Improving Pricing Recommendations Using Nearest Neighbors Retrieval Via Contrastive Learning and Hard Negatives Mining

Eyal Mazuz (✉)[1], Gilad Fuchs[2], Alex Nus[2], Lior Rokach[1], and Bracha Shapira[1]

[1] Ben-Gurion University of the Negev {mazuze, liorrk, bshapira}@post.bgu.ac.il
[2] eBay {gfuchs, alnus}@ebay.com

Abstract. Accurately determining selling prices for listings in online marketplaces poses a significant challenge due to the lack of universally recognized identifiers, such as Global Trade Item Numbers (GTIN) or Universal Product Codes (UPC). This lack of uniformity results in inconsistencies across product descriptions, titles, attributes, and features, complicating price prediction efforts. Traditional approaches for price prediction have predominantly relied on manually engineered features or direct price predictions from textual and image data, often failing to capture the nuanced differences between similar products. While transformer architectures have been widely used in e-commerce for item recommendation and retrieval tasks, these applications focus mainly on single-modal retrieval and do not address the complexities of pricing.

In this paper, we introduce a novel approach to price recommendation by leveraging item retrieval methods enhanced with hard negatives during training. Incorporating hard negatives improves the quality of the generated embeddings, enabling more effective differentiation between similar listings with significantly different prices. This methodology focuses on understanding the contextual relationships and characteristics of listings relative to one another, rather than solely focusing on direct price prediction.

By integrating contrastive learning with both price and aspects-based hard negatives, our approach better distinguishes between similar listings, significantly advancing price recommendation methods. Our research addresses this gap, aiming to significantly enhance the accuracy and effectiveness of pricing strategies. Extensive evaluations show that our method substantially improves pricing accuracy and enhances retrieval accuracy compared to existing approaches. We present extensive analysis and demonstrate successful deployment to production environment.

Keywords: Recommender Systems · Information Retrieval · Contrastive Learning · Hard-Negatives

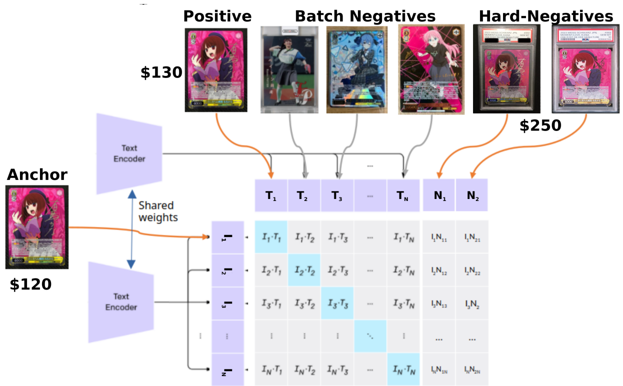

Positive
Batch Negatives
Hard-Negatives
$130
$250
Anchor
$120
Text Encoder
Shared weights
Text Encoder
T₁
T₂
T₃
...
T_N
N₁
N₂
I₁
I₂
I₃
I_N
I₁·T₁ I₁·T₂ I₁·T₃ ... I₁·T_N I₁N₁₁ I₁N₂₁
I₂·T₁ I₂·T₂ I₂·T₃ ... I₂·T_N I₂N₁₂ I₂N₂₂
I₃·T₁ I₃·T₂ I₃·T₃ ... I₃·T_N I₃N₁₃ I₃N₂
I_N·T₁ I_N·T₂ I_N·T₃ ... I_N·T_N I_NN₁N I_NN₂N

Fig. 1. Overview of the Training Process with Hard Negatives: We train a Siamese Neural Network to boost similarity among co-clicked pairs and reduce similarity between different listings. Using Multiple Negatives Ranking Loss with hard negatives refines the embeddings. The displayed prices show that listings with similar appearances can have drastically different prices.

1 Introduction

Providing explainable and accurate selling prices for listings, particularly in the online marketplace, is a complex task, largely due to the absence of universally recognized identifiers like Global Trade Item Numbers (GTIN) or Universal Product Codes (UPC). Moreover, there is no standardized process for sellers when posting listings online, leading to inconsistencies in how listings are described. Factors like item titles, attributes, and features can vary greatly between listings, or in some cases, may be completely omitted. These inconsistencies make it particularly challenging to identify similar listings based on a given seed item, complicating price estimation and comparison efforts. In this study, we focus on one of the most popular categories on the eBay marketplace: "Trading Cards." This category is known for its significant price volatility, making it particularly intriguing and valuable for analysis and improvement.

Previous research in price recommendation has primarily relied on manually extracted features, a process that is both time-consuming and requires domain expertise [23]. Some studies has explored direct recommendation of prices using textual and image features [15], but these approaches often suffer from limited coverage and a lack of explainability.

