Sharif-STR at SemEval-2024 Task 1: Transformer as a Regression Model for Fine-Grained Scoring of Textual Semantic Relations

Seyedeh Fatemeh Ebrahimi [♡], **Karim Akhavan Azari**[♡], **Amirmasoud Iravani**[♯]
Hadi Alizadeh[♢], **Zeinab Sadat Taghavi**[♡], **Hossein Sameti**[♡]
Ferdowsi University of Mashhad, Mashhad, Iran[♯]
Sharif University of Technology, Tehran, Iran[♡]
Iran Broadcasting University, Tehran, Iran[♢]

{sfati.ebrahimi, karim.akhavan, zeinabtaghavi, sameti}@sharif.edu
a.iravani@mail.um.ac.ir
alizadeh.hadi08@gmail.com

Abstract

Semantic Textual Relatedness holds significant relevance in Natural Language Processing, finding applications across various domains. Traditionally, approaches to STR have relied on knowledge-based and statistical methods. However, with the emergence of Large Language Models, there has been a paradigm shift, ushering in new methodologies. In this paper, we delve into the investigation of sentence-level STR within Track A (Supervised) by leveraging fine-tuning techniques on the RoBERTa transformer. Our study focuses on assessing the efficacy of this approach across different languages. Notably, our findings indicate promising advancements in STR performance, particularly in Latin languages. Specifically, our results demonstrate notable improvements in English, achieving a correlation of 0.82 and securing a commendable 19th rank. Similarly, in Spanish, we achieved a correlation of 0.67, securing the 15th position. However, our approach encounters challenges in languages like Arabic, where we observed a correlation of only 0.38, resulting in a 20th rank.

1 Introduction

STR delineates the meaningful association between linguistic units, showcasing conceptual proximity within a shared semantic frame (Taieb et al., 2019; Abdalla et al., 2021). For instance, "cup" and "coffee" are related in meaning, yet they are not synonymous (Jurafsky and Martin, 2009). Despite its crucial role in various NLP applications such as Spelling Correction, Word Sense Disambiguation, Plagiarism Detection, Opinion Mining, and Information Retrieval (Franco-Salvador et al., 2016; Chen et al., 2017; Taieb et al., 2019), STR has garnered less attention compared to Semantic Textual Similarity (STS) due to a scarcity of available datasets. Addressing this gap, Abdalla et al. (2021), and Ousidhoum et al. (2024a) contributed to the field by constructing the first sentence-level STR datasets. In this paper, we endeavor to tackle the STR problem within shared Task 1(Ousidhoum et al., 2024b), Track A, leveraging supervised data in English, Spanish, and Arabic languages provided by Ousidhoum et al. (2024a). Additionally, we briefly explore Track C and provide supplementary details in Appendix B as a secondary objective.

Building upon the findings of Abdalla et al. (2021), which underscore the superior performance of fine-tuning Transformer models in supervised tasks, our proposed system captures the relationship among sentences by fine-tuning the RoBERTa Transformer (Liu et al., 2019). At the core of our system, we employ a pre-trained RoBERTa model as a regression model and fine-tune it to generate a floating-point value for the input text. During the pre-training process of RoBERTa, the emphasis is placed on tasks related to NLU. This involves exposing the model to a diverse range of linguistic contexts and training it to comprehend the nuances of language. Furthermore, the integration of a Classifier Head enables sentence classification, a pivotal aspect of our system architecture elaborated upon in section 3.

Our experimental results showcase promising performance on English and Spanish datasets, achieving respective correlation rates of 0.82 and 0.67 on test data, surpassing the baseline correlation set by SemEval-2024 at Subtask A (Ousidhoum et al., 2024b). However, the model's performance on Arabic data falls short, yielding only a 38% correlation on development data. We attribute this discrepancy to differences in the underlying RoBERTa model and its training methodology across Latin and non-Latin languages, a topic further explored in section 5. To promote reproducibility and facilitate future research endeavors,

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 1043–1052

the complete codebase of our project has been shared on GitHub[1].

