The Elements of Visual Art Recommendation

Learning Latent Semantic Representations of Paintings

Bereket A. Yilma
Luis A. Leiva
name.surname@uni.lu
University of Luxembourg
Luxembourg

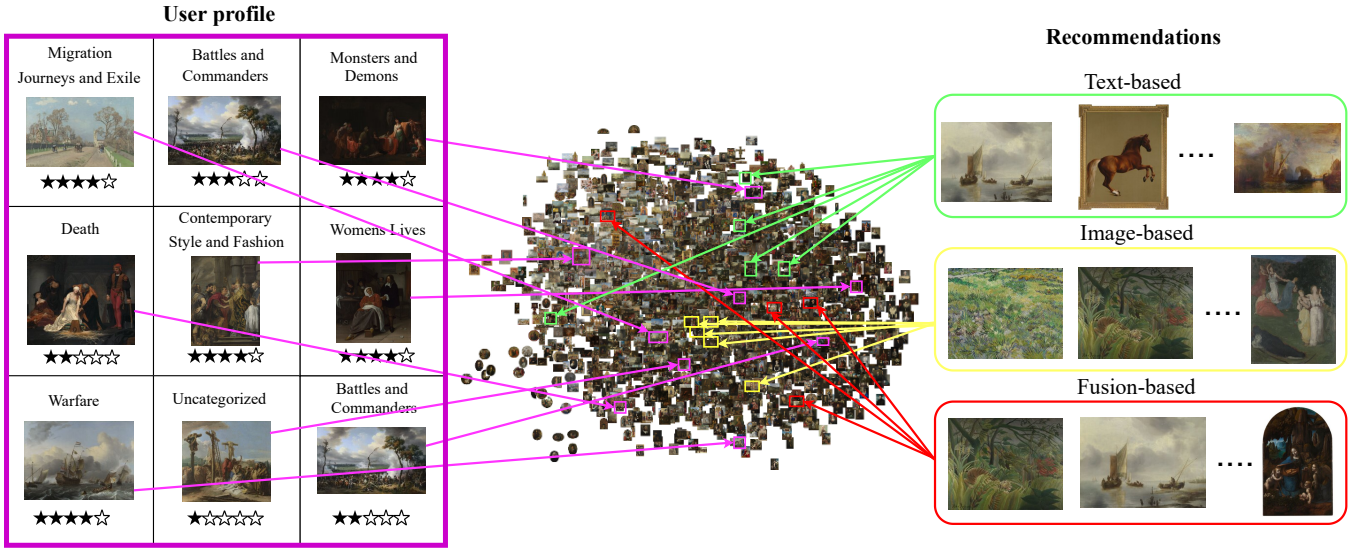


Figure 1: Overview of personalized visual art recommendation. Based on a small set of elicited preferences (left), the user is shown different paintings (right) according to different recommender systems.

ABSTRACT

Artwork recommendation is challenging because it requires understanding how users interact with highly subjective content, the complexity of the concepts embedded within the artwork, and the emotional and cognitive reflections they may trigger in users. In this paper, we focus on efficiently capturing the elements (i.e., latent semantic relationships) of visual art for personalized recommendation. We propose and study recommender systems based on textual and visual feature learning techniques, as well as their combinations. We then perform a small-scale and a large-scale user-centric evaluation of the quality of the recommendations. Our results indicate that textual features compare favourably with visual ones, whereas a fusion of both captures the most suitable hidden semantic relationships for artwork recommendation. Ultimately, this paper contributes to our understanding of how to deliver content that suitably matches the user's interests and how they are perceived.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Recommender systems**; • **Computing methodologies** → *Learning latent representations*; • **Applied computing** → *Media arts.*

KEYWORDS

Recommendation; Personalization; Artwork; User Experience; Machine Learning

ACM Reference Format:
Bereket A. Yilma and Luis A. Leiva. 2023. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581477

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581477

1 INTRODUCTION

In recent years, technology-mediated personalization and content recommendation has been areas of interest in Cultural Heritage environments such as museums, art galleries, and exhibitions [52]. Although, in many cases the primary motivation for designing

personalized services and recommender systems (RecSys) remains tightly linked to extrinsic motivation goals, such as maximizing revenue, increasing user engagement, and optimizing advertisement delivery. This approach to personalization may potentially overlook the very purpose of the cultural institutions as well as the users' quality of experience [68], who typically do it for their own pleasure, i.e., intrinsic motivation goals. Thus, to enhance the perceived utility of RecSys, it is of paramount importance to emphasize visitors' quality of experience. In this context, Visual Art (VA) recommendation is among the areas that has recently gained momentum [73]. Nevertheless, contrary to other application areas of RecSys where personalised content is delivered to users such as movies, music, news, etc., the domain of VA recommendation has not yet been sufficiently explored.

In the VA domain, paintings are important items that bring together complex elements such as drawings, gestures, narration, composition, or abstraction [46]. The task of personalized VA recommendation essentially entails suggesting paintings that are similar to what a user has already seen or previously expressed interest. The subjective nature of user's taste and the unique nature of their preferences, which are long-standing challenges in content personalization, are also salient issues in VA RecSys. Especially since paintings carry deeper semantics than their traditional metadata, i.e., categorizations based on their time period, technique, material, color, size, etc. Furthermore, the kind of emotional and cognitive reflections paintings may trigger in users are also diverse, depending on their background, knowledge, and several other environmental factors [52]. Hence, to enhance personalized VA recommendations, efficiently capturing latent semantic relationships of paintings is vital and yet remains an open research challenge.

Most VA RecSys usually infer similarities and relationships among paintings from high-level features derived from the abovementioned traditional metadata such as artist names, styles, materials, and so on. However, these features may not be expressive enough to capture abstract concepts that are hidden in paintings and that could better adapt the recommendations to the subjective taste of the users. For this, a high-quality representation of the data is crucial [7]. Unfortunately, research on machine-generated data representation techniques for VA RecSys has been often overlooked, as prominent works have largely relied on manually curated metadata [43].

Recent work has started to pay more attention to machine-generated data representations to drive better VA recommendations. He et al. [27] were among the first ones to use latent visual features extracted using Deep Neural Networks (DNN) and also use pretrained DNN models for VA recommendation. A study reported by Messina et al. [50] showed that DNN-based visual features perform better than leveraging textual metadata for VA recommendations. However, they were focused on the artwork market, which is driven by transaction data rather than enhancing the users' quality of experience. Therefore, it is unclear if their findings would transfer to a more **user-centric** setting, which essentially entails investigating the actual relevance of recommendations to users in terms of accuracy, novelty, diversity and serendipity. Furthermore, they did not explore the combination of visual and textual features.

Alternatively, Yilma et al. [73] proposed an approach to learn latent visual and textual features from paintings. Their study indicated that recommendations derived from textual features compare favorably with visual ones. Nonetheless, they also did not test hybrid approaches, therefore it remains unclear which data representation technique (text, image, or a combination of both) is more efficient to best capture the *elements* (i.e., latent abstract concepts) embedded within visual arts for recommendation tasks. A recent work by Liu et al. [42] have shown the benefit of jointly exploiting textual and visual features for recommendations. However, this has not been tested in the domain of VA RecSys. To this end, we set out to explore techniques to learn latent semantic representation of paintings for personalized VA RecSys, including the combination of each individual technique.

Overall, previous works showed that visual features tend to perform better than textual metadata [49, 73] and hence they argued for not considering text-based information in VA RecSys. In addition, it has not been explored yet whether hybrid approaches may yield better performance on VA recommendation tasks. Therefore, we formulate the following research hypotheses:

H1: Visual features result in higher-quality recommendations than textual features.
H2: Fusion of visual and textual features result in higher-quality recommendations than either could individually.

The first hypothesis is aimed at re-assessing our current understanding of the state of the art in VA RecSys research, whereas the second one, to the best of our knowledge, has never been assessed before in the domain of VA RecSys.

In this paper, we propose three different latent feature learning techniques leveraging both textual descriptions and images of paintings. To learn latent features from textual descriptions, we adopt Latent Dirichlet Allocation (LDA) [9] and Bidirectional Encoder Representations from Transformers (BERT) [17], whereas for visual feature learning we use the popular Residual Neural Network (ResNet) [26]. We also adopt a late fusion strategy proposed by Cormack et al. [14] which allows to combine different ranking techniques for information retrieval. We then conduct a small-scale and a large-scale study based on a user-centric evaluation framework [57]. Specifically, we evaluated how accurate, diverse, novel, and serendipitous were the generated recommendations for the users and derive valuable guidelines from our findings. In sum, this paper makes the following contributions:

- We develop and study five VA RecSys engines: LDA, BERT, ResNet, and their combinations.
- We conduct a small-scale ($N = 11$) and a large-scale study ($N = 100$) to assess VA RecSys performance from a user-centric perspective.
- We contextualize our findings and provide guidance about how to design next-generation VA RecSys.

2 RELATED WORK

RecSys are becoming more and more prevalent in Cultural Heritage environments such as museums and art galleries [36]. The huge potential and benefit of personalized recommendations, in particular in the field of visual arts, has been discussed by Esman [19]. In the following we review previous work on VA recommendation and feature learning approaches.

2.1 Recommending paintings

According to Falk et al. [20] the main motivation of museum visitors is to have fun, experience art, learn new things, feel inspired, and interact with others. When using digital museum guides, visitors' expectations are not only to be exposed to artwork that matches their interest but also learn more and have access to more information [29].

