E-commerce products
Project Proposal for NLP Course, Winter 2022

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

Abstract

This project's goal is to present an approach leveraging machine learning, especially Natural Language Processing (NLP), techniques to establish a robust similarity measure between products in e-commerce. The research aims to differentiate between identical, slightly different, and distinct products through taxonomy enrichment, information extraction from descriptions, and potential attribute generation from item descriptions. Hypotheses suggest that these methods will significantly enhance search accuracy and introduce new products effectively, promising a transformative impact on e-commerce user experiences by enhancing recommendations.

1 Introduction

The overarching goal of this project is to develop an innovative and robust framework leveraging machine learning and NLP techniques to create an automated and comprehensive system capable of measuring the similarity between products across multiple dimensions within e-commerce platforms. This research aims to achieve an advanced level of accuracy in defining and establishing similarity metrics between products based on various attributes, including taxonomy (categories), textual descriptions and titles. To be practical for the end users, calculations must be performed quickly. Therefore, the calculations need to work in real-time, so we set a limit of at least 0.01 seconds (10ms) per pair of products. We consider this a crucial requirement for the project.

The primary scientific objective is to construct a sophisticated similarity measurement that accurately captures the semantics and nuances of product relationships, distinguishing between identical, closely related (such as variations within the same product line differing in specifications like memory sizes), and entirely distinct products. This entails the exploration and development of algorithms that can comprehend and differentiate the diverse relationships among products, accommodating subtle variations in features while still recognizing commonalities.

Furthermore, the project will focus on innovating techniques for automatically extracting key information from item descriptions and titles using deep learning algorithms. These generated attributes will supplement existing data, enriching the product information available on e-commerce platforms.

Ultimately, the scientific aim is to significantly enhance e-commerce search functionalities, facilitate the simple yet efficient introduction of new products into online marketplaces, and provide users with access to a comprehensive range of available price ranges for similar products. This research endeavour will contribute to advancing the capabilities of e-commerce platforms by refining the accuracy and depth of product recommendations, thereby improving the overall user experience in online product search and comparison.

1.1 Research questions

In e-commerce, the sheer volume and diversity of products pose a significant challenge in accurately measuring their similarity across multiple dimensions. The current methodologies for comparing and categorizing products often fall short in capturing the nuanced differences and similarities between items. This leads to sub-optimal search experiences for users and hampers the efficiency of introducing new products into e-commerce platforms. The need for an automated and precise system to measure product similarity, considering varying attributes and features, is crucial for enhancing product search accuracy and user satisfaction.

Research Questions:

- What machine learning approaches can extract meaningful attributes from product descriptions, and how can these attributes be integrated into the similarity measurement process?
- What algorithms or models can best capture the semantics and nuanced relationships between products to differentiate between identical, slightly different, and entirely distinct items?
- How can the generated product attributes be integrated into existing e-commerce platforms to optimize search functionalities and introduce new products efficiently?
- How can we reach minimal pair comparison time while still using SOTA models with complex architectures?

Machine learning approaches for extracting meaningful attributes from product descriptions include NLP techniques, word embeddings, and Named Entity Recognition (NER). These attributes can be integrated into the similarity measurement process by converting them into feature vectors for products. Utilizing metrics such as cosine similarity or Euclidean distance on these vectors allows for an effective quantification of product similarity, enhancing the overall recommendation system.

Siamese Neural Networks, Graph Neural Networks (GNN), and Transformer models like BERT excel in capturing semantics and nuanced relationships between products. Siamese networks are adept at understanding subtle differences, GNNs model complex dependencies, and Transformers provide contextualized embeddings, collectively offering a robust framework for differentiating between identical, slightly different, and entirely distinct items.

The integration of generated product attributes into existing e-commerce platforms can be achieved by developing an attribute-based search functionality, enhancing recommendation systems, and optimizing the introduction of new products. By leveraging these attributes, platforms can offer more personalized search options, improve recommendation accuracy, and streamline the process of introducing new products efficiently into the market, ultimately enhancing the overall user experience.