2 Background

2.1 Dataset Overview

The SemEval-2024 Task 1 is structured into Tracks A, B, and C, each tailored to specific methodologies and objectives. Our focus lies on Track A (Supervised), which utilizes labeled data to train STR systems. The datasets for Task 1 encompass training, development, and test sets across 14 languages, each comprising sentence pairs (Ousidhoum et al., 2024a). Each sentence pair is annotated with a semantic relatedness score, ranging from 0 (indicating no relatedness) to 1 (suggesting strong relatedness). Participants are tasked with predicting the degree of semantic relatedness between sentence pairs, crucial for furthering research in NLP.

2.2 Related Work

The exploration of sentence-level STR has been hindered by the scarcity of available datasets (Abdalla et al., 2021). Existing datasets, such as those compiled by Finkelstein et al. (2002), Gurevych (2006), Panchenko et al. (2016), and Asaadi et al. (2019), predominantly focus on unigram and bigram STR. However, the seminal works of Abdalla et al. (2021), and Ousidhoum et al. (2024a) paved the way for further research by constructing the first sentence-level STR datasets. Traditionally, both STR and STS have been approached using knowledge-based and statistical methods (Sadr, 2020; Chandrasekaran and Mago, 2020). Notable efforts include the application of knowledge bases such as thesauri, ontologies, and dictionaries for STR, as surveyed by Salloum et al. (2020). Statistical methods, on the other hand, leverage features extracted from corpora, with prominent examples including Latent Dirichlet Allocation (LDA) by Blei et al. (2009) and Latent Semantic Analysis (LSA) by Landauer and Dumais (2008) for topic modeling.

In recent years, the application of deep learning methodologies has surpassed traditional approaches in STS tasks. Noteworthy advancements include the Tree-LSTM model proposed by Tai et al. (2015), which outperformed other neural network models in SemEval-2014. He and Lin (2016) introduced a hybrid architecture of Bi-LSTM and CNN, outperforming the Tree-LSTM model on the SICK dataset. Wang et al. (2016) achieved state-of-the-art results using the Word2Vec embeddings model in both the QASent and the WikiQA datasets, while Shao (2017) leveraged GloVe embeddings to achieve the third rank in SemEval-2017.

Several studies have demonstrated that fine-tuning transformer-based models achieves state-of-the-art in comprehending the semantics of textual data. The transformer model, first introduced by Vaswani et al. (2017), employs attention mechanisms to capture word semantics. Later on, Devlin et al. (2019) utilized it to create BERT word embeddings. Subsequently, XLNet, proposed by Yang et al. (2019), surpassed BERT in performance. Consequently, Lan et al. (2019) introduced ALBERT, which outperforms previous models. Additional transformer-based variations of BERT models include TinyBERT (Jiao et al., 2020), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). Also, Raffel et al. (2019) presented five distinct versions of the T5 transformer model, each varying in parameter size. Their work demonstrated that the performance of these pretrained models improves with larger datasets and enhanced computational resources.

Laskar et al. (2020) addressed sentence similarity modeling within an answer selection task. Through experiments conducted, they showed that fine-tuning RoBERTa model achieves state-of-the-art performance across datasets. Yang et al. (2020) showcased that the RoBERTa-based model achieved superior performance compared to the BERT and XLNET models in a clinical STS task, achieving a Pearson Correlation of 0.90. Similarly, Huang et al. (2021) conducted a comparison of TF-IDF combined with various models including ALBERT, BERT, and RoBERTa for word similarity detection in sentence pairs within Task 2 of SemEval-2021. Their experimental findings substantiated that RoBERTa yielded superior results by 0.846 on the test data. Nasib (2023) addressed reference validation task by employing BERT, SBERT, and RoBERTa. His study illustrated the efficacy of fine-tuning a RoBERTa-based model for text classification tasks, achieving state-of-the-art performance across multiple benchmark datasets. He emphasized that optimizing the model's performance involves activ-

[1] https://github.com/Sharif-SLPL/Sharif-STR

ities such as hyperparameter tuning, regularization, and data augmentation.

Abdalla et al. (2021) conducted an extensive investigation into semantic sentence representation methods, revealing that supervised methods utilizing contextual embeddings, particularly those fine-tuning BERT or RoBERTa, outperform other techniques, reaching a correlation of 0.83. Building upon these findings, we adopt fine-tuning RoBERTa as the primary strategy in this paper. Subsequent sections will detail our system architecture.

