# TECHSSN at SemEval-2024 Task 1: Multilingual Analysis for Semantic Textual Relatedness using Boosted Transformer Models

**Shreejith Babu G, Ravindran V, Aashika Jetti**
**Rajalakshmi Sivanaiah, Angel Deborah S**
Department of Computer Science and Engineering
Sri Sivasubramaniya Nadar College of Engineering
Chennai - 603110, Tamil Nadu, India
{shreejithbabu2213006, ravindran2213003, aashika2210193}@ssn.edu.in,
{rajalakshmis, angeldeborahs}@ssn.edu.in

## Abstract

This paper presents our approach to SemEval-2024 Task 1: Semantic Textual Relatedness (STR). Out of the 14 languages provided, we specifically focused on English and Telugu. Our proposal employs advanced natural language processing techniques and leverages the Sentence Transformers library for sentence embeddings. For English, a Gradient Boosting Regressor trained on DistilBERT embeddings achieves competitive results, while for Telugu, a multilingual model coupled with hyperparameter tuning yields enhanced performance. The paper discusses the significance of semantic relatedness in various languages, highlighting the challenges and nuances encountered. Our findings contribute to the understanding of semantic textual relatedness across diverse linguistic landscapes, providing valuable insights for future research in multilingual natural language processing.

## 1 Introduction

Semantic Textual Relatedness (STR) is a pivotal aspect of natural language processing (NLP) that underlies the foundation of various language-related tasks (Ousidhoum et al., 2024a,b). This task takes this challenge to a global scale by encompassing 14 languages, including Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Morrocan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu. This multilingual approach transcends linguistic boundaries, fostering collaboration and research within the NLP community. Exploring a diverse array of languages in this endeavor encourages the development of models adept at handling the distinct linguistic nuances inherent in each language. This progress fosters a more inclusive and universally applicable approach to natural language processing research.

SemEval-2024 Task 1 delves into the automated detection of semantic relatedness between pairs of sentences, a foundational aspect for unraveling meaning. The task's embrace of multiple languages is crucial, encompassing a spectrum of linguistic characteristics. This inclusivity fosters a collaborative atmosphere for researchers, pushing them to craft models adept at capturing semantic nuances across diverse linguistic landscapes. The task's importance extends to the assessment and benchmarking of sentence representation methods, pivotal for numerous NLP applications. STR evaluated through tasks, play a crucial role in areas such as question answering, summarization, and information retrieval. The task's outcomes serve as a benchmark, guiding the development and refinement of models that can effectively discern the relatedness of sentences, regardless of language.

When approaching this task, we concentrated on two languages: English and Telugu. Employing the Sentence Transformers library, we make use of the DistilBERT model for generating sentence embeddings in English and opt for a multilingual model for handling Telugu. The choice of state-of-the-art models and embedding techniques underscores our commitment to developing robust solutions capable of handling the linguistic diversity presented in this task.

Participating in this task has revealed crucial insights into the intricacies of semantic relatedness across languages. Our system demonstrated better performance, particularly in English, achieving noteworthy Spearman correlation coefficients on the test set. However, the challenges surfaced in capturing subtle nuances in semantic relations, especially in the context of Telugu. This highlights the necessity for specialized methodologies to address the complexities of multilingual semantic relatedness. As we explore the methodology, experiments, and results in the following sections, we delve deeper into the intricacies of our approach, providing a comprehensive understanding of how our model navigates the challenges posed by the



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 907–912

task.

## 2 Background

The task at hand revolves around Semantic Textual Relatedness (STR), focusing on evaluating the degree of semantic closeness between sentences in both English and Telugu. Unlike emotion recognition, this task delves into understanding the relationships between sentences rather than categorizing emotions in code-mixed interactions. In this scenario, the input comprises pairs of sentences in either English or Telugu, and the objective is to determine the relatedness score for each pair on a scale from 0 to 1.

For instance, consider the following English sentence pair:
Sentence 1: "The sun is setting over the horizon, casting a warm glow on the city."
Sentence 2: "As the day comes to an end, the sun sets, and the city is bathed in a warm glow."
Score Label: 0.85
In this example, the relatedness score of 0.85 indicates a high degree of semantic closeness between the two sentences, as they convey similar information about the sunset and the warm glow of the city.
Sentence 1: "The scientific method involves systematic observation and experimentation."
Sentence 2: "Bicycles are a popular mode of transportation in urban areas."
Score Label: 0.15

In this example, the relatedness score of 0.15 indicates a low degree of semantic closeness between the two sentences. Similarly, a dataset exists for Telugu.

