PEAR at SemEval-2024 Task 1: Pair Encoding with Augmented Re-sampling for Semantic Textual Relatedness

Tollef Emil Jørgensen
Norwegian University of Science and Technology
tollef.jorgensen@ntnu.no

Abstract

This paper describes a system submitted to the supervised track (Track A) at SemEval-24: *Semantic Textual Relatedness for African and Asian Languages*. Challenged with datasets of varying sizes, some as small as 800 samples, we observe that the PEAR system, using smaller pre-trained masked language models to process sentence pairs (**P**air **E**ncoding), results in models that efficiently adapt to the task. In addition to the simplistic modeling approach, we experiment with hyperparameter optimization and data expansion from the provided training sets using multilingual bi-encoders, sampling a dynamic number of nearest neighbors (**A**ugmented **Re**-sampling). The final models are lightweight, allowing fast experimentation and integration of new languages.

1 Introduction

The overall aim of the Semantic Textual Relatedness (STR) shared task (Ousidhoum et al., 2024b) is to correctly predict the relatedness between a given sentence pair on a scale from 0 to 1, described as closeness in meaning (Abdalla et al., 2023; Ousidhoum et al., 2024a), exemplified by *expressing the same views* and *one elaborating on the other*. This shared task covers a broader aspect of the well-established semantic textual similarity (STS) field, which fails to address the intuitive relatedness between two sentences.

From available STS data, such as from the SemEval-2012 task on similarity (Agirre et al., 2012), the sentences "A man is peeling a banana" and "A woman is peeling a potato" receive a normalized similarity of 0.3. In contrast, the two descriptions have a higher degree of relatedness, where *something is being peeled*. Relatedness tends to focus less on equivalence and paraphrasing and more on the broader case of entailment and the cause-effect relationship between two sentences. The task consists of three tracks: A (supervised), B (unsupervised), and C (cross-lingual). The system described here will only consider Track A, allowing the use of any training data. Refer to Ousidhoum et al. (2024a) for more details.

The System and Constraints This paper proposes a system for any language with an available pre-trained masked language model (MLM), such as BERT or RoBERTa, used to process pairs of sentences with full cross-attention. The constraint of using limited-size MLMs was set early in the project to study their performance compared to the impressive baselines observed through existing multilingual bi-encoders. However, following ideas of Thakur et al. (2021), the addition of weakly supervised labels from such bi-encoders was added as an optional step to inspect its impact on smaller datasets.

Being unfamiliar with most of the involved languages and thus being unable to verify the results, no language-specific rules were implemented. Consequently, no text manipulation (such as paraphrasing and replacing words), back-translation, or normalization steps were applied. While the task organizers permitted the use of any available data for the supervised track, in addition to large language models to a limited extent, the presented approach only uses the supplied training dataset per language. While performance suffers in some cases, we hope that the aforementioned constraints help to support as many future languages as possible with little to no modification. Continuing the idea of supporting lower-resourced languages, this system only uses *base* size transformer MLMs, ranging from 110M to 125M parameters.
All code is available on GitHub.[1]

2 Data

The full dataset for SemRel consists of 14 languages. However, only 9 of the 14 languages

[1] https://github.com/tollefj/SemRel-2024

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1405–1411

Language	ISO 639-2/3	Family	Selected Model	Train	Dev	Test	Total
Amharic	amh	Afro-Asiatic	Davlan/xlm-roberta-base-finetuned-amharic	992	95	171	1,258
Algerian Arabic	arq	Afro-Asiatic	CAMeL-Lab/bert-base-arabic-camelbert-da	1,262	92	584	1,938
Moroccan Arabic	ary	Afro-Asiatic	CAMeL-Lab/bert-base-arabic-camelbert-da	925	70	427	1,422
Hausa	hau	Afro-Asiatic	Davlan/xlm-roberta-base-finetuned-hausa	1,763	212	603	2,578
English	eng	Indo-European	FacebookAI/roberta-base	5,500	250	2,500	8,250
Spanish	esp	Indo-European	PlanTL-GOB-ES/roberta-base-bne	1,562	140	600	2,299
Marathi	mar	Indo-European	l3cube-pune/marathi-roberta	1,155	293	298	1,746
Kinyarwanda	kin	Niger-Congo	Davlan/xlm-roberta-base-finetuned-kinyarwanda	778	102	222	1,102
Telugu	tel	Dravidian	l3cube-pune/telugu-bert	1,146	130	297	1,573

