MBZUAI-UNAM at SemEval-2024 Task 1: Sentence-CROBI, a Simple Cross-Bi-Encoder-Based Neural Network Architecture for Semantic Textual Relatedness

Jesus-German Ortiz-Barajas
Mohamed Bin Zayed
University of
Artificial Intelligence
Jesus.OrtizBarajas@mbzuai.ac.ae

Gemma Bel-Enguix
Instituto de Ingeniería
Universidad Nacional
Autónoma de México
gbele@iingen.unam.mx

Helena Gómez-Adorno
IIMAS
Universidad Nacional
Autónoma de México
helena.gomez@iimas.unam.mx

Abstract

The Semantic Textual Relatedness (STR) shared task aims at detecting the degree of semantic relatedness between pairs of sentences on low-resource languages from Afroasiatic, Indoeuropean, Austronesian, Dravidian, and Nigercongo families. We use the Sentence-CROBI architecture to tackle this problem. The model is adapted from its original purpose of paraphrase detection to explore its capacities in a related task with limited resources and in multilingual and monolingual settings. Our approach combines the vector representation of cross-encoders and bi-encoders and possesses high adaptable capacity by combining several pre-trained models. Our system obtained good results on the low-resource languages of the dataset using a multilingual fine-tuning approach.

1 Introduction

Task 1 of SemEval 2024 (Ousidhoum et al., 2024b) focuses on Semantic Textual Relatedness (STR). Given two sentences, the semantic relatedness between them is defined as the degree of closeness between their meanings (Mohammad and Hirst, 2012). However, the traits that make two sentences to be understood as related entities can be of different order, such as the underlying syntactic structure, lexical affinity, or the author's style, among others.

The task organisers have chosen for this track a set of languages, among which English and Spanish stand out, two languages with numerous computational resources. The rest, are low-resourced languages from Africa (Algerian Arabic, Moroccan Arabic, Amharic, Hausa, Kinyarwanda) and Asia (Marathi, Telegu).

Three tracks were proposed in the task: supervised, unsupervised and cross-lingual. We participated in Track A, supervised. This is a regression problem since a relatedness coefficient must be given that ranges from 0 to 1 for each pair of sentences. Our solution is based on using the sentence-CROBI model, introduced in Ortiz-Barajas et al. (2022), which was designed for paraphrase detection with very good results in English. Our hypothesis is that the same methods used in paraphrase detection can be applied to the determination of the degree of relatedness.

The structure of the paper is the following. In section 2, we describe the related work on this task using pre-trained language models. Section 3 briefly describes the dataset. In section 4, we present our methodology. Finally, we present our results in the development and evaluation phases in section 5 and conclusions in section 6.

2 Related Work

The Sentence-BERT model (Reimers and Gurevych, 2019) is an approach that generates semantically meaningful sentence embeddings. By training BERT on siamese and triplet network structures, this approach is able to capture sentence similarity more effectively. It also reduces computational overhead compared to BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) while maintaining high accuracy in tasks such as semantic textual similarity and transfer learning.

Following this research line, there is an approach to improve BERT-based semantic embeddings for similarity tasks (Li et al., 2020). The authors propose a flow-based calibration method by transforming the original BERT embeddings into an isotropic latent space using flow. The proposed method aligns better with gold semantic similarity and reduces the influence of lexical similarity.

In this work, we use the Sentence-CROBI model, a simple architecture that combines bi-encoders and cross-encoders that was originally proposed to solve paraphrase detection. Due to its implementation facility, we adapt this model for the semantic relatedness task by only changing the task-specify

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1071–1079
June 20-21, 2024 ©2024 Association for Computational Linguistics

Language	train	dev	test
Amharic (amh)	992	95	171
Algerian Arabic (arq)	1,262	92	584
Moroccan Arabic (ary)	925	70	427
English (eng)	5,500	250	2,500
Spanish (esp)	1,562	140	600
Hausa (hau)	1,763	212	603
Marathi (mar)	1,155	293	298
Telugu (tel)	1,146	130	297
Kinyarwanda (kin)	778	102	222

Table 1: Number of instances in each train, dev and test language partition for the supervised learning track of the SemRel dataset.

block, the loss function and the pre-trained models for the cross-encoder and bi-encoder components.

3 Corpora

We briefly describe the corpora that we use to evaluate our model in the SemEval shared task 1 in this section.

