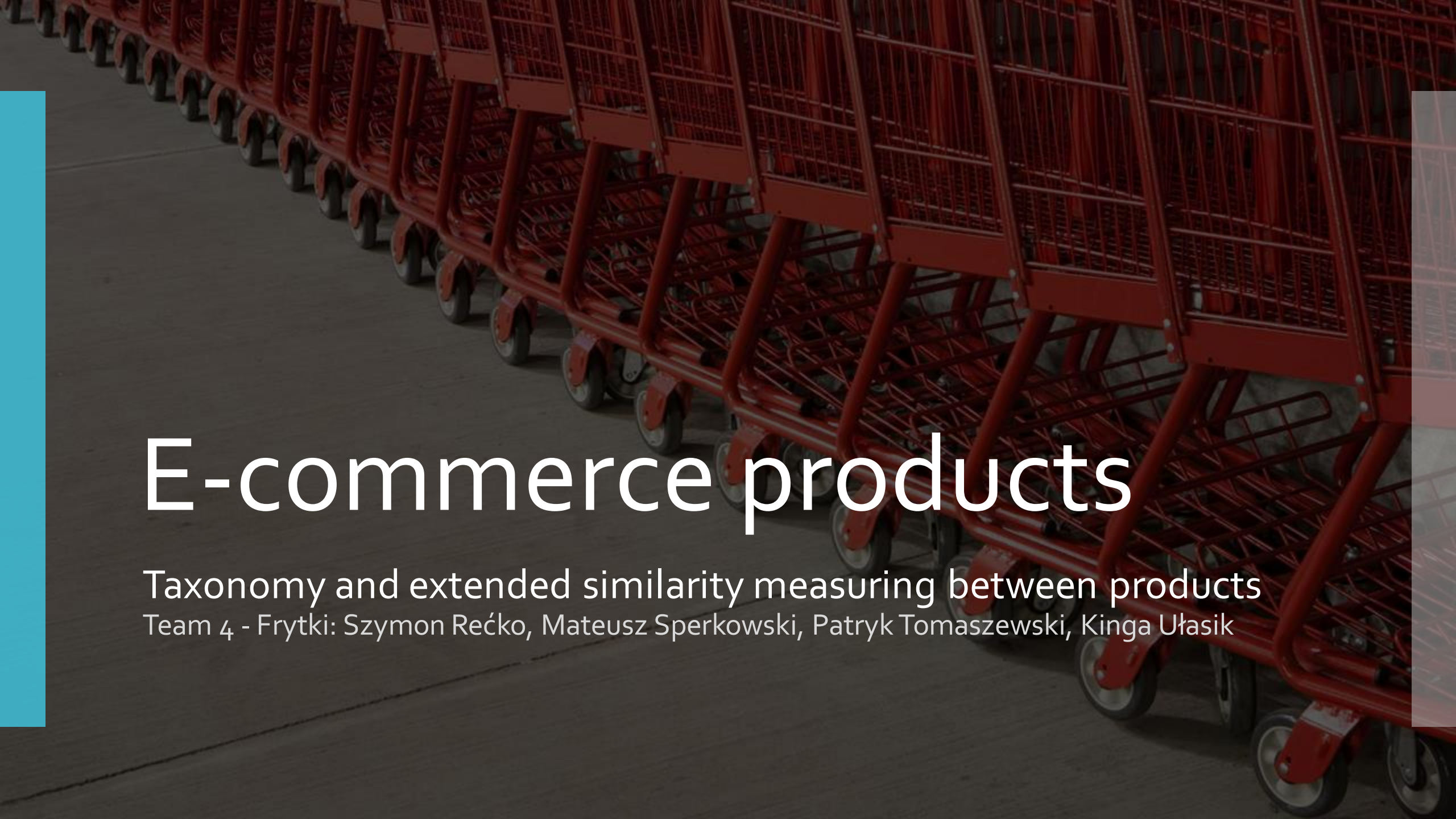


E-commerce products

Taxonomy and extended similarity measuring between products

Team 4 - Frytki: Szymon Rećko, Mateusz Sperkowski, Patryk Tomaszewski, Kinga Ułasik

Project Goals

- Extracting crucial information from descriptions and titles using LLM's
- Automatic methods for measuring similarity between products
- Incorporating taxonomy into the measurements