Table 1: Included languages and their respective families, along with data sources and data split size.

are included for Track A and have labeled relatedness scores between 0 and 1. Table 1 contains an overview of the languages, data sizes, and selected language models for experiments. Besides the differences in data size, the score distributions also vary greatly, as evident from the four examples in Figure 1. Moreover, when inspecting the

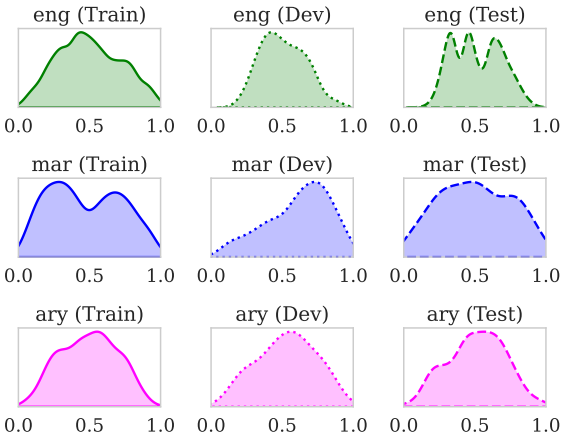

Figure 1: Examples of score distributions.

textual distributions through adversarial validation, modeled by adding a binary classification head to an XLM-R model (Conneau et al., 2020), most languages were seemingly sampled from the same distribution, with an expected ROC-AUC score of 0.5.[2] The English test split, however, had distributions deviating from the train split, shown in Figure 2. ROC curves for more languages are found in Appendix A. Attempts were made to iteratively sample the training set until a better distributional match with the test set was found, with little success in improving results. The more data, the better.

3 Related Work

Semantic Textual Relatedness (STR), in the context of language modeling and prediction, has considerably less research connected to it than Semantic Textual Similarity (STS), which has several datasets and evaluation benchmarks openly available (Muennighoff et al., 2023), many of which tied to the STS task within SemEval (Agirre et al., 2012). The included data is mainly monolingual English, but more recent additions added limited multilingual and cross-lingual tasks (Cer et al., 2017; Chen et al., 2022). The first STR dataset was introduced by (Abdalla et al., 2023), including a monolingual dataset of 5,500 English sentence pairs. New for this task is the inclusion of several low-resource languages not yet studied at the sentence level.

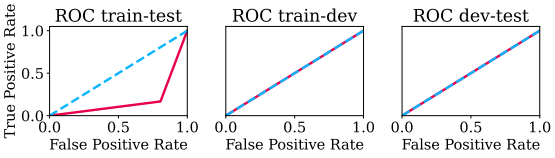

Figure 2: ROC curves for train/dev/test splits on the English data.

The field of natural language processing has drastically changed since the release of the majority of the datasets and shared tasks for semantic textual similarity, where the top-scoring methods typically included a significant amount of feature engineering based on methods like n-gram overlaps, edit distance, and longest common substrings, word alignments, and more, applied to both regression and deep learning models (Tian et al., 2017; Maharjan et al., 2017). Additionally, knowledge-informed systems included semantic information with WordNet and word frequency corpora (Wu et al., 2017). Applying the same efforts to new languages would require significant work, such as collecting new corpora.

Sentence Embeddings Modeling similarity between sentences is commonly associated with *sentence embedding* models, some of which include

[2] An ROC-AUC score of 0.5 indicates that a model cannot differ between samples in the provided data sources.

more than a billion gathered training pairs (Reimers and Gurevych, 2019; Wang et al., 2024). While applicable across many languages and domains, with models initialized from the XLM-Roberta models (Conneau et al., 2020), the included languages do not cover many of which are part of SemRel 2024.

Encoding Sentence Pairs This paper focuses on sentence-pair modeling, encoding the sentences with existing pre-trained language models to create a simpler model that allows fast and easy implementation for any language. This modeling scheme, referred to as cross-encoders, indicating full (cross) self-attention over the entire context, is well explained in previous work by (Wolf et al., 2019; Vig and Ramea, 2019; Humeau et al., 2020). Furthermore, cross-encoders have succeeded in supervised and unsupervised applications (Thakur et al., 2021; Liu et al., 2022). In addition to the sentence scoring, we follow the work by Thakur et al. (2021) to augment data with a bi-encoder, although on much smaller datasets, where the original work was carried out on data up to millions of samples. For details on bi-encoders and sentence embedding models, refer to the excellent implementations by (Reimers and Gurevych, 2019; Humeau et al., 2020; Liu et al., 2022). The baseline provided by the task organizers is LaBSE, a dual-encoder BERT-based sentence embedding model (Feng et al., 2022).

