E-commerce products

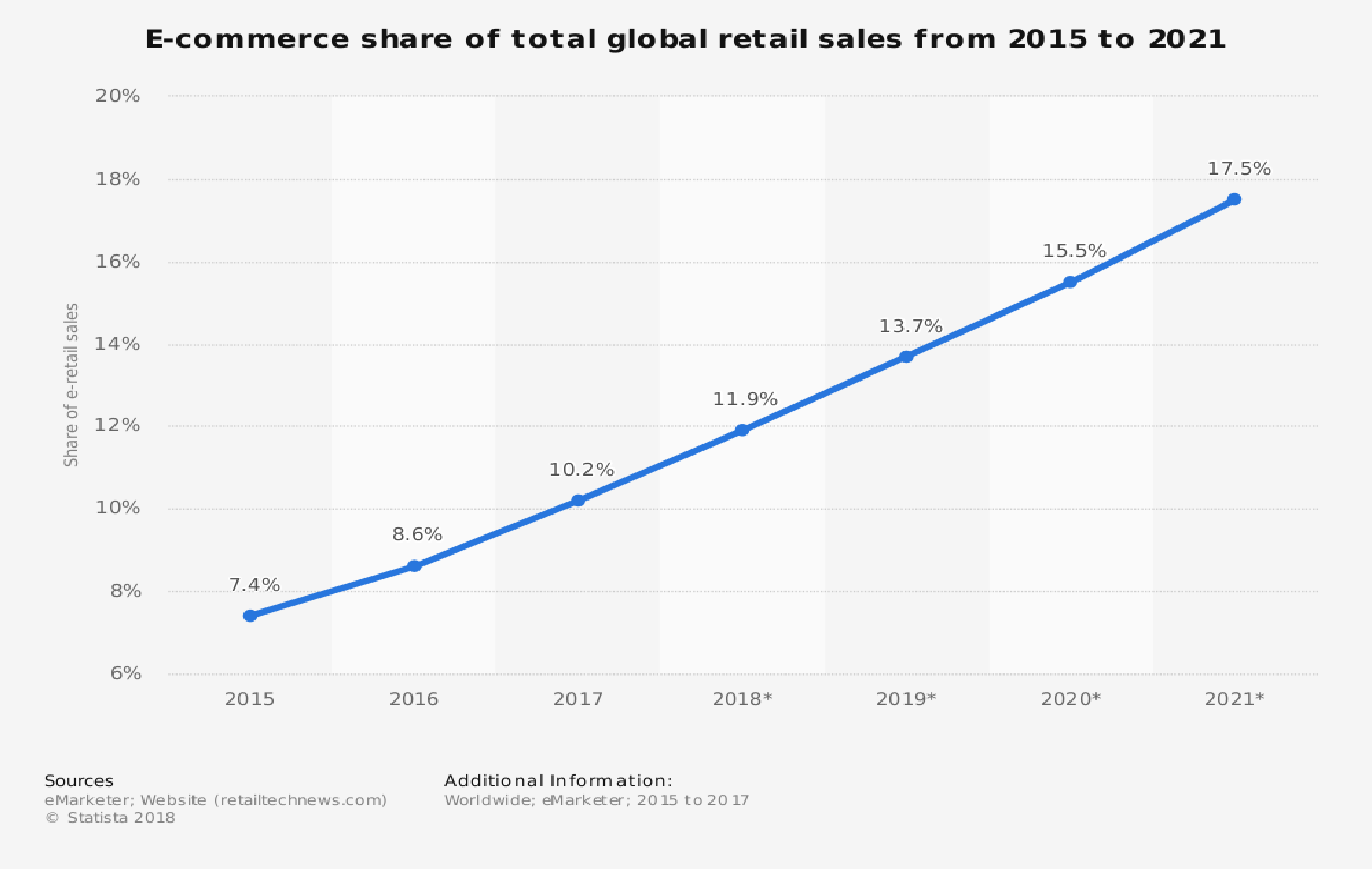
Taxonomy and extended similarity measuring between products

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

What is e-commerce?