Dataset

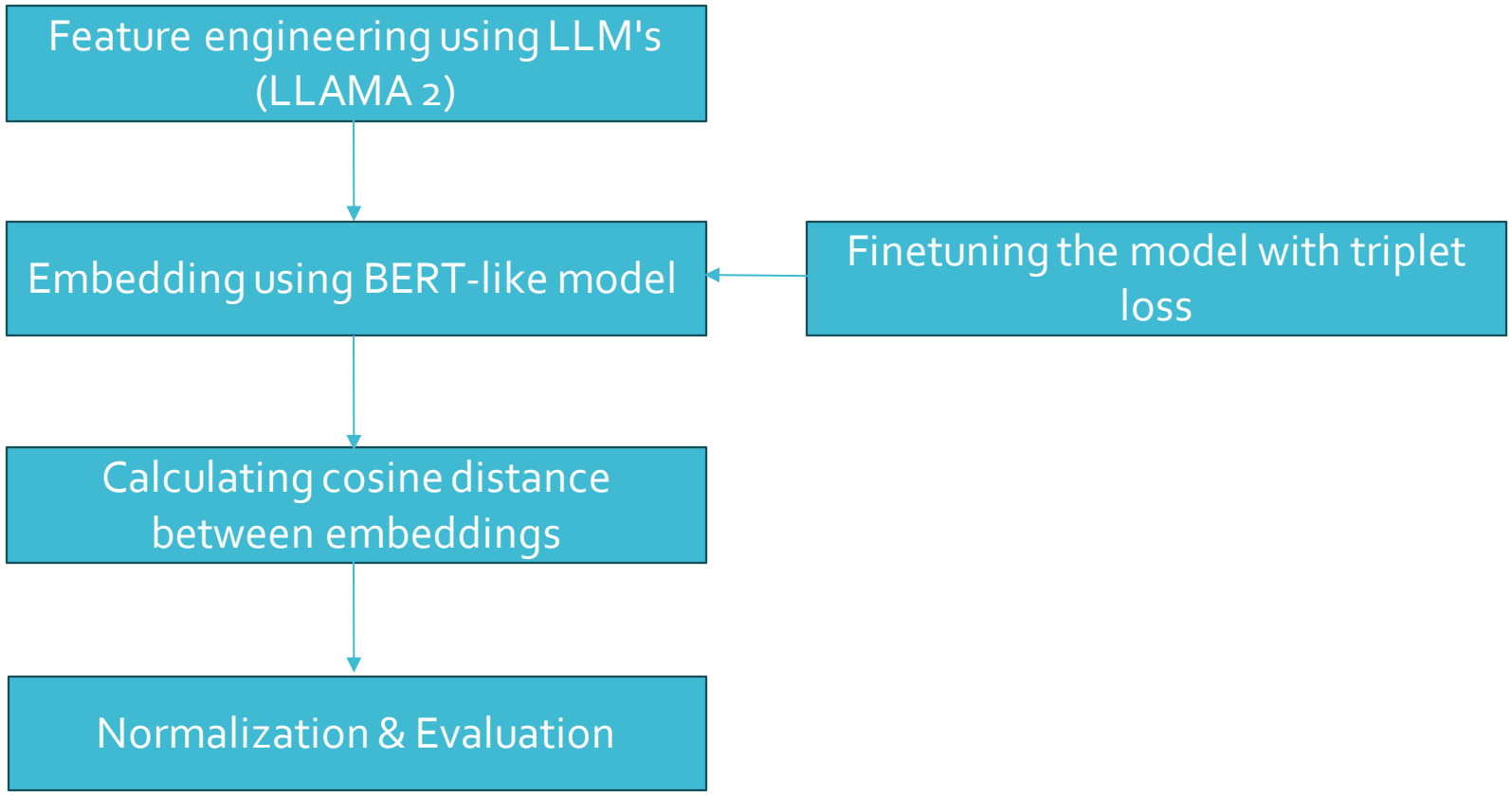
Web Data Commons - Training Dataset and Gold Standard for Large-Scale Product Matching"

- Full dataset was constructed from 16 million English-language offers sourced from a wide array of 79 thousand websites.
- Derived from the English product data corpus, set of 1,100 pairs of offers from each of the four product categories
- For each product included 2 matching pairs of offers and 5 or 6 non-matching pairs of offers

Category	# positive pairs	# negative pairs	% title	% description	% brand	% price	% specTableContent
Computers	300	800	100	82	42	11	22
Cameras	300	800	100	73	25	3	7
Watches	300	800	100	71	15	1	7
Shoes	300	800	100	70	8	1	2
All	**1200**	**3200**	**100**	**74**	**23**	**4**	**10**

Our contributions

Definition and implementation of pipeline for calculating similarity measure



Prompt engineering for extracting attributes from products

"Given a product title and description, generate a meaningful text representation that captures the essence of the product for effective similarity search. Consider relevant features, attributes, and contextual information to ensure the generated representation reflects the product's unique characteristics, allowing for accurate comparisons in a similarity search algorithm. Do not answer, just create a representation.

