# NRK at SemEval-2024 Task 1: Semantic Textual Relatedness through Domain Adaptation and Ensemble Learning on BERT-based models

**Nguyen Tuan Kiet**[1,2] and **Dang Van Thin**[1,2]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
21521042@gm.uit.edu.vn, thindv@uit.edu.vn

## Abstract

This paper describes the system of the team NRK for Task A in the SemEval-2024 Task 1: Semantic Textual Relatedness (STR). We focus on exploring the performance of ensemble architectures based on the voting technique and different pre-trained transformer-based language models, including the multilingual and monolingual BERTology models. The experimental results show that our system has achieved competitive performance in some languages in Track A: Supervised, where our submissions rank in the Top 3 and Top 4 for Algerian Arabic and Amharic languages. Our source code is released on the GitHub site[1].

## 1 Introduction

The SemEval-2024 Task 1 (Ousidhoum et al., 2024b) aims at detecting the degree of semantic relatedness between pairs of sentences across 14 different languages, encompassing Afrikaans, Algerian Arabic, Amharic, English, Hausa, Hindi, Indonesian, Kinyarwanda, Marathi, Moroccan Arabic, Modern Standard Arabic, Punjabi, Spanish, and Telugu. This shared task has three main tasks, each focusing on different aspects of predicting semantic textual relatedness within sentence pairs.

Semantic Textual Relatedness (STR) is a task in Natural Language Processing (NLP) that aims to measure the degree of semantic relatedness between two text passages, typically sentences. STR plays a crucial role in various NLP applications, as it allows computers to understand the relationships between different pieces of text. As mentioned in (Abdalla et al., 2023), it is also employed in chatbots and dialogue systems to understand the user's intent and in question-answering systems to identify answer passages that are semantically related to the question. Additionally, STR finds applications in text summarization, where it helps identify the most important and semantically relevant sentences to create a concise summary of a longer document. STR also plays a role in text generation tasks, such as machine translation and dialogue systems, by guiding the model to generate text that is semantically related to the input or context. However, accurately measuring STR presents several challenges. One key challenge lies in capturing the nuances of language, such as synonyms, paraphrases, and ambiguity. Another challenge is dealing with different languages and cultural contexts, where semantic relationships might not be directly translatable.

Our team only focuses on addressing Track A in the shared task. Our approach is based on the domain adaption for different transformer-based models, and then we continue to fine-tune the pre-trained transformer-based models on the task-specific training data. Therefore, our system is able to leverage domain-specific knowledge to improve performance. Subsequently, we train a cross-encoder model on the adapted transformer-based models, harnessing its ability to capture semantic relatedness between sentence pairs effectively. To further enhance the robustness and performance of our predictions, we adopt a weighted voting technique to combine the outputs of multiple models.

## 2 Background

### 2.1 Problem Description

This study investigates the task of predicting Semantic Textual Relatedness (STR) between sentence pairs across 14 languages. Each sentence pair will be associated with a human-annotated relatedness score ranging from 0 (completely unrelated) to 1 (maximally related). There are three Tracks for participants, however, in our work, we only focus on Track A: The first task entails a supervised approach, wherein participants are tasked with developing systems that leverage labelled training

[1] https://github.com/KiRzEa/Semeval2024-SemanticTextualRelatedness

*Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 76–81
June 20-21, 2024 ©2024 Association for Computational Linguistics

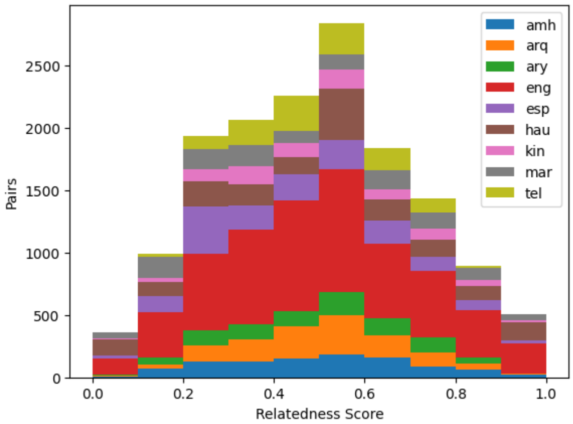

Figure 1: Relatedness Score distribution over languages on the training set.

datasets to infer the degree of semantic relatedness between sentence pairs.

