Fired_from_NLP at SemEval-2024 Task 1: Towards Developing Semantic Textual Relatedness Predictor - A Transformer-based Approach

Anik Mahmud Shanto, Md. Sajid Alam Chowdhury, Mostak Mahmud Chowdhury,
Udoy Das, Hasan Murad
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
u1904{049, 064, 055}@student.cuet.ac.bd, u1804109@student.cuet.ac.bd,
hasanmurad@cuet.ac.bd

Abstract

Predicting semantic textual relatedness (STR) is one of the most challenging tasks in the field of natural language processing. Semantic relatedness prediction has real-life practical applications while developing search engines and modern text generation systems. A shared task on semantic textual relatedness has been organized by SemEval 2024, where the organizer has proposed a dataset on semantic textual relatedness in the English language under Shared Task 1 (Track A3). In this work, we have developed models to predict semantic textual relatedness between pairs of English sentences by training and evaluating various transformer-based model architectures, deep learning, and machine learning methods using the shared dataset. Moreover, we have utilized existing semantic textual relatedness datasets such as the stsb multilingual benchmark dataset, the SemEval 2014 Task 1 dataset, and the SemEval 2015 Task 2 dataset. Our findings show that in the SemEval 2024 Shared Task 1 (Track A3), the fine-tuned-STS-BERT model performed the best, scoring 0.8103 on the test set and placing 25th out of all participants.

1 Introduction

Nowadays, there have been notable advancements in understanding and measuring pairwise semantic relatedness between texts within the domain of natural language processing. Predicting semantic relatedness plays a significant role in improving search engines, question-answering systems, text summarization tools, and machine translation.

However, previous works in natural language processing have mainly dealt with semantic similarity, a smaller aspect of relatedness, mainly due to the limited availability of relatedness datasets. Besides, dealing with ambiguous words or phrases that have multiple meanings can make semantic relatedness difficult. Understanding cultural context in language has been complex, and existing models have struggled to capture these variations. As language evolves, models struggle to adapt quickly to new linguistic patterns and expressions. To bridge these gaps, we need improved models that understand not just words but also context, cultural differences, and how language changes over time.

Semantic relatedness models have been developed using various transformer-based, deep learning, and machine learning techniques. Traditional machine learning methods (Buscaldi et al., 2015) have relied on predefined rules and features and offered moderate results. These approaches have often struggled with complex semantic relationships. Deep learning-based (Wang et al., 2018) approaches have surpassed traditional machine learning models in capturing complex relationships, particularly in tasks requiring a deep understanding of context. However, transformer-based approaches (Devlin et al., 2019) have outperformed others when it comes to capturing semantic relationships, particularly in understanding context, managing long-range dependencies, and handling contextual embeddings.

SemEval has arranged a shared task named SemEval 2024 Task 1: Semantic Textual Relatedness (STR) (Ousidhoum et al., 2024b), introducing a novel dataset called Shared Task 1 (Track A3) (Ousidhoum et al., 2024a) for determining the level of pairwise semantic relatedness between sentences based on the similarity score that ranges from 0.0 to 1.0.

The primary goal of this task is to build a robust and accurate model to predict the semantic relatedness between pairs of English sentences.

To accomplish this goal, we have used a variety of models, incorporating machine learning models (Linear Regression, Random Forest, XGBoost), models of deep learning (LSTM, BiLSTM), and pre-trained models based on transformer (RoBERTa, bert-base-uncased). We have named our approach of using the bert-base-uncased

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 859–864
June 20-21, 2024 ©2024 Association for Computational Linguistics

model as STS-BERT.

By training and assessing every model, we have carried out a comparison analysis on the Semeval 2024 Task 1 (Track A3) dataset (Ousidhoum et al., 2024a), STSB multilingual dataset (May, 2021), SemEval 2014 Task 2 dataset (Marelli et al., 2014) and dataset provided for Task 1 in SemEval 2015 (Agirre et al., 2015) and have finally come to a conclusion that the STS-BERT model has demonstrated better performance compared to others boasting an impressive Spearman correlation coefficient of 0.81033 on the test dataset.