4 System Overview

After restructuring the provided datasets into sentence pairs with their respective labels, they are passed to a MLM with an added regression head; using a sigmoid layer on top of the pooled output, the model is trained using a single-class binary cross-entropy loss, with mean reduction:

$$\ell(x, y) = \frac{1}{n}\sum_{i=1}^{n}\{l_1, \ldots, l_N\}^\top$$
$$l_n = -w_n \left[y_n \log \sigma(x_n) + (1 - y_n) \log(1 - \sigma(x_n))\right]$$

The models (Table 1) were chosen based on searches for existing models in the tasks' languages and closely related language families. In the development phase, scoring was based on 5-fold validation, benchmarked with language-specific and merged data. Experiments, including those presented in Section 6, are on the final release of labeled test datasets.

Augmented Re-sampling In a separate module, a bi-encoder (*multilingual-e5-base*) is employed to find the closest non-existing sentence pairs in the data by creating sentence embeddings and searching for nearest k neighbors with cosine similarity. Before initializing the bi-encoder, the cross-encoder is trained for $\mathcal{E}_{weak}$ epochs before predicting weak labels for the augmented pairs $(s_i, s_j, pred_{i,j})$, which are added to the training data. k determines the number of nearest neighbors to retrieve for each source sentence. A Figure outlining the weak supervision pipeline is found in Figure 3.

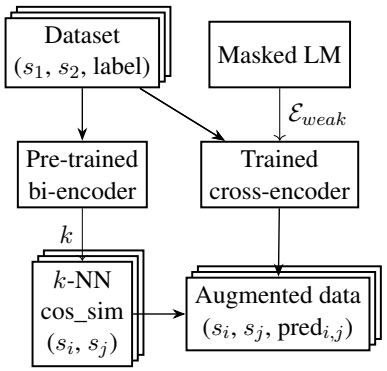

Figure 3: The weak supervision pipeline

Modularity A big focus in the development was to keep it as modular as possible. Models, parameters, data selection, and more are easily controlled through passed arguments. Furthermore, the cross-encoder is provided as a standalone module with varying levels of abstraction, e.g., calling *fit* directly or through a provided training pipeline, including optional weak supervision labeling.

5 Experimental Setup

All experiments and evaluations use the official train/dev/test data splits where applicable, and scores are presented by the Spearman rank correlation coefficient multiplied by 100. In the development phase, we studied the effect of combining or using only per-language data, working as an initial baseline before dev- and test labels were released. This was done by 5-fold validation. As stated in Section 1, no text manipulation or preprocessing was done to keep evaluations fair across languages. Moreover, upon manual inspection, the data seemed sufficiently preprocessed. The following definitions will be used to differ between model configurations:

- **`init`**: no training, only initial weights
- **`all`**: trained on all languages combined
- **`lang`**: trained on one language

Experiments were conducted in three parts:

1. Multilingual sentence embeddings with multilingual-e5-base (Wang et al., 2024)
2. Multi+monolingual MLMs as cross-encoders with XLM-R (Conneau et al., 2020) and models from Table 1.
3. Augmented data from bi-encoders

Augmentation and optimization As the data sizes and model configurations vary, Optuna (Akiba et al., 2019) is set up to search for parameter values for learning rates, k, epochs ($\mathcal{E}$), weak training epochs ($\mathcal{E}_{weak}$), and max gradient norm ($\mathcal{G}$) for clipping. With the augmentation being highly experimental for smaller datasets, we refrain from modifying the bi-encoder and use only its initial weights. Thus, this part of the system can easily be swapped with future models. While limiting the search, the learning rate, gradient clipping, and the k nearest neighbors for augmentation proved to be the most crucial parameters. Table 2 lists the parameters and ranges.

Hparam	Type	Search Space
lr	float	10^{-6} to 10^{-4} (log)
k	int	0 to 3
$\mathcal{E}$	int	1 to 5
$\mathcal{E}_{weak}$	int	0 to 2
$\mathcal{G}$	float	0.1 to 1.0

Table 2: Hyperparameter search space.

No External Data Given the readily available sentence similarity data, such as the STS-Benchmark dataset (Cer et al., 2017), experiments were done to include it in the training pipelines. However, we observed no benefits from this, likely affected by the diverging definitions and annotation styles of relatedness and similarity. Figure 4 shows scores with and without adding the STS-Benchmark dataset (Cer et al., 2017).

