Praktikum Natural Language Processing

Semantic Textual Relatedness (STR)

Shrineeth Vishwanath Kotian*[1]

[1]Julius-Maximilians-Universität Würzburg

March 03, 2024

Abstract:

Semantic Textual Relatedness (STR) plays a crucial role in numerous natural language processing tasks, including information retrieval, question answering, and document summarization. This research paper delves into the exploration and analysis of various methodologies and approaches aimed at quantifying the degree of relatedness between texts. By examining the semantic similarity between textual pairs, the study seeks to enhance our understanding of language representation and interpretation within computational frameworks. Leveraging state-of-the-art models and techniques, the research investigates the intricate interplay of linguistic features and context in determining textual relatedness. Through extensive experimentation and evaluation, the paper presents novel insights into the effectiveness and limitations of existing STR models, paving the way for advancements in language understanding and semantic processing. The findings of this study hold significant implications for the development of more accurate and robust natural language processing systems, thereby contributing to the advancement of artificial intelligence and human-computer interaction.

Introduction

Semantic Textual Relatedness (STR):

Semantic Textual Relatedness (STR) is a fundamental concept in natural language processing (NLP) that aims to quantify the degree of similarity or relatedness between two pieces of text. It plays a crucial role in various NLP tasks, including information retrieval, question answering, paraphrase detection, and sentiment analysis. By accurately measuring the semantic similarity between text fragments, STR models can enhance the performance of downstream applications and improve the overall user experience in human-machine interactions.

Textual similarity refers to the degree of likeness or resemblance between two pieces of text based on their underlying semantic content. Unlike surface-level measures of similarity that focus on lexical or syntactic features, textual similarity considers the meaning and context of the text, capturing the inherent semantic relationships between words, phrases, and sentences.

For example, consider the following pair of sentences:

Sentence 1: ”The cat sat on the mat.”
Sentence 2: ”The feline rested on the rug.”

While these sentences are not identical in terms of their lexical content, they convey a similar meaning or message – namely, that a cat is resting on a flat surface. Traditional measures of textual similarity would recognize the semantic equivalence between these sentences, despite differences in word choice or phrasing.

Semantic Textual Relatedness vs Semantic Text Similarity

Semantic Textual Relatedness (STR) and semantic text similarity are closely related concepts in natural language processing, but they capture different aspects of text comparison.

*shrineeth.kotian@gmail.com

Semantic Textual Relatedness (STR) focuses on measuring the degree of relatedness or similarity between two pieces of text, considering their underlying semantic content. It aims to quantify how closely two text fragments are semantically aligned, regardless of their surface-level differences in language or expression.

On the other hand, semantic text similarity specifically assesses the degree of similarity between two text fragments based on their semantic content. It seeks to determine how much overlap or resemblance exists between the meanings of two pieces of text, accounting for their shared semantic concepts and relationships.

In essence, while STR emphasizes the overall relatedness or connection between two texts, semantic text similarity zooms in on the specific level of similarity in their semantic content.

For example, consider the following pair of sentences:

Sentence 1: ”The dog chased the cat.”
Sentence 2: ”The feline pursued by the canine.”

In terms of STR, these sentences would likely be considered highly related or similar, as they both convey the idea of one animal pursuing another. However, in terms of semantic text similarity, the sentences may be judged to have moderate similarity, as they use different words (”dog” vs. ”canine,” ”cat” vs. ”feline”) but convey a similar action.

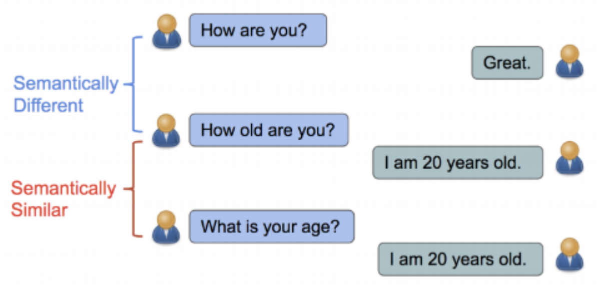


Figure 1: STR vs Semantic Text Similarity

Importance and Motivation of Semantic Textual Relatedness:

The importance of STR stems from its ability to capture such nuanced semantic relationships between text fragments, enabling applications to understand and interpret natural language more accurately. By leveraging advanced techniques from deep learning and natural language understanding, STR models can encode the rich semantic information embedded in textual data, facilitating more intelligent and context-aware NLP applications.