The datasets used for this task include pairs of sentences in English and Telugu, capturing the real-world scenario of diverse linguistic interactions. These datasets are annotated with relatedness scores, providing a basis for training and evaluating models effectively. The input parameters consist of the sentence pairs, and the output involves predicting the relatedness score for each pair. The multilingual nature of the task fosters collaboration and research across linguistic boundaries, contributing to a more inclusive and globally applicable approach in the NLP community.

## 3 Related Work

Palakorn Achananuparp et. al.presents an evaluation of fourteen existing text similarity measures (Achananuparp et al., 2008). The ability to accurately judge the similarity between natural language sentences is crucial for various applications such as text mining, question answering, and text summarization. The evaluation encompasses three different datasets: TREC9 question variants, Microsoft Research paraphrase corpus, and the third recognizing textual entailment dataset. The study explores three classes of measures: word overlap, TF-IDF, and linguistic measures. The goal is to judge sentence pairs based on the notion that they have identical meanings, considering factors such as paraphrase or entailment. They address the challenges of computing sentence similarity, highlighting the importance of recognizing semantic equivalence beyond surface form comparisons.

Pantulkar Sravanthi and B. Srinivasu, address the challenge of measuring sentence similarity, emphasizing the importance of semantic similarity over syntactic measures (Sravanthi and Srinivasu, 2017). They introduce three semantic similarity approaches—cosine similarity, path-based (Wu–Palmer and shortest path), and feature-based. The feature-based approach incorporates WordNet, tagging, and lemmatization, showing superior performance in generating semantic scores. This study contributes valuable insights into semantic similarity measures and can enhance the understanding of feature-based approaches based on WordNet in sentence categorization.

Syed S. Akhtar et al. (Akhtar et al., 2017) address the need for word similarity datasets in Indian languages, specifically Urdu, Telugu, Marathi, Punjabi, Tamil, and Gujarati. They introduce manually annotated monolingual word similarity datasets for these languages, created through translation and re-annotation of English datasets. The paper presents baseline scores for word representation models using state-of-the-art techniques for Urdu, Telugu, and Marathi, evaluated on the newly created datasets. This work contributes valuable resources for evaluating word representations in Indian languages, fostering the development of techniques leveraging word similarity.

## 4 System Overview

To optimize efficiency, we methodically integrated numerous critical algorithms and modeling decisions into our semantic textual relatedness model.

### 4.1 Data Preprocessing

#### 4.1.1 Text Cleaning

In the initial phase of our preprocessing pipeline, (Kadhim, 2018) we address the cleanliness of the textual data for both Telugu and English. For Telugu, we employ a language-specific approach, utilizing a tokenizing function tailored to the Telugu script. This ensures the proper segmentation of words while also excluding unwanted elements such as punctuation, special characters, and digits. Similarly, for English, we apply standard tokenization techniques to achieve a clean and well-structured representation, eliminating extraneous symbols and numerical values. This initial cleaning step lays the foundation for subsequent language-specific processing.

#### 4.1.2 Language-specific Tokenization

Recognizing the distinct linguistic features of Telugu and English, we implement language-specific tokenization methods. In the case of Telugu, we adapt tokenization to the unique script and structural characteristics of the language. This approach ensures the accurate representation of Telugu text for downstream tasks. Conversely, for English, we rely on conventional tokenization techniques suited for the Latin script. By tailoring tokenization to the linguistic attributes of each language, we pave the way for more effective and contextually rich representations in subsequent stages of the preprocessing pipeline.

#### 4.1.3 Stop Word Removal

Stopwords were removed from the combined tokens in order to enhance the model's focus on pertinent content. Through the removal of noise and refinement of the raw data, a deeper comprehension of the underlying sentiment was made possible.

#### 4.1.4 Data Splitting

The `train_test_split` function from the `scikit-learn` library was used to split the preprocessed data into training and testing sets. This made it possible to thoroughly assess the model's capacity for generalization using data that had never been seen before.

### 4.2 Model Architecture

#### 4.2.1 English

**Embedding with DistilBERT:** To capture semantic meanings, we utilized the DistilBERT model (version: distilbert-base-uncased) (Kici et al., 2021) from the Sentence Transformers library to generate sentence embeddings. DistilBERT is a distilled version of the BERT model, designed for faster inference while maintaining competitive performance. Sentences were encoded into embeddings using DistilBERT, facilitating the creation of robust representations. These embeddings served as the input features for subsequent relatedness score prediction.

**Gradient Boosting Regressor Model:**

To model the relatedness scores, we employed the Gradient Boosting Regressor algorithm. Specifically, we utilized the GradientBoostingRegressor class from the scikit-learn library (version: 0.24.2). The model was trained on the encoded sentences and evaluated on the test set.