3 System Overview

In this section, we present a comprehensive overview of our system's architecture, outlining the key algorithms and modeling decisions that underpin our model.

3.1 Core Algorithms and System Architecture

Our system harnesses the Transformer architecture for its ability to capture long-range dependencies. At its core, we harness the power of a pre-trained RoBERTa model (Liu et al., 2019) for regression analysis, tailoring its parameters to accurately predict a floating-point value from the input text. While RoBERTa isn't explicitly trained for sentence relatedness scoring, its training encompasses an understanding of the relatedness of sentences within discourse, rendering it suitable for our task.

During the pre-training process of RoBERTa, the emphasis is placed on tasks related to NLU. This involves exposing the model to a diverse range of linguistic contexts and training it to comprehend the nuances of language. Our word embeddings utilize an embedding matrix with a dimensionality of 768. Position embeddings and token type embeddings further contribute to the model's comprehension of sequential and contextual information within the input data.

The RobertaEncoder comprises a stack of 12 identical RobertaLayers, each employing a multi-head self-attention mechanism. This mechanism enables the model to concurrently absorb different parts of the input sequence, showing promise in analyzing similarities between various inputs. Following the attention mechanism are intermediate sub-layers and output sub-layers. The intermediate sub-layer employs a fully connected feed-forward network with a GELU activation function, while the output sub-layer is responsible for proper transformation and normalization of features.

The classification head, positioned after the encoder, is tasked with generating the final output for sequence classification. It consists of a linear layer with 768 input features, followed by a dropout layer to prevent over-fitting. An additional linear layer featuring a solitary output neuron enables binary classification. By viewing the problem as a regression task, the classifier yields a linear output designed for a singular class, producing a probabilistic value indicative of the relatedness between input sentences.

3.2 Resources

For training our model, we relied on the dataset provided for SemEval-2024 Task 1 (Ousidhoum et al., 2024a). In addition to the primary dataset, we augmented our training dataset using the T5 model (Raffel et al., 2019). By leveraging T5's paraphrasing capabilities, we explored data augmentation techniques for Track A on the training sets of our dataset but failed to achieve consistent results across experiments. While some experiments showed an increase in model accuracy, in other cases, it did not alter the results. Data augmentation consistently worked well only on the English dataset. More details about data augmentation results and our secondary investigation on Track C are provided in Appendix A and B.

By incorporating both the SemEval-2024 Task 1 dataset (Ousidhoum et al., 2024a) and augmented training data generated by T5, our approach benefits from a comprehensive and diverse set of resources, enabling robust training and evaluation of our STR model across multiple languages and textual domains.

3.3 System Challenges

Augmenting the dataset for training set using T5 paraphrases posed several challenges. Firstly, while the primary dataset was labeled through collaborative human judgment, the augmented data lacked this human validation. This absence of human labeling for the augmented data may potentially impact its quality. Moreover, the augmentation process introduced alterations to the diversity of the data, presenting a challenge to maintaining the original data variety.

The decision to employ data augmentation exclusively for testing purposes raises concerns re-

garding its potential impact on model quality. Addressing these challenges associated with data augmentation is crucial for improving the efficacy of our model. Exploring solutions to mitigate these issues can enhance our approach to tackling the task at hand.

4 Experimental Setup

4.1 Dataset

The dataset statistics utilized for each language are presented in Table 1:

As shown in Table 1, approximately 0.8 of the Task 1 dataset is allocated for system training, while the remainder is reserved for evaluation. The limited availability of training data necessitates cautious consideration during testing, as the model's performance may be influenced by the scarcity of training instances. Additionally, the entire development set is utilized for model selection.

4.2 Pre-processing and Hyper-Parameter Tuning

A crucial aspect of our pre-processing involves converting the labels (scores) of each data instance to float values, ensuring compatibility with the model's expected input format. Furthermore, the input texts undergo tokenization using the RoBERTa-tokenizer both during training and inference.