As the volume and diversity of textual data continue to grow exponentially, the demand for robust and scalable STR models becomes increasingly evident. Researchers and practitioners in the field are continually striving to develop more accurate, efficient, and domain-agnostic approaches to measure semantic relatedness effectively. Addressing this challenge requires innovative methodologies, large-scale datasets, and rigorous evaluation frameworks to assess the performance and generalization capabilities of STR models across different languages, domains, and contexts.

In this research paper, we aim to address these challenges by proposing a novel approach to Semantic Textual Relatedness (STR) that leverages state-of-the-art techniques from deep learning and natural language processing. Our objective is to develop a highly accurate, scalable, and domain-agnostic STR model capable of capturing the subtle semantic nuances present in diverse textual data. By advancing the state of the art in STR research, we seek to contribute to the broader goal of advancing human-machine interaction and enabling more intelligent and context-aware NLP applications.

Problem Statement:

Despite significant advancements in STR research, several challenges and opportunities remain to be addressed. One of the primary challenges is the development of STR models that can effectively handle the inherent ambiguity, variability, and diversity of natural language. Textual data exhibit complex semantic relationships that often defy simple categorization or classification, making it challenging to accurately capture the nuances of meaning and context.

Furthermore, existing STR models may exhibit biases, limitations, or deficiencies that affect their performance on specific tasks or datasets. These issues can arise due to factors such as dataset imbalance, domain-specific language patterns, or model architecture biases. Overcoming these challenges requires a comprehensive understanding of the underlying mechanisms driving semantic relatedness and the development of innovative strategies to mitigate bias, enhance model robustness, and improve generalization capabilities.

In this research paper, we aim to address these challenges by proposing a novel approach to Semantic Textual Relatedness (STR) that leverages state-of-the-art techniques from deep learning and natural language processing. Our objective is to develop a highly accurate, scalable, and domain-agnostic STR model capable of capturing the subtle semantic nuances present in diverse textual data. By advancing the state of the art in STR research, we seek to contribute to the broader goal of advancing human-machine interaction and enabling more intelligent and context-aware NLP applications.

Literature Review:

The paper titled "SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 14 Languages" presents a significant contribution to the field of natural language processing by providing a comprehensive collection of datasets for semantic textual relatedness (STR) across multiple languages. Led by Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, and other collaborators, the paper addresses the growing need for high-quality STR datasets to support research and development in multilingual NLP tasks.[1]

The key objective of the paper is to curate and release a diverse set of datasets covering 14 languages, including Arabic, English, French, Hindi, Telugu, and others. These datasets are meticulously constructed to capture semantic relatedness between text pairs across various domains and linguistic contexts. The authors employ a rigorous annotation process involving human annotators to ensure the quality and reliability of the collected data.[1]

By providing datasets in multiple languages, the SemRel2024 initiative aims to facilitate research in cross-lingual STR, enabling researchers to develop robust and language-agnostic models for tasks such as semantic similarity measurement, paraphrase detection, and textual entailment. The availability of diverse datasets empowers researchers to evaluate and benchmark their models across different languages and domains, fostering advancements in multilingual NLP and cross-lingual understanding.[1]

The paper highlights the significance of multilingualism in NLP research and underscores the importance of linguistic diversity in dataset construction. It emphasizes the need for inclusive and representative datasets that encompass a wide range of languages, dialects, and cultural contexts. By curating datasets for languages beyond the commonly studied ones, SemRel2024 contributes to promoting linguistic diversity and inclusion in NLP research.

Methodology:

Task Overview:

Objective:

The objective of this research is to develop a model capable of automatically detecting the degree of semantic relatedness between pairs of sentences. Semantic Textual Relatedness (STR) refers to the extent to which two sentences convey similar meanings or contexts.

Data Collection:

The research utilizes sentence pairs in multiple languages sourced from diverse datasets. These datasets contain annotated labels indicating the degree of semantic textual relatedness, typically ranging from 0 to 1, where 0 denotes no semantic similarity and 1 represents complete semantic equivalence.