Research studies such as the CHIP project [4], which implemented a RecSys for Rijksmuseum,[1] demonstrated the potential of personalization in such environments. Hence, over the years, different kinds of RecSys have been exploited to provide personalized experiences to museum visitors. For example, Aroyo et al. [24] proposed a semantically-driven RecSys and semi-automatic generation of personalized museum visits guided by visitor models. Deladiennee et al. [16] introduced a graph-based semantic RecSys that relies on an ontological formalisation of knowledge about manipulated entities. Similarly, Kuflik et al. [39] highlighted the benefits of graph-based recommendations. This work was based on the premise that parts of the underlying data in a museum context can be represented naturally by a graph that consists of typed entities and relations. On the contrary, Frost et al. [22] introduced an anti-recommendation approach called *"Art I don't like"* which exposes users to a variety of content and suggests artworks that are dissimilar to the ones the users selected, aiming to maximize serendipity and exploration. This method provides content that is aesthetically related in terms of low-level features, but challenges the implied conceptual frameworks, which are driven by the preferences elicited by the users. The very notion of this work was inspired by the work of Pariser [55] which states that removing access to opposing viewpoints can lead to *filter bubbles* in personalization. Pariser's idea describes a type of "intellectual isolation" issue that occurs as a result of personalization algorithms. These algorithms typically offer information to users that match previously viewed content and content viewed by similar users. Hence, users have little exposure to contradicting viewpoints and become unknowingly trapped in a digital bubble. This is a long-standing issue in RecSys and the community has explored different approaches to mitigate it, e.g. improving transparency by giving the user control over the settings of the personalization algorithms [10, 15] and making recommendations understandable to users [21]. However, there are several aspects that remain challenging in VA RecSys. Primarily, because paintings are both high-dimensional and semantically complex, we need a computationally efficient way of modelling both their content and their context. This essentially calls for efficient data representation techniques that are capable of capturing the complex semantics embedded in paintings. Secondly, it also demands a more accurate representation of user profiles such as modelling temporal and social dynamics in terms of users' tendency to interact with content more or less consistently, as well as their preferences towards individual artists, styles, colors, etc. However, these are rarely available or not directly accessible in practice, making the so called cold-start problem[2] a prevalent issue in VA RecSys.

[1] https://www.rijksmuseum.nl/en

[2] When the system has no information about the users, it cannot provide personalised recommendations.

2.2 Learning painting features

He et al. [27] proposed a visually, socially, and temporally-aware model for artistic recommendation. This was among the first works that utilized the power of DNNs to exploit latent representations for VA recommendation. Their work primarily builds upon two methods, factorized personalized Markov chains (FPMC) [62] and visual Bayesian personalized ranking (VBPR) [28]. On the one hand, FPMC was adopted to capture the fact that users tend to browse art with consistent latent attributes during the course of a browsing session, as FPMC models the notion of smoothness between subsequent interactions using a Markov chain. On the other hand, VBPR models the visual appearance of the items being considered. By combining the two models, He et al. tried to capture individual users' preferences towards particular VA styles, as well as the tendency of users to interact with items that are 'visually consistent' during a browsing session. They also proposed several extensions of these models to handle longer memory than simply previous actions. Unfortunately, their method is only applicable under the collaborative filtering scenario, for example matching products to users based on past purchases. However, collaborative filtering suffers from the above-mentioned cold-start problem. In addition, they did not investigate explicit visual features nor textual metadata.

Subsequently, Messina et al. [48–50] explored content-based artwork recommendation using images, keywords, and transaction data from the UGallery online artwork store.[3] Their work suggested that automatically computed visual features perform better than manually-engineered visual features extracted from images (i.e, texture, sharpness, brightness, etc.). Their work also indicated that a hybrid approach combining visual features and textual keyword attributes such as artist, title, style, etc., yields a further performance improvement. However, their hybrid approach was based on computing a score as a convex linear combination of the scores of individual methods (visual similarity and keyword similarity). Particularly, they did not explore feature learning approaches such as topic modeling techniques we study in this paper, which are more scalable and generalizable. Furthermore, their work was focused on predicting future purchases of artwork rather than enhancing personal experiences.

Recent works by Yilma et al. [72, 73] proposed a VA recommendation approach that leveraged topic modeling techniques from textual descriptions of paintings and performed a comparative study against visual features automatically extracted using DNNs. Their study demonstrated the potential of learning features from text-based data, especially when it comes to explaining the recommendations to the user. However, they never looked at the combination of text-based and image-based RecSys engines.

In sum, a number of VA Recsys strategies have been proposed over the years, but given that (i) user preferences are highly subjective and (ii) visual artwork is particularly complex to grasp, VA recommendation remains a rather challenging task. Thus, research effort in uncovering latent semantics of visual art is still considered a worthwhile endeavour, especially with regards to evaluating the quality of the recommendations from a user-centric perspective. To the best of our knowledge, this paper is the first to systematically shed light in this regard.

[3] https://www.ugallery.com/

3 BACKGROUND: LEARNING LATENT REPRESENTATIONS OF PAINTINGS

Data representation techniques play a great role in VA RecSys, as they can entangle and reveal interesting factors embedded within the artwork data, thereby eventually influencing the quality of the recommendations [63]. Specifically, the complexity of the concepts embodied within paintings makes the task of capturing semantics by machines far from trivial. To this end, we set out to study different representation techniques that can efficiently learn the elements (i.e., latent semantic relationships of paintings) of VA RecSys. Figure 2 summarizes the three painting representation learning approaches we propose and study in this paper.

3.1 Feature learning from Text-based representations of Paintings

In Natural Language Processing and Information Retrieval, vector space models have been used to represent documents efficiently [7]. However, this kind of representations has a limited ability to capture inter/intra-document relationships. It has been shown that, as data dimensionality increases, the distance to the nearest data point approaches the distance to the furthest data point [1]. Consequently, in high dimensional spaces the notion of spatial locality becomes ill-defined [8]. Hence, researchers have been proposing more advanced techniques aiming to tackle the curse of dimensionality reduction and to better capture hidden semantic structures in document modeling. Among these efforts, Latent Dirichlet Allocation (LDA), an unsupervised generative probabilistic model proposed by Blei et al. [9], has demonstrated superiority over several other models.

LDA has been applied in several text-based RecSys tasks such as scientific paper recommendation [2], personalized hashtag recommendation [74], and online course recommendation [3], among others. On the other hand, a more recent work by Devlin et al. [17] developed Bidirectional Encoder Representations from Transformers (BERT) and set a new state-of-the-art performance on sentence-pair related tasks like semantic textual similarity and question answering. However, BERT entails an important computational overhead due to the many possible combinations for prediction. For example, to find the most similar pairs in a collection of 10,000 sentences, BERT requires about 50 million inference computations. Sentence-BERT (SBERT) [61], a modification of the pre-trained BERT model, managed to reduce the computation time from 65 hours to 5 seconds in a single V-100 GPU. It uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using the tried-and-true cosine similarity measure. In the past few years, BERT models have proven to be powerful document embedding techniques and have been used for extracting latent semantic structures (i.e., topics) underlying a large set of documents [23]. As a result, BERT has gained tremendous popularity in the design of RecSys that exploit textual data [25, 33, 40].

We adopt LDA and BERT (actually SBERT) to learn painting feature representations from their associated textual metadata, where each painting is represented by a document containing detailed annotations such as title, format, or a curated description; see Figure 3 for an example. In essence, a painting can be described as a mixture of several concepts such as religion, nudity, portrait, etc. Thus, each document is a distribution of topics and each topic is a distribution of words. Prominent words in each latent topic explain the nature of the topic and prominent latent topics related to each document explain the nature of the document (i.e. paintings). For example, let us assume that latent topics are "religion", "still life", and "landscape". A painting may have the following distribution over the topics: 70% "religion", 10% "still life", and 20% "landscape". Moreover, each topic has a distribution over the words in the vocabulary. For the "religion" topic, the probability of the word "Saint" would be much higher than in the "landscape" topic. Hence, the employed LDA and BERT representation techniques will find high-dimensional vector representations that capture the topic proportions for each painting in such a way that semantically similar paintings are closer to each other in the feature representation space. In the following we briefly discuss each text-based feature learning approaches in more detail.

3.1.1 Latent Dirichlet allocation (LDA). LDA is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories. Each observation is a document, the features are the presence (or occurrence, or count) of words, and the categories are the topics. The topics themselves are not specified up-front, only their number, since they are learned as a probability distribution over the words that occur in each document. The procedure of building an LDA model for VA RecSys is described as follows. We start by constructing a collection of documents containing textual information about each painting. Then, a desired number of topics k is chosen and a topic is attributed to each word w in the collection of documents where $\theta_i \sim Dir(\alpha)$; θ is the topic distribution for a document d and α is the per-document topic distribution, with $i \in \{1, ..., k\}$ and $Dir(\alpha)$ is a Dirichlet distribution over the k topics. Subsequently, the learning is done by computing the conditional probabilities $P(t|d)$ (i.e., the likelihood of topic t given document d) and $P(w|t)$ (i.e., likelihood of word w given topic t). A detailed discussion on LDA topic modeling can be found in [9] and [32].

Once the LDA model is trained over the entire text dataset, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is produced where each entry $a(i, j)$ is the cosine similarity measure between document embeddings. This similarity matrix therefore captures the latent topic distribution over all documents, which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks, as explained in the next section.

