# BITS Pilani at SemEval-2024 Task 1: Using text-embedding-3-large and LaBSE embeddings for Semantic Textual Relatedness

**Dilip Venkatesh**[1] and **Sundaresan Raman**[1]

[1]Birla Institute of Technology and Science, Pilani, Rajasthan, India
{f20201203, sundaresan.raman}@pilani.bits-pilani.ac.in

## Abstract

Semantic Relatedness of a pair of text (sentences or words) is the degree to which their meanings are close. The Track A of the Semantic Textual Relatedness shared task aims to find the semantic relatedness for the English language along with multiple other low resource languages with the use of pretrained language models. We propose a system to find the Spearman coefficient of a textual pair using pretrained embedding models like **text-embedding-3-large** and **LaBSE**.

## 1 Introduction

**Semantic relatedness** is defined as the degree of closeness of textual units (sentences, words, paragraphs) (Mohammad, 2008, Mohammad and Hirst, 2012). This makes semantic relatedness an important metric to understand the meaning of text. A paragraph is a string of multiple related sentences and similar paragraphs in a sequential manner form passages or documents which provides valuable information. Understanding this cohesion among sentences (Bernhardt, 1980) and passages is critical for understanding meaning and therefore generating more powerful natural language processing systems. We consider text to be semantically close if there is some sort of similar meaning. We can see an example of textual relatedness in Table 1

We make an important differentiation between **Semantic relatedness** and **Semantic Similarity**. Semantic similarity is when two textual units are synonymous, hyponymous, antonymous, or troponymous relation between them (Abdalla et al., 2023). Semantic relatedness consists of when there is a lexical relation between two units of text *conductor-orchestra, teacher-book*.

Since semantic relatedness is crucial to understanding meaning, it has many use cases in various NLP tasks such as question answering and text generation to produce coherent statements (Abdalla et al., 2023). Other natural language challenges like machine translation or information retrieval can be reduced to a semantic distance problem. It is also a key factor for text summarization, the relation between sentences in text will allow for more accurate summaries without too much loss of context.

For Track A of the SemEval 2024 Task 1: *Semantic Textual Relatedness **(STR)*** (Ousidhoum et al., 2024b) on Codalab (Pavao et al., 2023), we aim to create a system to automatically detect the degree of semantic relatedness between pairs of sentences with the OpenAI *text-embedding-3-large* and LaBSE text embedding models. This is for languages like English, as well as multiple low resource languages like *Algerian Arabic, Amharic, Hausa, Kinyarwanda, Marathi, Moroccan Arabic, Spanish, Telugu*.

Our code can be found on GitHub at `https://github.com/dipsivenkatesh/SemEval-2024-Task-1`

## 2 Background

### 2.1 Task and Data Description

The Semantic Textual Relatedness shared task [1] consists of three tracks.

- **Track A**: Supervised
- **Track B**: Unsupervised
- **Track C**: Cross-lingual

In this paper we go through our team's system to solve the track A of the challenge.

For the first track, we must develop a system to automatically find the closeness of meanings (semantic relatedness) between two sentences. We need to generate a relatedness score between 0

[1]`https://codalab.lisn.upsaclay.fr/competitions/16799`



*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 865–868

| PAIRS | SENTENCE 1 | SENTENCE 2 |
|---|---|---|
| 1 | There was a lemon tree next to the house. | The boy enjoyed reading under the lemon tree. |
| 2 | There was a lemon tree next to the house. | The boy was an excellent football player. |

Table 1: Sentence relatedness: We can see that the sentences in pair 1 are more related than the sentences in pair 2

(completely unrelated) and 1 (maximum relations). For this track teams are allowed to submit systems that use the given datasets or any external datasets. The use of pre-trained language models are also allowed.

The datasets for training consisted of a pair of sentences and the 0 to 1 semantic relatedness scores graded through manual annotation. A comparative annotation approach was used for generating these gold label scores thereby avoiding biases of traditional rating and guaranteeing a high reliability.

### 2.2 Previous Work

In recent times the standard way to represent word meanings is as **vector semantics**. This comes from two major ideas, the idea to represent a word in three dimensional vector space (Osgood et al., 1957) and defining a word by the distribution of words around it (Harris, 1954 and Joos, 1950). Representing text as embeddings is an example of representation learning (Bengio et al., 2013).

The combination of term frequency (Luhn, 1957) and inverse document (Sparck Jones, 1972) frequency led to the use of **tf-idf** for representing word embeddings. Tf-idf had many faults, it did not represent contextual word relationships or word co-occurrence. This is fixed with in **Pointwise Mutual Information** (PMI) (Fano and Wintringham, 1961) a measure of how frequent two events occur, compared their occurrence if they were independent. The problem with tf-idf and PMI embeddings is that they are sparse vectors. Instead methods like **word2vec** (Mikolov et al., 2013) and **GloVe** (Pennington et al., 2014) produce dense vectors for word embeddings. Language models have gained a lot of traction due to their understanding of natural language. Language models like BERT have the ability to generate contextual embeddings. Contextual embeddings are used to represent the word in the context that it is used.