6 Results

Results from k-fold validation to quantify the differences between combining training sets vs. training per language show that combining data has a clear benefit. See table 3. However, important factors to consider are data size (e.g., 12,000 vs. 600 samples) and that we are validating in-domain. However,

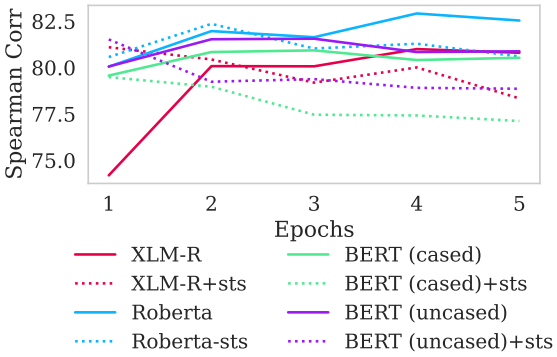

Figure 4: Scores on English-dev with and without STS data (dotted lines)

similar results on the development set show diminishing effects for merging data when predicting out-of-domain data. Development set scores are found in Appendix B. Furthermore, the results indicate how powerful MLMs are once trained, outscoring the *e5* model for most languages. Although the system performed well on the smaller dev sets, ranking in the top 2-4 in most languages in the competition, it struggled on the test sets. This is likely attributed to data distribution, overfitting, and thus failure to generalize relatedness. From the results with augmentation in Table 5, the observed change from default parameters is marginal for most languages. Increasing k without parameter optimization resulted in strictly negative results. Scores on the test set, including top scores and the LaBSE baseline, are shown in Table 4. Despite the lackluster improvement from augmentation and data expansion, the modeling scheme is still promising, outperforming the baseline (in official submissions) for 6/9 languages and 8/9 for the rerun without any changes to optimization configurations.

7 Conclusion

After testing the capabilities of commonly used models for masked language modeling and sentence embeddings, we find MLMs efficient at distinguishing relatedness with little training data. Although attempts at optimizing parameters for in-distribution data resulted in little to no performance gains, there are likely better-suited augmentation strategies for further improving performance with as little source data as possible. As a closing remark, we hope that the provided system may serve as a valuable tool for future developments in semantic textual relatedness.

	arq	amh	eng	hau	kin	mar	ary	esp	tel
XLM init	-2.64 (5.07)	-10.75 (7.67)	-15.49 (3.33)	-4.18 (6.07)	1.03 (3.90)	-7.65 (7.49)	-18.98 (4.88)	0.80 (4.94)	-11.98 (9.32)
XLM all	58.23 (5.30)	**84.56** (1.53)	83.63 (1.36)	**72.25** (0.66)	59.70 (4.09)	**83.44** (2.56)	**82.01** (3.04)	64.73 (3.23)	77.96 (3.87)
XLM lang	39.03 (4.49)	73.22 (3.02)	83.27 (0.89)	63.57 (1.96)	31.40 (7.34)	74.31 (3.02)	69.14 (4.16)	58.72 (8.00)	71.40 (3.99)
e5 init	50.41 (2.82)	75.86 (1.88)	80.72 (0.87)	52.38 (1.93)	46.20 (5.30)	77.00 (1.33)	36.03 (1.59)	60.30 (1.40)	75.28 (1.49)
e5 all	59.45 (2.34)	84.52 (0.88)	86.43 (0.55)	69.01 (0.19)	**69.08** (3.43)	**84.62** (1.35)	81.20 (1.44)	67.16 (2.44)	**80.14** (0.97)
e5 lang	**59.50** (3.25)	82.27 (2.35)	**86.72** (1.02)	68.43 (2.10)	63.04 (3.56)	82.89 (0.32)	75.73 (1.02)	**67.21** (0.39)	77.94 (1.23)

Table 3: 5-fold validation from training datasets using *multilingual-e5-base* (bi-encoder) and *XLM-Roberta base* (cross-encoder). Scores are the average correlation with standard deviations. Bold: best scores per language.

