DOI: 10.17323/2587-814X.2025.2.7.24

Product matching in digital marketplaces: Multimodal model based on the transformer architecture

Artem Yu. Varnukhov
E-mail: varnuhov_ayu@usue.ru

Dmitry M. Nazarov
E-mail: slup2005@mail.ru

Ural State University of Economics, Yekaterinburg, Russia

Abstract

In this paper we analyze the problem of intelligent product matching in digital marketplaces for which one requires evaluation of similarity of various records that describe products but may differ in format, content or volume of multimodal data. The subject area of this scientific research represents an intersection of entity resolution (ER) problem solving methods: record matching and multimodal data analysis. It is of extreme relevance in a fast-growing platform economy with the e-commerce market expanding exponentially. The main purpose of this research is to develop and test an intelligent multimodal model based on transformer architecture to improve the accuracy and robustness of product matching in digital marketplaces. The authors developed a model integrating textual, visual and tabular attributes which enables us to identify similar products, find competitive offers, detect duplicates and perform product clustering and segmentation in a more effective manner. The proposed approach is based on the self-

* The article is published with the support of the HSE University Partnership Programme

attention mechanism which enables contextual-semantic relations modeling of various-nature data. In order to extract the vector representation of text descriptions, language models are applied, in particular the Sentence-BERT architecture; for the graphical component Vision Transformer is used; and tabular data are processed using specialized learning mechanisms based on TabTransformer structured data. The experiment we carried out demonstrated that the developed multimodal model efficiently solves the task of product matching in digital marketplaces in an environment of significant variability of product items and data heterogeneity. Additionally, the results suggest that the model can be adapted successfully for application in other product categories. The results obtained confirm the efficiency and expediency to apply the multimodal approach for digital marketplace product matching implementation. This allows the e-commerce market participants to significantly improve the quality of inventory management, increase pricing efficiency and strengthen their competitive advantages.

Keywords: digital marketplace, contextual-semantic identification, competitive offers search, product matching, machine learning, deep learning, transformer architecture, data mining

Citation: Varnukhov A.Yu., Nazarov D.M. (2025) Product matching in digital marketplaces: Multimodal model based on the transformer architecture. *Business Informatics*, vol. 19, no. 2, pp. 7–24.
DOI: 10.17323/2587-814X.2025.2.7.24

Introduction

The rapid development of the platform economy in recent decades has been driven by the emergence and widespread adoption of digital platforms serving as intermediaries between buyers and sellers. Marketplaces as digital platforms enable sellers to access a broad audience, while providing buyers with convenient services for searching and comparing product offers from various sellers. At the same time, managing product data and matching have become a complex scientific challenge the resolution of which inextricably affects both the direct and indirect economic performance indicators of all participants in e-commerce platforms. Direct indicators include factors influencing the choice of pricing model, while indirect indicators involve such aspects as the effectiveness of product promotion, accuracy of recommendations and consumer satisfaction. It is also evident that these direct and indirect indicators are causally linked. The relevance of this research is determined by the growing volume of e-commerce markets and expanding assortment of goods on digital platforms, which create a need for more advanced mathematical and instrumental methods in economics capable of automated and even intelligent and reliable product matching within a large flow of heterogeneous information in digital marketplaces.

From a scientific point of view, the issue of product matching (identification of product records) is closely related to the task of entity resolution (hereinafter referred to as 'ER'), which is a process of aligning entities while accounting for potential duplicates. However, in the context of marketplaces, this problem is extended by the use of multimodal data in the product description. Indeed, product descriptions in marketplaces generally include not only textual information but also images and characteristics presented in a tabular form.

Below is a brief review of reference literature divided into two logical sections: personalization with regard to the specifics of e-commerce, and ER in the context of algorithmic approaches to matching.

A number of studies representing the scientific foundation of this research focus on analyzing user behavior, developing recommendation systems and evaluating the features of product categories in e-commerce (H. Angermann, M. Mao, F.T. Abdul Hussien, A. Fletcher, P. Ristoski, M. Cheung, et al.). Most researchers emphasize the importance of an integrated approach application to process the textual descriptions, images and metadata in order to improve recommendation accuracy and detect hidden relationships within large datasets (big data). Analysis of these studies in the context of matching highlights the need to apply multimodal data processing technologies to accurately match not only product titles and characteristics but also their visual attributes.

Another line of research within the subject area of this study involves classical tasks such as duplicate record detection, entity matching and the development of algorithms for accurate data integration (W.W. Cohen, S.S. Aanen, J. Devlin, A.K. Elmagarmid, N. Reimers, J. Wang, H. Köpcke, et al.). The works of these authors examine various similarity metrics, including Levenshtein distance and its analogs, language models and case-based learning methods enabling the construction of patterns for identification of potential matches. Additionally, the issue of scalability of solutions is often raised, from the development of batch-processing software to high-load systems processing dynamic data streams in real time. The rapid advancement of deep learning architectures, particularly those based on transformers caused a shift from automated feature search to intelligent methods detecting complex patterns. These approaches are especially relevant for digital marketplaces, where the promptness and reliability of processing large-scale data with various economic properties, including price dynamics represent crucial factors.

Thus, the existing research confirms the importance of applying advanced economic and mathematical methods in the development of product matching systems in marketplaces based on multimodal data processing. However, several questions remain open in the reviewed literature: how to optimally combine features, which approaches are most effective for analyzing multimodal data, and how to adapt the developed models to the new products and data sources. This study proposes a multimodal model based on a transformer architecture that can process data from various sources and enables the sequential and efficient integration of different modalities through an attention mechanism and multi-level representations for more accurate product matching. The approach proposed contributes to the mathematical and instrumental methods in economics aimed at building robust systems for big data analysis, while also supporting scalability and accommodating the high dynamism of e-commerce platforms. The scientific contribution of this paper includes the development and validation of an approach that allows for the simultaneous integration of textual, visual and tabular data within a unified modular system based on a transformer architecture.