- **E-commerce** is the activity of electronically buying or selling products on online services or over the Internet.
- Top 5 e-commerce companies in the world:
- The ones we might know and use:
  - Allegro (18th largest)
  - Etsy (20th largest)
  - Zalando (21st largest)

Rise of e-commerce



Recommendation systems

- "People don't know what they want until you show it to them" - Steve Jobs
- A fix for the lack of help by the shops staff or an annoying ad?
- Most commonly evaluates users shopping history and viewing behaviour to recommend things to buy.
- Cross-selling vs **upselling.**

Our Project

- Automatic methods for measuring similarity between products on multilevel dimensions
- Taxonomy
- Extracting crucial information from descriptions and titles

BERT-based similarity learning for product matching

Janusz Tracz[1], **Piotr Wójcik**[1], **Kalina Jasinska-Kobus**[1, 2],
Riccardo Belluzzo[1], **Robert Mroczkowski**[1], **Ireneusz Gawlik**[1, 3]
[1] ML Research at Allegro.pl
[2] Poznan University of Technology
[3] AGH University of Science and Technology
{janusz.tracz,piotr.wojcik,kalina.kobus,riccardo.belluzzo,
robert.mroczkowski,ireneusz.gawlik}@allegro.pl

BERT-based similarity learning for product matching

Generalized zero-shot multi-class classification

Categorizing instance into multiple classes, some of which have not been part of the model's training data.

Available data

- Title
- Description
- Category
- List of attributes (attribute name, value, and unit)

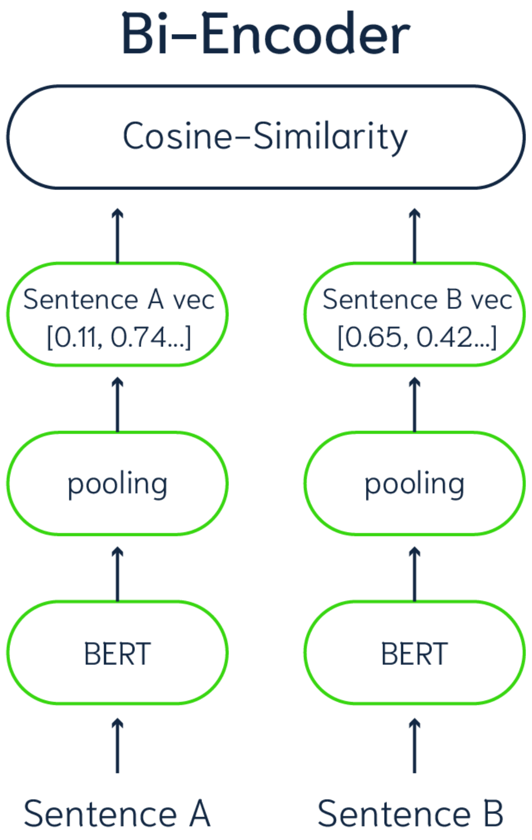
BERT-based similarity learning for product matching

Bi-Encoder architecture

- Products are represented as text
- A previously trained transformed is applied
- A distance between embeddings is calculated as similarity between instances

Used transformer: BERT

Used distance: cosine distance



BERT-based similarity learning for product matching

Textual representation

Each of the products was represented as a concatenation of:

- title
- attribute values
- attribute units

Description and attribute names were omitted as they deteriorated model performance.

Additionally, an assumption was made that only products from the same category were compared.

BERT-based similarity learning for product matching

Similarity learning with triplet loss objective

Training data consists of triples in the form of

$$(o, p^+, p^-)$$

with the elements being the anchor, a matching product, and a non-matching product.