3.1.2 Bidirectional Encoder Representations from Transformers (BERT). The second approach we study to learn latent feature representations of paintings from their associated textual metadata has been recently proposed by Grootendorst et al. [23] and is based on sentence Transformers. Similar to the LDA approach, we begin by constructing a collection of documents with textual metadata of paintings. Then, feature learning is done in three steps. First, each painting document is converted to an embedding representation using the pre-trained SBERT large language model,[4] which maps sentences and paragraphs to a 384-dimensional dense vector

[4]We used the "all-MiniLM-L6-v2" version, to optimize performance, but any other version can provide suitable painting embeddings.

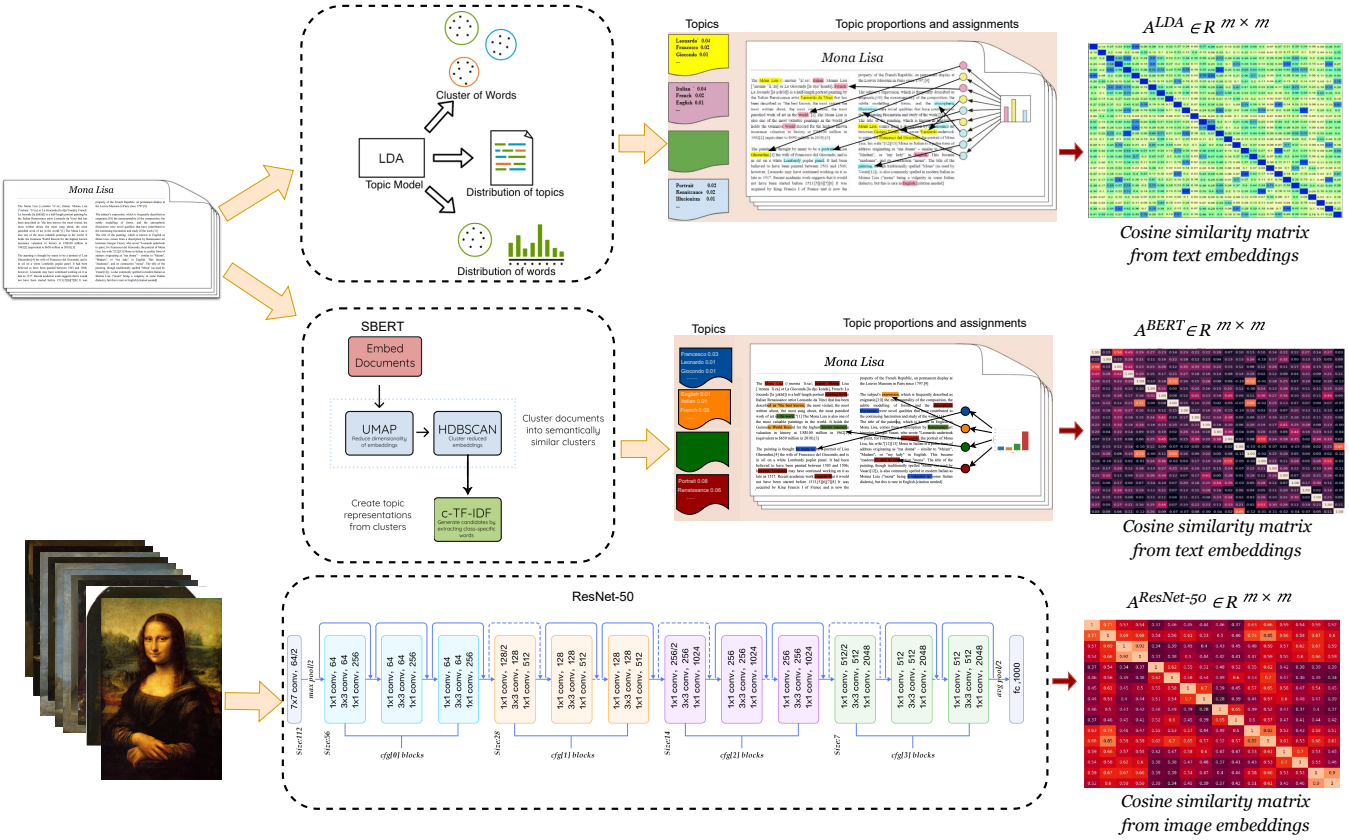


Figure 2: The elements of VA recommendation: Overview of our approaches to learn latent semantic representations of paintings.

space [61]. Second, the dimensionality of the embeddings is reduced using the uniform manifold approximation and projection (UMAP) algorithm [47]. This allows to learn a more efficient representation while at the same time preserving the global structure of the original embeddings. Third, the reduced embeddings are semantically clustered together using HDBSCAN [12], a soft-clustering algorithm that prevents unrelated documents to be assigned to any cluster. Finally, latent topic representations are extracted from the clusters using a custom class-based term frequency–inverse document frequency (c-TF-IDF) algorithm, which produces importance scores for words within a topic cluster. The main idea of c-TF-IDF is that extracting the most important words per cluster yields descriptions of topics. Hence, TF-IDF is adjusted and the inverse document frequency is replaced by the inverse class frequency to measure how much information a term provides to a class. Formally the c-TF-IDF of a word w in class C is given by:

$$\text{c-TF-IDF}(w, C) = f_{w,C} \cdot \log(1 + \frac{N}{f_w}) \tag{1}$$

where $f_{w,C} = \frac{|w|}{\sum_{c \in C} |c|}$ is the frequency of word w in class C, N is total number of words per class, f_w is the frequency of word w across all classes, and $|\cdot|$ denotes the number of items in a set. Words with high c-TF-IDF scores are selected for each topic t, thereby producing topic-word distributions for each cluster of documents d.

Once the BERT model is trained over the entire dataset, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is produced where each entry is the cosine similarity measure between all document embeddings. Again, this similarity matrix captures the latent topic distribution over all documents, which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks, as explained in the next section.

3.2 Feature learning from image-based representations of paintings

Visual feature extraction is critical to have a discriminative representation of images [45], and it is widely used in several tasks such as object detection, classification, or segmentation [64]. Traditional approaches to feature extraction include Harris Corner Detection [13], or the more advanced version Shi-Tomasi Corner Detector [5]. Other approaches have been proposed, such as SURF [44] or BRIEF [11], but they have been superseded by recent advances in Deep Learning, in particular in Convolutional Neural Networks (CNN).

Today, image feature extraction techniques are mostly based on pre-trained CNN architectures such as AlexNet [37], GoogLeNet [67], and VGG [66].The winner of the 2015 ImageNet

challenge, ResNet, proposed by He et al. [26] introduced the use of residual layers to train very deep CNNs, setting a world record of more than 100 layers. ResNet-50 is the 50-layer version of this architecture, trained on more than a million images from the ImageNet database.[5] Thus, it has learned rich feature representations for a wide range of images and has shown superiority over other pre-trained models as a feature extractor [6, 31, 41].

We used the ResNet-50 model pre-trained on ImageNet to extract latent visual features (image embeddings) from paintings. By passing each painting image through the network, a convolutional feature map (i.e., a feature vector representation) is obtained. Once we extract all image features from the entire dataset, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is produced where each entry is the cosine similarity measure between all image embeddings. This similarity matrix therefore captures the latent visual distribution over all images, which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks, as explained in the next section.

4 METHOD: PERSONALIZED RECOMMENDATION OF PAINTINGS

We consider approaches that, together, can learn features from both textual and visual information from paintings. We study two different techniques for learning text-based representations (LDA and BERT), as there are no exhaustive prior works of VA RecSys leveraging textual data. On the other hand, since visual features have been extensively explored in VA RecSys applications [27, 35, 48, 73], we study ResNet-50 for learning image-based representations, which it is considered the state of the art in prior work [35, 73]. Let $P = \{p_1, p_2, \ldots, p_m\}$ be a set of image paintings, $\mathcal{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_m\}$ be the associated embeddings of each painting according to LDA, BERT, or ResNet, and $P^u = \{p_1^u, p_2^u, \ldots, p_n^u\}$ be the set of paintings a user u has rated, where $P^u \subset P$ and $\omega^u = \{\omega_1^u, \omega_2^u, \ldots, \omega_n^u\}$ are the normalized ratings that u gave to a small set of paintings P^u.

Once the dataset embeddings (latent feature vectors) are learned using either model (LDA, BERT, or ResNet) we compute the similarity matrix for all the paintings $\mathbf{A}$. Next, the preferences of a user u are modelled by a normalized vector that transforms a simple 5-point scale rating into weights $\omega_i^u \in [0, 1]$ for every painting p_i^u the user has rated. Then, the predicted score $S^u(p_i)$ the user would give to each painting in the collection P is calculated based on the weighted average distance between the rated paintings and all other paintings:

$$S^u(p_i) = \frac{1}{n} \sum_{j=1}^{n} \omega_j^u \cdot \mathbf{A}_{ij} \tag{2}$$

where $\mathbf{A}_{ij} = d(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is the similarity between embeddings of paintings p_i and p_j in the computed similarity matrix. The summation in Equation 2 is taken over all user's rated paintings $n = |P^u|$. Once the scoring procedure is complete, the paintings are sorted and the r most similar paintings constitute a ranked recommendation list. In sum, the VA RecSys task consists of recommending the most similar paintings to a user based on a small set of paintings rated before, i.e., the elicited preferences.

[5] http://www.image-net.org

In this paper, we study five RecSys engines: three are based on LDA, BERT, and ResNet (non-fusion engines), whereas the other two engines (fusion engines) are hybrid combinations (text+image) of the first three engines. For the fusion engines, we adopted the "reciprocal rank fusion" strategy proposed by Cormack et al. [14] for combining rankings in information retrieval systems. It is a late fusion technique that is easily composable and simple to use. Late fusion (i.e. at post-hoc) is often preferred than early fusion (i.e. at the feature level) because the models involved are independent from each other, so each can use their own features, numbers of dimensions, etc. [60].