Key contributions of our research work are listed below -

- We have developed a fine-tuned-STS-BERT model that significantly helps in accurately predicting semantic textual relatedness across diverse sentence pairs.
- We have evaluated the model's performance through various tests conducted using the dataset and subsequently performed an in-depth evaluation of the outcomes.

The GitHub repository that follows has the implementation details available - https://github.com/Fired-from-NLP/SemEval-2024-task-1-track-A-eng.

2 Related Works

The associated works on semantic textual similarity can be generally categorized into three parts, approaches focused on machine learning, deep learning, and attention-based mechanism (transformer).

Among machine learning models, the Support Vector Regression model has been applied for calculating the semantic relationship between two short sentences (Sultan et al., 2013). In this system, three distinct measures, namely overlap in word n-gram, overlap in character n-gram, and semantic overlap, have been used for predicting similarity. In (Buscaldi et al., 2015), a Random forest-based approach has been utilized to find the semantic sentence similarity. The approach has relied on various similarity measures such as WordNet-based conceptual similarity, IC-based similarity, syntactic dependencies, and information retrieval-based similarity.

Traditional deep learning methods have depended on single or multiple granularity representations for detecting similarity. Apart from that, a different architecture that has focused on multiple positional sentence representations has been proposed (Wang et al., 2018). It has used Bi-LSTM for generating representations that enable the model to capture better context understanding. Another architecture has introduced a Siamese adaptation of LSTM (Mueller and Thyagarajan, 2016). Using a fixed-sized vector and a simple Manhattan metric, the model transforms sentence representation that represents semantic relationships. Another paper has described an architecture that has been built using deep learning paradigms (Zhao et al., 2015). This architecture has been trained using a combination of features like features based on a string, features based on a corpus, and features based on syntactic similarity, as well as newer matrices derived from distributed word embedding.

Transformer-based approaches have surpassed both machine learning and deep learning models in calculating semantic sentence relationships. Unlabeled text can be used to pre-train deep bidirectional representations using BERT (Devlin et al., 2019). BERT can be fine-tuned to do various NLP-related tasks like semantic analysis. A replication of BERT called RoBERTa (Liu et al., 2019), has focused on hyperparameters and training data size to improve model performance.

In this shared task, we have used BERT-based pre-trained models as they have been proven to be superior to other models available.

3 Dataset

We have employed the dataset made available as part of Shared Task 1 (Track A3) of the SemEval 2024: Semantic Textual Relatedness (STR) which contains 5500 samples in the training dataset and 250 samples in the dev dataset. Besides, the stsb multilingual benchmark dataset (May, 2021), the SemEval 2014 Task 1 dataset (Marelli et al., 2014) and the Semeval 2015 Task 2 dataset (Agirre et al., 2015) have been used.

Task	Sentence Pairs		
	Train	Validation	Test
SemEval 2014	4500	500	4928
SemEval 2015	2997	750	6729
stsb-multi-mt	5749	1500	1379

Table 1: Data sizes for external datasets

Table 1 shows the distribution of samples that we have used from external datasets. These datasets have been merged to get a total of 32508 samples and then divided further into two sets: train and validation comprising 25676 and 6732 samples respectively. We have replaced the similarity score of duplicate sentence pairs with the average value to avoid labeling biases among different datasets. For the test dataset, The dataset made available as part of Shared Task 1 (Track A3), which consists of 2600 samples, has been used. These datasets contain a pair of sentences in each row, which we have split into two separate sentences namely sentence1 and sentence2. Figure 1 and Figure 2 show that sentence1 contains an average of 6-12 words, while sentence2 contains 3-12 words.