Hyperparameter tuning plays a pivotal role in optimizing model performance. Our tuning process encompasses exploring various hyper-parameters, including learning rates in the range of [0.00001, 0.00003], dropout rates ranging from [0.1, 0.3], batch sizes spanning [4, 32], and token sizes from [32, 128]. Through iterative experimentation, we determined that a learning rate of 0.00003, a dropout rate of 0.1, a token size of 128, a batch size of 16, and a weight decay of 0.01 yield optimal results across all languages.

The selection of an appropriate token size is not solely based on computational considerations; rather, it is informed by dataset analysis. Upon examination, it became evident that the majority of data instances are predominantly short, aligning with our token size choice. Additionally, truncation during tokenization supports the chosen token size, ensuring efficient model training without sacrificing data representativeness.

4.2.1 Mean Squared Error (MSE)

Mean Squared Error quantifies the average of the squared differences between predicted and actual values. It is calculated using the formula:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (1)$$

Where N is the number of instances, y_i is the true label, and $\hat{y}_i$ is the predicted value.Additionally, Mean Absolute Error computes the average absolute differences between predicted and actual values.Moreover, the R-squared score assesses the proportion of variance in the dependent variable explained by the independent variable.

These evaluation measures collectively shed light on our regression model's performance in predicting the degree of relatedness between text samples. Using these metrics together enables the monitoring of the model's performance and, hence, facilitates decisions on hyper-parameters, model selection, etc. The evaluation method and hyper-parameter choices remain consistent across all models and languages. For the analysis of results presented in Section 5, the obtained scores were discretized and categorized into five distinct ranges to enhance visual understanding.

5 Results

5.1 Findings

A direct comparison with previous models and datasets similar to this task is challenging due to our specific focus on fine-tuning the RoBERTa model and utilizing the dataset provided by Ousidhoum et al. (2024a). Drawing from the insights of Raffel et al. (2019) working on the STS dataset, it is evident that the performance of transformer models improves with larger training corpora and enhanced computational resources. Raffel et al. (2019) demonstrated that the RoBERTa transformer-based model achieved a Pearson correlation of 0.922, surpassing ERNIE 2.0, DistilBERT, and TinyBERT on STS dataset benchmarks. Conversely, ALBERT, XLNet, and T5-11B outperformed RoBERTa on the same task, achieving a Pearson correlation of 0.925. Therefore, we recommend conducting a benchmark study of top-performing transformer models like RoBERTa, ALBERT, XLNet, and T5-11B in future research endeavors. Using the official metric of Spearman Correlation proposed in SemEval-2024 Task 1 (Ousidhoum et al., 2024b), our system achieves the following scores on different data splits and languages:

As shown in Table 2, Firstly, comparing the performance between English, Spanish, and Arabic

Language/Split	Dataset	Train	Testset	Devset
English	5752	4400	1101	251
Spanish	1702	1249	313	140
Arabic	1360	1009	252	97

Table 1: Dataset Statistics

Language/Split	Devset	Testset(Competition)
English	0.83	0.82
Spanish	0.71	0.67
Arabic	0.32	0.38

Table 2: Correlation Metric Scores

models, we observe varying degrees of success. The English model demonstrates the highest Spearman Correlation scores, both on the development and test sets, with scores of 0.83 and 0.82, respectively. This indicates that the English model performs relatively well in capturing the semantic relatedness between text pairs. Similarly, the Spanish model also achieves respectable scores, albeit slightly lower, with scores of 0.71 on the development set and 0.67 on the test set. However, the Arabic model lags significantly behind, exhibiting notably lower scores of 0.32 on the development set and 0.38 on the test set.

The disparity in performance between the Arabic model and the English and Spanish models could be attributed to several factors. One possible explanation is the availability and quality of training data. The Arabic dataset may suffer from a scarcity of labeled instances, resulting in a less robust model. Additionally, linguistic and structural differences between Arabic and Latin languages may pose challenges for the model in accurately capturing semantic relatedness. This discrepancy underscores the importance of adequately addressing language-specific characteristics and challenges in model development.