To reach minimal pair comparison time while utilizing State-of-the-Art (SOTA) models with complex architectures, various strategies can be employed. Techniques like model quantization and pruning help reduce the computational load, while hardware acceleration using specialized processors like GPUs or TPUs speeds up inference. Additionally, caching and batch processing can be implemented to precompute certain calculations and perform parallelized comparisons, ensuring efficient and real-time processing without compromising the sophistication of the underlying models. Distilling the knowledge of a bigger model to a smaller one, by enforcing similar outputs on the training data, is a another method commonly used for creating smaller networks.

2 Significance of the project

The rise of e-commerce has been an unprecedented change in the sales market. Over the past years, e-commerce has witnessed exponential growth, accelerated further by global shifts in consumer behavior, especially the increased adoption of online shopping. The convenience, accessibility, and wide array of products available online have transformed how people shop. The COVID-19 pandemic further expedited this shift, with many consumers transitioning to online platforms for everyday needs. As e-commerce continues to expand, the need for efficient recommendation systems becomes even more critical. With an ever-growing number of products available, these systems play a pivotal role in helping consumers navigate through the vast array of choices, making their shopping experiences more personalized, streamlined, and enjoyable.

This system harnesses the power of data, algorithms, and user behavior to offer personalized product recommendations, thereby creating a more engaging and tailored shopping experience.

By analyzing user preferences, purchase history, and behavior, an e-commerce recommendation system can offer personalized suggestions. This level of customization enhances user experience, making shopping more efficient and enjoyable.

Tailored recommendations lead to increased user engagement and extended time spent on the platform, potentially resulting in higher conversion rates and sales.

When users are presented with items that align with their interests, the likelihood of making a purchase increases. This directly impacts the conversion rates and revenue of the e-commerce platform.

Recommendation systems can suggest related or complementary products, effectively enabling upselling and cross-selling, which contribute to increased average order value.

Utilizing machine learning and algorithms, recommendation systems can predict future trends and customer preferences, assisting in inventory management and product development.

When customers find what they are looking for effortlessly, they tend to have a more satisfying shopping experience. This satisfaction leads to customer loyalty and retention.

Encouraging return visits and repeat purchases is crucial for the sustainability of any e-commerce platform. Tailored recommendations play a significant role in achieving this goal.

Building an e-commerce recommendation system involves cutting-edge technologies, fostering advancements in AI and machine learning applications. These advancements have broader implications for various industries beyond e-commerce.

3 Concept and work plan

In this section, we will first provide an overview of the dataset we will be using, followed by a detailed description of the proposed solution, offering insights into how we address the research problem.

3.1 Dataset

To create our solution, we have selected the Web Data Commons - Training Dataset and Gold Standard for Large-Scale Product Matching as the primary dataset for our project. Several key motivations drive the dataset selection. In recent years, entity resolution has shifted towards deep learning-based matching methods, necessitating large training data. Traditional benchmark datasets often prove inadequate for evaluating these methods due to their limited size and source diversity. The ”Web Data Commons” dataset addresses these challenges by offering a substantial volume of data, including 16 million English-language offers, sourced from 79 thousand websites. This diversity and scale make it an ideal choice for assessing deep learning-based matches and improving their evaluation and comparison. The dataset includes categorization using distant supervision from Amazon product data. Lexica containing terms and their TF-IDF scores for 26 product categories (look Figure 1) were created using publicly available Amazon product reviews and metadata. Each offer in the dataset is assigned the product category whose terms maximize the sum of overlapping TF-IDF scores. In cases with minimal overlap, the offer is categorized as ”not found”. We will exclusively utilize the ”Gold Standard” for the training of our product matching method. The gold standard, derived from the English product data corpus, comprises a set of 1,100 pairs of offers from each of the four product categories: Computers Accessories, Camera & Photo, Watches, and Shoes. For each product, the gold standard includes two matching pairs of offers (positives) and five or six non-matching pairs of offers (negatives).