The aim of this study is to develop and validate such a model that ensures high accuracy and robustness in product matching across large datasets of product listings, which are typically found in modern digital marketplaces.

The article is divided into four sections. The first section is devoted to the potential and applicability of product matching on digital marketplaces. The second section analyzes the existing approaches and formulates the product matching task for digital marketplaces. The third section presents the development of a multimodal product matching model for digital marketplaces (MMMP). The fourth section discusses the application of the model for evaluating product similarity in the Wildberries digital marketplace.

1. Potential and applicability of product matching in digital marketplaces

Product matching refers to the process of aligning (linking) products to determine which items are identical or essentially similar, ensuring that the same product is correctly recognized even if it appears under different titles, descriptions, or identifiers. For example, when equivalent models of a smartphone from the same brand are offered by different sellers in marketplaces, their specifications, images and even the naming format may vary significantly.

Product matching is a fundamental component widely used in various modern digital ecosystems, including marketplaces, classified ad platforms, e-commerce services and other online platforms. From the buyer's perspective, product matching significantly enhances the shopping experience by enabling quick and easy comparison of similar product offerings from different sellers [1]. For example, by implementing a product matching system that consolidates similar listings into a unified and well-structured collection of comparable offers, buyers can be spared the need to look through numerous variations within the same product category. Such a system would protect buyers from encountering duplicate or misleading offers that often make product selection a confusing and time-consuming task. Instead, buyers would be able to quickly compare prices, reviews, and seller ratings, which ultimately improves search relevance, enhances overall shopping satisfaction and supports more informed and rational decision-making. Moreover, product matching can also be used to generate personalized recommendations based on the analysis of consumer behavior [2], preferences and past purchases, helping to discover the most suitable products, increasing commitment and enhancing the overall value of the platform.

From the seller's perspective, product matching helps analyze competitive offers [3], enabling them to adjust and develop their pricing strategies based on real-time market trends [4]. Digital marketplaces accumulate vast amounts of product data; however, without proper processing and structural organization, much of this information remains fragmented, inconsistent and difficult to analyze effectively. For instance, many sellers list the same product on platforms with slight variations in the name, images, description, or characteristics. Without product matching, sellers are forced to manually track and compare thousands of listings, which is a highly time-consuming and costly process prone to errors. Automated identification and linking of similar products make it possible to consolidate scattered information into a unified coherent data repository which becomes a valuable source for further analysis and decision-making based on the current market environment. Possessing such a data source enables the application of algorithmic pricing models that rely on market assessments, as they allow for real-time monitoring of competitor prices and market demand fluctuations. This opens the door to dynamic pricing strategies. As a result, sellers can adapt their offerings in real time according to consumer behavior and the existing market environment, ensuring that product prices remain competitive and attractive to buyers without affecting margins. Such flexibility enables the identification of emerging trends and allows for proactive responses to seasonal fluctuations, promotional activities and economic shifts, ultimately contributing to long-term and sustainable development in the highly competitive digital platform market. Product matching also plays a crucial role in advertising and marketing campaigns. One of its main advantages in this context is the ability to optimize advertising costs. By correctly identifying similar products offered by other sellers, unnecessary competition can be avoided, helping to eliminate wasted expenditures and redirect the budget toward more profitable niches. In addition to advertising, product matching also enhances cross-selling opportunities. By intelligently linking complementary, frequently purchased, or otherwise related items, sellers can create more attractive offers for consumers.

Another important application of product matching algorithms lies in their role in fighting fraud and detecting counterfeit goods, thereby assisting platform operators in maintaining trust, safety and the integrity of ecosystems [5]. With the rapid development and growing influence of e-commerce platforms on public life, incidents involving fraudulent listings, counterfeit items, and deliberately misleading products are becoming increasingly common, posing a serious challenge to the industry. Such illicit activities not only harm end consumers but also undermine overall trust in digital platforms and the business of law-abiding sellers [6]. The need for counterfeit detection is especially crucial in many product categories, such as electronics, pharmaceuticals and cosmetics. In these segments, counterfeit goods can pose not only financial risks but also direct threats to consumer health and safety. In this context, product matching algorithms serve as a vital tool for preventing the appearance of listings and offers from unauthorized resellers or the sale of low-quality counterfeit products disguised as reputable brands.

Based on the above, it is reasonable to conclude that the potential and capabilities of product matching go well beyond the scope of a simple tool for solving a single task. It can be stated that product matching is a fundamental technology that enables the optimization of various business processes, supports fraud prevention and enhances the efficiency of pricing strategies which ultimately contributes to the creation of a more transparent and user-friendly ecosystem for all the participants in the e-commerce market, including both consumers and sellers.

2. Analysis of existing approaches and formulation of product matching problem in marketplaces

2.1. Product matching based on attribute value similarity and set-theoretic methods

Product attribute matching based on value similarity can be referred to fundamental matching approaches relying on comparison of textual and numerical fields across various characteristics to determine how closely they align [7]. Textual attributes are typically assessed using metrics such as Levenshtein distance, Jaro-Winkler distance or TF-IDF (Term Frequency – Inverse Document Frequency), which allows for determination of how similar two text fragments are while accounting for differences in spelling [8]. Numerical attributes (e.g., price, weight, dimensions, etc.) are usually evaluated by calculating absolute or relative deviations. Once individual similarity metrics are created and computed for each attribute, they are aggregated (commonly through a weighted sum) to produce an overall similarity evaluation between products being compared. If the indicator obtained exceeds a predefined threshold, the products are considered similar. Additionally, in the set-theoretic approach each product listing is designed as a set of atomic elements (of features, n-grams, tokens, etc.). Similarity between items is determined using classical metrics such as the Jaccard, Sørensen–Dice, or the Simkovich–Simpson index. Due to the computational simplicity, this approach provides high speed and easy scalability. However, it ignores word order and context, which significantly limits its accuracy. Key advantages of this approach include its relative simplicity, transparency and ease of implementation. However, its main limitations lie in handling large and diverse product categories, dealing with low-quality or noisy data and its severely restricted ability to capture semantic meaning.