Furthermore, the analysis of the Arabic model's performance on the test set reveals a noteworthy observation. Despite achieving a relatively low Spearman Correlation score, the model appears to disproportionately classify most inputs as highly related. This discrepancy suggests a potential limitation in the model's ability to discern varying degrees of relatedness accurately. It implies that while the model may perform adequately in certain aspects, such as overall correlation with human annotations, it may struggle with nuanced interpretations of relatedness levels in real-world scenarios. The output of the model is provided in Appendix D.

The scatter plots depicted in Figure 1, respectively for English, Spanish, and Arabic, illustrate the correlation between the model predictions and human annotations. The English model closely aligns with human annotations, while the Spanish model exhibits an even closer alignment on certain inputs. However, the Arabic model's performance varies, indicating discrepancies between predicted and actual relatedness scores. These findings underscore the importance of dataset size and linguistic nuances in model performance across different languages. Further investigation is warranted to elucidate the factors influencing model behavior and to improve performance, particularly in languages with limited training data.

5.2 Error Analysis

While confusion matrices are less commonly utilized in regression problems, discretizing the model's scores allows us to glean insights into its performance. Confusion matrix plots for English, Spanish, and Arabic are provided in Figure 2, respectively. Upon examining the confusion matrix of the English dataset, it becomes apparent that the model performs well within certain score ranges. However, there are notable areas, particularly within the highly related range (0.6-1.0), where our model could benefit from improvement.

A similar observation holds true for the Spanish dataset, where the model demonstrates proficiency in predicting less related sentences but encounters challenges with highly related ones. Conversely, the Arabic dataset presents a markedly different scenario. While the majority of predictions fall within the mid-range of relatedness, they are predominantly incorrect.

Based on the histogram and extracted statistics from the fine-tuning data in Figure 3 in Appendix C, it appears that the majority of the training data has a distribution centered around the median (Spanish