Transformer architectures have been widely employed in e-commerce for item recommendation systems [18, 28, 27] and have recently been adapted to item retrieval tasks [31, 8, 12]. However, most existing approaches focus on aligning embeddings across different modalities for the same product, rather than learning embeddings for similar products through image-to-text, text-to-image, or query-to-product tasks.

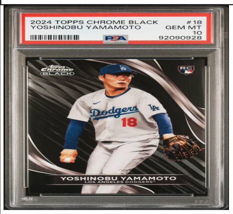
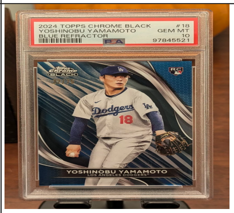
Type	Co-clicked		Hard-Negative
Image			
Title	2024 Topps Chrome Black - Yoshinobu Yamamoto #18 - GEM MINT 10	Graded 2024 Topps Chrome Black Yoshinobu Yamamoto #18 RC Baseball Card PSA 10	2024 Topps Chrome Black Yoshinobu Yamamoto RC **Blue Refractor** /75 PSA 10 Dodgers
Price	$134.59	$160.00	$299.99

Table 1. Hard Negative Mining Example: Although the listing titles differ, they represent the same player, team, year, and grading. In contrast, the hard-negative example is nearly identical except that it's a "Blue Refractor" (blue background instead of dark gray), a detail that significantly impacts its value compared to the co-clicked pair.

In this paper, we propose an enhanced approach to price recommendation using nearest neighbor retrieval by integrating hard negatives into the training process of existing models. Incorporating hard negatives improves the quality of embeddings generated during training [10], enabling more effective differentiation between similar listings with significantly different prices. This advancement has the potential to substaintially improve accuracy and reliability in online pricing strategies. Our approach addresses the pricing challenge through a streamlined, single-network training process, reducing development complexity and eliminating the need for manual feature extraction.

To the best of our knowledge, the combination of textual inputs with hard negatives that are jointly based on both price and listings' aspects has not been previously explored in the field of recommender systems. Thus, our research introduces a novel approach with the potential to significantly improve the accuracy and effectiveness of price reecommendation methods. By employing this unique strategy, we aim to uncover new insights and develop advanced methodologies in the domain of price recommendation and recommendation systems.

Moreover, we demonstrate how this method has been successfully deployed in a production environment, where it has delivered superior performance and proven effective in real-world applications.

2 Related Work

Assisting sellers in pricing their products is valuable across various marketplace domains, including e-commerce, tourism, and more. Traditional approaches, such as those used in the Kaggle Mercari Price Suggestion Challenge [1], relied heavily on extensive feature engineering and regression models for price prediction. By leveraging the BERT encoder, which effectively captures rich contextual and semantic information directly from textual data, the need for extensive feature engineering is eliminated, allowing listing title embeddings to be used instead.

Recent solution relate to E-commerce pricing include embeddings of different modalities, including textual, visual and structured data. The vector representation is then used as input for a trained model. For example, [15] built a price recommendation approach based on listings images and text descriptions. In [9] CNN and LSTM are combined and receive as input both text and visual features.

Although multi-modal data is a common approach, our experiments show that product titles alone provide the most accurate information for pricing, with images offering no performance improvement and, in some cases, even degrading model performance. This result supports an earlier finding by [21] which combined images and text features, and reached a similar conclusion.

Our work builds upon the study by [11], which identified a trade-off between semantic similarity and price accuracy. This trade-off was partially managed by a multi-task network that trained a BERT-based Siamese dual encoder in parallel with a BERT model incorporating a regression layer (Title2Price). The Title2Price model directly learns the pricing of listings, while the Siamese model does not have access to any price-related information. Our work focuses on improving the Siamese encoder by incorporating pricing information through data processing and hard-negative mining. Our approach significantly enhances the Siamese model's pricing accuracy while maintaining its semantic similarity capabilities. As a result, our approach enables the pricing challenge to be addressed with a single network training, simplifying both the development process and its subsequent use in production settings.

3 Methods

In this section, we present our proposed pipeline for product price guidance, which consists of four key stages: training product embeddings, using these embeddings to index previously purchased products in a KNN index, retrieving the k nearest neighbors of new products, and finally using the retrieved products for price prediction. We begin by outlining our data preparation approach followed by contrastive learning training, and a discussion on the hard-negative mining process. Finally, we describe our retrieval strategy and the implementation of the pricing mechanism.

3.1 Data Preparation

In recent years, contrastive learning has become a common approach for training large, generalized models applicable to many downstream tasks [19, 4, 24]. This method maps similar examples close together in the embedding space while learning to separate the dissimilar ones. With efficient training, it naturally organizes data into well-separated clusters, enabling simpler and more efficient applications in various downstream tasks.