The SemRel2024 dataset (Ousidhoum et al., 2024a) is a comprehensive collection of semantic textual relatedness datasets for 14 languages, predominantly spoken in Africa and Asia. These languages cover a wide range of language families and include both high-resource and low-resource languages. Each dataset consists of sentence pairs annotated by native speakers with relatedness scores ranging from 0 (completely unrelated) to 1 (maximally related). The datasets were curated by selecting pairs from various sources such as news data, Wikipedia, and conversational data to ensure diversity in topics and formality levels. The relatedness scores were generated through Best-Worst Scaling (BWS) annotations, enhancing the reliability of the rankings. Table 1 shows the SemRel dataset statistics for all languages in the supervised learning track.

It can be noticed it is a highly unbalanced dataset. Only English has more than 2,000 training examples, followed by Hausa, Spanish, Algerian Arabic, Marathi and Telugu with more than 1,000 instances and Amharic, Moroccan Arabic and Kinyarwanda with less than 1,000 examples.

4 Methodology

In this section, we describe the proposed architecture, the experimental configuration and the training details. For pre-processing the sentence pairs, we perform the same text pre-processing steps as mentioned in (Ortiz-Barajas et al., 2022).

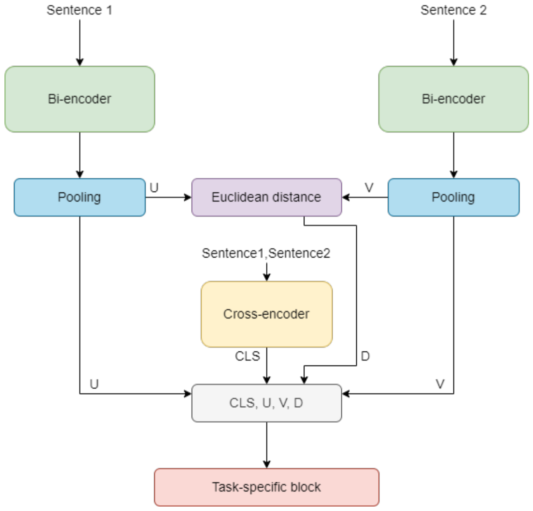

Figure 1: Diagram of the Sentence-CROBI model. U and V correspond to the individual vector representation of each text, CLS is the token classification obtained with the cross-encoder, and D is the Euclidean distance between U and V

4.1 Model

In this section, we present the Sentence-CROBI (Ortiz-Barajas et al., 2022) architecture and its implementation. The model has two main components: a bi-encoder and a cross-encoder. The bi-encoder is based on the Sentence-BERT model (Reimers and Gurevych, 2019); this is a BERT modification using a Siamese neural network that enables the model to obtain single vector representations for each text by applying a Pooling operation to the last hidden state of the bi-encoder model. We represent these vectors as u and v, respectively. The cross-encoder component receives the joint encoding of the sentence pair and is capable of capturing the relation between both texts. We use the classification token [CLS] as a final vector representation of the sequence.

We obtain a global representation of the sentence pair by concatenating the classification token [CLS] from the cross-encoder representation, the Euclidean distance D between u and v vectors, and the vectors u and v itself. This global vector is the input to a task-specific block composed of two fully connected networks with a single-neuron output. Figure 1 shows the structure of the Sentence-CROBI model.

The output of the bi-encoder component is a contextualised word embedding matrix obtained by taking the last hidden state of the component, where each row represents a word of the input sentence. In this work, we apply a mean Pooling operation, averaging all the matrix dimensions to obtain a vector representation.

Since we are working on a regression problem, the task-specific layer of our model is composed of a fully connected network featuring two layers. Initially, it accepts the global representation of sentence pairs as input, undergoing a Dropout (Hinton et al., 2012) layer with a probability of 0.1. This regularisation technique is implemented to prevent network over-fitting by randomly zeroing some input values. Subsequently, the input proceeds through a fully connected layer of 1793 units, employing a hyperbolic tangent as the activation function. Ultimately, the output layer is composed of one neuron.