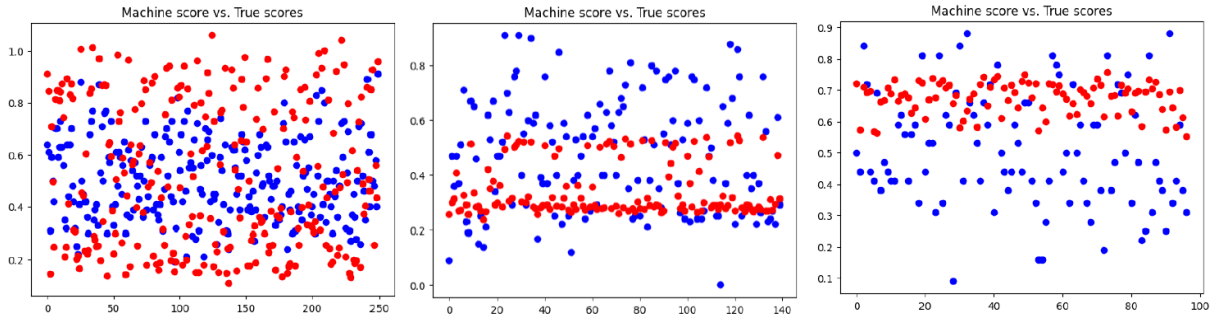

Figure 1: Scatter Plots of English, Arabic and Spanish Languages

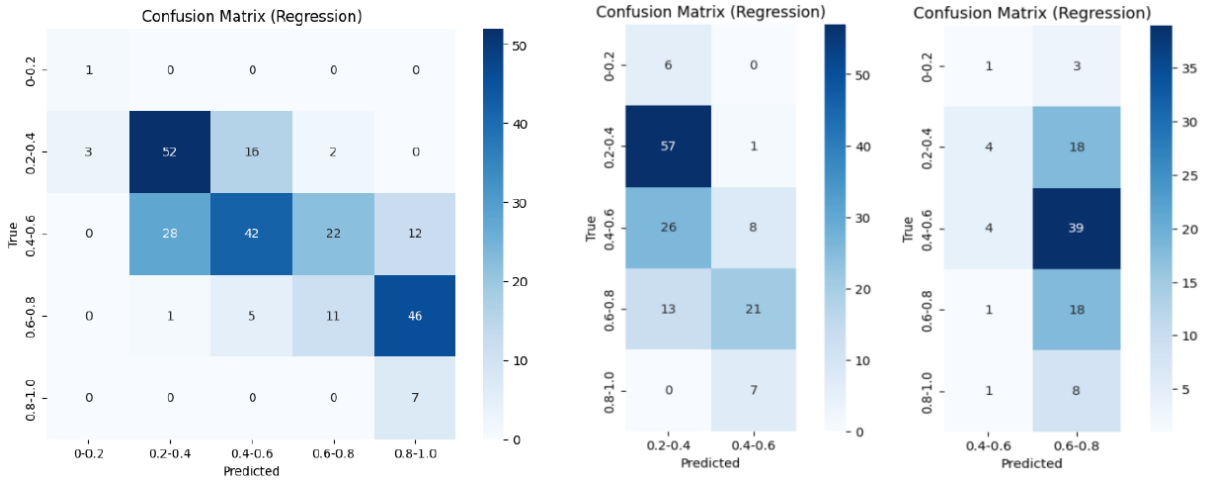

Figure 2: The Confusion Matrix Plot of English, Arabic and Spanish Languages

Mean Score: 0.43, Arabic Mean Score: 0.50). Consequently, finte-tuned Arabic and Spanish models seem to have less capability in understanding data on both ends of the spectrum.

These insights highlight the model's strengths and weaknesses across different datasets and underscore the need for further investigation into improving performance, particularly in accurately predicting highly related sentences across all languages. Further exploration of the factors contributing to model errors, such as dataset characteristics and linguistic nuances, is essential for refining the model's predictive capabilities.

6 Conclusion

In our investigation, we focused on fine-tuning RoBERTa for STR, primarily targeting Latin languages like English(0.82) and Spanish(0.67). While our approach showed promising results for these languages, particularly in achieving high correlation, the outlook was less favorable for Arabic(0.38). This echoes discussions in previous works, emphasizing the significant influence of the data on model performance. Our exploration into Track C, which is given in Appendix B, further enriched our understanding of the challenges and opportunities in STR system development. As a contribution to the field, we put forth several recommendations for enhancing STR systems. Firstly, we propose the development of additional Transformer models trained on diverse language families, focusing on languages that share similarities with Latin languages. Furthermore, a comprehensive benchmark of models on the STR dataset is essential, building on previous research that highlights the strong performance of models like ALBERT, XLNet, and T5-11B on the STS dataset. Moreover, the utilization of translation techniques and data augmentation methods could enhance model performance, particularly for languages with limited training data. In conclusion, our study sheds light on the nuances of STR system development and underscores the importance of considering language-specific factors and domain characteristics. By pursuing the avenues outlined in this paper, we aim to contribute to the advancement of STR research and facilitate the development of more robust and accurate models for NLU tasks.

Acknowledgments

We express our gratitude to the Speech and Language Processing Laboratory at Sharif University of Technology[2] for offering us the opportunity for collaborative work.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif M. Mohammad. 2021. What makes sentences semantically related? a textual relatedness dataset and empirical study. *ArXiv*, abs/2110.04845.

Shima Asaadi, Saif M. Mohammad, and Svetlana Kiritchenko. 2019. Big bird: A large, fine-grained, bigram relatedness dataset for examining semantic composition. In *North American Chapter of the Association for Computational Linguistics*.

David M. Blei, A. Ng, and Michael I. Jordan. 2009. Latent dirichlet allocation.

Dhivya Chandrasekaran and Vijay Mago. 2020. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54:1 – 37.

Fuzan Chen, Chenghua Lu, Harris Wu, and Minqiang Li. 2017. A semantic similarity measure integrating multiple conceptual relationships for web service discovery. *Expert Syst. Appl.*, 67:19–31.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20:116–131.

Marc Franco-Salvador, Paolo Rosso, and Manuel Montes y Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. *Inf. Process. Manag.*, 52:550–570.

Iryna Gurevych. 2006. Thinking beyond the nouns - computing semantic relatedness across parts of speech.

Hua He and Jimmy J. Lin. 2016. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. In *North American Chapter of the Association for Computational Linguistics*.