Recently, similar approaches have been proposed for language models [26, 13]. Building upon previous work [25], our approach involves training Siamese Neural Networks [2, 20] based on the BERT architecture [6].

To construct our training dataset, we first filter search queries to remove those with fewer than six words, ensuring co-clicked listings are similar (e.g., a query like "sports cards" might yield dissimilar listings). A pair is considered positive if the same user clicks both listings under the same query. These positive pairs are then aggregated and further filtered based on aspects, price ratio (i.e., ensuring the price difference stays within a threshold), entity extraction, and other criteria to minimize false positives.

Since attributes are user-defined and often incomplete, many listings lack full attribute details. However, sellers usually include key information (such as player name, team, year, rarity, and grading) in the title. Therefore, we focus on generating rich embedding representations solely for the listings' titles.

3.2 Model Architecture

Given mini-batch of n positive pairs:

$$B = \{(a_i, b_i)\}_{i=1}^{n} \tag{1}$$

Let a and b be a pair of positive listings. We feed each into a BERT model and extract the final hidden layer:

$$E_{a_i} = BERT(a_i), \quad \text{with} \quad E_{a_i} = \langle e_1, e_2, \ldots, e_n \rangle.$$

Let e_i (and similarly b_i) denote the embedding for the i-th token; we then pool these to form a single title embedding.

$$u_i = Pooler(E_{a_i});\ v_i = Pooler(E_{b_i}) \tag{2}$$

Since our dataset contains only positive pairs, we generate pseudo-negatives by considering mismatched pairs within each batch as negatives. This enables the use of the Multiple Negative Ranking Loss [3, 16] during training. Specifically, we compute cosine similarity between all pairs (a_i, b_j) for $i, j = 1, \ldots, n$:

$$s_{i,j} = \frac{\langle u_i, v_j \rangle}{\|u_i\|_2 \|v_j\|_2}$$

This yields an $n \times n$ similarity matrix. The loss function is defined as:

$$\mathcal{L}_{Siamese} = -\sum_{i=1}^{n} \log \frac{\exp(s_{i,i})}{\sum_{j=1}^{n} \exp(s_{i,j})}$$

3.3 Hard-Negatives Mining

Contrastive learning depends on in-batch negatives, which can be limiting when the negatives are too dissimilar. For example, if our goal is to learn detailed animal embeddings, a batch with two dog images and several car images won't challenge the model. With only obvious negatives like cars, the model may fail to capture the subtle differences between similar animal images.

To address this issue, we avoid using only in-batch negatives and enrich the data with "hard negatives" [14].

With this modification, our training batches are now structured as follows:

$$B = \{(a_i, b_i, h_{i_1}, \ldots, h_{i_m})\}_{i=1}^{n} \tag{3}$$

Now when constructing our similarity matrix we end up with a matrix that its shape is $n * (n + m)$ where m is the number of hard-negatives added to each example in our training data.

$$\begin{array}{cc} & \begin{pmatrix} b_1 & b_2 & b_3 & h_1 & h_2 \end{pmatrix} \\ \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} & \begin{pmatrix} s_{11} & s_{12} & s_{13} & s_{1h_1} & s_{1h_2} \\ s_{21} & s_{22} & s_{23} & s_{2h_1} & s_{2h_2} \end{pmatrix} \end{array}$$

Given that our input data consists of listings' titles, relying on hard negatives based solely on this data is not optimal. Traditional hard-negative mining processes usually focus on finding examples that are "close enough" in similarity to the positive pairs [14, 29]. However, similar titles do not always correspond to significant price differences. For example, a "Red iPhone" and a "Green iPhone" may differ in color but typically share the same price. If we base our hard-negative mining on the titles alone, we risk introducing many false positives into our model. This approach is inefficient in our e-commerce pricing context.

Instead, we opt for a different mining approach that is more suitable for price recommendation. During the dataset generation, in addition to the positive pairs we mine "negative pairs" by applying two relaxation rules to our preprocessing pipeline. First, we remove some of the aspect constraints; instead of requiring that pairs match in all aspects such as team, player, year, grader, graded status, variety, etc., we allow mismatches in some attributes.

We remove predicates that could yield similar listings with significant price differences; for example, the grade of a card—a measure of its condition by a professional grading service. Two cards may be identical except that one is graded PSA 10 (mint condition) and the other PSA 6 (fair condition), leading to price differences of hundreds or thousands of dollars.

Secondly, we invert the price ratio requirement. Instead of requiring the price ratio between two cards to be below a threshold for positive pairs, we set a threshold that requires it to be above a specified value. These augmentations help capture listings that are nearly identical except for a minor detail making one card worth $200 and another $1,000.