2.2. Product matching based on a rule-based expert system

A natural extension of the attribute value comparison approach is the development of rule-based systems [9, 10]. Rule-based product matching is built using expert-defined logical constructs that describe the functioning of a specific domain and help determine the degree of similarity between products being

assessed. Each rule typically evaluates a subset of well-defined attributes and applies simple logical conditions or threshold values (e.g., "If the brand name is identical and the Levenshtein distance between model names is less than two, then the two products can be considered similar"). Since this approach relies on system of rules directly encoding the expert knowledge, it is generally quick and easy to understand. However, when applied to broad and complex product categories, the rule system can become cumbersome and difficult to maintain. It also requires constant manual updates and enhancement. Moreover, because each rule must be determined manually by domain experts, the system is prone to human error and ultimately turns out to be insufficiently adaptive and excessively overloaded.

2.3. Product matching based on taxonomies and ontologies

These methods rely on in-depth contextual and semantic analysis of domain-specific information and use structured representations to map relationships between products and their attributes in the form of a knowledge graph [11, 12]. These relationships can indicate shared characteristics (e.g., belonging to a certain brand or product family), hierarchical inheritance or descriptive associations. Modeling data through a dependency graph allows for deeper exploration of available information by incorporating context, internal relationships between entities and hidden patterns. Instead of relying solely on pairwise attribute comparison, this approach enables the identification of complex dependencies, the use of intricate logical relationships and the evaluation of structural features in aggregate. Additionally, the constructed ontologies and taxonomies can make a significant contribution to the standardization or alignment of fragmented information, which is particularly important when integrating data from multiple sources [13]. Key advantages of this approach include its holistic perspective, which enables the discovery of clusters and related components of equivalent products by analyzing entire subgraphs sometimes revealing matches missed by simpler methods. However, building and maintaining a comprehensive knowledge graph as a rule represents a labor-intensive and resource-demanding process that requires regular updates as a product assortment expands or new data sources emerge. Furthermore, incomplete or inconsistent taxonomies and ontologies can dramatically affect matching accuracy, which potentially negates the benefits of this approach.

2.4. Product matching based on machine learning

Machine learning and deep neural network approaches treat the product matching as an objective functional optimization task [14, 15]. In this formulation, a pre-trained model evaluates a set of product attributes and determines whether they correspond to an equivalent item. Classical machine learning methods begin with choice and engineering a feature collection to construct a representative description of a product. These engineered features are then passed into a binary classifier (such as logistic regression, random forest or support vector machines (SVM), etc.) which is trained to distinguish similar products based on dependencies in the original attribute space. Application of these methods requires a labeled dataset with indication of target class labels, enabling the model to learn how similarities and differences in features affect the probability of a match. While classical machine learning methods can achieve high predictive accuracy, they rely on labels, strong domain expertise and meticulous engineering of initial attributes, which necessitate ongoing adaptation in case of changes in data or their distribution. Deep learning methods significantly reduce or even eliminate the need for costs related to manual feature engineering by automatically learning patterns from raw

data. Common base models include recurrent neural networks (RNNs), long short-term memory networks (LSTMs) and convolutional neural networks (CNNs). These models process the input attributes to generate embedding representing characteristic vector which is then used to assess product similarity. Key advantages of deep learning models include their ability to more accurately handle noisy or heterogeneous data, absence of need for manual feature engineering, and greater robustness when processing previously unknown inputs. However, they also face challenges in capturing and interpreting semantic meaning and still require availability of labeled data.

2.5. Formulation of the task for product matching in marketplaces

Based on the analysis and the operational specifics of digital marketplaces, it becomes clear that typical challenges in product matching include: ambiguous descriptions and the use of marketing terms, data duplication and missing information, multi-format inputs (text descriptions, images, tabular attributes) and varying levels of details. All these factors increase the risk of incorrect matches or, conversely, missed potential matches. Therefore, it is necessary to develop a multimodal model capable of processing both the visual component and product characteristics while accounting for their semantic meaning. This task formulation reflects the heterogeneous nature of the data and the high variability of the digital environment. Relying solely on attributes is often insufficient and can result in matching errors, especially when key product differences are visible in images. In a similar manner, images alone can be ambiguous or appear almost identical. Thus, by integrating these components into a single system, the model will be able to perform product matching more effectively and comprehensively.

3. Multimodal product matching model for marketplaces (MMMP)

Let there be a set of product information cards $\Omega = \{p_1, p_2, p_3, ..., p_n\}$ containing information about each product. The task is to design a model Ψ (1) that maps the set Ω into a space $\mathbb{R}^d$ such that the resulting vector representation enables the assessment of similarity between any pair of products from Ω according to (2), where S is a similarity measure:

$$\Psi(p_i) = (x_1, x_2, x_3, ..., x_d); p_i \in \Omega, \quad (1)$$

$$S : \mathbb{R}^d \times \mathbb{R}^d \to [0, 1]. \quad (2)$$

Since the model must process multiple modalities, its conceptual representation can be decomposed into component parts as shown in (3):

$$\Psi(p_i) = \text{Concat}(\psi_{title}(p_i), \psi_{image}(p_i), \psi_{metadata}(p_i)); p_i \in \Omega, \quad (3)$$

where ψ_{title} encodes the product name into a vector representation;

ψ_{image} encodes the product image into a vector representation;

$\psi_{metadata}$ encodes the product attributes into a vector representation.

Since it is necessary to account for context and semantic meaning, this paper proposes using the transformer architecture introduced in [16] as the foundation. It is known that this architecture can represent textual data in vector form and, due to its attention mechanism combined with deep neural networks, it effectively captures semantic and contextual features [17]. Unlike traditional embedding generation models such as Word2Vec [18] and GloVe [19], which assign a fixed vector to each word, the transformer generates dynamic, context-dependent embeddings. This means that the same word can have different representations depending on the context in which it appears in a sentence. Given that this paper focuses on vector repre-

sentation, only the encoder block of the transformer is relevant. The encoder architecture consists of three key components: a multi-head self-attention mechanism, a residual connection and normalization layer and a feedforward neural network. Let us examine the main transformations involved.