### 2.2 Data Description

The dataset (Ousidhoum et al., 2024a) typically contains pairs of text along with their corresponding relatedness score, which indicates how semantically related the two fragments are.

Figure 1 shows the distribution of relatedness score over languages. Among the languages included in the dataset, English comprises the largest subset of sentence pairs. The remaining languages also contribute sentence pairs, albeit with varying degrees of representation. It is notable that while most languages exhibit relatedness score distributions spanning the entire range of 0 to 1, some languages demonstrate more limited distributions.

## 3 Related Work

STR is a fundamental concept which has been considered as an important role in language understanding tasks. Historically, many previous studies focused on semantic similarity, which aims to measure the likeness or resemblance between linguistic elements based on their meaning (Abdalla et al., 2023). Unlike semantic similarity, which often involves assessing the degree of overlap or similarity in meaning between words or phrases, STR involves determining the overall relatedness or closeness in meaning between pairs of sentences or longer textual units (Mohammad and Hirst, 2012). (Gabrilovich et al., 2007) proposed a novel method called Explicit Semantic Analysis (ESA) for fine-grained semantic representation of unrestricted natural language texts. The effectiveness of ESA is evaluated by automatically computing the degree of semantic relatedness between fragments of natural language text. Hussain et al. (2023) proposed a novel vector space model for computing semantic similarity and relatedness between concepts by aggregating taxonomic features from WordNet and Wikipedia.

With the emergence of deep learning models, Gu et al. (2023) introduced a novel Siamese Manhattan LSTM-SNP approach (SiMaLSTM-SNP) which combines Word2Vec and a 10-layer Attention strategy to represent and extract sentence pairs. The multi-head self-attention layer identifies text associations and redistributes hidden state weights. The last hidden state is extracted, and the relatedness score is calculated using the Manhattan distance. Hany et al. (2023) employed a two-layered approach. Firstly, embedding similarity techniques were utilized, leveraging seven different transformers to obtain vectors for each pair of sentences. Secondly, a classical machine learning regressor was trained on these seven vectors. This research highlights the potential of combining embedding similarity techniques with machine learning methods to enhance relatedness score assessment and other NLP tasks.

## 4 System Description

### 4.1 Approach

The diagram in Figure 2 illustrates our ensemble approach for Task A. The framework consists of two main layers: a layer of cross-encoder model, and a voting ensemble layer. Firstly, the input sentence pair is passed through a single encoder to produce a joint representation which captures the semantic relationship between the two sentences in the pair and produces a number ranging from 0 to 1. Following this, the predictions of chosen models are combined using the weighted voting technique with each weight determined by its performance in the development phase.

Our approach commences with domain adaptation on masked language modeling (MLM) task (3) which has been shown a powerful training strategy for learning sentence embeddings (Gururangan et al., 2020). To achieve this, we leverage each sentence in the sentence pairs of the training dataset to train MLM which is called In-domain corpus in Figure 2. This process involves masking certain tokens within the input sentences and training the model to predict the masked tokens based

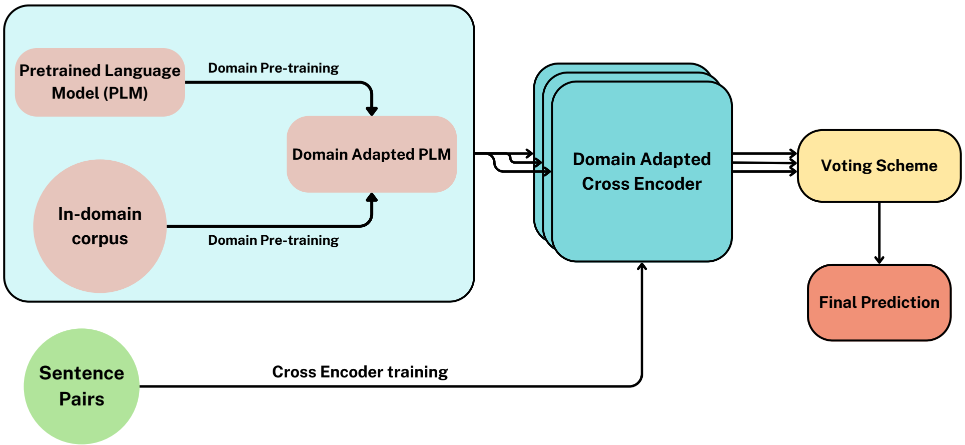

Figure 2: The overall framework of our system for the Track A: Supervised in the Semantic Textual Relatedness shared task.