After generating the negative pairs dataset, we have two datasets—one for positive and one for negative pairs. Given their largely overlapping preprocessing,

Algorithm 1: Hard-negatives mining

```
Data: $P$- positive-pairs dataset, $N$- negative-pairs dataset, $K$- hard negative size
Result: $D$- Final dataset with k hard negatives per example
$J \leftarrow pd.DataFrame()$;
for $(a,\ b)\ \ in\ \ [(a_i, a_i), (a_i, b_i), (b_i, a_i), (b_i, b_i)]$ do
    $J_i \leftarrow pd.merge([P,\ N],\ how = inner,\ left\_on = a,\ right\_on = b)$;
    $J \leftarrow pd.concat([J,\ J_i],\ axis = 0)$;
end
$J \leftarrow J.groupBy(by = [query, a, b])$;
$J[b] \leftarrow J.apply(list, J[b])$;
$J \leftarrow J[J[b].len >= K]$ # Drop rows with $\leq$ k negatives
$J[b] \leftarrow random.sample(J[b],\ K)$ # pick k random negatives
$Negatives \leftarrow pd.DataFrame(J[b],\ columns = [Neg_1, \ldots, Neg_k])$;
$D \leftarrow pd.concat([P,\ Negatives],\ axis = 1)$;
return $D$
```

some listings appear in both. As described in Algorithm 1, we first perform an inner join between the datasets to create (anchor, positive, negative) triples, then group all negatives for each positive pair.

Because our training requires a fixed number of hard negatives, we filter out rows with fewer than K negatives, randomly select K negatives per positive pair, and combine them into a single dataset.

3.4 Retrieval and Pricing

To identify similar listings for pricing, we use a k-nearest neighbors (KNN) approach, searching for listings with cosine similarity that meets a predefined threshold. This threshold was determined by a grid search process - multiple candidate thresholds were evaluated using a train-validation-test split. The grid search involved testing various thresholds on the training set, tuning the threshold based on performance on the validation set, and finally choosing the optimal threshold based on its ability to generalize to the test set.

For efficient nearest neighbor search and retrieval, we utilized the Faiss library [17], a high-performance tool specifically designed for fast similarity search in large datasets. Faiss allows efficient indexing and retrieval of embeddings. The Faiss implementation we used is an optimized algorithms for both exact and approximate nearest neighbor search, depending on the data characteristics and performance requirements. This enable to quickly and accurately fetch nearest neighbors that are most similar to a given listing, making the pricing recommendation process both fast and scalable.

For the final predicted price, several pricing strategies were evaluated, with the best strategy selected through a grid search (similar to the approach used above for determining the optimal similarity threshold). Specifically, we explored various price percentiles, the exponential moving average, and a modified version of k-nearest neighbors (KEN), which allows soft thresholding and a larger recall set per query [11]

4 Experiments and Results

AMP@5	**CLIP-Image**	**CLIP-Text**	**Resnet**	**Title2Price**	**ViT**	**ViLT**	**eBERT 1-HN**
Graded ↓	11.2	14.7	12.1	9.4	**3.9**	6.7	7.2
Sport ↓	3.3	3.3	3.15	12.2	5.5	3.2	**3.1**
Season ↓	31.8	29.4	24.6	35.1	31.2	**21.0**	26.4
Manufacturer ↓	15.5	11.2	12.1	19.2	15.2	**10.6**	12.9
Set ↓	46.8	44.5	**33.2**	45.5	43.0	37.3	36.2
Player Athlete ↓	10.3	6.3	9.6	25.5	20.1	**5.9**	10.6
Team ↓	21.0	16.9	**15.8**	30.2	30	17.5	17.7
Grade ↓	59.0	45.0	27.7	20.6	39.7	12.6	**11.2**
Variation ↓	76.7	50.8	52.1	60.7	55.3	70.1	**49.7**
Professional Grader ↓	22.9	7.3	**4.1**	15.0	8.2	8.7	9.5
Autographed ↓	6.9	6.5	6.2	7.3	**4.8**	5.2	5.3

Table 2. Aspect Mismatch retrieval results. X-HN signifies the number of hard-negatives used during training. ↓ indicates lower is better.

Price Accuracy	**CLIP-Image**	**CLIP-Text**	**Resnet**	**Title2Price**	**ViT**	**ViLT**	**eBERT 1-HN**
P(20) ↑	32.8%	38.4%	47.1%	49.1%	39.2%	42.4%	**53.1**%
MAE ↓	853.8	672.5	468.7	**264.0**	685.9	652.3	325.8
Recall ↑	69.7%	67.6%	71.2%	**71.7%**	68.5%	68.1%	70.5%

Table 3. Retrieval-based pricing accuracy results. X-HN signifies the number of hard-negatives used during training. ↑ indicates higher is bette.