Let there be a set of input tokens $t = \{t_1, t_2, t_3, ..., t_n\}$, each of which is associated with its own embedding as defined in (4):

$$\text{Embed: } t_n \to e_i \in \mathbb{R}^{d_{model}}, \tag{4}$$

where d_{model} corresponds to the dimensionality of the model's embeddings.

To account for the order of elements, positional encoding of the form (5, 6) is used:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \tag{5}$$

$$PE\left(pos, 2i+1\right) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right), \tag{6}$$

where *pos* is the position in sequence;

i is the dimensionality.

Then, the input to the first encoder block is given by (7):

$$X = \left(e_1 + \rho e_1, e_2 + \rho e_2, \ldots, e_n + \rho e_n\right) \in \mathbb{R}^{n \times d_{model}}. \tag{7}$$

Each encoder block contains an attention mechanism component that uses query (Q), key (K) and value (V) matrices, which are defined as follows in (8):

$$Q = XW^Q,\ K = XW^K, V = XW^V;$$
$$W^Q,\ W^K,\ W^V \in \mathbb{R}^{d_{model} \times d_k}, \tag{8}$$

where W^Q, W^K and W^V are the trainable parameters of the model;

d_k is typically defined as the ratio of d_{model} to the number of attention heads.

The attention-based importance score is computed according to (9):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{9}$$

The transformer architecture uses multiple attention heads, which are concatenated as shown in (10) and (11):

$$head_i = \text{Attention}\ (Q_i, K_i, V_i), \tag{10}$$

$$\text{MultiHead}(X) =$$

$$= \text{Concat}\ (head_1, head_2, head_3, \ldots head_n)W^O, \tag{11}$$

where the matrix W^O is also a trainable parameter of the model.

An important component of the transformer architecture is the residual connection [20] and normalization layer [21], applied after the attention mechanism as shown in (12):

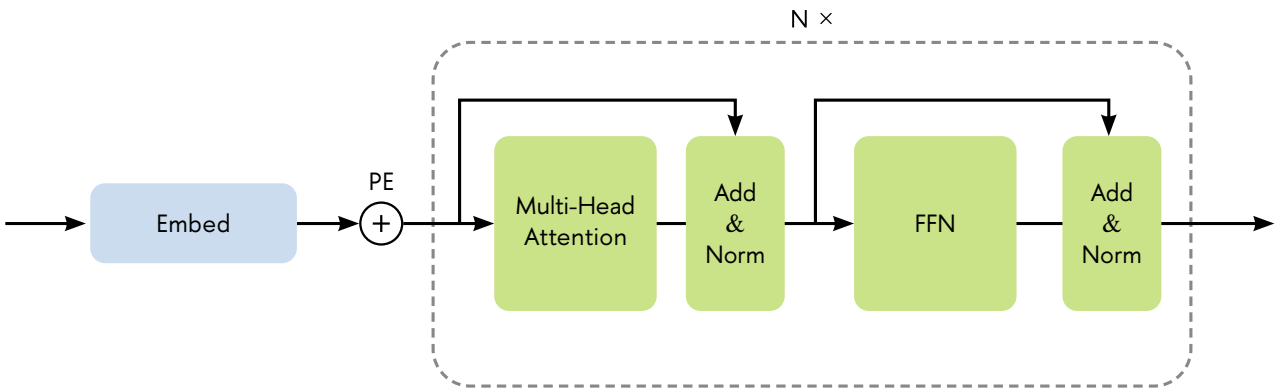
$$X_{attn} = \text{LayerNorm}\,(X + \text{MultiHead}(X)). \tag{12}$$

Each encoder block also includes a feedforward neural network, typically in the form of a two-layer MLP (13) with a nonlinearity such as ReLU (or GeLU):

$$\text{FFN}(X_{attn}) = \max(0, X_{attn} W_1 + b_1)W_2 + b_2, \tag{13}$$

after which, the residual connection and normalization layer is applied again. The output from one encoder block is then transferred sequentially through N similar encoder blocks, producing the final output. A conceptual representation of the encoder architecture is shown in *Fig. 1*.

The task of determining the similarity between product titles requires a reliable representation of sentence semantics that allows for effective and accurate comparison of textual content. Traditional transformer-based models, such as BERT (Bidirectional Encoder Representations from Transformers) [22], have dem-



onstrated significant success across a wide range of natural language processing (NLP) tasks. However, BERT is optimized for token-level tasks and does not generate fixed-size, holistic sentence-level embeddings. To overcome this limitation, Sentence-BERT [23] is proposed, which is specifically designed to generate sentence-level embeddings and employs a Siamese neural network architecture. Unlike BERT, which processes each pair independently, Sentence-BERT encodes input into a dense fixed-size vector enabling efficient computation of cosine similarity between sentences in vector space. Moreover, Sentence-BERT can learn using contrastive loss functions of the form (14), which explicitly optimize the model for identifying semantic similarity between sentences rather than just capturing contextual token-level relationships:

$$L = (1-y_k)\left\|f(x_i)-f(x_j)\right\|_2^2 + \\ +y_k\left(\max\left(0, m-\left\|f(x_i)-f(x_j)\right\|_2\right)\right)^2, \quad (14)$$

where x_i and x_j are a pair of items;

y_k is a binary label that takes the value 0 if the objects are similar, and 1 if they are different;

m defines the margin that separates the items.