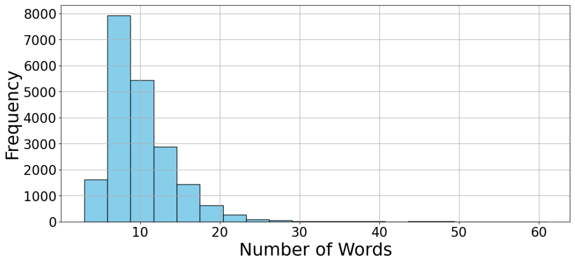

Figure 1: Word distribution of sentence1

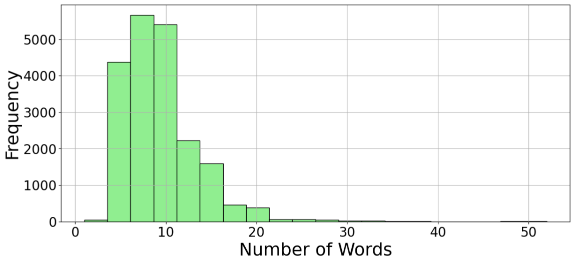

Figure 2: Word distribution of sentence2

4 System Overview

In this section, we have outlined our methodology to develop models for determining sentence relatedness. First, we have used various extraction strategies to extract characteristics and then utilized a variety of machine learning and deep learning algorithms. Moreover, we have employed different transformer models to develop the system. Figure 3 provides a summary of our working methods.

4.1 Machine Learning-based Approaches

For determining sentence relatedness, we have applied traditional Machine learning-based methods such as Linear Regression and Random Forest. Moreover, To increase the performance, we have employed an ensemble classifier called XGBoost.

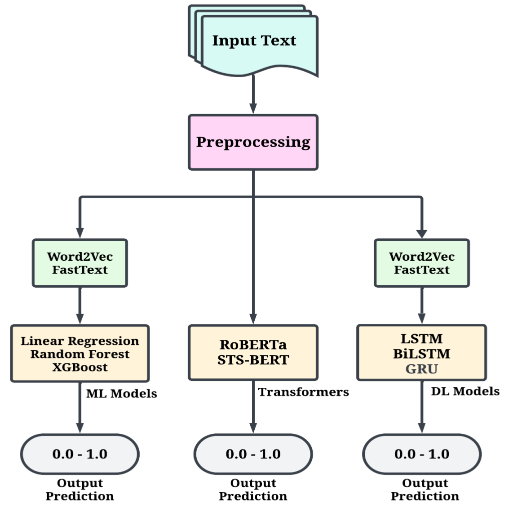

Figure 3: An outline of our approach

Here, we have tokenized the dataset using NLTKTokenizer, and then have we used Word2Vec to extract features. We also have used FastText for feature extraction as it not only captures semantic meaning like Word2Vec but also encodes subword information, allowing it to handle out-of-vocabulary words and morphologically rich languages more effectively. We have set the number of decision trees or boosting rounds to n_estimators for the ensemble approach at 100.

4.2 Deep Learning-based Approaches

Deep learning-based models have been utilized for determining sentence relatedness. We have implemented both models based on LSTM and Bi-LSTM. Two LSTM layers with various numbers of LSTM cells have been applied to the LSTM model. Each of the two directional layers has 50 or 100 LSTM cells in it. We have employed two Bi-LSTM layers, each with 100 and 50 Bi-LSTM cells, in the Bi-LSTM model.

4.3 Transformer-based Approaches

Methods based on transformers are now widely employed in many different contexts. We have employed STS-BERT (Devlin et al., 2019) and RoBERTa to tackle this task. As the sentences can be diverse, having a single representation and better understanding of the sentences is very important. For this reason, we have used the feature vector of the pooling layer as shown in Figure 4.