In the process of hyperparameter tuning, the following parameters were optimized:
Learning rate: 0.05
Number of estimators: 200
Maximum depth of each estimator (max depth): 3
Subsample ratio of the training instances (subsample): 0.8

These hyperparameters were chosen based on a grid search conducted to maximize the model's effectiveness in predicting semantic textual relatedness for English sentences. Specifically, the learning rate controls the contribution of each tree in the ensemble, while the number of estimators determines the number of boosting stages. Additionally, the maximum depth of each tree and the subsample ratio influence the depth of the individual trees and the sampling strategy, respectively.

**Model Persistence and Reporting:** The trained Gradient Boosting Regressor model is saved for future use. Spearman correlation on the test set provides a quantitative measure of the model's ability to predict sentence relatedness. The model's performance, including the correlation coefficient, is printed for further analysis.

#### 4.2.2 Telugu

**Multilingual Sentence Embeddings:** For Telugu, we opt for a pre-trained multilingual model (paraphrase-multilingual-MiniLM-L12-v2) to generate sentence embeddings. The Telugu dataset is encoded into embeddings using this multilingual model. The Telugu model is trained using a Gradient Boosting Regressor, and its performance is

evaluated on the test set using the Spearman correlation coefficient.

**Advanced Model Tuning:** We used an approach similar to one used for the English dataset where hyperparameter tuning is performed using a grid search for the Gradient Boosting Regressor on Telugu data. The best model is selected based on the optimal combination of hyperparameters, leading to improved performance. The chosen model is subsequently applied for prediction and evaluation.

### 4.3 Model Evaluation

#### 4.3.1 Performance Metrics

To gauge the performance of the English-relatedness detection model, we utilized a comprehensive set of metrics, incorporating the Spearman correlation coefficient. These metrics collectively offered a well-rounded understanding of the model's accuracy, precision, and capacity to capture the subtleties of relatedness among English sentences. Following a similar evaluation approach for the Telugu-relatedness detection model, we subjected it to a thorough assessment using performance metrics. The Spearman correlation coefficient served as a valuable measure to assess the accuracy and precision of the model in capturing relatedness between Telugu sentences.

#### 4.3.2 Prediction on the Test Set

We employed the trained model to predict textual relatedness on a test set. This involved passing the preprocessed test data through the model and deciphering the anticipated labels for subsequent analysis.

## 5 Experimental Setup

### 5.1 Data Preprocessing for Subtask 1a (English)

For the English dataset, we adopted a two-step preprocessing approach. Initially, sentences underwent precise tokenization using the Sentence Transformers library. This process harnessed the advanced capabilities of DistilBERT to encode sentences into dense embeddings, establishing the foundation for subsequent model training. The selected Gradient Boosting Regressor model underwent training on these embeddings, contributing to improved semantic textual relatedness.

### 5.2 Data Preprocessing for Subtask 1b (Telugu)

For Telugu, we implemented dedicated preprocessing, involving nuanced tokenization facilitated by the Natural Language Toolkit (NLTK). This process was complemented by a careful removal of stopwords, specifically tailored for Telugu text. Subsequently, the (Gillioz et al., 2020) Sentence Transformer's multilingual model came into play, encoding Telugu sentences into high-dimensional embeddings. These embeddings formed the basis for training a Gradient Boosting Regressor model. Importantly, hyperparameter tuning played a pivotal role in fine-tuning the model's performance specifically for Telugu.

### 5.3 Hyperparameter Tuning

For Subtask 1a, our focus on hyperparameter tuning aimed to optimize the performance of the Gradient Boosting Regressor model on the English dataset. The primary goal was to fine-tune the model for improved semantic textual relatedness prediction.

In Subtask 1b, specifically for Telugu, we conducted a thorough hyperparameter tuning process using GridSearchCV. Key hyperparameters such as estimators, learning rate, max-depth, and subsample were carefully explored to enhance the model's efficacy. This step was intended to fine-tune the Gradient Boosting Regressor model for optimal performance on the Telugu dataset.

### 5.4 External Tools / Libraries

External tools and libraries played a pivotal role in our experimentation. Sentence Transformers (v2.0.0) with its sophisticated capabilities was instrumental in encoding sentences into high-dimensional embeddings. NLTK (v3.6.3) facilitated precise tokenization and stopword removal, contributing to meticulous linguistic preprocessing. Scikit-Learn (v0.24.2) emerged as the preferred library for machine learning models and hyperparameter tuning, providing a standardized and comprehensive experimental framework.