Labeling:

Each sentence pair is labeled with a score representing the degree of semantic relatedness between them. These labels serve as ground truth annotations for training and evaluating the model's performance.

Task Track:

The research is conducted within the context of two task tracks: Supervised (Track A) and Cross-lingual

(Track C). The SemRel2024 Task 1 provides the framework for evaluating models in both tracks.

Model Selection:

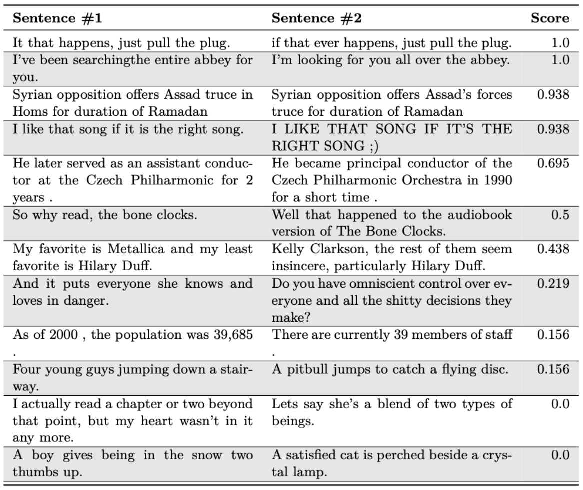
The chosen model architecture for this task is XLM-RoBERTa, a state-of-the-art transformer-based model known for its effectiveness in capturing cross-lingual semantics and contextual information.

Evaluation Metric:

The primary evaluation metric for assessing the performance of the model is the Spearman Rank correlation coefficient. This metric measures the strength and direction of association between the predicted scores and the ground truth labels, providing insights into the model's ability to capture semantic relatedness effectively across different languages.

Data Collection:

The process of collecting data for Semantic Textual Relatedness (STR) involved several intricate steps, particularly in dealing with languages that have limited resources, such as Hausa, Kinyarwanda, and Algerian Arabic. The primary challenge was to select sentence pairs that demonstrated semantic relatedness while ensuring that unrelated instances were avoided. This necessitated the application of various heuristics, which were customized for each language and corpus. These methods were determined through collaboration with native speakers to maintain linguistic authenticity and relevance.



Sentence #1	Sentence #2	Score
It that happens, just pull the plug.	if that ever happens, just pull the plug.	1.0
I've been searchingthe entire abbey for you.	I'm looking for you all over the abbey.	1.0
Syrian opposition offers Assad truce in Homs for duration of Ramadan	Syrian opposition offers Assad's forces truce for duration of Ramadan	0.938
I like that song if it is the right song.	I LIKE THAT SONG IF IT'S THE RIGHT SONG ;)	0.938
He later served as an assistant conductor at the Czech Philharmonic for 2 years .	He became principal conductor of the Czech Philharmonic Orchestra in 1990 for a short time .	0.695
So why read, the bone clocks.	Well that happened to the audiobook version of The Bone Clocks.	0.5
My favorite is Metallica and my least favorite is Hilary Duff.	Kelly Clarkson, the rest of them seem insincere, particularly Hilary Duff.	0.438
And it puts everyone she knows and loves in danger.	Do you have omniscient control over everyone and all the shitty decisions they make?	0.219
As of 2000 , the population was 39,685 .	There are currently 39 members of staff .	0.156
Four young guys jumping down a stairway.	A pitbull jumps to catch a flying disc.	0.156
I actually read a chapter or two beyond that point, but my heart wasn't in it any more.	Lets say she's a blend of two types of beings.	0.0
A boy gives being in the snow two thumbs up.	A satisfied cat is perched beside a crystal lamp.	0.0

Figure 2: Examples of sentence pairs and their corresponding scores (from 0 to 1)

For languages like English and Spanish, which have abundant resources, a diverse array of sentence structures, formalities, and grammatical variations were sourced from multiple outlets. In English, for example, sentence pairs were gathered from datasets containing both formal and informal text, book reviews, paraphrases, and semantic similarity scores. Similarly, Spanish datasets included samples from semantic similarity datasets, entailment datasets, summarization datasets, and various question sets.