Based on the above and in line with the stated task, this study proposes using Sentence-BERT to generate high-quality sentence embeddings that enable effective assessment of similarity on the basis of semantic relationships between product titles presented in textual form. In this way, we obtain a robust implementation of the ψ_{title} component. Using the generated embeddings, we apply cosine similarity as the similarity metric (15):

$$\mathrm{Sim}\left(e_i, e_j\right) = \frac{e_i \cdot e_j}{\left\|e_i\right\|\left\|e_j\right\|}. \quad (15)$$

To define the visual component of the model ψ_{image} it is important to consider that digital platforms host a large variety of products with significant differences in image quality, angle, lighting, shadows, and the positioning of elements. These factors collectively render traditional similarity assessment methods insufficiently effective. It is known that traditional convolutional neural networks, such as ResNet [24] and EfficientNet [25] are widely used for image similarity search. However, these models rely on local convolutions and therefore have a fixed receptive field, which can limit their ability to capture distant or disjoint relationships between elements within an image. In product matching, fine-grained relationships between elements such as logos, labels, and shapes are crucial, and CNNs often struggle to effectively process such dependencies. Furthermore,

CNNs tend to be highly sensitive to the aforementioned variations in images, which raises further concerns about their suitability for solving the current task. To address this issue, this study proposes the use of the vision transformer (ViT) architecture, as described in [26]. The vision transformer (ViT) processes images as a sequence of patches and uses an attention mechanism to capture both local and global relationships within the image. Unlike CNNs, which extract features in a hierarchical manner, the vision transformer analyzes the entire image simultaneously, allowing it to dynamically focus on the most relevant areas. This distinctive property makes ViT particularly useful for assessing the similarity of clothing items, electronics and consumer goods, where differences in texture, brand placement, and various distortions can significantly impact the final similarity score. Another important advantage of the vision transformer in the context of this task is its high robustness to occlusions in images. For example, ViT can correctly identify a product model even if the brand logo is partially obscured or the image is captured from a different angle, something traditional CNNs typically struggle with. It is also worth noting that vision transformers offer strong potential for integrating image and its textual description to compute semantic similarity between them [27]. This is especially valuable in product search scenarios, where different textual attributes (such as title, description, brand, model, etc.) and the corresponding image can be associated to check compliance.

Based on the above, this study proposes defining ψ_{image} as a multistage pipeline consisting of pretraining, fine-tuning and subsequent evaluation.

In the first stage, we perform initial training on a large dataset to learn generalized representations and extract fundamental features. Let f_θ denotes a vision transformer (ViT) parameterized by θ and let there be a dataset containing images of all products $I = \{x_1, x_2, x_3, ..., x_n\}$. Using the self-supervised distillation method DINO [28], we optimize according to (16) and generate an initial embedding for each product image as shown in (17):

$$\theta^* = \arg\min_\theta L_{SSL}(f_\theta, I). \quad (16)$$

$$e_i = f_{\theta^*}(x_i) \forall x_i \in I. \quad (17)$$

In the second stage, we adapt the vision transformer. To do this, we use subsets of product images obtained by dividing I into categories and price segments, and by performing fine-tuning according to (18):

$$L = \max\left(\begin{array}{l}\left\|f_{\theta^*}(x_a) - f_{\theta^*}\left(x_p\right)\right\|_2^2 - \\ -\left\|f_{\theta^*}(x_a) - f_{\theta^*}(x_n)\right\|_2^2 + \alpha, 0\end{array}\right), \quad (18)$$

where x_a is the anchor object;

x_p denotes the object similar to the anchor;

x_n represents the object that is different from both x_a and x_p.

After completing the fine-tuning process, we use (15) to evaluate the similarity between any pair of vector representations e_i and e_j. The described pipeline ensures high efficiency and accuracy of the ψ_{image} component.

When defining the third component of the model $\psi_{metadata}$ it is important to consider that product characteristics in digital marketplaces include various types of data, such as numerical variables (e.g., weight, volume, or dimensions) and categorical variables (e.g., configuration, color, or composition). Traditional approaches often rely on multilayer perceptrons (MLPs) or ensemble methods, which frequently struggle to detect complex interactions in a multitude of diverse features, in the absence of prior feature engineering. Given the nature and specifics of the domain, it is reasonable to assume that certain features within the tabular representation of product attributes may interact with each other in a non-

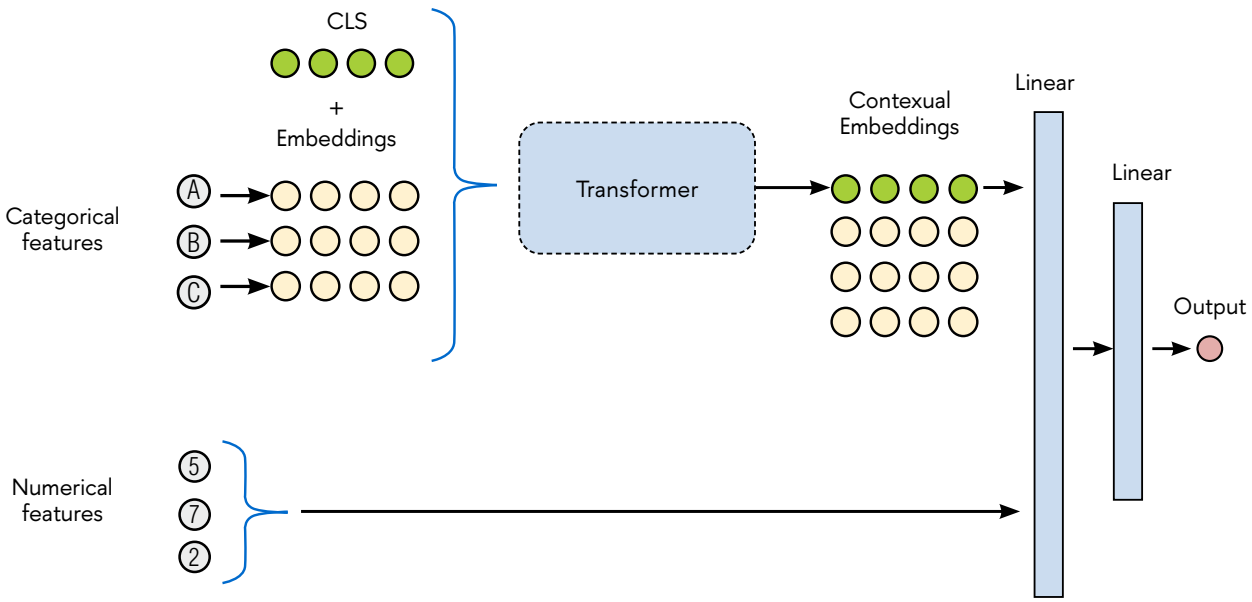


Fig. 2. Architecture of the $\psi_{metadata}$ component.