TEXT TO REPRESENT:

<product title>

<product description>"

Definition of a metric for determining quality of a multi-hierarchical similarity measure

Kendall Tau Distance

$$K_d(\tau_1, \tau_2) = |\{(i,j) : i < j, [\tau_1(i) < \tau_1(j) \wedge \tau_2(i) > \tau_2(j)] \vee [\tau_1(i) > \tau_1(j) \wedge \tau_2(i) < \tau_2(j)]\}|.$$

where $\tau_1(i)$ and $\tau_2(i)$ are the rankings of the element i in τ_1 and τ_2 respectively.

Custom-defined metric

$$METRIC(a, a', b, c, d) = \alpha KDT$$

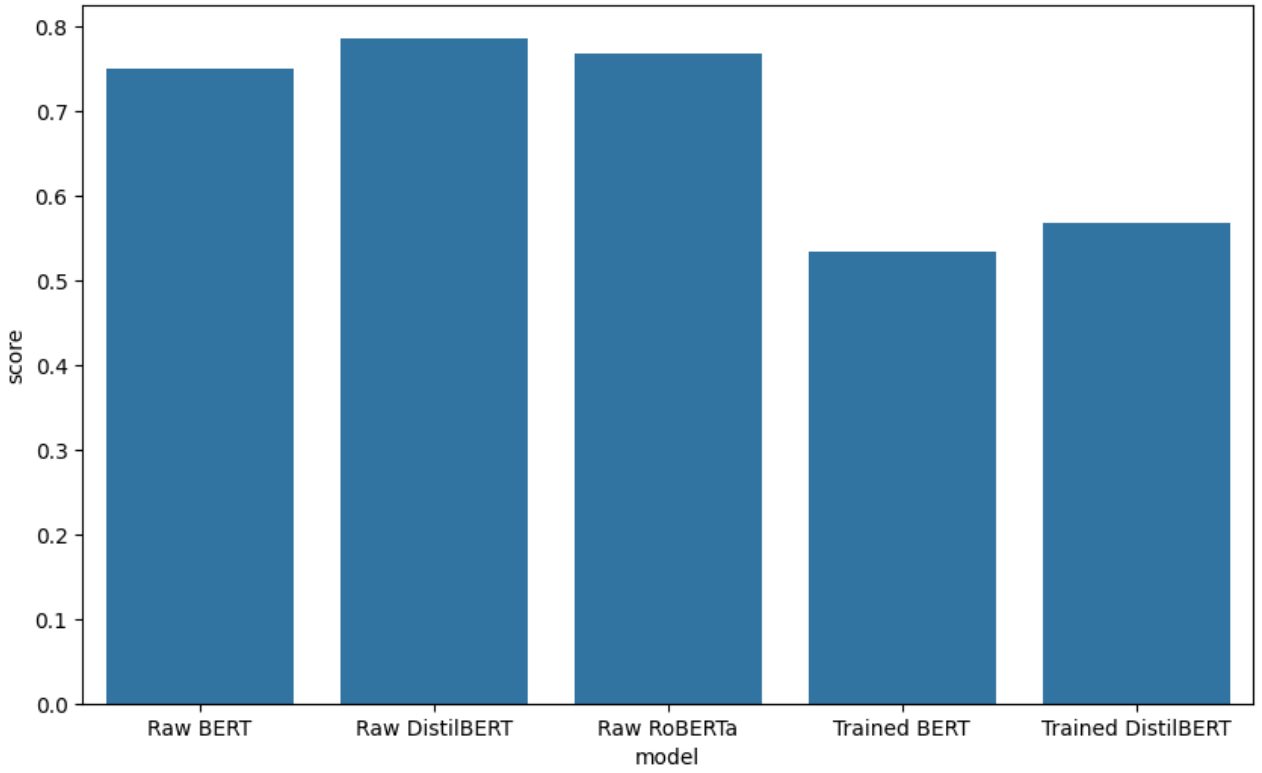
$$+MSE([\cos(a, a'), \cos(a, b), \cos(a, c), \cos(a, d)], [1, 0.66, 0.33, 0])$$