on their context. In the next stage, we employ a cross-encoder architecture from Sentence-BERT (Reimers and Gurevych, 2019) which is a variant of the BERT model specifically designed for generating fixed-size sentence embeddings that capture semantic similarity between sentences. The cross-encoder architecture of SBERT processes sentence pairs jointly, encoding them into dense fixed-size vectors while considering their contextual information and semantic relationships. After obtaining the logits, we apply the sigmoid function to transform the logits into scores ranging from 0 to 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

This transformation ensures that the output scores are normalized and represent the degree of semantic relatedness between sentence pairs. To optimize the model during training, we utilize Binary CrossEntropy loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{2}$$

**Fine-tuning Language Model:** As can be seen in Figure 2, we utilize the power of pre-trained contextual language models, encompassing BERT-based models which are BERT (**?**), DeBERTa-V3 (He et al., 2022), XLM-RoBERTa (Conneau et al., 2019) and E5 (Wang et al., 2022). To fine-tune the language models, we followed the approach of (Devlin et al., 2019), which is presented in detail below.

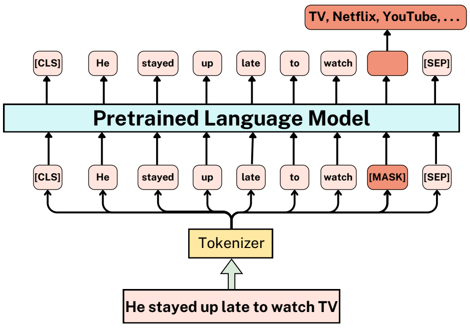
TV, Netflix, YouTube, . . .
[CLS]
He
stayed
up
late
to
watch
[SEP]
Pretrained Language Model
[CLS]
He
stayed
up
late
to
watch
[MASK]
[SEP]
Tokenizer
He stayed up late to watch TV

Figure 3: Masked language modelling task illustration for BERT-based models.

**Voting Scheme:** Our motivation for applying an ensemble approach is to take advantage of the performances of various models. Given predictions $\{\hat{y}_{\theta_1}, \hat{y}_{\theta_2}, .., \hat{y}_{\theta_n}\}$ of the $n$ base regressors. We applied the weighted voting technique to merge the predictions of the base models. In our case, the individual regressors are treated based on their performance in the evaluation phase. We compute the weighted sum of the output of n regressors as the final prediction.

### 4.2 Pre-trained Contextual Language Models

We briefly explain the pre-trained language models used in this paper.

- **mBERT**: we use the multilingual version of BERT (Devlin et al., 2019) which is trained

Table 1: Results of our best submission compared with two top systems on 9 languages for Track A.

| **Track A1: Algerian Arabic** | | **Track A2: Amharic** | | **Track A3: English** | |
|---|---|---|---|---|---|
| **Team** | **Score** | **Team** | **Score** | **Team** | **Score** |
| Top 1 | 0.6823 | Top 1 | 0.8886 | Top 1 | 0.8596 |
| Top 2 | 0.6788 | Top 2 | 0.8878 | Top 3 | 0.8532 |
| **Ours (Top 3)** | **0.6736** | **Ours (Top 4)** | **0.8641** | **Ours (Top 14)** | **0.8352** |

| **Track A4: Hausa** | | **Track A5: Kinyarwanda** | | **Track A6: Marathi** | |
|---|---|---|---|---|---|
| **Team** | **Score** | **Team** | **Score** | **Team** | **Score** |
| Top 1 | 0.7642 | Top 1 | 0.8169 | Top 1 | 0.9108 |
| Top 2 | 0.7472 | Top 2 | 0.8134 | Top 2 | 0.8968 |
| **Ours (Top 8)** | **0.6719** | **Ours (Top 6)** | **0.7568** | **Ours (Top 6)** | **0.8792** |

| **Track A7:Moroccan Arabic** | | **Track A8: Spanish** | | **Track A9: Telugu** | |
|---|---|---|---|---|---|
| **Team** | **Score** | **Team** | **Score** | **Team** | **Score** |
| Top 1 | 0.8625 | Top 1 | 0.7403 | Top 1 | 0.8733 |
| Top 2 | 0.8596 | Top 2 | 0.7310 | Top 2 | 0.8643 |
| **Ours (Top 6)** | **0.8269** | **Ours (Top 12)** | **0.6898** | **Ours (Top 8)** | **0.8341** |

on the top 104 languages with the largest Wikipedia using a masked language modelling (MLM) objective with case sensitivity.