We use the Mean Squared Error (MSE) as a loss function during the training of the Sentence-CROBI model. MSE quantifies the average squared difference between the predicted values and the ground truth across a dataset, which is widely used in deep learning (Bishop, 2006; Goodfellow et al., 2016). For a dataset with N samples, MSE is defined as the mean of the squared differences between predicted $\hat{y}_i$ and actual y_i values as shown in 1.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{1}$$

Notably, our task-specific block and the loss function differ from that proposed in (Ortiz-Barajas et al., 2022) as paraphrase detection entails a binary classification task. In contrast, semantic relatedness is defined as a regression task.

One of the advantages of the Sentence-CROBI model is its implementation facility that only relies on using two pre-trained models, one as a bi-encoder and the other as a cross-encoder. The selection of these models depends on the specific task and available computational resources. The implementation facility also allows the performing of fast experimentation with minor changes. These model features enable us to build solutions for all languages in Track A following the same methodology.

4.2 Data splitting

We perform K-fold cross-validation to create training and validation subsets in the development phase of the shared task. The process entails iteratively designating one of the K folds as the validation set while the remaining $K - 1$ folds collectively form the training set. This procedure is repeated K times, with each of the K folds serving as the validation set exactly once. K-fold cross-validation mitigates the impact of data partitioning on model assessment and aids in obtaining a more reliable estimate of a model's performance (James et al., 2013). We set $K = 5$ for all languages in the dataset.

4.3 Fine-Tuning

In this section, we describe our fine-tuning approaches. All approaches use a small number of epochs and a small learning rate. We train our models with a batch size of 32, a learning rate in the range $\{1 \times 10^{-5}, 2 \times 10^{-5}, 3 \times 10^{-5}\}$, and the Adam optimizer (Kingma and Ba, 2014), with a warm-up ratio of 0.06 and a linear decay to zero. We train all models for a maximum of 10 epochs and perform pseudo early stopping to use the model with the best performance on the validation data. The maximum length is 35 for individual texts and 128 for text pairs. The tokenization method differs between sentence pairs and individual texts, resulting in varying length representations. Hence, the length of each representation does not align. We use HuggingFace's Transformers library (Wolf et al., 2020) to implement the Sentence-CROBI model. Our implementation is publicly available on GitHub[1].

The first experimental setting that we use follows a monolingual approach, which means we fine-tune a model for each language of the dataset. We leveraged the HuggingFace Hub platform[2] to select bi-encoder and cross-encoder components for each model. To constrain the search space, we exclusively focused on encoder-only architectures that were either pre-trained or fine-tuned for the specific language of interest and possessed an associated paper describing the employed dataset and training details. In case there are no specific-language models, we use a multilingual model. We provide further details for the bi-encoder and cross-encoder combinations for each language in the dataset to fine-tune our model in Appendix A.

We also follow a multilingual approach to fine-tune our model. We group the languages based on their linguistic family. We consider two families.

[1] https://github.com/jgermanob/Sentence-CROBI

[2] https://huggingface.co/models

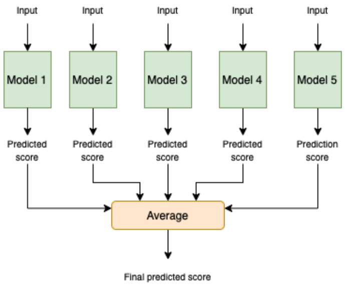

Figure 2: Bagging method diagram to obtain the final predicted score. Each model is fine-tuned using a different random seed and the final prediction is the average of all predictions.

The first one is the Semitic family, which includes Algerian and Moroccan Arabic as well as Amharic. The second one is the Indoeuropean family, which includes English, Spanish and Marathi. Telugu, Kinyarwanda and Hausa languages belong to different families; therefore, we do not include them in this approach. We concatenate each training and validation split to create each family-based split to train the models. For both families we use XLM-RoBERTa base (Conneau et al., 2020) as a cross-encoder and the multilingual uncased base version of BERT (Devlin et al., 2018) as a bi-encoder.

4.4 Ensemble Learning

In order to enhance the performance of the model in the Semantic Textual Relatedness task, we employ the Bagging method (Breiman, 1996), a strategy that mitigates generalisation errors by combining multiple models. This approach involves training different models independently and combining each output set to vote on test data and obtain the final prediction.

In the case of neural networks, differences in random initialisation or in batch generation cause independent errors in each member of the ensemble; therefore, the ensemble will perform significantly better than its members (Goodfellow et al., 2016).

We compute the final similarity score by averaging the output of each fine-tuned model with a different fold from the cross-validation splitting. Therefore, we use five distinct and independent models to obtain a final prediction. Figure 2 shows a diagram of how this method is used in this work.