In the case of Arabic, which exhibits variations across different contexts and dialects, datasets were carefully curated to reflect these distinctions in language usage. For Modern Standard Arabic (MSA), datasets were sourced from TED Talk subtitles and news articles, with additional preprocessing steps to ensure grammatical correctness. Algerian Arabic datasets were compiled from YouTube comments and the Parallel Arabic Dialect Corpus (PADIC), utilizing lexical overlap and contiguous sentence selection. Moroccan Arabic datasets were derived from headlines, focusing on lexical overlap for sentence pairing.

For languages with fewer resources like Afrikaans, Amharic, Hausa, and Kinyarwanda, datasets were primarily sourced from news articles. Sentence pairs were created based on lexical overlap, with constraints on length and language purity to maintain dataset quality and linguistic coherence.

Indonesian datasets were collected from Wikipedia texts and the IndoSum dataset, with attention to parsing errors and criteria for sentence-level selection. Similarly, datasets for Hindi, Marathi, Telugu, and Punjabi were sourced from news headlines and articles. Techniques such as paraphrasing, contiguous sentence selection, and random sentence pairing were employed to enrich the datasets.

Training Procedure:

The training procedure involves optimizing the model parameters to minimize a predefined loss function while maximizing the predictive accuracy of the model. We employ the AdamW optimizer with a learning rate scheduler to update the model weights iteratively. During training, the model processes batches of text pairs, computes the loss between predicted and ground truth scores, and backpropagates the gradients to update the parameters. Regularization techniques such as dropout are applied to prevent overfitting, ensuring generalization to unseen data. The training process continues for a fixed number of epochs, with model performance monitored on a validation set to prevent overfitting and determine the optimal stopping point.

Experimental Setup:

The datasets consist of pairs of text sequences annotated with human-assigned similarity scores. We preprocess the data by tokenizing, encoding, and splitting it into training, validation, and test sets. Hyperparameters such as batch size, maximum sequence length, and learning rate are tuned through grid search or random search to optimize model performance. Model evaluation is performed using standard metric such as Spearman correlation coefficient to assess the model's ability to capture semantic textual relatedness accurately. The experiments are conducted on hardware with sufficient computational resources, including GPUs, to expedite model training and evaluation.

Project Implemetation:

Semantic Textual Relatedness (STR) is all about understanding how similar or related two pieces of text are. To make this happen, we're using advanced models called RoBERTa and XLM-RoBERTa.

RoBERTa:

Think of RoBERTa as a super-smart language understanding tool. It's like a huge library that's read tons of books, newspapers, and articles in different languages. RoBERTa has learned to understand the meaning behind words and sentences by reading all these texts. When we give RoBERTa two sentences to compare for relatedness, it uses its vast knowledge to figure out how similar they are. It's like having a very well-read friend who can tell if two stories have similar plots or themes.

XLM-RoBERTa:

Now, imagine RoBERTa's smartness, but with the ability to understand many languages. That's what XLM-RoBERTa does! It's like a multilingual genius that can understand and compare text in different languages. So, whether it's English, Spanish, Hindi, or any other language, XLM-RoBERTa can handle it. This makes it incredibly useful for projects like ours, where we're dealing with text in various languages and want to measure how related they are.

What is XLM-RoBERTa?

XLM-RoBERTa stands for "Cross-lingual Language Model - RoBERTa." It is a type of language model developed by Facebook AI that builds upon the RoBERTa (Robustly optimized BERT approach) architecture. XLM-RoBERTa is designed to understand and process text in multiple languages simultaneously, making it particularly useful for tasks involving multilingual data analysis.

How Does it Work?

XLM-RoBERTa is based on the Transformer architecture, which is a deep learning model specifically designed for processing sequential data like text. It consists of multiple layers of self-attention mechanisms that enable it to capture dependencies between words in a sentence, regardless of their positions.

For our project on Semantic Textual Relatedness (STR), XLM-RoBERTa offers several advantages:

Multilingual Support: XLM-RoBERTa is trained on text data from multiple languages, allowing it to understand and generate representations for words and sentences in various languages. This capability enables us to analyze semantic relatedness across different languages, which is essential for our multilingual STR task.