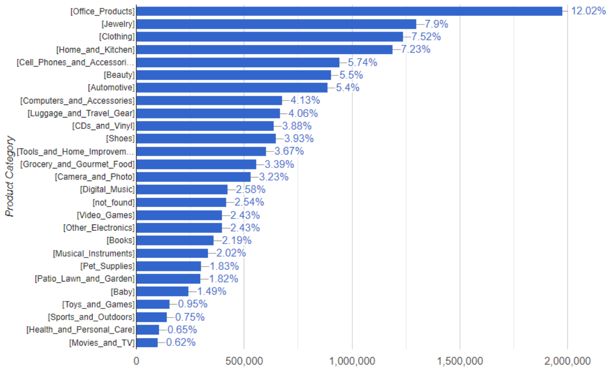

Figure 1: Distribution of offer entities per category in the English Training Set.

Additionaly we want to test or solution on Polish dataset for product matching created by the authors of (Michał Mozdzonek, 2022). The dataset was derived from popular Polish stores, involving data collection, subsequent cleaning, and transformation into a tabular format compatible with the WDC dataset. It consist of three training set sizes denoted as small (1), medium (3), and large (7). Within each set, the positive to negative sample ratio is maintained at 1:3. While the dataset size and distribution align closely with WDC datasets, a departure exists in large datasets, exhibiting a size ratio of 7:1 compared to the 15:1 ratio in the WDC project. This adjustment was prompted by a scarcity of positive samples in the Polish datasets. Notably, unlike the WDC project, the test dataset pairs were not manually validated but were generated following the aforementioned process.

3.2 Proposed Solution

Our main aim would be to test a method for product similarity matching using reliable language models. An important part of the project is the limitation of the execution time set to 10 ms. Because of that, we decided to test out BERT (Bidirectional Encoder Representations from Transformers) as bi-encoder and its smaller, faster, cheaper and lighter version - DistilBERT. Despite of having 40% less parameters that original model, DistilBERT achieves competative results performance with the benefit of running 60% faster. Additionaly, as it is not stated otherwise in the project description, we assumed that the imposed time limitation doesn't include feature extraction, we will test LLMs (Large Language Models) as they have demonstrated remarkable capabilities in understanding context, capturing complex patterns, and nuances in language.

The primary goal is to extract informative embeddings from product attributes. In related literature for example, in (Michał Mozdzonek, 2022) the authors achieved the best results when using only product titles. Still, we also want to test various prompts for feature extraction from other product attributes, focusing mainly on description. Finally, we want to calculate the cosine similarity of these embeddings to determine product similarity.

We will employ BERT-based model or embedding model, like BGE to serve as our encoders. Transformers has proven highly effective in capturing semantic and contextual information in text data. Our bi-encoder will consist of one main model for encoding both product attributes. The encoder will convert the product attribute pairs into dense, fixed-length vectors (embeddings) that represent the semantic information of the products.

After obtaining embeddings for both products in a pair, we will calculate the cosine similarity between these embeddings. Cosine similarity is a widely used metric for measuring the similarity between vectors, providing a score that quantifies the degree of similarity between the products. Higher cosine similarity scores indicate greater similarity between the products, while lower scores indicate dissimilarity.

Our approach has two major advantages:

- Contextualized Representations: Transformers contextualized word representations enables the model to capture the nuanced and contextual meaning of product attributes, leading to more accurate similarity measurements.

- Scalability: The use of smaller transformer model enables us to handle a wide range of product categories and attributes, making the method highly scalable and adaptable to various e-commerce domains.

4 Methodology and literature review

Our proposed solution is based on the pipeline presented in (Tracz et al., 2020). In this article, the authors propose usage of a transformer architecture, specifically a fine-tuned version of BERT, to embed and then compare a pair of products. As the pre-embedding product representation, a concatenation of the title, attributes values, and units was used. For the fine-tuning process they used a triplet loss objective with the cosine distance. Additionally, various batch construction strategy for selecting the triplets used in training were analysed.