5 Results

We present the results of our proposed model in the following section in the development and evaluation phases of the SemEval 2024 Task 1: Semantic Textual Relatedness.

Lang	**Val ρ (avg)**	**Dev ρ**
amh	0.4828	0.6230
arq	0.4784	0.6370
ary	0.7308	0.8030
eng	0.8709	0.8440
esp	0.5861	0.6900
hau	0.6076	0.6740
mar	0.7913	0.8470
tel	0.7290	0.8112

Table 2: Results of the proposed model in the development phase using a monolingual fine-tuning approach. We report an average of 5 runs in the validation splits used for cross-validation. We obtain the final score predictions in the development set using the bagging technique.

5.1 Development Phase

We report the average Spearman rank correlation coefficient in the validation dataset corresponding to each fold and the performance score in the development dataset reported in the Codalab page of the shared task for the development phase. We obtain the final score for each instance in the development dataset using the bagging technique and the average predictions of the five independent models for each fold.

In the case of the monolingual fine-tuning approach, we use a different model for each language in the dataset. Table 2 shows the results for each language. Half of our results in this approach achieve a performance higher than 0.80 in the performance metric, while the remaining models obtain a result above 0.60. The best performance is for the English language, with a Spearman correlation coefficient of 0.844. In contrast, the lowest performance is for the Amharic language, with a Spearman correlation coefficient of 0.623.

Due to the imbalance present in the dataset, we employed a multilingual fine-tuning approach by grouping languages into linguistic families. In this approach, we considered two groups: the Semitic (Sem) languages and the Indoeuropean (IE) languages.

Table 3 shows the results using the multilingual fine-tuning approach. There is a performance decrease in 6 of 8 considered languages. In the case of the Indoeuropean family, our model obtains a Spearman correlation coefficient of 0.8191 for En-

Lang	Fam	Val ρ (avg)	Dev ρ
eng	IE	0.8079	0.8191
esp	IE	0.8079	0.6874
mar	IE	0.8079	0.8290
amh	Sem	0.6926	0.8223
arq	Sem	0.6926	0.4727
ary	Sem	0.6926	0.8519

Table 3: Results of the proposed model in the development phase using a multilingual fine-tuning approach. We report an average of 5 runs in the validation splits used for cross-validation. We obtain the final score predictions in the development set using the bagging technique.

glish, which represents a 0.0249 decrease; in the case of Spanish and Marathi, our model decays 0.0026 and 0.018, respectively. In the case of the Semitic family, the multilingual fine-tuning approach improves the model performance in 2 of 3 considered languages: Moroccan Arabic (Moroc. A.) and Amharic. The model increases its performance from 0.803 to 0.8519 in Moroccan Arabic and from 0.623 to 0.8223 in Amharic, which represents a 0.1993 performance improvement in terms of Spearman correlation coefficient.

We must mention that we did not report any results for the Kinyarwanda language in the development phase because it was added to Track A later (December 12, 2024). Therefore, we were unable to conduct any experiments prior to the evaluation phase.

5.2 Evaluation Phase

We select the best-performing model for each language in the evaluation phase of the shared task. We use a monolingual fine-tuning approach for Algerian Arabic, English, Spanish, Hausa, Marathi, Telugu and Kinyarwanda, as well as a multilingual approach for Amharic and Moroccan Arabic. We create a new training set for each language and family by adding the development subset and its gold scores released by the shared task organisers. We train five independent models for each language and obtain the final score predictions using the bagging technique.

Table 4 shows the results of our proposed model in the evaluation test, its comparison with the baseline and the final ranking in the shared task for each language. We add a * to denote a multilingual-fine-tune-based approach. Our model outperforms the baseline in English and Moroccan Arabic with a difference from the leaders of 0.0183 and 0.0765, respectively. The lowest performance of the proposed model is in the Kinyarwanda language, with a Spearman correlation coefficient of 0.4585 and a difference from the leader of 0.3584.