In this section, we present our evaluation. We begin by presenting our datasets, followed by definition of our evaluation metrics. Finally we present our results, discuss their implication and perform an ablation study to further investigate our approach.

4.1 Data

In our experiments we use two types of datasets:

1. Co-clicked dataset: listings pairs that were clicked in the same eBay search session. It consists of 240,000 co-clicked listing pairs.
2. Vault dataset: Simulates eBay's vault listings, which are high-value items stored and secured by eBay. It includes a sample of 20k listings with known sold prices for validation and hyper-parameter tuning, as well as a 10k listing sample for testing over a two-week period.

Both datasets consist of structured key-value pairs representing listing attributes (e.g., "Player Name": "LeBron James"). We use the co-clicked dataset to train our models using the "Multiple Negative Ranking Loss". The Vault dataset was segmented into validation, and test sets using a time-ordered split and is used to evaluate our models on their retrieval and pricing performance. The overlap between the co-clicked data and the Vault test set is negligible ($< 0.001\%$) The Vault dataset, which simulates eBay's high-value, secured listings, is crucial for evaluating our models' retrieval and pricing performance. By using a dataset of real-world, high-value items, we can ensure that our models are robust and accurate in a production setting where precise pricing and efficient retrieval of these expensive cards are paramount. This is especially important for Vault listings, as these items are often rare and unique, making accurate pricing a challenging task. In addition, our actual production use-case required the development of a model capable of accurately pricing these high-value items, which is a complex problem due to the rarity and variability of such products.

4.2 Metrics

Our primary metric for price guidance is P(20), Additionally we also calculate P(20)-Smooth, "Attribute Mismatch Percentage" (AMP), Mean Absolute Error (MAE), and Recall. The seed items used for evaluation are taken from the Vault dataset, as outlined in Section 4.1.

1. **P(k)**- This calculates the percentage of successfully retrieved cards that their listing price is at most a certain percentage away from the seed item. Since card prices have small variations in them (not all sellers sell at the exact same price), this metric captures the ability of the model to retrieve similar priced cards.
   for each query listing $a \in A$, and a set of seed listings S and their prices, P(k) is then defined as:
   $$P(k) = \frac{\mathbb{1}\{|\frac{P_a - P_s}{P_a}| \leq \frac{k}{100} + \delta\}}{|A|} \tag{4}$$
   where P_a and P_s are the prices of a and s respectively, and δ is a smoothing factor when dealing with low-priced listings.
2. **MAE**- The average absolute difference between the seed item's price and the prices of the retrieved listings
   $$\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

3. **AMP@K-** measures the percentage of successfully retrieved cards whose listing attributes match the seed item attributes.
   Since retrieval is based on the listing title embedding, this metrics captures the model's ability to retrieve correct cards.
   For each attribute a, and a set of seed listings S that contain a value for this attribute, we take the k nearest neighbors $NN_k(s)$ for each listing $s \in S$. AMP is then defined as:

$$AMP@K(S,a) = 100 * \frac{\sum_{s\in S}\sum_{r\in NN_k(s)} \mathbb{1}\{a(s)\neq a(r), a(r)\neq \emptyset\}}{\sum_{s\in S}\sum_{r\in NN_k(s)} \mathbb{1}\{a(r)\neq \emptyset\}}$$

4. **Recall-** The percentage of listings that a successful retrieval could be performed.

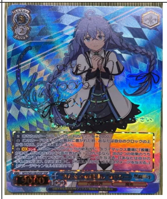
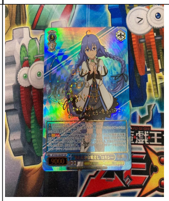
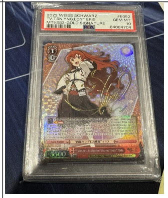
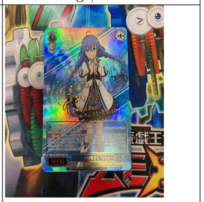
Type	Seed Listing	Co-Clicked	Same Title/Random Image	Same Image/Random Title
Image				
Title	Weiss Schwarz Roxy Migurdia Mushoku Tensei Card FOIL	Weiss Schwarz Mushoku Tensei Roxy Signed Card S83-082SSP	Weiss Schwarz Mushoku Tensei Roxy Signed Card	Weiss Schwarz Violent Tsundere Young Lady Eris MUSHOKU TENSEI
Cosine Similarity		0.98	0.89	0.42

Table 4. Effect of Modalities in Multi-Modal Models: The model correctly identifies similarity between a seed listing and its co-clicked pair. However, when the title is kept but the image is replaced with a random one, similarity remains high, while using the correct image with an incorrect title causes a sharp drop. This indicates that the model primarily relies on text over visual cues.