Then, the transformer is adjusted to minimize the following loss function:

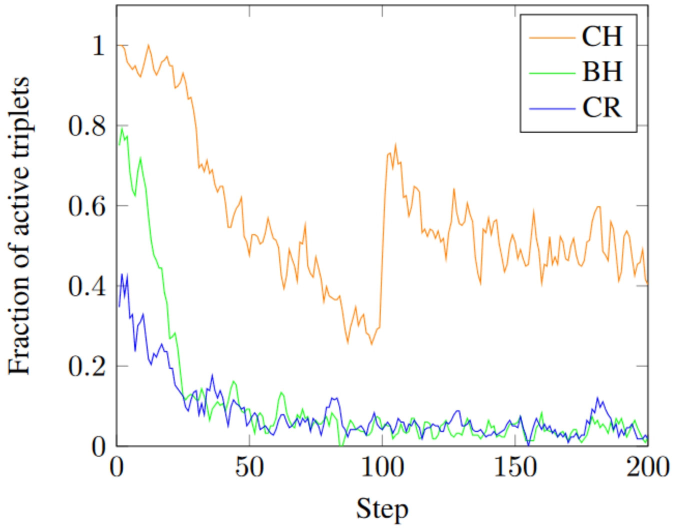
$$\mathcal{L}(o, p^+, p^-) = max(0, m + d(\mathcal{E}_\theta(o), \mathcal{E}_\phi(p^+)) - d(\mathcal{E}_\theta(o), \mathcal{E}_\phi(p^-)))$$

BERT-based similarity learning for product matching

Batch construction strategy

In the article, different strategies were considered when selecting the negative match to minimize inactive triplets:

- randomly from a category (CR)
- most similar product from the non-matching products in the sampled batch (BH)
- most similar product from the non-matching products in the entire category (CH)



(a) Active triplet fraction for HerBERT initialised model for different negative item selection strategies.

BERT-based similarity learning for product matching

Results

	Available matches	Products
CULTURE	300K	800K
ELECTRONICS	200K	400K
BEAUTY	300K	200K

Table 1: Datasets used for our experiments.

	CULTURE	ELECTRONICS	BEAUTY
BOW	0.8863	0.8032	0.7687
HerBERT-NFT	0.8206	0.6716	0.5542
HerBERT	0.9550	0.8580	0.9064
eComBERT-NFT	0.8208	0.6755	0.6127
eComBERT	**0.9777**	**0.8840**	**0.9219**

Table 2: Test accuracy per each dataset. NFT stands for *non-finetuned*.

Multilingual Transformers for Product Matching – Experiments and a New Benchmark in Polish

A Preprint

Michał Możdżonek[1], **Anna Wróblewska**[1], **Sergiy Tkachuk**[2], **Szymon Łukasik**[2,3]
[1]Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
Email: {michal.mozdzonek,anna.wroblewska1}@pw.edu.pl
[2]Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
Email: {stkachuk,slukasik}@ibspan.waw.pl
[3]Faculty of Physics and Applied Computer Science, AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków, Poland
Email: slukasik@agh.edu.pl

mBERT and XML-RoBERTa transfer learning

- Product matching problem
- Transfer learning for data in different languages
- Web Data Commons dataset (4 categories, sizes: small, medium, large)
- Own Polish product matching dataset
- Running pre-trained models and perfomance comparison