Furthermore, with late fusion it is possible to precisely control the contribution of each model (e.g. 25% text and 75% image). Although in our work both text and image features contribute equally (i.e., 50% each) when it comes to producing the recommendations. Our proposed VA RecSys engines are outlined in Algorithms 1 to 3, respectively.

5 DATASET

We used a dataset containing 2,368 paintings from The National Gallery, London.[6] This curated set of paintings belongs to the Cross-Cult Knowledge Base.[7] Each painting image is accompanied by a set of text-based metadata, which makes this dataset suitable for testing the proposed feature learning approaches. A sample data point is shown in Figure 3. For our text-based RecSys engines (LDA and BERT) we use all available painting attributes, such as artist name, painting title, technique used, etc. as well as a description provided by museum curators. These descriptions carry complementary information about the paintings such as stories and narratives that can be exploited to better capture the painting semantics. The image-based RecSys engine (ResNet) uses the convolutional feature maps automatically extracted from the painting images.

The dataset also provides curated stories that we study to sample initial user preferences in the profiling phase. In the following subsections we present a detailed analysis of the dataset to better understand the behavior and implementation of our RecSys engines.

5.1 Story groups

The dataset provides 8 curated stories (categories) linked to a few of the paintings, namely: *'Women's lives', 'Contemporary style and fashion', 'Water, Monsters and Demons', 'Migration: Journeys and exile', 'Death', 'Battles and Commanders', and 'Warfare'*. Figure 4 shows a 2D projection map of the story groups in the dataset using the non-linear projection t-SNE algorithm [69]. We can see that the majority of the paintings belong to the 'uncategorized' class. These story groups are meant to provide context to a selected group of paintings, according to the museum experts who created the dataset. We can observe from the latent space projections that the story groups are scattered across the entire dataset, suggesting that museum curators considered them to be representative examples of the collection. The map projection also surfaces the complex latent semantic relationships among the paintings.

[6] https://www.nationalgallery.org.uk/
[7] https://www.crosscult.lu/

Algorithm 1: Dataset preprocessing.

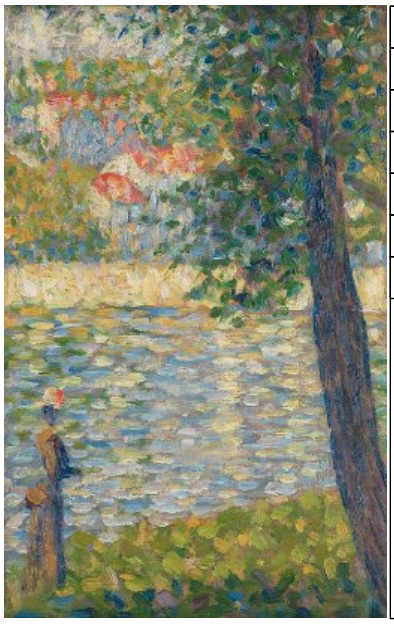
procedure PREPROCESS(Model, P) ☞ Model ∈ {LDA, BERT, ResNet}
  $\mathcal{P} \leftarrow$ FEATURIZEPAINTINGS(Model, P) ☞ Compute painting embeddings
  $\mathbf{A} \leftarrow \emptyset$ ☞ Initialize $\mathbf{A} \in \mathbb{R}^{m \times m}$ and $m = |\mathcal{P}|$
  for $\boldsymbol{p}_i$ and $\boldsymbol{p}_j \in \mathcal{P}$ **do** ☞ Iterate over collection
    $\mathbf{A}_{ij} \leftarrow$ COSINESIMILARITY($\boldsymbol{p}_i, \boldsymbol{p}_j$)
  return $\mathbf{A}$

Algorithm 2: Non-fusion VA RecSys.

Precondition: Similarity matrix $\mathbf{A}$ of featurized paintings
procedure RECOMMENDPAINTINGS(Model, P^u, ω^u, r) ☞ Model ∈ {LDA, BERT, ResNet}
  $\mathcal{P}^u \leftarrow$ FEATURIZEPAINTINGS(Model, P^u) ☞ Load computed embeddings, as $P^u \subset P$
  $S^u \leftarrow \emptyset$ ☞ Initialize recommendations list
  for $\boldsymbol{p}_i \in \mathcal{P}^u$ and $\boldsymbol{p}_j \in \mathcal{P}$ **do** ☞ Iterate over collection
    $S^u(\boldsymbol{p}_i) = \frac{1}{n} \sum_{j=1}^{n} \omega_j^u \cdot \mathbf{A}_{ij}$ ☞ Bias similarity matrix towards user's preferences
  SORT(S^u) ☞ Descending sort, as higher cosine distance means higher similarity
  return SLICE(S^u, r) ☞ Pick the first r elements in the ranking

Algorithm 3: Fusion-based VA RecSys.

procedure FUSERECOMMENDATIONS($\text{Model}_1, \text{Model}_2, P^u, \omega^u, r$) ☞ Models 1 & 2 ∈ {LDA, BERT, ResNet}
  $\mathcal{R}_1 \leftarrow$ RECOMMENDPAINTINGS($\text{Model}_1, P^u, \omega^u, r$) ☞ Get recommended lists from each model
  $\mathcal{R}_2 \leftarrow$ RECOMMENDPAINTINGS($\text{Model}_2, P^u, \omega^u, r$)
  $F(\boldsymbol{p} \in \mathcal{R}_1 \bigcup \mathcal{R}_2) = \sum_{i \in \mathcal{R}_1, j \in \mathcal{R}_2} \frac{1}{n(i)n(j)}$ ☞ Fuse lists: $n(\cdot)$ is the nth ranking position of each painting
  SORT(F) ☞ Descending sort, as higher cosine distance means higher similarity
  return SLICE(F, r) ☞ Pick the first r elements in the fused ranking



painting_id	000-018P-0000
title	The Morning Walk
artist	Georges Seurat
publication_date	19th_century
size_format	Portrait
size	Very Small
technique	oil painting
description	A woman, silhouetted against the shimmering water, strolls along a riverbank. The red roofs of houses can be made out along the opposite bank. Between 1882 and 1886 Seurat painted numerous such landscape studies on small wooden panels, some as independent works and others in preparation for his large-scale compositions. This sketch provided the starting point for a painting of 1885, 'The Seine at Courbevoie' (private collection).

Figure 3: Sample painting and associated metadata from the National Gallery dataset.

5.2 Preprocessing

On the one hand, to learn textual features with LDA and BERT models, the painting metadata were pre-processed: text fields concatenation, removal of punctuation symbols and stop-words, lowercasing, and lemmatization. On the other hand, to learn visual features with the ResNet model, we used the actual images

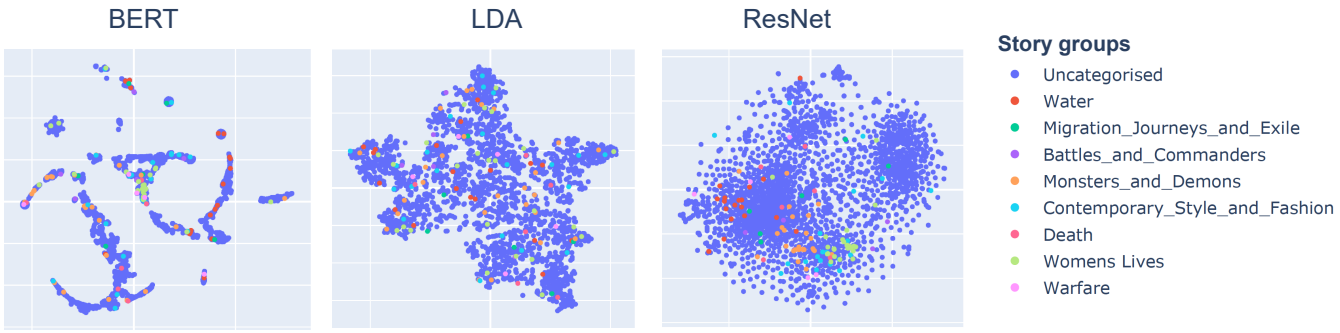


Figure 4: Latent space projection (t-SNE) of the curated story groups.

of paintings[8] to extract the convolutional feature maps with the pre-trained ResNet-50 model discussed in Section 3.2.

5.3 Text source analysis

In topic modeling, "topic coherence" is a commonly used technique to evaluate topic models. It is defined as the sum of pairwise similarity scores on the words $w_1, ..., w_n$ that describe each topic, usually the most frequent n words according to $p(w|t)$ [32]:

$$\text{TopicCoherence} = \sum_{i<j}^{n} \text{CosineSimilarity}(w_i, w_j) \quad (3)$$

Ideally, a good model should generate coherent topics; i.e the higher the coherence score the better the model is [53]. Figure 5 shows topic coherence in LDA as a function of the number of topics when using two text sources: 'description-only' (using only the curated stories from the `description` metadata, see Figure 3) and 'all-metadata' (using all the available metadata shown in Figure 3).

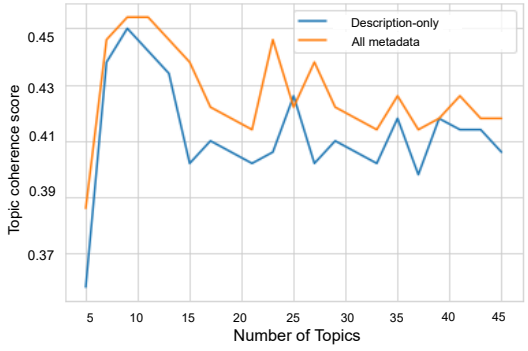


Figure 5: Comparative topic coherence analysis of the two text sources.