trivial way. These interactions are crucial for the task at hand, as even minor differences (such as between "cotton" and "poly-cotton" in composition) can significantly impact the perceived identity of characteristics. As previously discussed, the core of the transformer architecture is the attention mechanism, enabling detection of how the attributes are interrelated. The context-dependent embeddings generated by the transformer form vector representations that position similar elements close to each other in the feature space, taking their interrelationships into account. This enables more accurate differentiation of similar products based on their available attributes. Moreover, the attention mechanism effectively identifies feature interactions in an automated manner, eliminating the need for domain experts and manual feature engineering. In addition to its high flexibility with respect to input features, the transformer is also effective at handling missing or distorted data, which is a common issue on large e-commerce platforms. Furthermore, several studies [29, 30] exploring the application of transformers to tabular data have shown that such architectures can outperform traditional approaches.

Based on the above, the $\psi_{metadata}$ component of the model is defined using the concept of a tabular transformer. The architecture adapted for the task at hand, as proposed in this study, is shown in *Fig. 2*. To evaluate the similarity between any pair of products using the tabular transformer, we apply (15).

The proposed tabular transformer architecture effectively generates vector representations of product attributes by applying context-dependent embeddings for categorical features, specialized processing of numerical variables and an attention mechanism to capture complex interactions. This is particularly relevant in large-scale and heterogeneous catalogs, which are typical of digital marketplaces.

After defining all components, the final similarity measure is specified using a weighted scoring method as shown in (19):

$$S\left(p_i, p_j\right) = w_1 \mathrm{Sim}\left(\psi_{title}(p_i), \psi_{title}(p_j)\right) + \\ + w_2 \mathrm{Sim}\left(\psi_{image}(p_i), \psi_{image}(p_j)\right) + \\ + w_3 \mathrm{Sim}\left(\psi_{metadata}(p_i), \psi_{metadata}(p_j)\right), \sum w_i = 1. \quad (19)$$

Here the weighting coefficients w_1, w_2 and w_3 can be flexibly adjusted to account for the characteristics and specific requirements of the matching task.

4. MMMP application for product similarity assessment in Wildberries marketplace

As part of the experiment, a decision was made to conduct a study on products listed on the Wildberries platform. The primary focus is on the product category "Smartphones," which was chosen due to several important factors. First, smartphones are complex technical devices with a wide range of features, including both numerical attributes (e.g., memory size, battery capacity, etc.) and categorical attributes (e.g., brand, operating system, etc.). This makes them a suitable object for testing in heterogeneous features environment. Second, the assortment of such devices in marketplaces is quite extensive and diverse, allowing the model operation to be examined on highly variable data. Third, this category is one of the most popular and in-demand in marketplaces, which adds practical significance to the results of the study. In combination these factors provide an effective setting for testing the model's ability to assess product similarity. Thus, selecting the "Smartphones" category enables evaluation of the proposed model under real-world conditions, offering sufficient complexity and diversity to assess its performance. It follows that a successful application of the model in this product group can be extended to other categories, including technically complex goods.

During the experiment, a dataset was collected consisting of 12233 product listings. Each listing contains information about a specific item available on the platform and includes the following features: X1 – product image, X2 – product title, X3 – color, X4 – model, X5 – SIM card type, X6 – operating system, X7 – operating system version, X8 – warranty period, X9 – dust and water protection rating, X10 – display type, X11 – screen size, X12 – screen resolution, X13 – screen refresh rate, X14 – screen protective coating, X15 – internal storage capacity, X16 – RAM, X17 – main camera, X18 – front camera, X19 – lens features, X20 – built-in flash, X21 – battery capacity, X22 – processor model, X23 – number of processor cores, X24 – network standard, X25 – wireless interfaces, X26 – satellite navigation, X27 – connector type, X28 – additional features, X29 – package contents, X30 – product description, X31 – country of manufacture, X32 – number of SIM cards, X33 – product condition, X34 – processor clock speed, X35 – product weight with packaging, X36 – service life.

A sample of the collected data is shown in *Table 1*.

Before using the collected data to train the model, thorough preprocessing was carried out to ensure data quality, consistency and suitability for further analysis. These procedures included cleaning, duplicate removal, handling of missing values, correction of inconsistencies, normalization of numerical data and tokenization of textual information. Additionally, outlier detection and removal were performed to minimize the impact of anomalous values.

After training the model, similarity scores were obtained for the products shown in *Table 1*, with the results presented in *Table 2*. The effectiveness of the proposed method was also evaluated by comparing it with a naive baseline approach based on Jaccard index, applied without prior tokenization.

The experimental part of the study confirmed the effectiveness of the proposed approach. Specifically, the data in *Table 2* illustrate the example of similarity scores calculated by the MMMP model for four product listings (Product 1, Product 2, Product 3, and

Table 1.