Model/Language	Multiling	k	arq	amh	eng	hau	kin	mar	ary	esp	tel
Baseline (LaBSE)	y	0	60.00	85.00	83.00	69.00	72.00	88.00	77.00	70.00	82.00
Best result	-	-	**68.23**	**88.86**	**85.96**	**76.43**	**81.69**	**91.09**	**86.26**	**74.04**	**87.34**
e5 init (multiling)	y	0	45.32	72.56	80.39	51.23	51.38	77.37	40.14	58.75	77.43
e5 all (multiling)	y	0	59.28	82.06	83.53	68.36	71.61	87.27	78.27	69.16	83.25
e5 lang (multiling)	y	0	60.68	81.46	83.55	69.97	71.87	87.91	77.75	69.02	82.24
e5 init	n	0	43.94	9.02	82.69	40.79	48.23	52.76	15.41	65.22	28.69
e5 all	n	0	59.32	14.48	82.88	61.87	68.15	69.59	77.30	71.26	43.64
e5 lang	n	0	55.30	13.70	83.54	63.63	63.60	67.88	36.11	70.86	34.21
xlm-r init	y	0	-1.10	12.45	-4.23	-0.75	1.93	-10.24	-28.12	1.73	-14.68
xlm-r all	y	0	59.88	83.42	83.69	70.74	67.48	85.99	83.04	71.39	85.75
xlm-r lang	y	0	47.66	81.90	83.46	70.17	56.76	85.84	82.23	69.73	80.78
custom init	n	0	-10.97	20.40	10.09	9.52	14.35	-3.35	-1.91	-3.35	8.40
custom lang	n	0	40.04	83.86	83.31	68.79	72.09	86.10	81.15	72.05	83.46
custom lang	n	1	44.56	81.99	83.42	66.56	72.75	85.83	80.74	71.95	84.42
custom lang	n	2	43.75	81.89	83.27	65.66	70.27	85.72	80.49	71.79	85.05
custom lang	n	3	42.28	81.13	83.39	64.67	71.47	85.36	80.37	72.29	84.73
PEAR$_{test}$	n	+	46.33	83.42	84.79	69.41	77.22	85.60	81.53	71.01	82.75
PEAR$_{rerun}$	n	+	48.58	85.72	83.95	70.68	73.92	88.81	81.68	72.52	86.82

Table 4: Performance on the test set, ordered by languages as presented on the task website. *Multiling* indicates whether the model was pre-trained on multilingual data. k indicates the k-NN resamples used (+: different k per language). *custom*: monolingual models as listed in Table 1. Bold: best score (from all submissions to Track A). Underline: second best from the experiments.

lang	lr	k	$\mathcal{E}$	$\mathcal{E}_{weak}$	$\mathcal{G}$	score	Δ
arq	9.80e-5	1	4	2	0.82	48.58	+8.54
amh	8.42e-5	3	5	2	0.80	85.72	+1.86
eng	3.34e-5	2	2	2	0.12	83.95	+0.64
hau	4.87e-5	0	3	2	0.65	70.68	+1.89
kin	2.47e-5	0	5	1	0.67	73.92	+1.83
mar	5.32e-5	3	2	2	0.42	88.81	+2.71
ary	9.01e-5	1	4	2	0.99	81.68	+0.53
esp	2.40e-5	1	5	2	0.85	72.52	+0.47
tel	3.66e-5	3	3	1	0.66	86.82	+3.36

Table 5: Parameters found from the search space in Table 2. Δ indicates change vs. default parameters.

8 Limitations

Few models were tested per language for the competition. Alternative multi- and monolingual models could provide much better results, especially for Algerian Arabic. This limitation is also influenced by the lack of understanding of most involved languages, e.g., to inspect the source datasets used for pretraining. Finally, grouping specific languages for training, such as merging Indo-European and Afro-Asiatic languages, was not explored.

9 Ethical Considerations

The final system performs predictions of input texts. Predictions may impose ethical concerns, e.g., when used for public-facing applications. Furthermore, automating *relatedness* has possible side effects in bias and fairness towards specific nationalities. For further details about the data and annotation, refer to Ousidhoum et al. (2024a).

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework.

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

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

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

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring.

Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations.

Nabin Maharjan, Rajendra Banjade, Dipesh Gautam, Lasang J. Tamang, and Vasile Rus. 2017. DT_Team at SemEval-2017 task 1: Semantic similarity using alignments, sentence-level embeddings and Gaussian mixture model output. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 120–124, Vancouver, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 14 languages.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks.

Junfeng Tian, Zhiheng Zhou, Man Lan, and Yuanbin Wu. 2017. ECNU at SemEval-2017 task 1: Leverage kernel-based traditional NLP features and neural networks to build a universal model for multilingual and cross-lingual semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 191–197, Vancouver, Canada. Association for Computational Linguistics.

