Multi-lingual Semantic Text similarity on 3 BERT variants

Shravan Gadbail
Masumi Desai
Rajath Gadagkar

April 21, 2025

1 Introduction

The Semantic Textual Similarity (STS) task is crucial for understanding semantic relationships between sentences across languages. With the advent of multilingual transformer models, our project aims to evaluate and compare the performance of three state-of-the-art models: XLM ROBERTa, DistilBERT, and mBERT, when fine-tuned on a multilingual STS benchmark. This interim report covers the work completed thus far and proposes directions for future improvements.

2 Methodology

2.1 Models and Training Setup

In this project, we have fine-tuned three transformer models: XLM ROBERTa, DistilBERT, and mBERT. Each model was trained on the multilingual STS benchmark dataset. The following table summarizes the hyperparameter settings used for each model.

Table 1: Hyperparameters for Each Model

Model	**Learning Rate**	**Batch Size**	**Epochs**	**Optimizer**
XLM ROBERTa	2e-5	16	3	AdamW (Default)
DistilBERT	2e-5	16	3	AdamW (Default)
mBERT	2e-5	16	3	AdamW (Default)

2.2 Dataset and Preprocessing

The multilingual STS benchmark dataset consists of pairs of sentences annotated with similarity scores. Both the dev and test sets were used to evaluate the models. Preprocessing steps such as tokenization were uniformly applied across all languages to ensure consistency in training and evaluation.

3 Experiments and Results

For the evaluation, we computed the following metrics on both the dev and test sets:

- **Spearman Correlation**
- **Cosine Similarity**
- **Mean Squared Error (MSE)**
- **Pearson Correlation**

3.1 Comprehensive Results Table

The table below summarizes the performance of XLM ROBERTa, DistilBERT, and mBERT across all the above metrics.

Table 2: Comprehensive Evaluation Metrics for All Models on Dev and Test Sets

Model	**Spearman**		**Cosine**		**MSE**		**Pearson**	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
XLM ROBERTa	0.87	0.82	0.96	0.96	0.57	0.75	0.87	0.83
DistilBERT	0.807	0.76	0.949	0.947	0.79	0.949	0.81	0.77
mBERT	0.85	0.80	0.96	0.96	0.65	0.84	0.85	0.80

4 Discussion and Analysis

Our evaluation compared three multilingual transformer models—XLM ROBERTa, DistilBERT, and mBERT—fine-tuned on the STS benchmark dataset. The evaluation metrics (Spearman and Pearson correlations, Cosine similarity, and Mean Squared Error) on both the development (dev) and test sets reveal several interesting trends and insights.

- XLM ROBERTa: XLM ROBERTa achieved the highest Spearman and Pearson correlations on the dev set (0.87 and 0.87, respectively) and maintained strong performance on the test set (0.82 and 0.83). Additionally, its cosine similarity values were among the highest (0.96 on both splits), and it had the lowest dev-set MSE (0.57). This suggests that XLM ROBERTa has excellent linear and rank correlation with human-annotated similarity scores and generates robust sentence embeddings, potentially due to its larger model size and sophisticated cross-lingual training regime.
- DistilBERT: DistilBERT, as a compact and lightweight model, showed lower Spearman and Pearson correlation values (around 0.807–0.81 on dev and 0.76–0.77 on test) compared to XLM ROBERTa and mBERT. Its cosine similarity scores were slightly lower (0.949–0.947), and the MSE was the highest among the three (0.79 on dev and 0.949 on test). While DistilBERT offers faster inference and lower computational cost, its trade-off appears to be in capturing the fine-grained semantic nuances, which impacts its regression performance on STS.
- mBERT: Our mBERT-based model achieved intermediate results. It recorded Spearman and Pearson correlations of 0.85 (dev) and 0.80 (test), which are quite competitive with XLM ROBERTa, although it exhibited a slightly higher MSE (0.65 on dev and 0.84 on test) compared to XLM ROBERTa. The cosine similarity values for mBERT were equivalent to those of XLM ROBERTa (0.96). These results indicate that mBERT performs robustly across multiple languages, capturing semantic relationships effectively even if it sometimes incurs a minor increase in error magnitude (MSE).