4.3 Training

Our main approach and model is eBERT [5], a domain-adapted version of BERT that achieves better performance in e-commerce settings. Encoder-based models have shown great success in dominating embedding tasks [22], as they offer great balance between efficiency and performance. The decision to use a BERT-based architecture is also motivated by the high traffic volume on our marketplace, necessitating a system that can efficiently handle large-scale data in real time.

We compared our eBERT with hard negatives to different approaches and architectures across various modalities (image/text) and tasks (contrastive learning, regression, multi-class classification). All models, except ResNet and Title2Price, were trained using the same contrastive learning approach (see Section 3), hyperparameters, and datasets. Title2Price was adapted from [11], and

ResNet was trained using multi-class classification [30]. Each model was trained on 8 GPUs for 120 epochs using the Adam optimizer with a weight decay of 0.01, a linear scheduler with warm-up, and an initial learning rate of $2 \cdot 10^{-5}$.

4.4 Results

As can be seen in Table 2 and Table 3, Title2Price achieved the lowest MAE due to its regression-based training that directly predict the prices from titles. However, Title2Price achieved poor performance on aspect matching compared to eBERT. CLIP-text also achieves subpar results because it isn't trained on titles and can't capture the hidden relationships that exist between the tokens in titles. Notably, Our eBERT approach with hard-negatives outperformed all other models on the P(20) metric at the same recall range, suggesting that supplementing the training process with price-based hard negatives allows the model to identify price-sensitive words in titles and leads to more accurate price predictions.

When comparing the performance of image-based models (such as CLIP-Image, ResNet, and ViT) to text-based models (like BERT and Title2Price), we observe a decline in performance for the image-based models as seen previously [7]. This degradation is likely due to the weaker signal from images, which obstructs the training process compared to using the listings' titles. When a new listing is created on eBay, sellers typically include comprehensive information to stand out and help buyers make informed decisions. This information often encompasses details such as player, team, year, rarity, grade, and more. Essentially, a listing title is a collection various aspects and attributes, and the inclusion or exclusion of specific terms can be easily distinguished, aiding in understanding their impact on pricing—for example, "graded" versus "ungraded" or "PSA 10" versus "PSA 4". Conversely, extracting this information from images is more challenging. Image models lack the domain knowledge necessary to grasp certain subtleties solely from visual data, such as the year a card was printed and its effect on price, or the relationship between the player and the team. These limitations make retrieval-based pricing a more difficult task for image-based models compared to their text-based models. In Appendix A, we explore a setting that image-based models are on par with text-based models.

In our experiments, the multi-modal vision-text model ViLT outperforms purely image-based models. This is primarily because ViLT ignores the image when embedding the listing and relies mainly on the signal that the titles provide. Table 4 shows a representative example: when ViLT is fed the images and titles of co-clicked listings, it recognizes them as similar (cosine similarity = 0.98). In contrast, replacing the title in one listing with a random title causes the cosine similarity to drop significantly (0.42), whereas replacing an image with a random image only modestly reduces the similarity (0.89). These results, consistently observed across multiple examples, indicate that ViLT's superior performance over image-based models like CLIP or ViT is largely driven by its reliance on textual information.

4.5 Hard-Negatives Ablation

We conducted an ablation study to evaluate the contribution of the hard-negatives to our model's performance. As shown in the results presented in Table 5, we observed that incorporating a single hard-negative into our training scheme enhances pricing performance.

When increasing the number of hard-negatives during training we experience a decrease in our metrics performance. This effect is similar to those reported in [10] (We refer the readers to appendix B in [10]) due to decreased batch size needed to train with more hard-negatives. Furthermore, we also observe that aspects not used in creating the negative set (See Subsection 3.3), such as, "Parallel Variety", "Professional Grader", and "Autographed", have better aspect matching. This aligns with the idea that hard-negative training allows for better separation between related examples that result in notable price difference.

Since k hard-negatives adds a $k \times n \times d_k$ additional operations between embedding vectors of size $u \in d_k$, and due to the fact that most system have limited amount of memory, the only option is to reduce the batch size, which negates the benefits of having very large batch sizes when applying contrastive learning methods. This trade-off might explain the decrease in performance when increasing the number of hard-negatives from one to four.