Cross-lingual Transfer Learning: XLM-RoBERTa leverages transfer learning, where knowledge learned from one task or language is transferred to another. By pretraining on a large corpus of text data from diverse languages, XLM-RoBERTa can learn general language patterns and semantics, which can be fine-tuned for specific downstream tasks like STR.

Robust Performance: XLM-RoBERTa is trained using large-scale datasets and sophisticated training procedures, resulting in robust and high-quality representations of text. This robustness helps improve the performance of our STR model, especially when dealing with diverse linguistic contexts and data sources.

Applications of XLM-RoBERTa:

XLM-RoBERTa, short for Cross-lingual Language Model - RoBERTa, is a powerful language model developed by Facebook AI. It offers various applications across different domains, including natural language processing (NLP) and multilingual text analysis. Below are some key applications of XLM-RoBERTa:

Paraphrase Detection: XLM-RoBERTa can be used for detecting whether two sentences convey the same meaning, even if they are worded differently. This application is useful in tasks such as duplicate content detection, plagiarism detection, and sentence similarity scoring.

Question Answering: XLM-RoBERTa can aid in question answering tasks by comprehending the context of a question and providing relevant answers from a given text corpus. It can understand the nuances of language across different languages, making it suitable for multilingual question answering systems.

Text Summarization: XLM-RoBERTa can generate concise summaries of lengthy text documents by identifying the most important information and condensing it into a shorter form. This application is valuable for quickly extracting key insights from large volumes of text data.

Sentiment Analysis: XLM-RoBERTa can analyze the sentiment expressed in text data across multiple languages. It can distinguish between positive, negative, and neutral sentiments, enabling businesses to gauge customer feedback, social media sentiment, and brand reputation in diverse linguistic contexts.

Named Entity Recognition (NER): XLM-RoBERTa can identify and classify named entities such as people, organizations, locations, and dates mentioned in text data. This application is useful for tasks such as information extraction, entity linking, and content categorization.

Machine Translation: XLM-RoBERTa can improve machine translation systems by providing better representations of text in different languages. It can capture cross-lingual semantic similarities and differences, leading to more accurate and contextually appropriate translations.

Training and Model Evaluation:

For training the model, we first load an English training dataset containing pairs of sentences along with their relatedness scores. It then splits the dataset into training and validation sets. To facilitate training, a custom dataset class called SemanticRelatednessDataset is defined. This class preprocesses the data and prepares it for training. The RoBERTa tokenizer is initialized, and the pretrained model is loaded for sequence classification. Data loaders are created to iterate over batches of data during training. The model is trained for a specified number of epochs using the AdamW optimizer. Training loss is monitored and printed for each epoch. After training, the model and tokenizer are saved for future use.

Visual Inspection:

In Visual Inspection, the trained model's predictions on the validation set are visually inspected. The model is set to evaluation mode to disable dropout and batch normalization layers. Predictions are made on batches of validation data using the trained model. Predicted scores and actual scores are printed for each case in the validation set in batches of 8 or 16 in our case. This manual inspection allows researchers to assess the model's performance and identify any patterns or discrepancies.

Spearman Correlation:

Then we calculate the Spearman correlation coefficient between the predicted scores and the actual scores on the validation set. The model is set to evaluation mode, and predicted scores are accumulated for the entire validation set. The Spearman correlation coefficient is calculated using the "spearmanr" function from the scipy.stats module. This coefficient provides a measure of the monotonic relationship between the predicted and actual scores, indicating the model's performance.

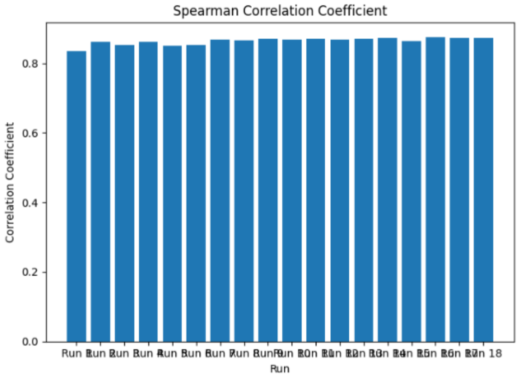


Figure 3: Spearman correlation coefficient between the predicted scores and the actual scores across multiple runs

Training with Hyperparameter Tuning:

Here, hyperparameter tuning is performed by training the model with different combinations of learning rates (5e-5, 3e-5, 2e-5), epochs, and random seeds. The code iterates over predefined lists of learning rates, epochs, and seeds. For each combination of hyperparameters, the model is trained on the training set and evaluated on the validation set. Training loss, accuracy, and correlation coefficients are monitored and recorded for each run. After training, the model and tokenizer are saved, and the results are plotted for analysis. This process helps researchers identify optimal hyperparameters for training the model effectively.