Ablation study of quality of simialrity measure for three different BERT-like models

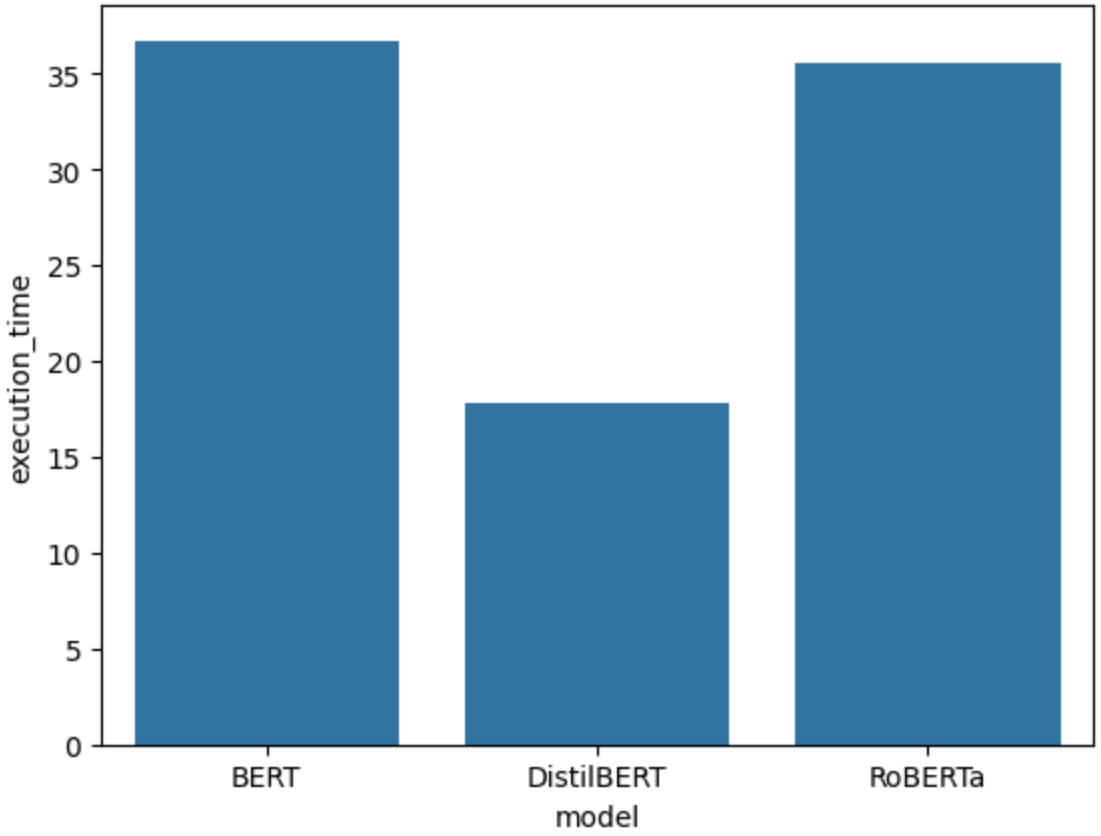
- Finetuning of a pretrained model
- Training on the golden standard with the triplet loss

$$\mathcal{L}(o, p^+, p^-) = max(0, m + d(\mathcal{E}_\theta(o), \mathcal{E}_\phi(p^+)) - d(\mathcal{E}_\theta(o), \mathcal{E}_\phi(p^-)))$$

- Evaluation done on pentalets with our custom metric



Ablation study of execution time for three different BERT-like models



Main takeaways

Main takeaways

- Gained experience in dealing with the practical challenges, such as computational constraints, time limitations, and the impact on project decisions.
- Deepened our knowledge on NLP models like BERT, DistilBERT and RoBERTa.
- Learned to recognize the importance of prompt engineering for guiding LLMs towards desired outputs.
- Learned that the good computation time not always comes with loss on accuracy.
- Observed that sometimes a task that is very intuitive for a human is a hard task for machines.
- Improved project managements skills and coding skills (clean code, importance of comments, reproducibility).

Thank you for attention

Feel free to ask questions