Reproducibility Appendix
Project Report for NLP Course, Winter 2023/4

S. Rećko
WUT
01151399@pw.edu.pl

M. Sperkowski
WUT
01151430@pw.edu.pl

P. Tomaszewski
WUT
01151442@pw.edu.pl

K. Ułasik
WUT
01151444@pw.edu.pl

supervisor: A. Wróblewska
WUT
anna.wroblewska1@pw.edu.pl

Reproducibility checklist

The description below applies to the 3 models used in a project: BERT, DistilBERT and Roberta.

- MODEL DESCRIPTION
  BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model developed by Google, designed to understand context in natural language processing tasks by pretraining on large text corpora. It utilizes a bidirectional approach, capturing context from both directions, which enhances its ability to comprehend nuances in language.
  DistilBERT is a distilled version of BERT, developed by Hugging Face, aimed at reducing computational resources while maintaining performance. It achieves this through knowledge distillation techniques, resulting in a more compact model suitable for deployment in resource-constrained environments.
  RoBERTa (Robustly optimized BERT approach) is another variant of BERT, introduced by Facebook AI, which employs enhanced pretraining techniques and larger training corpora to improve performance across various NLP tasks, including text understanding and generation.
- LINK TO CODE
  Source Code: `https://github.com/grant-TraDA/NLP-2023W/tree/main/4.%20E-commerce%20products/P1_Final/code`
  Requirements Instalation: `https://github.com/grant-TraDA/NLP-2023W/blob/main/4.%20E-commerce%20products/P1_Final/code/installation.ipynb`
- INFRASTRUCTURE
  Processor Intel Core i7-13700KF with clock speed of 3.4GHz and 16 threads. Computer has 32GB of RAM.
- RUNTIME PARAMETERS
  Approximately 2 hours per 3 epochs.
- PARAMETERS
  BERT (Base): Around 110 million parameters.
  DistilBERT: Approximately 66 million parameters (40% fewer parameters than the base BERT model).
  RoBERTa (Base): About 125 million parameters.
- VALIDATION PERFORMANCE
  Not applicable
- METRICS
  RankedSimilarity is our custom hierarchical similarity metric. It is a sum of Kendall Tau Distance (KDT) and Mean Square Error (MSE). KDT measures the dissimilarity between two rankings by counting the number of pairwise disagreements in ordering elements. MSE measures how much the vector of ordered cosine similarities between product pairs differs from the vector of values from 1 to 0 with equal differences between them.

$$RankedSimilarity(x) = KDT(x) \quad (1)$$
$$+MSE(x, [1, 0.66, 0.33, 0])$$

  where x is a vector of cosine similarity.

Multiple Experiments:

- NO TRAINING EVAL RUNS
  3 Epochs
- HYPER BOUND

Optimizer	Adam
Beta1 (Adam)	0.9
Beta2 (Adam)	0.999
Learning Rate	0.001
Batch Size	1
Epochs	3

Table 1: Hyperparameters used.

- HYPER BEST CONFIG
  Not applicable
- HYPER SEARCH
  Not applicable
- HYPER METHOD
  Not applicable
- EXPECTED PERF

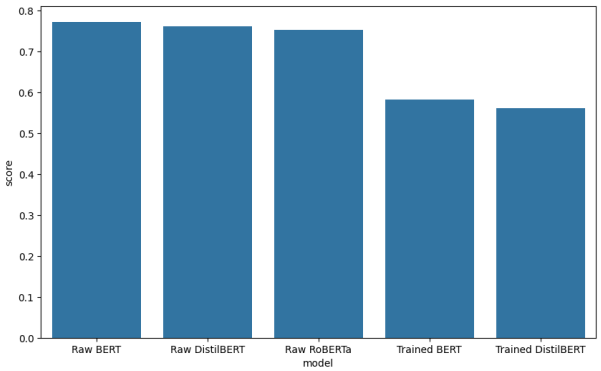

Figure 1: Evaluation metric score for pre-trained models before and after our training

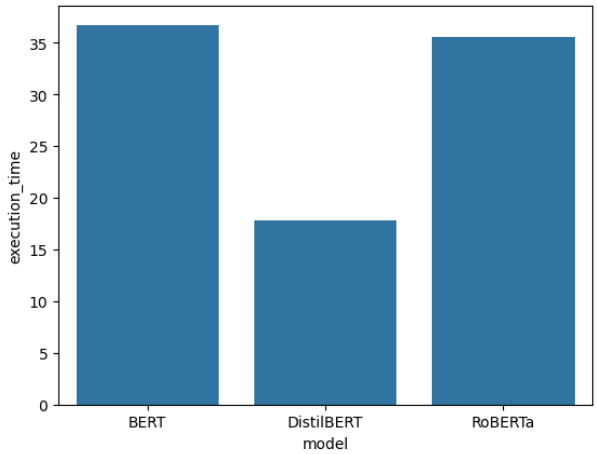

Figure 2: Inference time for chosen models.

Datasets – utilized in the experiments and/or the created ones:

- DATA STATS

Golden Standard		
Category	positive pairs	negative pairs
Computers	150	400
Cameras	150	400
Watches	150	400
Shoes	150	400

- DATA SPLIT
  80% of the data containing augumented text representations were used for train set. The rest was used as a test set.
- DATA PROCESSING
  Text Processing Steps:
  - Extraction of title and description
  - Creating augumented text represenation using LLM
  - Combining products representations into vector of five from the most similar to least similar
  - Creating embeddings from augumented text representations of products
- DATA DOWNLOAD `https://data.dws.informatik.uni-mannheim.de/largescaleproductcorpus/data/v2/goldstandards/all_gs.json.gz`
- NEW DATA DESCRIPTION
  Not applicable
- DATA LANGUAGES
  English