mBERT and XML-RoBERTa transfer learning

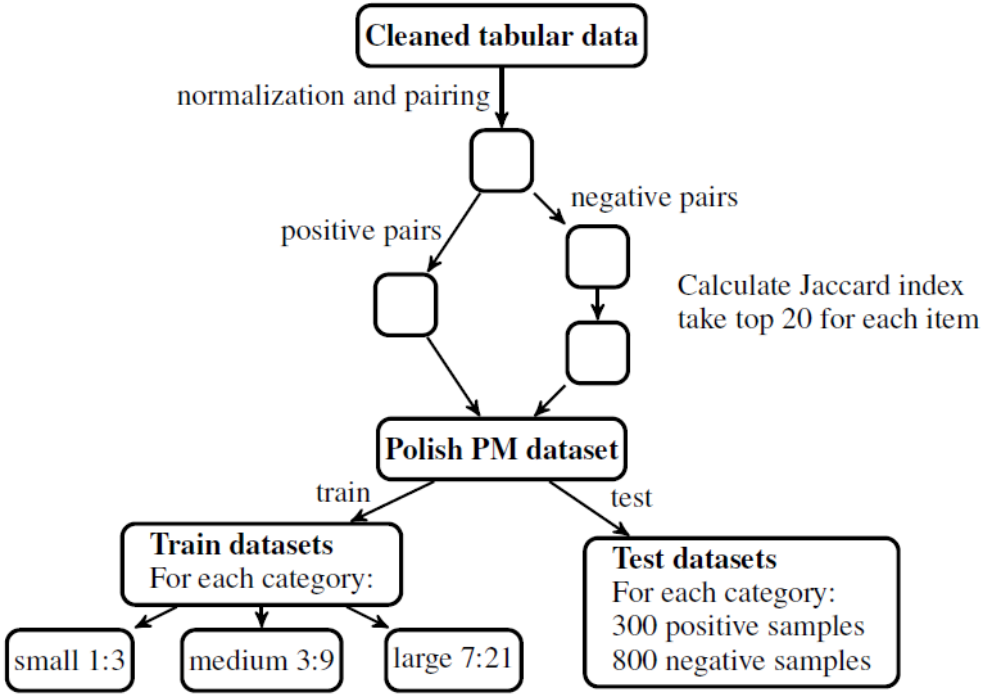


Figure 2: The process of creating the Polish PM datasets. In each training set, the ratio of positive to negative samples is 1:3.

mBERT and XML-RoBERTa transfer learning

- Selecting the title column only and concatenating it with token markers
- HuggingFace Transformers library
- Two types of models: : mBERT and XLM-RoBERTa
- Pre-trained models on Wikipedia articles in about 100 languages
- Models run on both WDC and Polish datasets

mBERT and XML-RoBERTa transfer learning

Table 7: F1 scores for models trained on English WDC datasets. Mean value and standardized error (confidence level 95%) for each dataset were calculated from **4** samples. For further information on how standardized error was calculated see Section 5.

dataset type	dataset size	mBERT	XLM-RoBERTa	Ditto (reported in Li et al. [2020])	WDC-Deepmatcher (reported in Peeters et al.)
Cameras	small	**82.13**(±**4.70**)	81.96(±7.75)	80.89	68.59
	medium	87.86(±2.04)	**88.11**(±**4.22**)	88.09	76.53
	large	90.88(±2.28)	**92.36**(±**0.76**)	91.23	87.19
	xlarge	-	-	**93.78**	89.21
Computers	small	**86.43**(±**3.69**)	81.10(±13.40)	80.76	70.55
	medium	**90.13**(±**1.89**)	88.69(±2.19)	88.62	77.82
	large	92.48(±2.33)	**93.71**(±**0.77**)	91.70	89.55
	xlarge	-	-	**95.45**	90.80
Shoes	small	**79.20**(±**7.89**)	74.98(±13.36)	75.89	73.86
	medium	**84.11**(±**3.40**)	81.30(±8.21)	82.66	79.48
	large	90.28(±2.36)	**91.26**(±**2.09**)	88.07	90.39
	xlarge	-	-	90.10	**92.61**
Watches	small	**87.31**(±**1.64**)	83.78(±4.38)	85.12	66.32
	medium	**91.17**(±**4.21**)	89.50(±3.69)	91.12	79.31
	large	93.52(±2.63)	93.62(±0.67)	**95.69**	91.28
	xlarge	-	-	**96.53**	93.45

mBERT and XML-RoBERTa transfer learning

Table 6: F1 scores for mBERT and XLM-RoBERTa trained on Polish datasets. Mean value and standardized error (confidence level 95%) for each dataset were calculated from **4** samples. For further information on how standardized error was calculated see Section 5.