### 5.5 Evaluation Metric

The evaluation metric of choice was the Spearman correlation coefficient. Renowned for its ability to discern the monotonic relationship between predicted scores and gold standard scores, this metric offered a nuanced assessment of semantic textual relatedness.

| Approach | Accuracy |
|---|---|
| distilbert-base-uncased (English) | 0.57 |
| paraphrase-multilingual-MiniLM-L12-v2 (Telugu) | 0.527 |

Table 1: Comparison of Accuracy for Different Approaches

## 6 Results

The system's performance was assessed using a regression model that predicts similarity scores between English sentences. The Spearman Correlation Coefficient is the major quantitative finding, measuring the monotonic relationship between expected and actual similarity scores. The Spearman correlation coefficient, often denoted as (rho), is a statistical measure used to assess the strength and direction of the monotonic relationship between two variables. Specifically, in the context of evaluating models, the Spearman coefficient is employed to quantify the association between predicted scores and actual scores.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (1)$$

Table 1 shows the Spearman score of the models used for English and Telugu. The English model, employing advanced natural language processing techniques and utilizing DistilBERT embeddings, demonstrated noteworthy performance on the test set. The Spearman correlation coefficient, a key indicator of the model's ability to predict relatedness between English sentences, was calculated. Our model achieved a Spearman correlation coefficient of 0.57 on the test set, showcasing its effectiveness in capturing semantic nuances and relationships within English sentence pairs.

The model's performance is influenced by the quality of the preprocessing applied to the text data. Any limitations or challenges encountered during the preprocessing stage, such as handling rare words or specific language nuances, should be discussed. The performance of our Telugu model in predicting semantic relatedness scores using a fine-tuned Gradient Boosting Regressor with hyperparameter tuning and embeddings from the Sentence Transformer model 'paraphrase-multilingual-MiniLM-L12-v2' is evaluated here. On the test set, our Telugu model achieved a Spearman correlation coefficient of 0.527. This result signifies a strong positive monotonic relationship between the predicted relatedness scores and the actual scores. The Spearman correlation coefficient is a crucial indicator of the model's ability to capture the underlying trends in sentence similarity within the Telugu language. The hyperparameter tuning process identified the following optimal hyperparameters for the Gradient Boosting Regressor on the Telugu dataset: learning rate : 0.05, max depth: 3, n_estimators: 200, subsample: 0.8 . These hyperparameters represent the configuration that maximizes the model's effectiveness in predicting relatedness scores for Telugu sentences.

## 7 Conclusion

Our system, anchored in the robust capabilities of Sentence Transformers and Gradient Boosting Regressor models, showcased a better performance in predicting semantic textual relatedness. The fusion of advanced tokenization, embeddings, and hyperparameter tuning resulted in a model finely attuned to the intricacies of the English and Telugu languages.

The results on the evaluation metrics, particularly the Spearman correlation coefficient, underscore the efficacy of our approach in capturing the nuances of semantic relatedness across diverse sentence pairs. The successful adaptation to multiple languages, as evident in the Telugu experiments, showcases the versatility of our system.

For future work, delving deeper into language-specific processing techniques and exploring more sophisticated models may unlock additional performance gains. Additionally, expanding the system's applicability to handle a broader array of languages and domains could further enhance its utility in real-world applications. Overall, our endeavors open avenues for continuous refinement and exploration in the realm of semantic textual relatedness prediction.

## References

Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. 2008. The evaluation of sentence similarity measures. In *Data Warehousing and Knowledge Discovery: 10th International Conference, DaWaK 2008 Turin, Italy, September 2-5, 2008 Proceedings 10*, pages 305–316. Springer.

Syed Sarfaraz Akhtar, Arihant Gupta, Avijit Vajpayee, Arjit Srivastava, and Manish Shrivastava. 2017.

Word similarity datasets for indian languages: Annotation and baseline systems. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 91–94.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE.

Ammar Ismael Kadhim. 2018. An evaluation of preprocessing techniques for text classification. *International Journal of Computer Science and Information Security (IJCSIS)*, 16(6):22–32.

Derya Kici, Garima Malik, Mucahit Cevik, Devang Parikh, and Ayse Basar. 2021. A bert-based transfer learning approach to text classification on software requirements specifications. In *Canadian Conference on AI*, volume 1, page 042077.

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. *Preprint*, arXiv:2402.08638.

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.

Pantulkar Sravanthi and B Srinivasu. 2017. Semantic similarity between sentences. *International Research Journal of Engineering and Technology (IRJET)*, 4(1):156–161.