Jesse Vig and Kalai Ramea. 2019. Comparison of transfer-learning approaches for response selection in multi-turn conversations. In *Workshop on DSTC7*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents.

Hao Wu, Heyan Huang, Ping Jian, Yuhang Guo, and Chao Su. 2017. BIT at SemEval-2017 task 1: Using semantic information space to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 77–84, Vancouver, Canada. Association for Computational Linguistics.

A Adversarial Validation

Figures 5 and 6 show the ROC curve for a selection of languages where the AUC value deviated from the norm. English was an outlier here, where the test set is seemingly out-of-distribution. An XLM-Roberta base model set up as a cross-encoder was used for classification. 5 epochs, learning rate 2×10^{-5}. All languages not shown in the figures have an expected ROC-AUC close to 0.5.

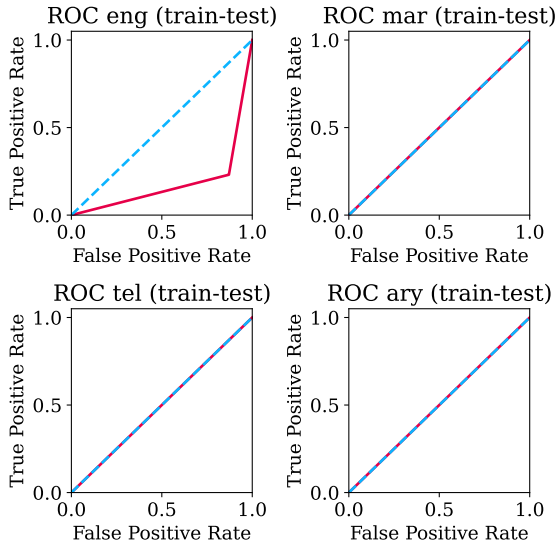

Figure 5: Adversarial validation for Train vs Test. English (eng), Marathi (mar), Telugu (tel) and Moroccan Arabic (ary).

B Development Set Results

Table 6 shows the results on dev sets for *multilingual-e5-base*, *XLM-Roberta-base*, and language-specific masked language models (as defined in Table 1). Modeling configurations are the same as listed in Section 5 – repeated below:

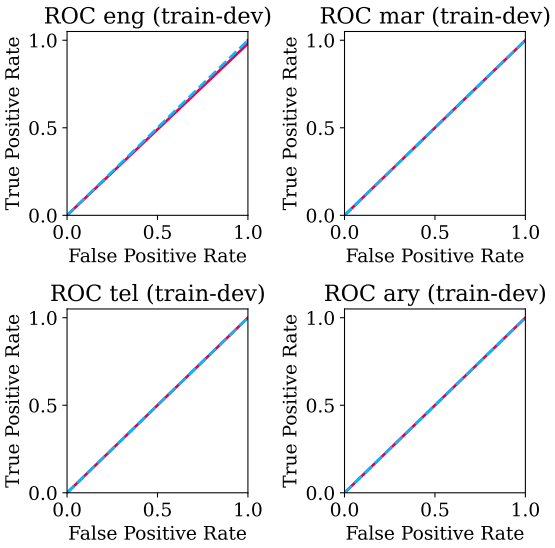

Figure 6: Adversarial validation for Train vs Dev. English (eng), Marathi (mar), Telugu (tel) and Moroccan Arabic (ary).

- **`init`**: no training, only initial weights
- **`all`**: trained on all languages combined
- **`lang`**: trained on one language

Lang	e5 multilingual			XLM-Roberta			MLM	
	init	all	lang	init	all	lang	init	lang
arq	39.70	54.26	**59.71**	-11.11	59.22	57.32	4.76	38.90
amh	61.82	78.47	77.79	-5.30	**86.57**	83.38	19.65	85.90
eng	78.31	81.44	82.10	-12.06	80.88	81.05	10.98	**82.79**
hau	45.01	73.27	72.91	-9.19	**76.39**	75.41	12.95	75.99
kin	27.80	62.55	**65.63**	-18.13	59.90	48.67	7.09	64.73
mar	72.48	81.56	80.56	-15.13	84.24	82.86	-9.91	**84.73**
ary	44.21	78.84	73.79	-28.55	**83.96**	82.61	-19.75	81.43
esp	62.63	68.54	63.16	22.38	**71.24**	65.01	12.80	68.23
tel	77.35	**82.27**	79.75	-16.67	80.34	80.57	18.36	80.74

Table 6: Results on the development sets. Bold: best score per language.