AMP@5	**eBERT**	**eBERT 1-HN**	**eBERT 4-HN**
Graded ↓	**5.4**	7.2	7.8
Sport ↓	3.1	3.1	**3.0**
Set ↓	37.5	**36.2**	36.6
Player Athlete ↓	**9.6**	10.6	10.0
Professional Grader ↓	**10.2**	11.2	11.3
Variation ↓	53.5	49.7	**47.9**
Autographed ↓	5.7	5.3	**5.2**
Price Accuracy	**eBERT**	**eBERT 1-HN**	**eBERT 4-HN**
P(20) ↑	51.7%	**53.1**%	52.6%
MAE ↓	404.8	**325.8**	358
Recall ↑	**71.7%**	70.5%	71.4%

Table 5. Effect of hard-negatives on price guidance in text models. X-HN signifies the number of hard-negatives used during training. ↑ indicates higher is better, while a ↓ indicates lower is better.

4.6 Production Deployment

One of the services eBay provides to its users is the 'eBay Price Guide', specifically developed for collectible and sports trading cards enthusiast. We applied the hard-negative mining approach to develop production-ready models for both sports and collectible cards. During inference, the listing title is sent to a service node, which generates an embedding and performs an index lookup to retrieve a shortlist of similar listings. These candidates are then used to calculate the recommended price based on a defined similarity threshold. As shown in Table6, a comparison of our models with the existing production models (which rely on keyword search) demonstrates improvement in both price accuracy and recall. Based on these results, our models were deployed to production in 2024, replacing the previous models. Given this improvement, the models are expected to be also deployed in 2025 to assist sellers during the listing process.

Sport Cards	**P(20)**	**Recall**
Production	55%	54.2%
Our Model	**64%**	**62%**
Collectible Cards	**P(20)**	**Recall**
Production	62.7%	48%
Our Model	**63.8%**	**73%**

Table 6. Comparing our nard-negative based approach to eBay Price Guide production models.

5 Discussion and Conclusion

In this paper, we propose a retrieval-based pricing method for trading cards that enhances training using hard-negative mining. Unlike prior approaches that rely solely on input data (e.g., comparing similar texts or applying image augmentations), our method leverages the idea that small item variations can lead to significant price differences. We generate hard negatives based on price and aspect differences, resulting in pairs nearly identical to the original co-clicked pair except for one trait that causes large price variations. While we currently use domain knowledge to identify these aspects, this process can be automated by analyzing their values and correlations with prices.

Allowing for better pricing of trading cards, improve the experience of both shoppers and sellers, as accurate prices would allow seller to know what prices to set in order to guarantee selling their trading card, and for shoppers knowing the current price trends for their desired purchases will help making better purchases.

We evaluated our approach against text-based (CLIP-text), regression (Title2Price), and image-based models (CLIP-image, ViT, Resnet) using a common test set and multiple pricing and aspect metrics. Image-based models underperformed compared to title-based ones, emphasizing that explicit details in listing titles offer a stronger signal for pricing than images, which often provide less clear information. This underscores the challenge of relying on images for price prediction in e-commerce.

Furthermore, we determined that adding additional hard-negatives to our training, without changing the batch size improve pricing accuracy, and partial aspect accuracy compared to training without them.

Future work can explore different strategies or settings in which models can benefit from incorporating images into their training process.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, A.A.S., Seker, H., Farnie, S., Elliott, J.: Extensive data exploration for automatic price suggestion using item description: Case study for the kaggle mercari challenge. In: Proceedings of the 2nd International Conference on Advances in Artificial Intelligence, ICAAI 2018, Barcelona, Spain, October 06-08, 2018. pp. 41–45. ACM (2018). https://doi.org/10.1145/3292448.3292458, https://doi.org/10.1145/3292448.3292458
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. Advances in neural information processing systems **6** (1993)
3. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: Proceedings of the 24th international conference on Machine learning. pp. 129–136 (2007)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org, Vienna (2020)
5. Dahlmann, L., Lancewicki, T.: Deploying a bert-based query-title relevance classifier in a production system: a view from the trenches. arXiv preprint arXiv:2108.10197 (2021)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Di, W., Sundaresan, N., Piramuthu, R., Bhardwaj, A.: Is a picture really worth a thousand words? -on the role of images in e-commerce. In: Proceedings of the 7th ACM international conference on Web search and data mining. pp. 633–642 (2014)
8. Dong, X., Zhan, X., Wu, Y., Wei, Y., Wei, X., Lu, M., Liang, X.: M5product: A multi-modal pretraining benchmark for e-commercial product downstream tasks. arXiv preprint arXiv:2109.04275 **4** (2021)
9. Fathalla, A.E., Salah, A., Li, K., Li, K., Francesco, P.: Deep end-to-end learning for price prediction of second-hand items. Knowledge and Information Systems pp. 1 – 28 (2020)
10. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852 (2020)
11. Fuchs, G., Petrov, P., Ben-Shaul, I., Mandelbrod, M., Zinman, O., Basin, D., Arshavsky, V.: Pricing the nearly known-when semantic similarity is just not enough. In: eCom@ SIGIR (2023)