Bo Huang, Yang Bai, and Xiaobing Zhou. 2021. hub at semeval-2021 task 2: Word meaning similarity prediction model based on roberta and word frequency. In *International Workshop on Semantic Evaluation*.

[2] https://github.com/Sharif-SLPL

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing*, 2. ed., [pearson international edition] edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, London [u.a.].

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Thomas K. Landauer and Susan T. Dumais. 2008. Latent semantic analysis. *Scholarpedia*, 3:4356.

Md Tahmid Rahman Laskar, Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *International Conference on Language Resources and Evaluation*.

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*, abs/1907.11692.

Abdullah Umar Nasib. 2023. References validation in scholarly articles using roberta. Project report, Brac University.

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.

Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia V. Loukachevitch, and Chris Biemann. 2016. Human and machine judgements for russian semantic relatedness. *ArXiv*, abs/1708.09702.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Hossein Sadr. 2020. Exploring the efficiency of topic-based models in computing semantic relatedness of geographic terms.

Said A. Salloum, Rehan Khan, and Khaled F. Shaalan. 2020. A survey of semantic analysis approaches. In *International Conferences on Artificial Intelligence and Computer Vision*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yang Shao. 2017. Hcti at semeval-2017 task 1: Use convolutional neural network to evaluate semantic textual similarity. In *International Workshop on Semantic Evaluation*.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *ArXiv*, abs/1503.00075.

Mohamed Ali Hadj Taieb, Torsten Zesch, and Mohamed Ben Aouicha. 2019. A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, 53:4407 – 4448.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *International Conference on Computational Linguistics*.

Xi Yang, Xing He, Hansi Zhang, Yinghan Ma, Jiang Bian, and Yonghui Wu. 2020. Measurement of semantic textual similarity in clinical texts: Comparison of transformer-based models. *JMIR Medical Informatics*, 8.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

A Data Augmentation Results

As we describe data augmentation in section 3.2, we use T5 model to augment some training data and use them in training of model.So in this section we show results of data augmentation effect on Pearson Correlation for English language in Table 3.

Model hyper parameters		without data augmentation	with data augmentation
Learning rate	3e-5	0.79	0.81
Max length	128		
Batch size	16		
Epoch	4		

Table 3: Data Augmentation Affect on Pearson Correlation

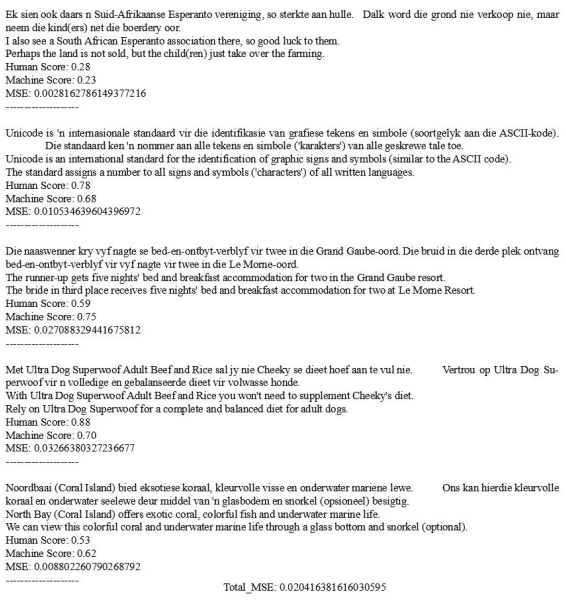

Figure 3: Output of Afrikaans

B Track C - Cross-Lingual

Using the translation method in Track C, we employed our Track A model trained on English language. The input sentences were first translated into English using the Google Translate API, followed by the utilization of the trained Track A model. The evaluation results demonstrate promising performance across some languages with this approach. However, errors might arise from either the Google Translate API or the model itself. Exploring alternative translation APIs could potentially enhance the overall performance. Figures 3, 4, and 5 display the outputs in Afrikaans, Amharic, and Modern Standard Arabic. Additionally, the high-quality output images are provided in our GitHub project.