From the analysis presented in Figure 5 we can make two important observations. First, a topic model using all the available metadata consistently gives better topic coherence scores, therefore it is considered a better model. Using only the curated stories is a close contender, however we should note that it is very time consuming to produce these, as they require human expertise. Second, the maximal topic coherence is obtained at 10 topics when using all the available metadata. Having too many topics requires more resources as well as more computation time, therefore it is important to find a reasonable balance. On the other hand, for topic modelling using BERT we rely on HDSCAN to choose the optimal number of initial clusters and then latent topics are determined based on c-TF-IDF over those clusters.

Figure 6 shows the generated topics by LDA and BERT for our dataset. The size of each circle represents the prevalence of a topic, i.e, the popularity of a topic among the paintings. The distance between the circles represents the similarity between topics. The objective here is to have topics that are overlapping as little as possible. For LDA, the 10 automatically identified topics are evenly popular while being sufficiently distinct from each other with some overlaps between topic 9 & 7 and 5 & 10. For BERT, the four automatically identified topics are significantly distinct from each other while they are also laid out in descending order or popularity across the dataset. This indicates that after clustering, similar topics are merged together to reduce the total number of topics, thereby creating more cohesive topic models.

Hence, the number of topics in our implementation were set to 10 and 4 for LDA and BERT respectively.

6 EVALUATION

The main goal of our evaluation was to understand the user's perception towards the quality of our studied RecSys engines, and ultimately to assess which feature learning approach best captures the semantic relatedness of paintings. We conducted two user studies, to be described later, that were approved by the Ethics Review Panel of the University of Luxembourg with ID 22-031.

6.1 Apparatus

We created a web application that first collected preference elicitation ratings from users and then showed a set of VA recommendations based on their elicited preferences. As shown in Figure 7, participants were provided with one set of recommendations from each VA RecSys engine at a time.

Since participants could use any device (desktop computer, laptop, or mobile) to complete the study, we used a responsive design;

[8]All paintings are available under a Creative Commons (CC) license.

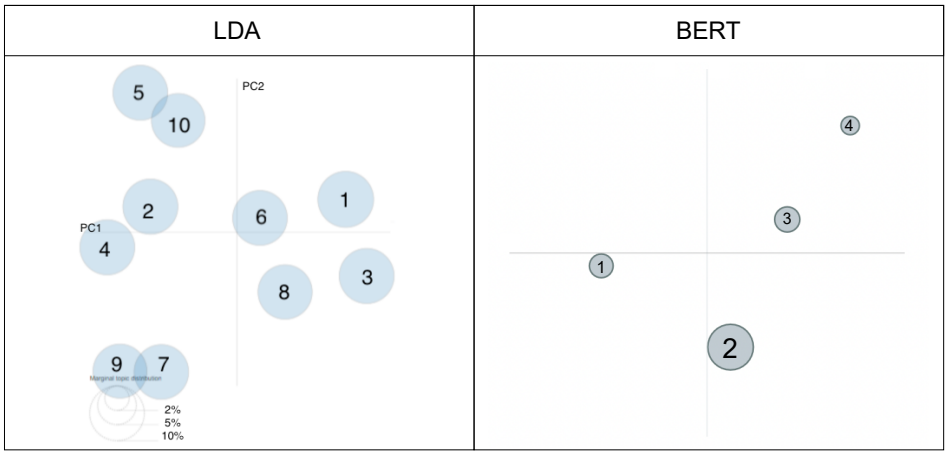


Figure 6: Inter-topic distance map of LDA and BERT in a projected 2-dimensional space.

see Figure 7. By clicking or tapping on any image, in both the elicitation and rating screens, a modal window displays an enlarged version of the image.

6.2 Participants

As described in the next section, we first conducted a small-scale study ($N = 11$) with museum visitors, to gather insights from real-world usage of our application, and then we conducted a large-scale study ($N = 100$) with a carefully selected pool of crowdworkers.

6.3 Design

Participants were exposed to all VA engines exactly once (within-subjects design) and rated the provided recommendations in a 5-point Likert scale. Our dependent variables are widely accepted proxies of recommendation quality [57]:

Accuracy: The paintings match my personal preferences and interests.
Diversity: The paintings are diverse.
Novelty: I discovered paintings I did not know before.
Serendipity: I found surprisingly interesting paintings.

6.4 Procedure

Participants accessed our web application and entered their demographics information (age, gender) on a welcome screen. There, they were informed about the purpose of the study and the data collection policy. They also indicated their visiting style, for which we adopted the framework proposed by Veron et al. [70] to classify museum visitors into four visiting style metaphors [38], related to the time they spend during visits:

Ant: I spend a long time observing all exhibits and move close to the walls and the exhibits avoiding empty space.
Fish: I walk mostly through empty space making just a few stops and see most of the exhibits but for a short time.
Grasshopper: I see only exhibits I am interested in. I walk through empty space and stay for a long time only in front of selected exhibits.
Butterfly: I frequently change the direction of my tour, usually avoiding empty space. I see almost all exhibits, but time varies between exhibits.

Then, participants advanced to the preference elicitation screen, where they were shown one painting at random from each of the nine curated story groups. They rated each painting in a 5-point numerical scale (5 is better, i.e. the user likes the painting the most). Finally, users advanced to the RecSys assessment screen, where they were shown a set of nine painting recommendations drawn from each VA RecSys engine. Note that each user initially rated nine paintings (one from each story group) but recommendations may come from only one or a few story groups, depending on their elicited preferences.

6.5 Museum study

We physically advertised our call for participants in the museum Centre Pompidou-Metz, France with a flyer that had a QR code for people to scan in order to access the study. A small sample of $N = 11$ participants (6 female, 5 male) aged 36 years (SD=20.8) voluntarily took part in the study. The study took 4.7 min on average to complete (SD=4.3).

Figure 8 shows the distributions of user ratings for each of the dependent variables considered. Figure 9 segregates the results by the different visiting profiles. We can see that participants perceived each VA engine differently for each of the evaluation metrics considered. For example, LDA was rated the highest in terms of Accuracy and Novelty, whereas the fusion of BERT+ResNet was rated higher in terms of Diversity. Interestingly, ResNet was rated the lowest in terms of Serendipity.

We investigated whether there is any difference between any of the five RecSys engines, for which we use a linear mixed-effects (LME) model where each dependent variable is explained by each VA RecSys engine. The visiting profile is considered an interaction effect (model covariate) and participants are considered random effects. An LME model is appropriate here because the dependent variables are discrete and have a natural order. In addition, LME

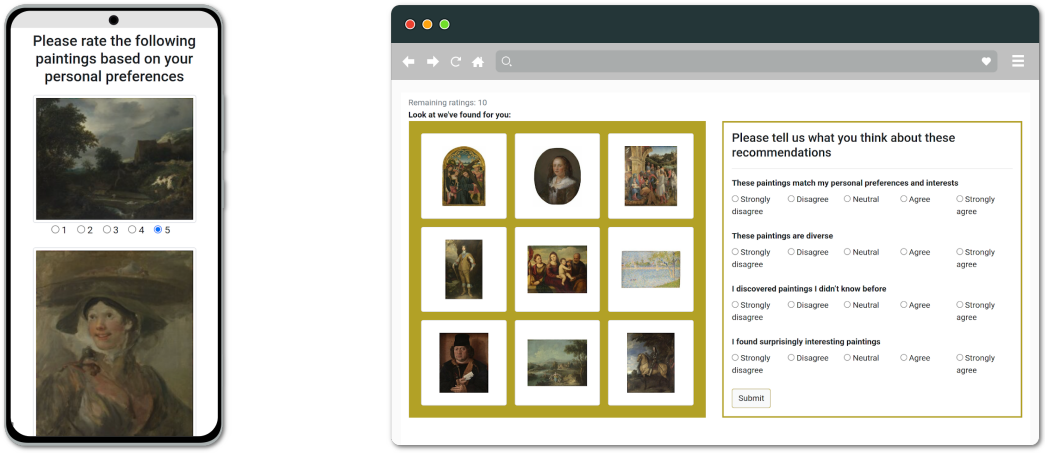
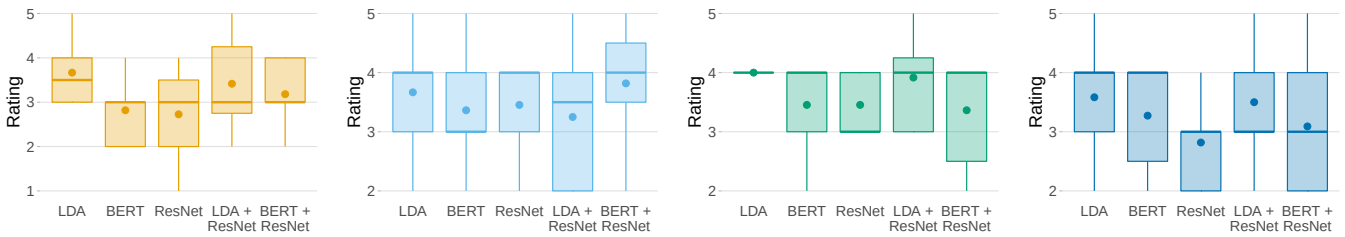


Figure 7: Screenshots of our web application for evaluation. Left: elicitation screen in mobile mode. Right: Recommendation evaluation screen in laptop mode.