Similar methodology has been described in (Peeters et al., 2020), where a cross-encoder structure was used instead of a bi-encoder. Additionally, textual representation has been replaced by a concatenation of product brand, title, truncated description and truncated specification table.

Alternative approach was presented in (Peeters and Bizer, 2023), where the similarity score was generated with a LLM under proper prompt construction, however due to performance limitations present in our problem this solution was disregarded.

Authors of the (Michał Mozdzonek, 2022) article focused not only on the Product matching problem, but also on the idea of using transfer learning for data in different languages.

As English is one of the most popular languages in the world, the Web Data Commons dataset focused on four categories: computers, cameras,

Week	Work Plan
15-22.11.2023	Corrections to the project proposal. EDA of the datasets. Finding all possible options to be used/tested during the development. Proof of Concept for the project, initial simplest example of the architecture.
22-29.11.2023	Attempts at fine-tuning the models using the two datasets. Ablation study - coming up with better prompts to extract most important information from descriptions. Trying to optimize the calculations.
29.11-6.12.2023	Finishing the development. Testing different thresholds for the similarity. Ablation study - comparing performance with the added tags and without them. Ablation study - comparing different BERT models.
6-13.12.2023	Finishing the project, mostly clean up of code and last tests. Creating the presentation. Coming up with the future works. Plan for project 2.

Table 1: Work plan for the first project.

watches, and shoes. Putting the WDC data as an example, the authors managed to gather data from popular stores and created own Polish product-matching dataset. The cleaning procedure consisted of dealing with missing data, extracting the main category, name unification, concatenating proper columns, removing the duplicates, final filtering, and renaming. There were two categories distinguished from the data: drinks (pl. ”napoje”) and household chemistry (pl. ”chemia”).

To create training and testing sets for the PM problem, offers were paired within each category. Pairs with identical EAN for the first and second offers were labeled as positive pairs, and this information was added as a target column. Finally, three training set sizes were created (small, medium, and large) with a 1:3 ratio of positive to negative samples.

Data preparation for the model involved selecting the title column and concatenating it with token markers. This resulted in a single input string for the model, which was then tokenized using a pre-trained model-specific tokenizer.

For the experimental part of the work, by using HuggingFace Transformers library, two types of pre-trained models were used: mBERT and XLM-RoBERTa. The models were pre-trained on Wikipedia articles in about 100 languages. The autors run the models on both WDC and Polish datasets. The F1 score was used as a metric to compare the models.

The mBERT and XLM-RoBERT models consistently outperformed other models. Notably, mBERT excelled, particularly in smaller-sized datasets (small, medium), a trend also observed in the Polish dataset. In the ”Shoes” category, results were slightly lower, with the mBERT model. However. XLM-RoBERT performed exceptionally well in ”large” datasets.

These results demonstrate that multilingual models effectively address the product matching problem, often yielding comparable or superior results to prior studies.

References

[Michał Mozdzonek2022] Sergiy Tkachuk Szymon Łukasik Michał Mozdzonek, Anna Wróblewska. 2022. Multilangual transformers for product matching – experiments and a new benchmark in polish.

[Peeters and Bizer2023] Ralph Peeters and Christian Bizer. 2023. Using chatgpt for entity matching. *arXiv preprint arXiv:2305.03423*.

[Peeters et al.2020] Ralph Peeters, Christian Bizer, and Goran Glavaš. 2020. Intermediate training of bert for product matching. *small*, 745(722):2–112.

[Tracz et al.2020] Janusz Tracz, Piotr Iwo Wójcik, Kalina Jasinska-Kobus, Riccardo Belluzzo, Robert Mroczkowski, and Ireneusz Gawlik. 2020. Bert-based similarity learning for product matching. In

Proceedings of Workshop on Natural Language Processing in E-Commerce, pages 66–75.