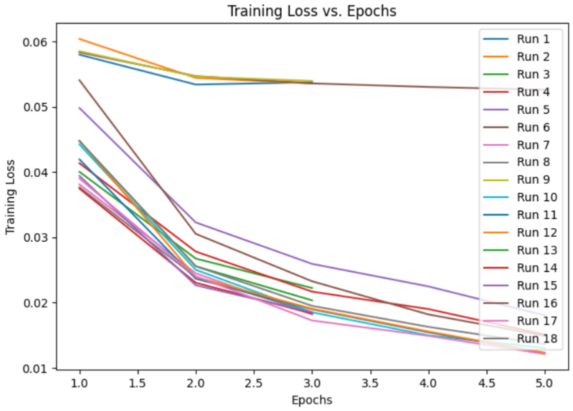


Figure 4: Comparison of training loss during model training across multiple runs

Multilingual Testing:

To evaluate the performance of the Semantic Textual Relatedness (STR) model across multiple languages, a multilingual testing approach was employed. Initially, preprocessing steps were undertaken to prepare the multilingual datasets for evaluation. The datasets consisted of paired sentences in various languages, including Telugu, Arabic, Spanish, Marathi, and others. Each sentence pair was split into two separate sentences, representing the first and second sentences, respectively, to facilitate the evaluation process. Subsequently, the pretrained STR model, based on RoBERTa architecture, was utilized for the evaluation. This model had been trained on an English training dataset, enabling it to capture semantic relationships. For each language, the model's performance was assessed using the Spearman correlation coefficient, which measures the strength and direction of association between predicted and actual scores of textual relatedness. The correlation coefficients obtained were indicative of the model's ability to accurately capture semantic textual relatedness across diverse languages. Additionally, scatter plots were generated to visualize the predicted scores against the actual scores, providing insights into the model's predictive capabilities. Overall, the multilingual testing approach allowed for a comprehensive assessment of the STR model's performance across various languages, highlighting its potential for cross-lingual text understanding applications.

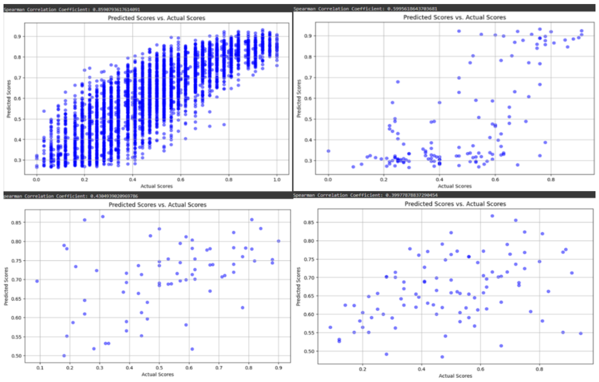


Figure 5: Predicted Scores vs Actual Scores and Spearmen Correlation Coefficient of English, Spanish, Arabic and Amharic respectively

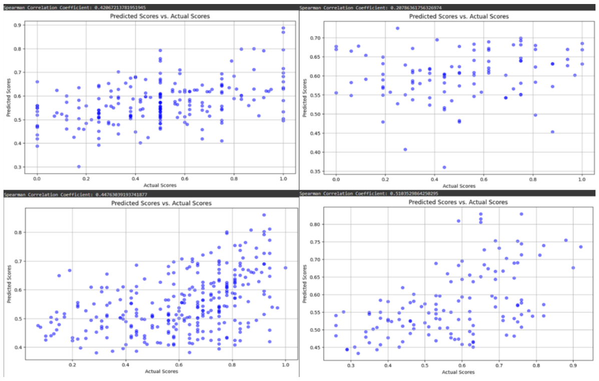
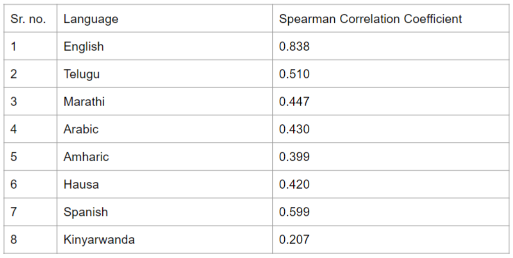