(a) Accuracy ratings **(b) Diversity ratings** **(c) Novelty ratings** **(d) Serendipity ratings**

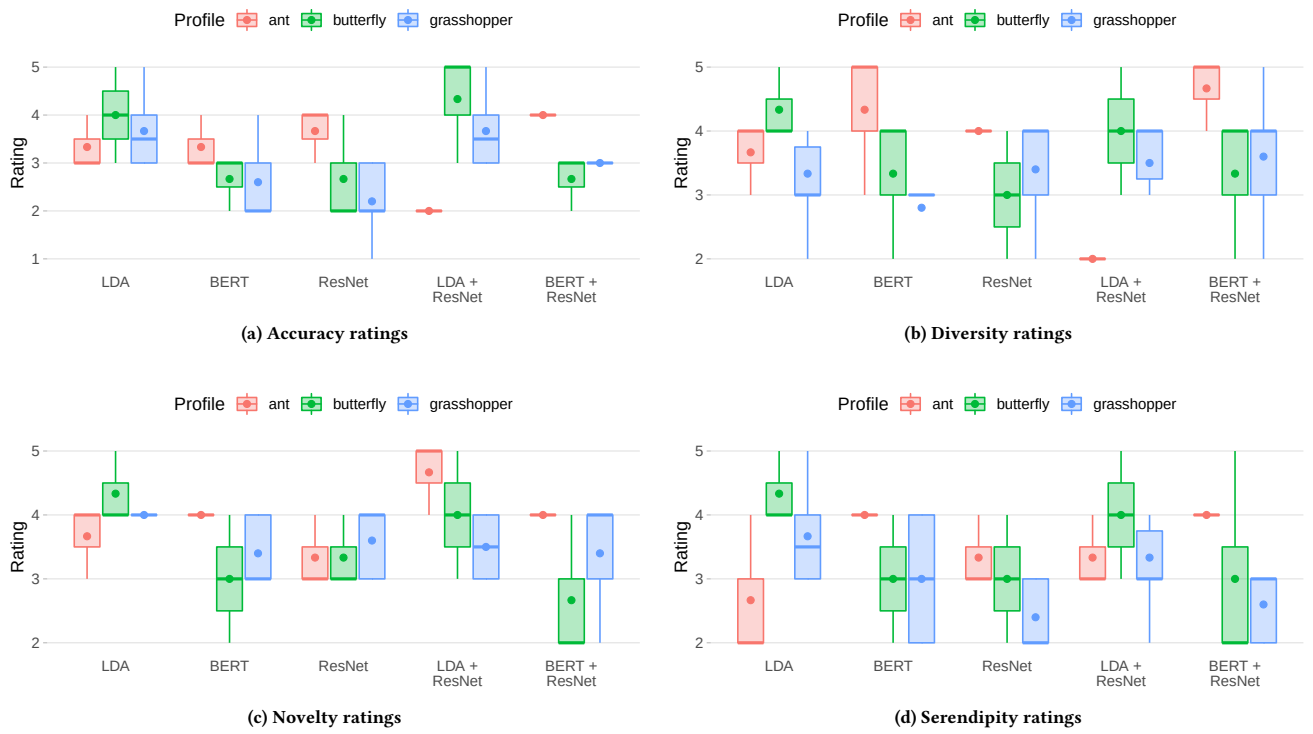
Figure 8: Distribution of ratings from museum users. Dots denote mean values.

models are quite robust to violations of several distributional assumptions [65].

We fit the LME models (one per dependent variable) and compute the estimated marginal means for specified factors. We then run pairwise comparisons (also known as *contrasts* in LME parlance) with Bonferroni-Holm correction to guard against multiple comparisons.[9] We observed that LDA was significantly preferred over BERT ($p = .028, r = 0.449$) and ResNet ($p = .028, r = 0.459$) engines in terms of Accuracy. LDA was preferred over ResNet ($p = .048, r = 0.459$) as well as over the fusion of BERT+ResNet ($p = .046, r = 0.409$) in terms of Novelty. The LDA+ResNet engine outperformed BERT ($p = .048, r = 0.379$) and ResNet ($p = .048, r = 0.404$) as well the fusion of BERT+ResNet ($p = .046, r = 0.439$) in terms of Novelty. All other comparisons were not found to be statistically significant. However, effect sizes (r, analogous to Cohen's d) suggest a moderate importance of the differences between RecSys engines in practice. For example, LDA was preferred over BERT in terms of Novelty ($r = 0.347, p = .060$) and the fusion of LDA+ResNet was preferred over BERT+ResNet in terms of Diversity ($r = 0.338, p = .357$). ResNet was less preferred than LDA or LDA+ResNet in terms of Serendipity ($r = 0.335, p = .223$).

If we take closer look at the results per visiting profiles (Figure 9), we can observe that Ant users prefer BERT topics over LDA topics, and this is also reflected in the fused rankings. For example, in terms of Accuracy, Diversity and Serendipity, Ant users ranked BERT-based recommendations higher than Butterfly and Grasshopper users. On the other hand, Grasshopper users did not like BERT-based recommendations overall. Instead, in terms of Novelty, the fusion of LDA+ResNet was preferred over BERT+ResNet. We observed a statistically significant correlation between visitor profiles and ratings in terms of Diversity ($\rho = 0.26, p < .01$) and Serendipity ($\rho = 0.3, p < .001$). This can potentially be an indication that the visiting style of the user, to a certain extent, reflects their preferences towards art content. Hence, it could be leveraged to parameterise different aspects of RecSys (e.g, Diversity, Novelty, etc.) in future work.

[9]The Bonferroni-Holm correction method sorts p-values from lowest to highest and compares them to nominal alpha levels of $\frac{\alpha}{m}$ to α. Then, it finds the index k that identifies the first p-value that is not low enough to validate rejection of the null hypothesis.



(a) Accuracy ratings

(b) Diversity ratings

(c) Novelty ratings

(d) Serendipity ratings

Figure 9: Distribution of ratings from museum users, segregated by visiting profiles. Dots denote mean values.

6.6 Crowdsourcing study

We recruited a large sample of $N = 100$ participants via the Prolific crowdsourcing platform.[10] We enforced the following screening criteria for any participant to be eligible:

- The primary language is English.
- Art is listed among their interests/hobbies.
- Minimum approval rate of 99% in previous crowdsourcing studies in the platform.
- Registration date before January 2022.

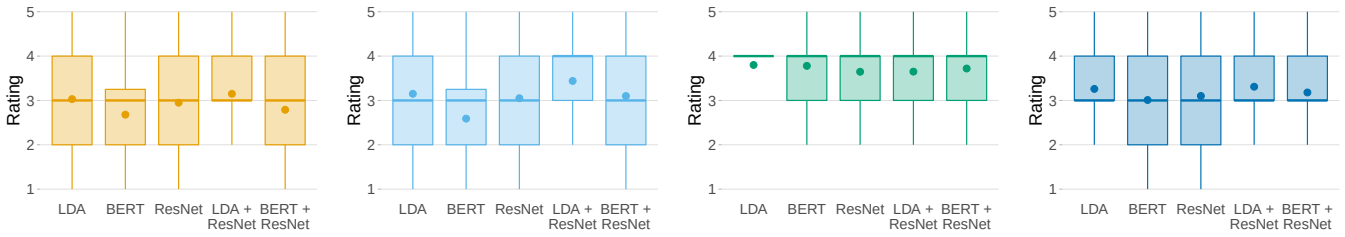
Our recruited participants (75 female, 25 male) were aged 39.7 years (SD=14.1) and could complete the study only once. Most of them had UK nationality (59%) or were living in the UK (64%). The study took 5.7 min on average to complete (SD=2.3) and participants were paid an equivalent hourly wage of $10/h.

Figure 10 shows the distributions of user ratings for each of the dependent variables considered. Figure 11 segregates the results by the different visiting profiles. We can see that, overall, crowdworkers tended to rate the VA RecSys engines slightly higher than museum users. We observed that the fusion of LDA+ResNet delivered the highest-quality results, as the ratings received had the narrower inter-quartile difference. This was systematically so for all the four evaluation metrics considered; see Figure 10.

As in the previous study, we fit the LME models and compute the estimated marginal means for specified factors. We then run pairwise comparisons with Bonferroni-Holm correction to guard against over-testing the data because of the multiple comparisons. We observed that BERT was significantly less preferred than LDA ($p = .033, r = 0.131$) and LDA+ResNet ($p = .003, r = 0.18$) in terms of Accuracy. BERT+ResNet was outperformed by LDA+ResNet in terms of Accuracy ($p = .014, r = 0.151$). In terms of Diversity, BERT was rated significantly lower than any other approach ($p < .001, 0.196 < r < 0.36$) and the fusion of LDA+ResNet outperformed BERT+ResNet ($p < .01, r = 0.154$) as well as the individual LDA ($p = .013, r = 0.132$) and ResNet ($p < .001, r = 0.181$) engines. All other comparisons were not found the be statistically significant. We can conclude therefore that the fusion of text and image features is the most beneficial approach to deliver more adequate recommendations to the user.