Sample of collected data

	Product 1	Product 2	Product 3	Product 4
Image				
Title	Apple iPhone 16 Pro Max Gold/Desert 512 GB	Apple iPhone 16 Pro Max White 1 TB	Galaxy S24 Ultra 512 GB Yellow	Xiaomi 14 12/256 GB 5G White RST
Price	168 750	182 267	159 800	77 585
Model	iPhone 16 Pro Max	iPhone 16 Pro Max	S24	Xiaomi 14
OS version	iOS 18	IOS 18	Android 14	Android 14
Warranty period	1 year	1 year	1 year	1 year
Display type	Super Retina XDR OLED ProMotion	Super Retina XDR OLED ProMotion	Dynamic AMOLED 2X	LTPO AMOLED
Screen size	6.9”	6.9”	6.8”	6.36”
Screen resolution	2868x1320	2868x1320	3120x1440	2670x1200
Internal storage capacity	512 GB	1 TB	512 GB	256 GB
RAM	8 GB	8 GB	12 GB	12 GB
Wireless interfaces	Wi-Fi 7; NFC; Bluetooth 5.3	Wi-Fi 7; NFC; Bluetooth 5.3	Wi-Fi; NFC; Bluetooth	Wi-Fi; IR-Port; NFC; Bluetooth
…	…	…	…	…

Table 2.

Product similarity assessment

		Product 1	Product 2	Product 3	Product 4
Product 1	MMMP	1	0.98657	0.877564	0.544109
	Jaccard	1	0.308	0.11	0.077
Product 2	MMMP	0.98657	1	0.883621	0.549463
	Jaccard	0.308	1	0.10	0.11
Product 3	MMMP	0.877564	0.883621	1	0.670598
	Jaccard	0.11	0.10	1	0.13
Product 4	MMMP	0.544109	0.549463	0.670598	1
	Jaccard	0.077	0.11	0.13	1

Product 4). The highest similarity score (0.98657) is observed between Product 1 and Product 2, indicating successful identification of nearly identical items despite differences in descriptions and data modalities. Simultaneously less similar items (such as Product 1 and Product 4) received significantly lower similarity score (0.544109), demonstrating the model's ability to correctly distinguish between groups of products with different sets of attributes and descriptions. Similar results (0.549463 between Product 2 and Product 4, and 0.670598 between Product 3 and Product 4) show that the model is capable of recognizing even subtle differences in product characteristics while also detecting meaningful patterns in their descriptions. Thus, the proposed approach reliably identifies both clear and nuanced matches and avoids incorrect grouping of products that are, in fact, different.

The results obtained confirm that the transformer-based multimodal model is capable of effectively solving the product matching task in the presence of a wide variety of product listings and heterogeneous data sources. Accurate identification of product records in marketplaces reduces the risk of duplication and analytical errors, positively impacting all stages of economic activity within the digital platform ecosystem. It also helps both sellers and buyers to navigate the product assortment more accurately, simplifying the decision-making process. Moreover, the results obtained demonstrate the model's correct functioning in relation to the defined task making possible its application in intelligent product matching in marketplaces.

Conclusion

In the course of this study, the task of intelligent product matching in digital marketplaces was proposed and solved. This task requires comprehensive analysis and processing of multimodal data, as well as the application of modern instrumental methods within the framework of an economic and mathematical model (MMMP). The first section of the article highlighted the potential of product matching in marketplaces as typical representatives of digital platforms and demonstrated the relationship between intelligent matching capabilities and key economic performance indicators within the e-commerce market. The second section analyzed existing approaches to entity resolution (ER) and multimodal data analysis, revealing the core challenge of product record matching in marketplaces: the fact that products are described by different sellers with varying levels of details and in different formats (text, images, tabular attributes, etc.). The third section described the detailed development process of the multimodal product matching model (MMMP) whose core is composed of transformer-based modules for processing textual, visual, and tabular data. The proposed architecture takes into account the flexibility of the attention mechanism and is capable of self-learning as product assortments expand and new unstructured descriptions emerge with consideration of visual data components. MMMP enables effective integration of diverse input data modalities and captures complex contextual relationships which are critical for accurate determination of product similarity or dissimilarity, considering pricing characteristics.

Finally, in the fourth section, the application of the proposed model for assessing product similarity in the Wildberries marketplace was demonstrated, where the results confirmed the high accuracy and stability of the MMMP operation.

Thus, the main conclusions of the article can be summarized in an axiomatic form:

- the high relevance of multimodal analysis for product matching tasks has been substantiated in the context of the growing diversity of product listings in marketplaces;
- the potential of transformer architectures for the integrated processing of textual, visual and tabular features has been established;
- the need for further adaptation of such models to evolving market conditions has been identified;
- the MMMP model has been developed, which can be used both in academic research of intelligent product record identification in marketplaces and for practical purposes by participants in the e-commerce market — where matching accuracy is crucially important, for instance, in pricing, and thus for establishing an equilibrium model of supply and demand.

The results of this work demonstrate that a well-designed deep learning architecture that integrates multiple data modalities provides a significant advantage over simpler and more narrowly focused existing solutions. Moreover, the application of such models contributes to enhancing the transparency and effectiveness of analytical tools, which ultimately helps strengthen the trust of sellers and more importantly, the trust of buyers in a given marketplace as a whole.

Future research directions within the chosen subject area may include expanding the set of processed modalities (e.g., incorporating video reviews or audio information) and developing self-learning mechanisms that allow the model to automatically reconfigure itself in response to changes in data structure. Additionally, the continued development of the MMMP model suggests the integration of active learning methods (in real time), as this would enable faster accumulation and processing of relevant examples to refine matching criteria, including in the environment of marketplaces. ■

References

1. Fletcher A., Ormosi P. L., Savani R. (2023) Recommender systems and supplier competition on platforms. *Journal of Competition Law & Economics*, vol. 19, no. 3, pp. 397–426. https://doi.org/10.1093/joclec/nhad009

2. Hussien F.T.A., Rahma A.M.S., Abdulwahab H.B. (2021) An e-commerce recommendation system based on dynamic analysis of customer behavior. *Sustainability*, vol. 13, no. 19, article 10786. https://doi.org/10.3390/su131910786

3. Chen F., Liu X., Proserpio D., et al. (2020) Studying product competition using representation learning. Proceedings of the *43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, pp. 1261–1268. https://doi.org/10.1145/3397271.3401041

4. Hu S., Wei M.M., Cui S. (2023) The role of product and market information in an online marketplace. *Production and Operations Management*, vol. 32, no. 10, pp. 3100–3118. https://doi.org/10.1111/poms.14025