Figure 6: Predicted Scores vs Actual Scores and Spearmen Correlation Coefficient of Hausa, Kinyarwanda, Marathi and Telugu respectively

Analysis and Interpretation of results:

In Figure 7, the table displays the Spearman correlation coefficient values obtained across a spectrum of languages, providing insight into the model's proficiency in semantic textual relatedness across diverse languages. A higher correlation coefficient signifies a more robust alignment between the model's predicted scores and the actual textual relatedness scores. Notably, the model demonstrates its most substantial performance in English, achieving a Spearman correlation coefficient of 0.838. This observation suggests the model's adeptness at capturing semantic textual relatedness nuances within the English language. However, the model's efficacy varies across different languages, as evidenced by comparatively lower correlation coefficients for languages such as Kinyarwanda (0.207) and Amharic (0.399). These findings imply potential areas for further model refinement to improve its performance in languages exhibiting lower correlation coefficients. Figure 8. is a Bar graph depicting the values of the Spearman correlation coefficient values obtained by the model for a better visualization.



Sr. no.	Language	Spearman Correlation Coefficient
1	English	0.838
2	Telugu	0.510
3	Marathi	0.447
4	Arabic	0.430
5	Amharic	0.399
6	Hausa	0.420
7	Spanish	0.599
8	Kinyarwanda	0.207

Figure 7: Multililingual Testing Results

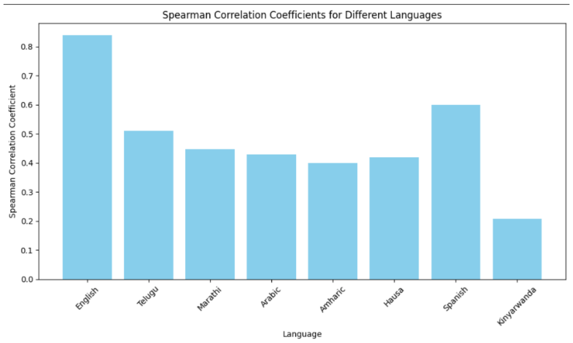
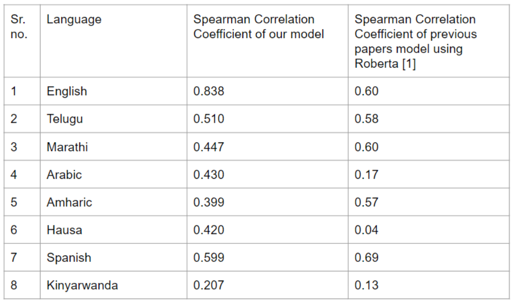


Figure 8: Multililingual Testing Results Graph

Comparison with Previous Work:

In Figures 9 and 10, we see a comparison between our model and the SemRel2024 research paper [1] regarding the Spearman Correlation Coefficient. The comparison reveals that our model performs well in predicting values accurately for languages like English, Arabic, Hausa, and Kinyarwanda. However, for languages such as Telugu, Marathi, Amharic, and Spanish, the SemRel2024 model tends to predict values more accurately than our model. This indicates that while our model shows proficiency in predicting textual relatedness for English, there's room for improvement. It's worth noting that our model sometimes overestimates or underestimates the relatedness of sentences, suggesting areas where refinement could enhance its performance.