In our approach, we have first split the pair of sentences in the dataset into two. We have used two bert-based-uncased (Devlin et al., 2019) for

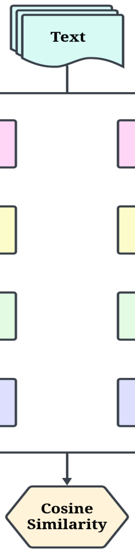

Figure 4: STS-BERT: Transformer-based model architecture for predicting semantic textual relatedness

two sentences. We have obtained the feature vector from the pooling layer of these two Bert models. After obtaining the pooled embeddings, we feed them to the cosine similarity for performing the relatedness task and compare them with the ground truth relatedness score. Then, we compute the loss using MSE (Mean Squared Error) based on the predicted relatedness and the actual relatedness. After the loss is calculated to improve the performance and minimize the loss, we have updated the model parameters using gradient descent.

5 Experimental Setup

This section gives a summary of our experimental setup while training and evaluating our model architectures for semantic textual relatedness.

5.1 Environment Setting

The simulation was executed on a personal computer featuring an Intel Core i7-9700 CPU clocked at 3.00 GHz and an NVIDIA GeForce GTX 2060 GPU. Additionally, to ensure ample processing capability, a Kaggle Notebook equipped with a P100 GPU was utilized.

5.2 Data Preparation

Besides the dataset provided in this competition, we have used three external datasets. We have used the stsb multilingual benchmark dataset (May, 2021), the SemEval-2014 Task 1 dataset (Marelli et al., 2014), and the SemEval-2015 Task 2 dataset (Agirre et al., 2015). We have combined all three datasets. The similarity score of external datasets ranges from 0.0 to 5.0. However, the provided dataset for this competition holds the relatedness between sentences ranging from 0.0 to 1.0. We multiplied the relatedness score of the dataset offered in the competition by 5.0 to match the similarity score in the combined dataset. We have replaced the similarity score of duplicate sentence pairs with the average value. Then, we have split the combined dataset into the training dataset and the validation dataset. The final size of the training dataset is 25676, whereas the overall size of the validation dataset is 6732. We have used the test dataset provided in the competition. The test set contains 2600 samples.

5.3 Parameter Settings

Table 2 shows the parameter settings used in LSTM. BiLSTM, and RoBERTa models.

Model	lr	optim	bs	epoch
LSTM	$1e^{-6}$	Adam	32	10
BiLSTM	$1e^{-6}$	Adam	32	10
RoBERTa	$1e^{-6}$	Adam	32	12

Table 2: Parameter configurations for various models

In Table 2, learning rate, optimizer, batch size, and number of epochs are represented by the variables lr, optim, bs, and epoch, in that order.

Table 3 summarizes the parameter settings used in our proposed STS-BERT model.

Parameter	Value
Learning Rate	1×10^{-6}
Optimizer	AdamW
Batch Size	8
Number of Epochs	12
Loss Function	Mean Squared Error (MSE)
Pooling	Mean Pooling

Table 3: Model parameter settings.

5.4 Evaluation Metrics

The instruction of Shared Task 1 of SemEval 2024 has been to use the Spearman correlation to evaluate the performance of our model using the test dataset. The mathematical representation of the Spearman correlation is provided in equation 1. Besides, we have used Cosine similarity in our model to predict similarity between sentences. Equation 1 presents the mathematical representation for Cosime similarity.

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

In this context, ρ denotes the Spearman correlation coefficient. d_i represents the difference between the ranks of corresponding observations in the two variables, while n indicates the total number of observations.

$$\text{cos\_sim}(A, B) = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

Where, cos_sim(A, B) is the cosine similarity between vectors A and B. A_i and B_i denote the components of vectors of A and B respectively. n indicates the dimensionality of the vectors.

6 Experimental Results

In this section, we have showcased the experimental findings obtained during the training and evaluation stages of the proposed model for semantic textual relatedness prediction.

Table 4 presents a comparative analysis of different types of models, evaluating their performance using the Spearman correlation coefficient on the test dataset.