Lang	Score	Baseline	Rank	Highest score
amh *	0.8398	0.85	7/18	0.8886
arq	0.5407	0.6	11/24	0.6823
ary *	0.7861	0.77	13/23	0.8625
eng	0.8316	0.83	16/36	0.8499
esp	0.6968	0.7	11/25	0.7403
hau	0.6702	0.69	9/21	0.7642
mar	0.8669	0.88	11/25	0.9108
tel	0.7847	0.82	17/25	0.8733
kin	0.4585	0.72	16/21	0.8169

Table 4: Results of the proposed model in the evaluation phase using monolingual and multilingual fine-tuning approaches compared with the baseline and the highest score. We obtain the final score predictions in the development set using the bagging technique. * Denotes a multilingual approach.

We perform an error analysis of our model's performance in the evaluation dataset for each language in Appendix B. The analysis suggests that the global vector representation of the sentence pair has a limited capacity to capture other semantic relationships between the texts apart from similarity, and future work should follow this direction. Nevertheless, it is essential to highlight that only the task-specific block should change, which illustrates the high adaptability capacity of the model.

6 Conclusions

This work presents the Sentence-CROBI model and its adaptation to the SemEval 2024 Task 1: Semantic Textual Relatedness. We evaluate the model's capacities in monolingual and multilingual fine-tuning approaches to measure its performance and adaptability across diverse linguistic families, yielding acceptable performance in low and mid-resource languages. Ensemble techniques further enhance the robustness and reliability of the model's predictions. Overall, the findings underscore the model's capacity for solving relatedness detection tasks, emphasising its versatility in accommodating linguistic variations and resource constraints.

References

Christopher Bishop. 2006. Pattern recognition and machine learning. *Springer google scholar*, 2:5–43.

Leo Breiman. 1996. Bagging predictors. *Machine learning*, 24(2):123–140.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. Deep learning, volume 1.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An introduction to statistical learning*, volume 112. Springer.

Raviraj Joshi. 2022. L3cube-mahacorpus and mahabert: Marathi monolingual corpus, marathi bert language models, and resources. In *Proceedings of The WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 97–101, Marseille, France. European Language Resources Association.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif M. Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *CoRR*, abs/1203.1858.

Jesus-German Ortiz-Barajas, Gemma Bel-Enguix, and Helena Gómez-Adorno. 2022. Sentence-CROBI: A simple cross-bi-encoder-based neural network architecture for paraphrase identification. *Mathematics*, 10(19):3578.

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.

Hariom A. Pandya, Bhavik Ardeshna, and Dr. Brijesh S. Bhatt. 2021. Cascading adaptors to leverage english data to improve performance of question answering for low-resource languages.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ali Safaya, Moutasem Abdullatif, and Deniz Yuret. 2020. KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2054–2059, Barcelona (online). International Committee for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Abinew Ali Ayele, Gopalakrishnan Venkatesh, Ibrahim Gashaw, and Chris Biemann. 2021. Introducing various semantic models for amharic: Experimentation and evaluation with multiple tasks and datasets. *Future Internet*, 13(11).

A Monolingual approach

We use a monolingual fine-tuning approach for Track A, which means we fine-tune a model for each language in the dataset as described in section 4.3. We only consider publicily avaliable models in the HuggingFace Hub [3] that were either pre-trained or fine-tuned for the specific language and possess and associated paper describing the dataset as well as the training details.

Table 5 shows the bi-encoder and cross-encoder combinations for each language in the dataset to fine-tune our model following the monolingual approach. Following (Ortiz-Barajas et al., 2022) methodology, we choose a RoBERTa-based model for the cross-encoder and a BERT-based model for the bi-encoder. Only in the case of the Hausa language do we use a multilingual combination of bi-encoder and cross-encoder models because there are no available pre-trained or fine-tuned models that made our criteria.

B Error Analysis

We perform an error analysis of our model's performance in the evaluation dataset for each language. It is essential to mention that Spanish is excluded because the organisers do not provide the gold scores for this language.

Table 6 shows the differences between our model's predictions and the gold scores for each language in the evaluation dataset. We compute the difference by subtracting each example's predicted score from the gold score. Therefore, a negative difference means that our model predicts a higher score than the gold score, whereas a positive difference means that our model predicts a lower score than the gold score. The negative differences are higher than the positive differences in all languages. This result indicates that our model predicts a higher semantic textual relatedness score than the actual relatedness score in all cases.

Table 7 shows the top-5 negative differences predicted by the Sentence-CROBI model in the English evaluation dataset; that is, the model predicts a higher score than the gold score. It is possible to observe a high semantic similarity between the texts in the first four examples, and they can be considered paraphrases. Therefore, our model captures only one kind of semantic relatedness in these examples.