Sr. no.	Language	Spearman Correlation Coefficient of our model	Spearman Correlation Coefficient of previous papers model using Roberta [1]
1	English	0.838	0.60
2	Telugu	0.510	0.58
3	Marathi	0.447	0.60
4	Arabic	0.430	0.17
5	Amharic	0.399	0.57
6	Hausa	0.420	0.04
7	Spanish	0.599	0.69
8	Kinyarwanda	0.207	0.13

Figure 9: Comparison with Previous Work Table

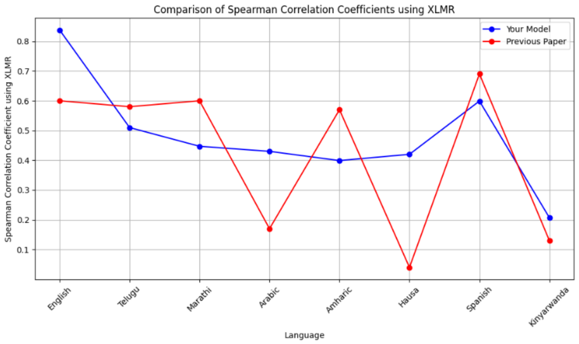


Figure 10: Comparison with Previous Work Graph

Conclusion:

In this study, we explore the concept of Semantic Textual Relatedness (STR) and its significance in natural language processing tasks. Through an extensive literature review and experimentation with models like RoBERTa and XLM-RoBERTa, we showcase the effectiveness of advanced neural network architectures in capturing semantic relationships between text pairs. Our research significantly contributes to advancing the understanding of STR and its wide-ranging applications across domains such as sentiment analysis, question-answering, and information retrieval.

The experimental findings underscore the promising performance of our proposed model in estimating textual relatedness across diverse languages and datasets. Through the utilization of pre-trained language models and fine-tuning specific STR tasks, we achieve competitive results compared to baseline models and prior approaches. The insights derived from our experiments underscore the potential of deep learning techniques in addressing intricate language understanding tasks, thereby enhancing the accuracy and efficiency of semantic textual relatedness assessment.

Limitations:

Despite the promising results obtained in our study, several limitations should be acknowledged. Firstly, the performance of the model may vary depending on the characteristics of the dataset and the complexity of the textual relationships. Limited availability of annotated data for certain languages or domains can also pose challenges in training robust STR models. Additionally, the computational resources required for training and fine-tuning large-scale language models may limit the scalability of the approach in resource-constrained environments. The lack of computational resources e.g. GPU was one of the major challenges faced during the study.

Future Scope:

Moving forward, several avenues for future research can be explored to enhance the capabilities and applicability of STR models. Firstly, expanding the scope of the study to include a wider range of languages and domains can help improve the cross-lingual and cross-domain generalization of the model. Incorporating additional contextual information, such as user preferences, temporal dynamics, and domain-specific knowledge, can further enhance the model's performance in real-world applications.

Moreover, investigating ensemble methods and multi-task learning approaches can help leverage complementary information from different sources and tasks to improve the robustness and accuracy of STR models. Exploring interpretability techniques and explainable AI methods can also enhance the transparency and trustworthiness of the model predictions, enabling users to better understand and interpret the underlying textual relationships.

Overall, the research presented in this paper lays the foundation for future advancements in semantic textual relatedness assessment and paves the way for developing more sophisticated and context-aware language understanding models tailored to specific applications and domains.

References

[1] Ousidhoum, N., Muhammad, S. H., Abdalla, M., Abdulmumin, I., Ahmad, I. S., Ahuja, S., Aji, A. F., Araujo, V., Ayele, A. A., Baswani, P., Beloucif, M., Biemann, C., Bourhim, S., De Kock, C., Dekebo, G. S., Hourrane, O., Kanumolu, G., Madasu, L., Rutunda, S., Shrivastava, M., Solorio, T., Surange, N., Tilaye, H. G., Vishnubhotla, K., Winata, G., & Yimam, S. M. (2024). SemRel2024: A Collection of Semantic Textual Relatedness Datasets for 14 Languages. *Journal/Conference Name*, Volume(Issue), Page Range.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirec-

tional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171-4186.

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Approach. *arXiv preprint arXiv:1907.11692*.

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440-8451.

[5] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 33.

[6] Yin, D., Yu, T., & Roth, D. (2016). End-to-end Relation Extraction using LSTMs on Sequences and Tree Structures. *arXiv preprint arXiv:1601.00770*.

[7] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982-3992.

[8] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic Textual Similarity—Multilingual and Cross-lingual Focused Evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1-14.

[9] Chollampatt, S., Alam, F., Subramanian, S., & Tetreault, J. (2018). A Multilingual Evaluation Benchmark for Semantic Textual Similarity. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3125-3135.