Category	Model	Embedding	Score
ML	Linear Regression	word2vec	0.0507
		fasttext	0.0507
	Random Forest	word2vec	0.1298
		fasttext	0.1198
	XGBoost	word2vec	0.3178
		fasttext	0.2072
DL	LSTM	word2vec	0.445
		fasttext	0.420
	BiLSTM	word2vec	0.4990
		fasttext	0.429
BERT	RoBERTa	-	0.749
	STS-BERT	-	**0.810**

Table 4: Results of different models on the test dataset

Among the machine learning models, we have found that the XGBoost model with word2vec embedding has achieved the highest score of 0.3178. In the deep learning category, we have seen better performance as both LSTM and BiLSTM models have higher scores than the machine learning models. The BiLSTM model achieved a score of 0.499, slightly outperforming the LSTM model, which obtained a score of 0.445.

In some cases, Fasttext word embedding has obtained the best results compared to word2vec (Meden, 2022). Therefore, we have also tested the performance of the model using Fasttext embedding. However, the transformer-based models have clearly outperformed other models based on machine learning and deep learning. For instance, the RoBERTa model achieved a score of 0.749, while our proposed STS-BERT model demonstrated exceptional performance with an impressive score of 0.810.

7 Error Analysis

In the development phase external datasets, SemEval 2014 Task 1 (Marelli et al., 2014), SemEval 2015 Task 2 (Agirre et al., 2015) and multilingual benchmark dataset (May, 2021) along with the competition dataset have been utilized. Hence the training set becomes more diverse and our model fails to learn about the relatedness between the sentences. The similarity scores of the external datasets ranged between 0.0 to 5.0. To make all the scores similar we have multiplied the scores of the competition dataset by 5.0 and normalized the whole training set by dividing all the scores by 5.0. Due to multiple conversions of the range of scores, precision loss has occurred. Sentence transformation has been another key reason for the poor performance of the model. When the second sentence is the transformation of the first sentence, our model can not detect it. For example, if the first sentence is in simple form and the second sentence is in the complex form of the first sentence, the model shows poor performance in that case. As a result, the overall performance of our proposed system has degraded.

8 Conclusion

In this research, we have conducted a comparative performance analysis, assessing a range of machine learning, deep learning, and transformer-based models to predict the semantic textual relatedness between pairs of English sentences. We have utilized the Task 1 (Track A3) dataset provided in the shared task, along with additional external datasets, for training various models. Our results indicate that the STS-BERT model has outperformed all other models, achieving an impressive score of 0.810. However, after analyzing errors, we have discovered that the slight score decrease is due to the integration of large external STS datasets with varying output ranges. To address this in future work, we plan to implement alternative strategies. Moreover, we will work on Task 1 (Track B and C) to have more comprehensive findings.

9 Ethical Considerations

To advance semantic text relatedness, we commit to emphasizing privacy through informed consent, reducing biases, as well as transparent modeling. Our ethical position prioritizes responsibility, accessibility, and privacy to build a positive and open technology environment.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *SemEval 2015*, Denver, Colorado. ACL.

Davide Buscaldi, Jorge García Flores, Ivan V. Meza, and Isaac Rodríguez. 2015. SOPA: Random forests regression for the semantic textual similarity task. In *SemEval 2015*, Denver, Colorado. ACL.

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

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*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *SemEval 2014*, Dublin, Ireland. ACL.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Katja Meden. 2022. Semantic Similarity of Parliamentary Speech using BERT Language Models & fastText Word Embeddings.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. *Proceedings of the AAAI Conference on Artificial Intelligence*.

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.

Md. Sultan, Steven Bethard, and Tamara Sumner. 2013. DLS@CU-CORE: A simple machine learning model of semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. ACL.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2018. A deep architecture for semantic matching with multiple positional sentence representations. In *ACL*.

Jiang Zhao, Man Lan, and Jun Feng Tian. 2015. ECNU: Using traditional similarity measurements and word embedding for semantic textual similarity estimation. In *SemEval 2015*. ACL.