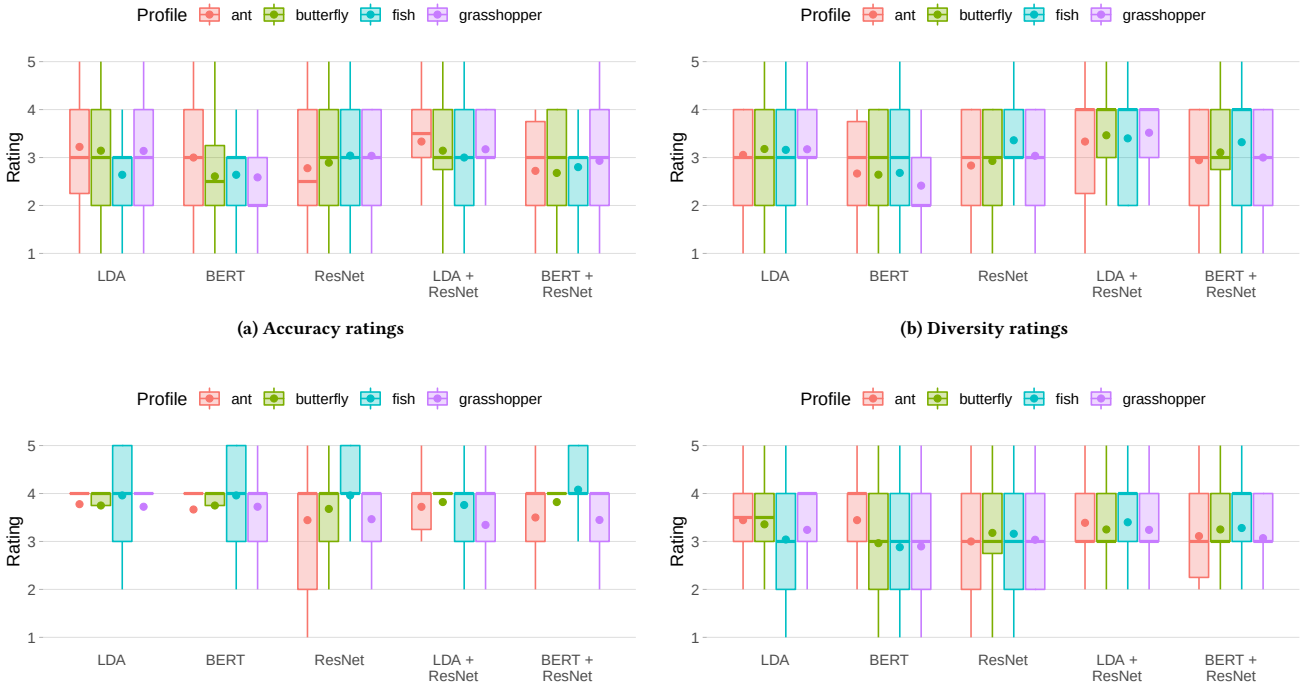
In this crowdsourcing study we did not observe strong correlations between user profiles and ratings. However, a few interesting observations can be made. For example, from the results per visiting profiles (Figure 11) we can see that Fish users did not like BERT-based recommendations, which was also reflected in the fused ranking BERT+ResNet. In terms of Diversity, Grasshopper users prefer LDA over BERT. This can be attributed to the larger topic size in LDA (10 topics) compared to BERT (4 topics). Hence, we hypothesise that users who preferred LDA are most likely interested in diverse VA content, especially it we take into account that Grasshopper profiles have a clear expectation of what to find in a museum. In terms of Novelty, Butterfly users showed more agreements in their rankings, as the interquartile range is much smaller as compared to the other visiting profiles. Finally we observed that

[10] https://www.prolific.co/



(a) Accuracy ratings (b) Diversity ratings (c) Novelty ratings (d) Serendipity ratings

Figure 10: Distribution of ratings from crowdsourcing users. Dots denote mean values.



(a) Accuracy ratings (b) Diversity ratings

(c) Novelty ratings (d) Serendipity ratings

Figure 11: Distribution of ratings crowdsourcing users, segregated by visiting profiles. Dots denote mean values.

Fish users tended to provide higher ratings than the other user profiles, especially for ResNet and BERT+ResNet recommendations. As discussed in the previous section, these observations could potentially inform novel ways of operationalising different aspects of RecSys in future work.

6.7 Ranking overlap analysis

We conducted an additional analysis that checked whether the users were receiving truly personalized recommendations. Otherwise, our VA RecSys engines would have been recommending the same contents to every user. To account for this, we compute the Intersection over Union (IoU) and Rank-Biased Overlap (RBO), which are widely used measures in information retrieval [71]. RBO and IoU were calculated in a pairwise manner among all users exposed to the same engine and averaged. Table 1 presents the results of this analysis. As shown in the table, there is no substantial overlap in the rankings produced by each engine. This analysis indicates that each user indeed was shown a personalized set of recommendations.

7 DISCUSSION

From a conceptual point of view, this paper has advanced our understanding of how users perceive and evaluate VA RecSys. In recent years, the research community has shifted to include a wider range of "beyond accuracy" objectives [34], such as the user-centric dependent variable we have used in our studies, however the field of VA personalization has remained largely unexplored in this regard.

Table 1: Ranking overlap results, showing Mean ± SD of IoU and RBO measures.

		LDA	BERT	ResNet	LDA+ResNet	BERT+ResNet	All
Crowdsourcing study	IoU	0.09 ± 0.15	0.09 ± 0.15	0.09 ± 0.15	0.09 ± 0.15	0.09 ± 0.15	0.07 ± 0.11
	RBO	0.10 ± 0.16	0.10 ± 0.16	0.10 ± 0.17	0.09 ± 0.16	0.09 ± 0.16	0.07 ± 0.12
Museum study	IoU	0.27 ± 0.26	0.33 ± 0.26	0.32 ± 0.26	0.27 ± 0.27	0.33 ± 0.26	0.11 ± 0.16
	RBO	0.26 ± 0.27	0.31 ± 0.27	0.31 ± 0.27	0.26 ± 0.27	0.31 ± 0.26	0.09 ± 0.16

We have found that text-only and vision-only RecSys compare similarly in terms of recommendation quality, but the fusion of these two approaches delivers the best results. We also have observed that different visiting style profiles may benefit differently from each type of recommendations, although fusion-based recommendations are systematically preferred overall.

Previous work suggested that visual features are preferred over textual features when it comes to delivering high-quality VA recommendations to the users [48–50]. However, our experiments have demonstrated that they provide similar results. This was so for the small-scale and the large-scale study. Therefore, we reject **H1** and conclude that visual features perform no better than textual features. This is somehow understandable, since each type of latent representation provides a different understanding about the paintings. Furthermore, the improved performance observed in the fusion approaches indicates that both visual and textual features complement each other to efficiently capture the elements of VA RecSys, which leads us to validate **H2**. In the following, we provide a critical and in-depth discussion about our results and what they imply for the HCI community.

7.1 Visual similarity does not entail semantic similarity (and vice versa)

Nowadays, with the recent advances in computer vision, capturing visual similarity of images is relatively an effortless task. Hence, finding visually similar paintings to what users previously saw or expressed interest seems straightforward. However, as discussed in Section 2, understanding users' perception of artwork is an extremely challenging task due to the complexity of concepts embedded within the artworks as well as the reflections they may trigger on users. Contrary to most prominent work in VA RecSys that leveraged only visual features to derive recommendations, we explored textual features as well as hybrid approaches combining the learned text-based and image-based features. Interestingly, our work provides compelling evidence that visual similarity does not necessarily entail semantic relatedness.

In Figure 12 we illustrate this phenomenon with examples. We show a target painting (top) and its most similar painting (bottom) according to the three VA RecSys engines. For LDA and BERT we additionally show the paintings' topic distributions and their descriptions. For LDA (first column) we can see that paintings have very similar topic distribution and topic 8 stands out. This implies that words in topic 8 are more likely to be found in the paintings descriptions than the words from the other topics. Actually topic 8 is very well defined as there is high coherence between the words. In fact, topic 8 can be described as a "Christian" topic of the collection, since many of the words in this topic are usually found in christian corpora such as biblical texts. When looking at the paintings, there are many references to Christianity, therefore we can assume that their descriptions contain vocabulary that refers to a religious context. The ground-truth from the National Gallery documentation also supports this claim, as both paintings are from the panels of the high altarpiece of the church of Sant'Alessandro Brescia, painted by Girolamo Romanino in the 16th century. Then, the target painting[11] shows Saint Filippo Benizzi, who was the fifth general of the Servites, the order to whom the church belonged. The most similar painting according to LDA[12] is a portrait of Saint Gaudioso, who was the bishop of Brescia in the 5th century, and was buried in the church.

For BERT (second column), the target painting is "Calm: A Dutch Ship coming to Anchor and Another under Sail"[13] by Willem van de Velde, and the most similar one according BERT is "Dutch Ships and Small Vessels Offshore in a Breeze"[14] by the same artist. When we look at how BERT represents these two paintings, we can observe that they have very similar topic distributions. Particularly topic 3 is very prominent in both paintings. Taking a closer look at the topic descriptions, we can understand that BERT created a coherent representation. Observing the actual images of the paintings, we can also tell that the paintings are visually very similar. Overall, both examples of LDA and BERT demonstrate that similarities of visual features can be captured from semantic similarities of textual features. However, our analysis on ResNet shows that the inverse is not necessarily true.