5. Cheung M., She J., Sun W., Zhou J. (2019) Detecting online counterfeit-goods seller using connection discovery. *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 2, article 35. https://doi.org/10.1145/3311785

6. Sun J., Zhang X., Zhu Q. (2020) Counterfeiters in online marketplaces: Stealing your sales or sharing your costs. *Journal of Retailing*, vol. 96, no. 2, pp. 189–202. https://doi.org/10.1016/j.jretai.2019.07.002

7. Köpcke H., Thor A., Rahm E. (2010) Evaluation of entity resolution approaches on real-world match problems. *Proceedings of the VLDB Endowment*, vol. 3, nos. 1–2, pp. 484–493. https://doi.org/10.14778/1920841.1920904

8. Cohen W.W., Ravikumar P., Fienberg S.E. (2003) A Comparison of string distance metrics for name-matching tasks. Proceedings of *Workshop on Information Integration (IJCAI-03)*, pp. 73–78.

9. Singh R., Meduri V.V., Elmagarmid A., et. al. (2017) Synthesizing entity matching rules by examples. *Proceedings of the VLDB Endowment*, vol. 11, no. 2, pp. 189–202. https://doi.org/10.14778/3149193.3149199

10. Wang J., Li G., Yu J.X, Feng J. (2011) Entity matching: How similar is similar. *Proceedings of the VLDB Endowment*, vol. 4, no. 10, pp. 622–633. https://doi.org/10.14778/2021017.2021020

11. Angermann H. (2022) TaxoMulti: Rule-based expert system to customize product taxonomies for multi-channel e-commerce. *SN Computer Science*, vol. 3, article 177. https://doi.org/10.1007/s42979-022-01070-8

12. Mao M., Chen S., Zhang F., et. al. (2021) Hybrid ecommerce recommendation model incorporating product taxonomy and folksonomy. *Knowledge-Based Systems*, vol. 214, article 106720. https://doi.org/10.1016/j.knosys.2020.106720

13. Aanen S. S., Vandic D., Frasincar F. (2015) Automated product taxonomy mapping in an e-commerce environment. *Expert Systems with Applications*, vol. 42, no. 3, pp. 1298–1313. https://doi.org/10.1016/j.eswa.2014.09.032

14. Ristoski P., Petrovski P., Mika P., Paulheim H. (2018) A machine learning approach for product matching and categorization: Use case: Enriching product ads with semantic structured data. *Semantic Web*, vol. 9, no. 5, pp. 707–728. https://doi.org/10.3233/SW-180300

15. Shah K., Kopru S., Ruvini J. D. (2018) Neural network based extreme classification and similarity models for product matching. Proceedings of the *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans – Louisiana*, vol. 3, pp. 8–15. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-3002

16. Vaswani A., Shazeer N., Parmar N., et. al. (2017) Attention is all you need. Proceedings of the *31st International Conference on Neural Information Processing Systems (NIPS'17), Long Beach, CA, USA*, pp. 6000–6010. https://dl.acm.org/doi/pdf/10.5555/3295222.3295349

17. Zhang H., Shafiq M.O. (2024) Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, vol. 11, article 25. https://doi.org/10.1186/s40537-023-00842-0

18. Mikolov T., Chen K., Corrado G., Dean J. (2013) Efficient estimation of word representations in vector space. *arXiv:1301.3781*. https://doi.org/10.48550/arXiv.1301.3781

19. Pennington J., Socher R., Manning C. D. (2014) GloVe: Global vectors for word representation. Proceedings of the *2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar*, pp. 1532–1543.

20. He K., Zhang X., Ren S., Sun J. (2016) Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

21. Ba J.L., Kiros J.R., Hinton G.E. (2016) Layer normalization. *arXiv:1607.06450.* https://doi.org/10.48550/arXiv.1607.06450

22. Devlin J., Chang M. W., Lee K., Toutanova K. (2019) Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota*, vol. 1, pp. 4171–4186. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

23. Reimers N., Gurevych I. (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China*, pp. 3982–3992. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

24. Wu Z., Shen C., van den Hengel A. (2019) Wider or deeper: Revisiting the ResNet model for visual recognition. *Pattern Recognition*, vol. 90, pp. 119–133. https://doi.org/10.1016/j.patcog.2019.01.006

25. Tan M., Le Q. (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the *36th International Conference on Machine Learning*, vol. 97, pp. 6105–6114.

26. Dosovitskiy A., Beyer L., Kolesnikov A., et al. (2021) An image is worth 16x16 words: transformers for image recognition at scale. *arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929

27. Radford A., Kim J. W., Hallacy C., et. al. (2021) Learning transferable visual models from natural language supervision. Proceedings of the *38th International Conference on Machine Learning*, vol. 139, pp. 8748–8763.

28. Caron M., Touvron H., Misra I., et. al. (2021) Emerging properties in self-supervised vision transformers. *arXiv:2104.14294*. https://doi.org/10.48550/arXiv.2104.14294

29. Huang X., Khetan A., Cvitkovic M., et. al. (2020) TabTransformer: Tabular data modeling using contextual embeddings. *arXiv:2012.06678*. https://doi.org/10.48550/arXiv.2012.06678

30. Gorishniy Y., Rubachev I., Khrulkov V., et. al. (2021) Revisiting deep learning models for tabular data. Proceedings of the *35th International Conference on Neural Information Processing Systems (NIPS'21)*, article 1447, pp. 18932–18943.

About the authors

Artem Yu. Varnukhov

Assistant, Department of Business Informatics, Ural State University of Economics, 62, 8 Marta St., Yekaterinburg 620144, Russia;

E-mail: varnuhov_ayu@usue.ru

Dmitry M. Nazarov

Doctor of Sciences (Economics);

Head of Department, Department of Business Informatics, Ural State University of Economics, 62, 8 Marta St., Yekaterinburg 620144, Russia;

E-mail: slup2005@mail.ru

ORCID: 0000-0002-5847-9718