Test Data	Pearson Correlation	MSE
afr_test_with_labels.csv	0.8	0.0204
amh_test_with_labels.csv	0.73	0.0309
arb_test_with_labels.csv	0.51	0.0431

Table 4: Track C Results

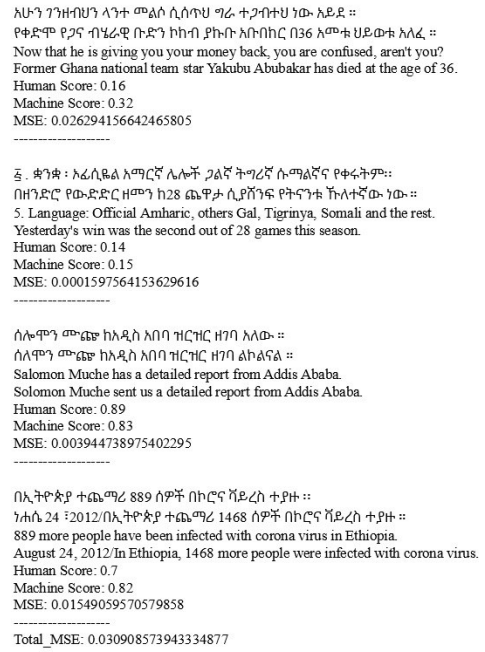

Figure 4: Output of Amharic Language

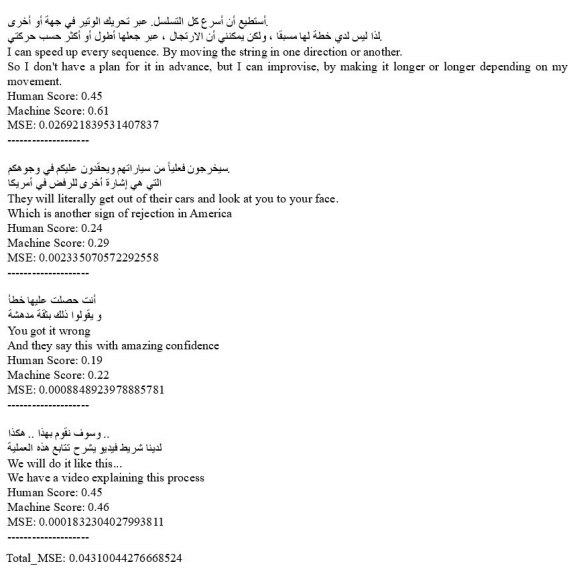

Figure 5: Output of Modern Standard Arabic

C Histogram of Spanish and Arabic Languages

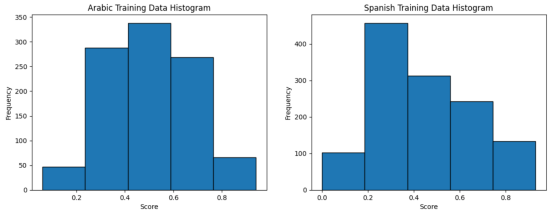

Figure 6: Histogram of Spanish and Arabic Training Dataset

D Outputs of Track A (Supervised)

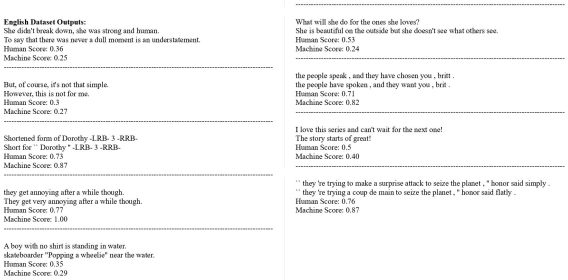

Figure 7: Output of English Language

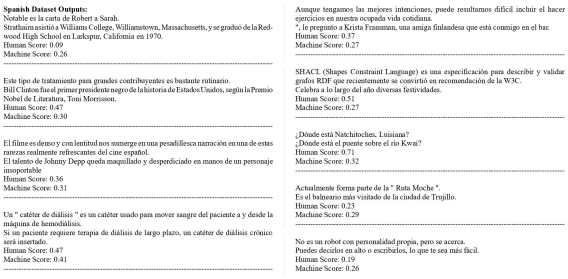

Figure 8: Output of Spanish Language

Figure 9: Output of Arabic Language