.77	3359.48
Critically endangered	315.94	1542.20
Extinct	251.69	969.93

Table 14: The average number of translations found for UNESCO-classified vulnerable and endangered languages in PanLex. Existing documentation for endangered languages is generally low. The high standard deviation may be attributed to outliers (e.g. certain vulnerable languages may contain significantly more documentation than others in the category) as shown in Figure 12.

	German	Turkish	Arabic	Mandarin
best-performing sentence embedding	0.371	0.488	0.516	0.046
average MTEB score	0.391	0.466	0.439	0.588
minimum MTEB score	0.082	0.038	0.052	0.048
25th percentile MTEB score	0.266	0.370	0.304	0.600
50th percentile MTEB score	0.418	0.473	0.524	0.654
75th percentile MTEB score	0.506	0.582	0.571	0.668
maximum MTEB score	0.609	0.688	0.598	0.749

Table 15: Comparison of our models' Spearman correlation scores to MTEB models'. Data compiled from Muennighoff et al. (2023). Our models perform average (with the exception of Mandarin) compared to models in this leaderboard.