Factors Contributing to the Differences: Several factors might explain the differences in performance across these models

- Model Architecture and Capacity: XLM ROBERTa, typically larger and equipped with advanced cross-lingual pre-training, is better at understanding complex language patterns. In contrast, DistilBERT is a distilled version with fewer parameters that sacrifices some representational power for speed. mBERT strikes a balance, offering good cross-lingual performance with a relatively standard architecture.
- Training Data and Pre-training Objectives: Although all models were fine-tuned on the same multilingual STS dataset, the initial pre-training differs significantly. XLM

ROBERTa's extensive pre-training on multiple languages might give it an edge in capturing subtle semantic variations, while DistilBERT's distillation process may lead to some loss in these detailed representations.
- Embedding Quality and Regression Head Calibration: The high cosine similarity scores for both mBERT and XLM ROBERTa suggest that the underlying sentence embeddings are well-aligned with human judgment of similarity. However, slight differences in the calibration of the regression head (i.e., the mapping of embeddings to the scalar similarity score) might contribute to variations in MSE and correlation values.
- Language Coverage and Balance: The dataset is composed of translations across multiple languages. mBERT's relatively balanced performance across languages might contribute to its stable performance metrics, whereas models with specific pre-training focuses (or those that are more parameter-efficient, like DistilBERT) may not capture this multilingual diversity as effectively.

Implications for Future Work: The analysis suggests that while XLM ROBERTa currently leads in terms of correlation and low MSE, mBERT offers competitive performance and could be further improved, for example, through methods such as LoRA or other parameter-efficient fine-tuning strategies. On the other hand, despite DistilBERT's slightly lower performance, its inference speed and lightweight design make it a strong candidate for applications where computational resources are limited.

5 Future Work

Building on our current findings, we plan to expand our project in the following directions:

1. **Semantic Search Interface:** Develop a user-friendly interface that leverages our pretrained models for semantic search between words. This interface will allow users to input queries and receive semantically similar words or phrases based on model embeddings.
2. **Plagiarism Checker:** Develop an automated plagiarism detection tool that leverages our pretrained STS model. This tool will compare text documents to identify similar content and potential plagiarism by analyzing semantic similarity between sentences and paragraphs.
3. **Linguistically Motivated Features:** Enhance our models by integrating at least three linguistically motivated features. These may include:
   (a) **Part-of-Speech (POS) Tagging:** Provides grammatical context by identifying the role of each word in a sentence, which aids in aligning similar syntactic structures.
   (b) **Dependency Parsing:** Captures the syntactic relationships between words, offering insights into the structural composition of a sentence.
   (c) **Named Entity Recognition (NER):** Detects and classifies key entities such as names, locations, and organizations, which is valuable for understanding semantic content.
   (d) **Semantic Role Labeling (SRL):** Identifies the predicate-argument structure in sentences, clarifying the roles of words and phrases in conveying meaning.
   (e) **Word Sense Disambiguation:** Determines the intended meaning of words in context, essential for accurately comparing semantic content in sentences.
4. **Cross-Lingual STS Model:** Extend our work to develop a cross-lingual STS model that effectively transfers knowledge between languages. This will involve exploring transfer learning techniques and possibly incorporating additional multilingual data.

6 Conclusion

This interim report summarizes the work done so far in fine-tuning three transformer models for the multilingual STS task. With preliminary experiments conducted and results summarized in terms of Spearman correlation, Cosine similarity, MSE, and Pearson correlation, the groundwork is laid for further experiments. The future work outlined above will help in addressing current limitations and pushing the state-of-the-art further.

7 Links

Link to full XLM model: Click here
Link to full DistilBERT model: Click here
Link to full mBERT model: Click here