- **XLM-R**: XLM-R (Conneau et al., 2020) is another multilingual language model. It is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

- **mDeBERTa-V3**: a DeBERTa (He et al., 2020) version improved the efficiency of original DeBERTa using ELECTRA-Style pre-training with Gradient Disentangled Embedding Sharing (He et al., 2022). In our case, we choose the multilingual version of DeBERTa-V3 which was pre-trained only on the ConmmonCrawl dataset and other versions, which are fine-tuned on the XNLI dataset and multilingual-NLI-26lang-2mil7 dataset (Laurer et al., 2024), respectively.

- **E5**: E5 (Wang et al., 2022) is trained in a contrastive manner with weak supervision signals from our curated large-scale text pair dataset. We chose monolingual (which is trained only in English) and multilingual versions for our task.

## 5 Experimental Setup

**Data and Pre-processing**: We utilized the official training set for training models. The development set was used to determine the weights for each model chosen to apply the voting technique based on their performance.

**Configuration Settings**: We implemented our models using the Trainer API from the Hugging Face library (Wolf et al., 2020) for the MLM task and employed the Cross Encoder architecture from SBERT (Reimers and Gurevych, 2019) for the Cross Encoder task.

- **MLM Task**: The maximum input length is set to 512 tokens, and the number of epochs is set to 10 with a batch size of 16 for all languages. During the training phase of the MLM, we set the MLM probability to 0.15, which means a token will be replaced with the [MASK] token in the input sequence with a probability of 0.15.

- **Cross Encoder Task**: The maximum input length is set to 512 tokens, and the number of epochs is set to 10 with a batch size of 16 for all languages.

We used the AdamW optimizer with a linear schedule warm-up technique for both the MLM task and the Cross Encoder task.

**Submission Systems**: We submitted the performance of the ensemble weighted voting model for all languages for both the development phase and evaluation phase and as mentioned above, the weights of each model based on its performance in the development phase and determined manually.

## 6 Results and Discussion

In this section, we present the official results of our final submission model for Track A in the SemEval

Table 2: Results of all the base models and our ensemble models on the development dataset.

| **Track A1: Algerian Arabic** | | **Track A2: Amharic** | | **Track A3: English** | |
|---|---|---|---|---|---|
| **Model** | **Score** | **Model** | **Score** | **Model** | **Score** |
| XLMR-large | 0.570 | XLMR-large | 0.878 | XLMR-large | 0.818 |
| mBERT | 0.566 | mBERT | 0.257 | mBERT | 0.798 |
| mE5-base | 0.559 | mE5-base | 0.828 | mE5-base | 0.805 |
| mE5-large | 0.523 | mE5-large | 0.889 | mE5-large | 0.824 |
| mDeBERTa-v3-base | 0.561 | mDeBERTa-v3-base | 0.859 | mDeBERTa-v3-base | 0.821 |
| mDeBERTa-v3-xnli | **0.664** | mDeBERTa-v3-xnli | 0.878 | mDeBERTa-v3-xnli | 0.823 |
| - | - | - | - | E5-v2-large | 0.828 |
| **Ensemble** | 0.659 | **Ensemble** | **0.891** | **Ensemble** | **0.840** |

| **Track A4: Hausa** | | **Track A5: Kinyarwanda** | | **Track A6: Marathi** | |
|---|---|---|---|---|---|
| **Model** | **Score** | **Model** | **Score** | **Model** | **Score** |
| XLMR-large | 0.785 | XLMR-large | 0.641 | XLMR-large | 0.858 |
| mBERT | 0.741 | mBERT | 0.651 | mBERT | 0.822 |
| mE5-base | 0.747 | mE5-base | 0.664 | mE5-base | 0.825 |
| mE5-large | 0.752 | mE5-large | 0.652 | mE5-large | 0.860 |
| mDeBERTa-v3-base | 0.718 | mDeBERTa-v3-base | 0.646 | mDeBERTa-v3-base | 0.829 |
| mDeBERTa-v3-xnli | 0.759 | mDeBERTa-v3-xnli | 0.662 | mDeBERTa-v3-xnli | 0.839 |
| **Ensemble** | **0.791** | **Ensemble** | **0.665** | **Ensemble** | **0.862** |