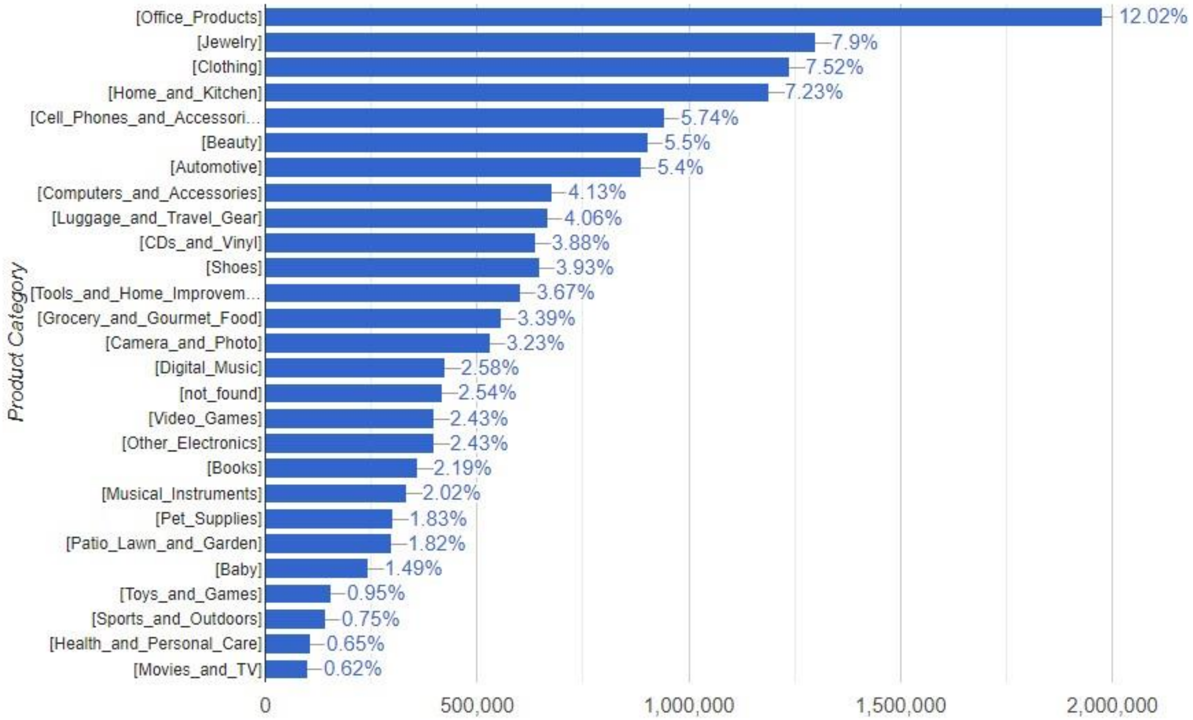
dataset type	dataset size	mBERT	XLM-RoBERTa
Household chemistry (pl. chemia)	small	**85.73**(±**1.89**)	83.15(±4.15)
	medium	**90.78**(±**3.03**)	89.03(±5.96)
	large	**93.25**(±**1.77**)	92.52(±1.77)
Drinks (pl. napoje)	small	**85.17**(±**1.61**)	84.43(±7.16)
	medium	**88.98**(±**2.63**)	88.44(±2.88)
	large	89.39(±2.12)	**89.93**(±**3.99**)
All	small	**85.73**(±**1.96**)	84.67(±9.03)
	medium	**90.78**(±**1.13**)	88.63(±2.79)
	large	91.41(±3.17)	**91.61**(±**1.39**)

Solution Concept

Dataset

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

- 16 million English-language offers sourced from a wide array of 79 thousand websites.
- Includes product categorization based on Amazon product data and TF-IDF scores for 26 product categories.
- Each offer was assigned to one of 26 categories



Dataset

GOLD STANDARD

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**

Proposed approach

Products data processing

- Exploratory Data Analysis
- Select and/or augument product attributes that will be concatenated to create input for model.

Implementation of modified product similarity pipeline

- Fine-tune BERT-based model (RoBERTa, DistilBERT) or different embedding model, like BGE (#1 on HF MTEB) using bi-encoder framework.
- Experiment with loss/distance functions.
- Performance tests and evaluation

THANK YOU FOR ATTENTION