More recently, state of the art embedding models use pre-trained transformers by fine tuning the to a certain task. This is used in text embdding models like **Sentence-BERT** (SBERT) (Reimers and Gurevych, 2019) that uses BERT along with siamese and triplet network structures to generate sentence embeddings.

For the Semantic Textual Relatedness shared task track A our system, uses OpenAI's text-embedding-3-large to generate text embeddings. We also use the Language-agnostic BERT Sentence Embedding model (LaBSE) (Feng et al., 2022) to generate embeddings for the other languages as this model generates better representations for these languages.

### 2.3 Evaluation Metrics

There are multiple ways to evaluate relatedness using the vector embeddings of the text, dot product is one such metric. However it favors longer vectors, therefore normalized dot product or the cosine of the angle between the two vectors is used

The evaluation metric for this challenge was the Spearman rank coefficient which compares the the sentence relatedness predictions of the system against the gold truth human judgements. The Spearman rank coefficient ($\rho$) can be calculated as:

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

where:

- $d_i$ is the difference between the ranks of corresponding variables $x_i$ and $y_i$,
- $n$ is the number of observations.

## 3 System Overview

### 3.1 text-embedding-3-large

We use OpenAI's latest large text embedding model **text-embedding-3-large**[2] to generate embeddings. The state of the art *text-embedding-3-large* creates embeddings of 3072 dimensions. The embedding model achieves a score of 54.9% MIRACL benchmark (Zhang et al., 2023) and 64.6% on the MTEB benchmark (Muennighoff et al., 2022).

[2] https://openai.com/blog/new-embedding-models-and-api-updates

| Model / Language | amh | arq | ary | eng | esp | hau | kin | mar | tel |
|---|---|---|---|---|---|---|---|---|---|
| LaBSE | 0.79 | 0.46 | 0.41 | N/A | 0.72 | 0.48 | 0.50 | 0.80 | 0.78 |
| text-embedding-3-large | 0.68 | 0.56 | 0.45 | 0.86 | 0.70 | 0.47 | 0.52 | 0.78 | 0.74 |

Table 2: Performance comparison of sentence-transformers/LaBSE and text-embedding-3-large on training set

### 3.2 LaBSE

We use the **Language-agnostic BERT Sentence Embedding** (LaBSE) model (Feng et al., 2022) to generate the embeddings for most of the non-English languages. We use the model for inferencing using the HuggingFace Transformers library (Wolf et al., 2020).

#### 3.2.1 Model Architecture

The model architecture of LaBSE is similar to BERT (Devlin et al., 2019) and uses self attention to process input text. This is then pre-trained on a large corpus that of multiple languages. After this pre-training the model can generate fixed length sentence embeddings. These embeddings are designed to be language-agnostic.

## 4 Experimental Setup

### 4.1 Dataset

We use the *SemRel* datasets (Ousidhoum et al., 2024a), a semantic relatedness dataset annotated by native accross 14 languages 14 languages: *Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish* and *Telugu*. These datasets consists of multiple records of sentence pairs along with their manually annotated relatedness score. All the languages of Track A of the dataset consist of a train-test split.

### 4.2 Embedding

We propose a system where we take a zero-shot approach to the test set with the pre-trained embedding models. We don't train or fine-tune the models used on the training data. We use the training data for evaluation of model performance on the languages in this track. We can find the evaluation of the models on the training set in table 2. Based on this performance we use text-embedding-3-large for Algerian Arabic, Moroccan Arabic, English, and Kinyarwanda and LaBSE for Amharic, Spanish, Hausa, Marathi and Telugu.

## 5 Results

For evaluation, the organizers rank the system based on Spearman rank correlation coefficient with the golden labels. The performance of the models on all the languages can be found in Table 3. Our system to identify the relatedness scores uses a zero shot method and achieves scores similar to the baseline scores. The score for English surpasses the baseline score.

## 6 Conclusions and Limitations

In our paper for the SemEval Task 1: Semantic Textual Relatedness we propose a zero-shot approach for relatedness using the text-embedding-3-large and LaBSE embedding models. It is importand to consider that text-embedding-3-large is not an open-source model and that these models may contain inherent biases in them.

## A Spearman Correlation on test dataset

| Language | Our scores |
|---|---|
| **Algerian Arabic (arq)** | 0.5097117963 |
| **Amharic (amh)** | 0.8000962937 |
| **English (eng)** | 0.8323738277 |
| **Hausa (hau)** | 0.5083993463 |
| **Kinyarwanda (kin)** | 0.5183340316 |
| **Marathi (mar)** | 0.8415291711 |
| **Moroccan Arabic (ary)** | 0.4441887719 |
| **Spanish (esp)** | 0.6557116114 |
| **Telugu (tel)** | 0.814199637 |

Table 3: Spearman Correlation on test dataset

## 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.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new

perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Stephen A. Bernhardt. 1980. *Style*, 14(1):47–50.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Robert M Fano and WT Wintringham. 1961. Transmission of information.

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

Zellig S. Harris. 1954. Distributional structure.

Martin Joos. 1950. Description of language design. *Journal of the Acoustical Society of America*, 22:701–707.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Saif Mohammad. 2008. *Measuring semantic distance using distributional profiles of concepts*. University of Toronto.

Saif M. Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Charles E. Osgood, George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of Meaning*. University of Illinois Press.

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.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2023. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.