| **Track A7: Moroccan Arabic** | | **Track A8: Spanish** | | **Track A9: Telugu** | |
|---|---|---|---|---|---|
| **Model** | **Score** | **Model** | **Score** | **Model** | **Score** |
| XLMR-large | 0.833 | XLMR-large | 0.665 | XLMR-large | 0.803 |
| mBERT | 0.831 | mBERT | 0.673 | mBERT | 0.790 |
| mE5-base | 0.840 | mE5-base | 0.666 | mE5-base | 0.797 |
| mE5-large | 0.851 | mE5-large | 0.691 | mE5-large | 0.809 |
| mDeBERTa-v3-base | 0.816 | mDeBERTa-v3-base | **0.729** | mDeBERTa-v3-base | 0.805 |
| mDeBERTa-v3-xnli | 0.818 | mDeBERTa-v3-xnli | 0.701 | mDeBERTa-v3-xnli | 0.810 |
| **Ensemble** | **0.860** | **Ensemble** | 0.728 | **Ensemble** | **0.827** |

2024 Task 1, comparing them with the results of the two top-performing teams for each sub-track.

Table 1 showcases the performance of our ensemble model alongside that of the top two teams across nine tracks. Our system demonstrates competitive performance across four sub-tracks: Track A1 (Algerian Arabic), Track A2 (Amharic), Track A3 (English), and Track A7 (Moroccan Arabic). Additionally, we provide the results of both base models and ensemble systems on the development set. As indicated in Table 2, the ensemble gives better performance in most of the sub-tracks. Notably, we observe a decline in the performance of the ensemble on certain tracks (e.g., Track A1, Track A8) attributed to the presence of a base model that significantly outperforms the others and when this superior model is combined with the rest, it leads to a degradation in the overall performance of the ensemble that underscores the complexity of ensemble. In Track A2, the mBERT model was excluded from the ensemble due to its poor performance, the ensemble was thus formed using only the remaining models. Consequently, we opted for the ensemble model as the final submission system over the best model identified on the development set.

## 7 Conclusion

This paper introduces a straightforward yet effective ensemble architecture for Track A in the SemEval-2024 Task 1: Semantic Textual Relatedness. Our system leverages fine-tuning of pre-trained transformer-based language models as base regressors, coupled with a weighted voting technique to amalgamate predictions from diverse base models. Experimental results demonstrate its competitive performance across select languages in Track A without any additional resources. For future works, we propose enhancing our system by integrating African transformer-based models and exploring data augmentation techniques to improve the overall performance.

## Acknowledgements

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

## References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What makes sentences semantically related? a textual relatedness dataset and empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611.

Xu Gu, Xiaoliang Chen, Peng Lu, Xiang Lan, Xianyong Li, and Yajun Du. 2023. Simalstm-snp: novel semantic relatedness learning model preserving both siamese networks and membrane computing. *The Journal of Supercomputing*, pages 1–30.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

Mena Hany, Mostafa Mohamed Saeed, Rana Reda Waly, Abdelrahman Ezzeldin Nagib, and Wael H Gomaa. 2023. Enhancing textual relatedness assessment with combined transformers-embedding similarity techniques and machine learning regressors. In *Proceeding of IMSA*, pages 13–18. IEEE.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2022. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Muhammad Jawad Hussain, Heming Bai, Shahbaz Hassan Wasti, Guangjian Huang, and Yuncheng Jiang. 2023. Evaluating semantic similarity and relatedness between concepts by combining taxonomic and non-taxonomic semantic features of wordnet and wikipedia. *Information Sciences*, 625:673–699.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Saif M Mohammad and Graeme Hirst. 2012. Distributional measures of semantic distance: A survey. *arXiv preprint arXiv:1203.1858*.

Nedjma Ousidhoum, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, et al. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.