Table 8 shows the top-5 positive differences predicted by the Sentence-CROBI model in the English evaluation dataset; that is, the model predicts a lower score than the gold score. It is possible to observe different types of semantic relatedness that differ from semantic similarity between the texts. In the first example, the texts are semantic contrastive; the first text hints at excitement, while the second portrays boredom. The texts in the second example describe similar situations where a person performs some public activity. In the third example, both texts offer insights into events or situations concerning government or administration within a specific historical context. The semantic relatedness between the texts in the fourth example is their shared focus on the reading experience and the consideration of delving into further books within a series. Finally, the semantic relatedness in the fifth example lies in their depiction of situations involving young children, albeit with distinct tones and activities.

[3] https://huggingface.co/models

Lang	cross-encoder	bi-encoder
amh	Am-RoBERTa (Yimam et al., 2021)	mBERT-base FT on amharic-CC100 (Conneau et al., 2020)
arq	XLM-RoBERTa-base Arabic (Pandya et al., 2021)	BERT-base Arabic (Safaya et al., 2020)
ary	XLM-RoBERTa-base Arabic (Pandya et al., 2021)	BERT-base Arabic (Safaya et al., 2020)
eng	RoBERTa-large (Liu et al., 2019)	BERT-base (Devlin et al., 2018)
esp	BERTIN (la Rosa et al., 2022)	BETO (Cañete et al., 2020)
hau	XLM-RoBERTa-base (Conneau et al., 2020)	mBERT-base (Devlin et al., 2018)
mar	Marathi-RoBERTa (Joshi, 2022)	Marathi-BERT (Joshi, 2022)
tel	XLM-RoBERTa-base (Conneau et al., 2020)	Telugu-BERT (Joshi, 2022)

Table 5: Bi-encoder and cross-encoder model combinations for each language in the dataset using a monolingual fine-tuning approach.

Lang	Negative difference	Positive difference
amh	111	60
arq	335	246
ary	225	201
eng	1604	996
hau	314	289
kin	314	289
mar	238	60
tel	167	130

Table 6: Negative and positive differences in the scores predicted by our model concerning the gold score in the evaluation dataset for each language. A negative difference means that our model predicts a higher score than the gold score, whereas a positive difference means that our model predicts a lower score than the gold score.

Text 1	Text 2	Pred score	Gold score	abs diff
In general conversation , aerosol usually refers to an aerosol spray can or the output of such a can	When they say aerosol most people mean an aerosol spray can or the spray it makes	0.8610	0.44	0.4210
Ciampi was born in Livorno(Province of Livorno)	Carlo Azeglio Ciampi was born in 1920 in Livorno , Italy	0.7860	0.39	0.3960
TAKE A Shower then talk to her	I advise you to have a shower before speaking with her	0.8354	0.44	0.3954
if there 's a reason , we 'll discuss it	if you have a legitimate reason , we will discuss it	0.9060	0.52	0.3860
Forget that this is YA lit and READ IT	It's OK for what it is but you definitely won't forget you're reading a YA novel	0.7010	0.32	0.3810

Table 7: Top-5 negative differences predicted by the Sentence-CROBI model in the English evaluation dataset; that is, the model predicts a higher score than the gold score.

Text 1	Text 2	Pred score	Gold score	abs diff
A lot of this book is setting up the last book	This book is beige wallpaper	0.2798	0.64	0.3602
A man with glasses is playing his instrument in a small crown of people that includes another man in a suit with a trumpet	A man holding his arms out horizontally, and gripping a fencing sword in his right hand, as people in the background do the same thing	0.3780	0.69	0.3120
This date was January 3, 1867, which was two weeks before the beginning of the first administrative year of Governor Gove Saulsbury	Currently the distribution of the Senate Assembly seats was made to three senators for each of the three counties	0.2890	0.60	0.3110
i found it different from many other books i've read	I am trying to decide whether to read the other books in the series	0.4192	0.072	0.3008
A young boy wearing a red winter coat is eating and holding up a candy bar	A young baby boy crying while wearing a shirt that says ""I am the BOSS	0.3433	0.63	0.2867

Table 8: Top-5 positive differences predicted by the Sentence-CROBI model in the English evaluation dataset; that is, the model predicts a lower score than the gold score.