12. Gao, D., Jin, L., Chen, B., Qiu, M., Li, P., Wei, Y., Hu, Y., Wang, H.: Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2251–2260 (2020)
13. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021)
14. Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Abrego, G.H., Stevens, K., Constant, N., Sung, Y.H., Strope, B., et al.: Effective parallel corpus mining using bilingual sentence embeddings. arXiv preprint arXiv:1807.11906 (2018)
15. Han, L., Yin, Z., Xia, Z., Tang, M., Jin, R.: Price suggestion for online second-hand items with texts and images (2020)
16. Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.H., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R.: Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652 (2017)
17. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. arXiv preprint arXiv:1702.08734 (2017)
18. Kang, W.C., McAuley, J.: Self-attentive sequential recommendation. In: 2018 IEEE international conference on data mining (ICDM). pp. 197–206. IEEE (2018)
19. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
20. Koch, G., Zemel, R., Salakhutdinov, R., et al.: Siamese neural networks for one-shot image recognition. In: ICML deep learning workshop. vol. 2, pp. 1–30. Lille (2015)
21. Li, B., Liu, T.: An analysis of multi-modal deep learning for art price appraisal. 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom) pp. 1509–1513 (2021)
22. Muennighoff, N., Tazi, N., Magne, L., Reimers, N.: Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 (2022). https://doi.org/10.48550/ARXIV.2210.07316, https://arxiv.org/abs/2210.07316
23. Pal, N., Arora, P., Sundararaman, D., Kohli, P., Palakurthy, S.S.: How much is my car worth? a methodology for predicting used cars prices using random forest (2017)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
25. Reimers, N.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
26. Shen, D., Zheng, M., Shen, Y., Qu, Y., Chen, W.: A simple but tough-to-beat data augmentation approach for natural language understanding and generation. arXiv preprint arXiv:2009.13818 (2020)
27. Si, Z., Sun, Z., Zhang, X., Xu, J., Zang, X., Song, Y., Gai, K., Wen, J.R.: When search meets recommendation: Learning disentangled search representation for recommendation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 1313–1323. SIGIR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3539618.3591786, https://doi.org/10.1145/3539618.3591786

28. Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., Jiang, P.: Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In: Proceedings of the 28th ACM international conference on information and knowledge management. pp. 1441–1450 (2019)
29. Tan, W., Heffernan, K., Schwenk, H., Koehn, P.: Multilingual representation distillation with contrastive learning. arXiv preprint arXiv:2210.05033 (2022)
30. Yang, F., Kale, A., Bubnov, Y., Stein, L., Wang, Q., Kiapour, H., Piramuthu, R.: Visual search at ebay. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 2101–2110 (2017)
31. Yao, S., Tan, J., Chen, X., Zhang, J., Zeng, X., Yang, K.: Reprbert: distilling bert to an efficient representation-based relevance model for e-commerce. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4363–4371 (2022)

A Collection Set Performance

AMP@5	**eBERT**	**ViT**
Graded ↓	7.06	**4.0**
Sport ↓	12.25	**9.4**
Set ↓	**33.6**	40.1
Player Athlete ↓	**11.8**	41.1
Professional Grader ↓	**17.0**	21.7
Variation ↓	**53.3**	71.4
Autographed ↓	10.1	**5.5**
Price Accurracy	**eBERT**	**ViT**
$P(20)$ $\delta = 2$ ↑	60.8%	**63.8%**
MAE ↓	20.9	**18.3**
Recall ↑	57.5%	**57.9%**

Table 7. Performance of ViT model compared to BERT in a low-priced and short title setting. a smoothing factor of $2 was added due to low prices in collection cards.

As observed, text-based models achieve top results in retrieval-based price guidance due to the strong signal provided by detailed titles. In contrast, image-based models have underperformed, even though images can offer useful, complementary information. We aim to investigate when image-based models—specifically ViT—can match text-based models like BERT.

For this experiment, we use the same training settings but switch our evaluation dataset to a "Collection set." Unlike the vault dataset, which features high-priced listings with detailed titles, the Collection set has no lower price bound and includes cards priced as low as $1. We select a seed test set containing titles with five words or fewer, then evaluate BERT and ViT, both trained with the same contrastive learning loss, on this new dataset.

As shown in Table 7, ViT achieves results comparable to BERT when only short titles are used. Sellers usually employ descriptive titles to attract buyers, providing a strong signal for text-based models like eBERT. However, with short titles, eBERT has fewer tokens and less information to extract, while the image signal remains constant. Although listings with short titles represent only about 0.65% of the Collection set, these results suggest that image-based models can be particularly valuable in domains where text is less informative.