The last column in Figure 12 illustrates a sample target painting and its most similar painting according to ResNet. The target is a painting from the 18th century titled "Time orders Old Age to destroy Beauty" [15] by Pompeo Girolamo Batoni. In this case, the most similar painting is from 16th century titled "The Donor and Saint Mary Magdalene"[16] by Marten van Heemskerck. Looking at the two paintings, without further context, one can easily tell that ResNet manages to capture visual features such as colors, edges, and corners among the paintings. However, the two paintings are not very semantically related. The target painting depicts "time" by

[11] https://www.nationalgallery.org.uk/paintings/girolamo-romanino-saint-filippo-benizzi
[12] https://www.nationalgallery.org.uk/paintings/girolamo-romanino-saint-gaudioso
[13] https://www.nationalgallery.org.uk/paintings/willem-van-de-velde-a-dutch-ship-coming-to-anchor
[14] https://www.nationalgallery.org.uk/paintings/willem-van-de-velde-dutch-ships-and-small-vessels-offshore-in-a-breeze
[15] https://www.nationalgallery.org.uk/paintings/pompeo-girolamo-batoni-time-orders-old-age-to-destroy-beauty
[16] https://www.nationalgallery.org.uk/paintings/marten-van-heemskerck-the-donor-and-saint-mary-magdalene

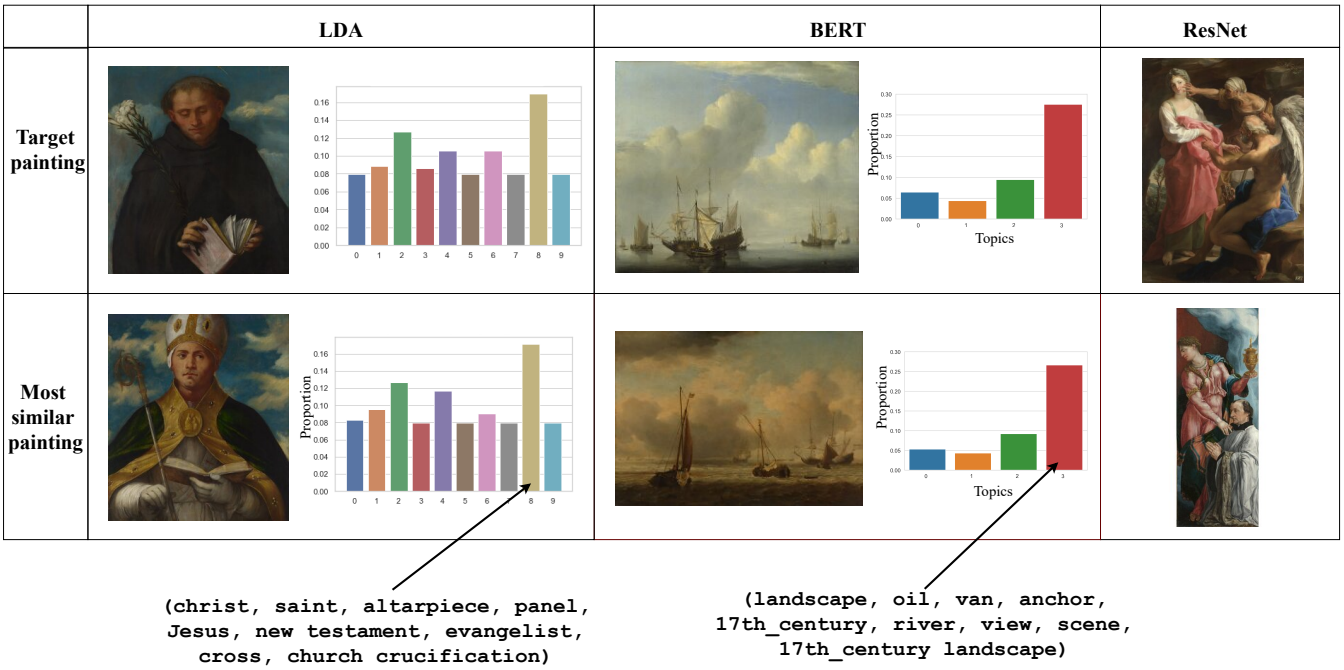


Figure 12: Examples of target paintings (top) and most similar paintings (bottom) according to LDA, BERT and ResNet.

the winged figure holding an hourglass, ordering his companion Old Age to disfigure the face of a young woman, the personification of Beauty. The National Gallery documentation states: *With this painting, Batoni intends to encourage considering the brevity of youth and the inevitable passing of time.* On the other hand, *"The Donor" depicts a statuesque Mary Magdalene, one of Christ's followers, resting her fingers on the shoulder of a kneeling donor, and with the other hand she is nonchalantly lifting a large golden vessel. This is the pot containing the precious ointment with which she anointed Christ's feet (Luke 7:37). In sharp contrast to her colourful opulence, the donor is a serious-looking middle-aged man dressed as a canon.* The National Gallery documentation also mentions that this is one of two shutters from a triptych (a painting made up of three sections), the central part of which is lost.

Given the above discussion, we can deduce that visual similarity does not necessarily entail semantic relatedness. Especially for VA RecSys applications, relying only on visual features can have a negative impact on the quality of recommendations. For example, a user who is not at all interested in religion or Christianity receiving " The Donor" as a recommendation just because they liked or previously expressed interest for "Time orders Old Age to destroy Beauty" may not be desirable. Thus, although visual features are important in describing an artwork they alone can not represent the underlying complex semantic relationships.

7.2 Not all topics are created equal

In general, the topic distributions learned by a topic model can be used as a semantic representation, which can be used in several downstream tasks such as document classification, clustering, retrieval, or visualisation [75]. Particularly, our topic models for VA RecSys have demonstrated the power of exploiting textual data to understand semantic relationships of paintings. This was reflected in the improved performance when combining LDA and BERT with ResNet. Although LDA and BERT bring in statistical analysis of abstract concepts from the textual data, each technique has its own uniqueness and relies on different assumptions.

Topic models learn from documents in an unsupervised way and usually measured using a single metric (e.g., topic coherence), which can reflect just one aspect of a model. However, documents are usually associated with rich sets of metadata at both the document and word levels. Overall, evaluating topic models is challenging due to the variety of current frameworks and architectures. It is also evident that quantitative methods are limited in their ability to provide in-depth contextual understanding [18]. Thus, the interpretation of topic models still relies heavily on human judgment.

7.3 High-quality recommendations emerge from high-quality latent representations

The elements of VA recommendation are the latent features of paintings, therefore is clear that high-quality latent representations must be learned in order to provide high-quality recommendations. We have shown that each type of RecSys engine (text or image based) is capturing one dimension of the user/painting latent space. Since paintings are made of visual and textual data, it is beneficial to consider both aspects when generating recommendations to the user. As mentioned in previous work, the community has been arguing for ignoring textual features [27, 28, 48] in favor of visual features, however we have shown that doing so will ignore an important dimension of paintings and therefore an important element of visual art recommendation.

7.4 Knowing the user preferences is key, but it comes at a cost

Recommender systems require interactions from users to infer personal preferences about new items [30]. It is paramount to know as much information as possible from the users, in particular in the form of ratings, however we should not burden the users by asking the users to rate every painting they have visited. Therefore, we must seek a balance between how many ratings we want the user to provide and how much quality we aim to achieve.

In our study, each participant rated one painting from each of the nine categories of the collection we analyzed. Because the number of categories is small, we could collect one observation from the user for each group of paintings. However, when the number of categories is too large this approach becomes unfeasible. To alleviate this, we could explore agglomerative clustering techniques [59] to select the most interesting groups of paintings to elicit the user's preferences, based e.g. on dispersion-aware metrics such as cluster intra-variance.

7.5 Optimizing for real-time performance is important

We implemented several real-time RecSys engines, where computing performance is critical. In web applications, it is argued that if users do not receive a response by the system in 1 second, they will perceive that they do not have control over the system [54] and quite often they will quit the application if it remains unresponsive [51]. To ensure our engines will reply in such a constrained scenario, we implemented several optimizations, such as using a lightweight version of SBERT with a small memory footprint instead of the fully-fledged pre-trained model, and adopting a late fusion technique to merge the contributions of two engines instead of considering early fusion approaches.

8 LIMITATIONS AND FUTURE WORK

We acknowledge that our crowdsourcing users were not really intrinsically motivated, or at least not as much as our museum participants, since they had a monetary incentive to take part in the study. This might have influenced the results, however to mitigate this we collected a large sample of participants interested in artwork and considered the user as a random effect in our statistical analysis.

On the other hand, we consider our museum participants intrinsically motivated, as they were actually visiting a museum, had to scan a QR code with their phones, and all of them fully completed the study without any monetary compensation. Also, the correlation coefficient between profile type and recommendation ratings was higher (and sometimes statistically significant) for museum users. However, we acknowledge that the sample size is very small to derive general conclusions from that user sample. The small-scale study, however, agrees with the large-scale crowdsourcing study in the sense that visual a textual features result in same-quality VA recommendations. It is therefore advised to consider both approaches when deploying VA RecSys, as both approaches complement well each other in terms of uncovering different painting semantics. We believe that, in order to improve the quality of recommendations further, future work should incorporate more user feedback on artwork, if available (e.g. in the form of reviews or even the elicited ratings themselves), as part of our model training pipelines.

As discussed in Section 7.1 an interesting takeaway from our study is that the elements of VA recommendation (i.e, key explanatory factors for semantic relatedness of visual arts) lie not only in visual but also in textual features. We were able to uncover this thanks to our late fusion engines. Particularly the late fusion approach is advantageous as it allows to control the contribution of each fused engine compared to an early fusion approaches in multimodal feature learning such as [42] and a more recent work CLIP [58] by Open AI. We should note that we used the same backbone architectures as state-of-the-art approaches like CLIP and others [42,58], i.e. Transformers (BERT) for computing text embeddings and ResNet for computing image embeddings. The only difference is that we adopt a late fusion approach since it provides a clear way of understanding the contribution of each modality (image or text) to the generated recommendations. On the contrary, an early fusion approach such as CLIP prevents us from controlling the exact contribution of text and image embeddings because they are entangled, thereby CLIP behaves like a black box model. In our studies, we set exactly 50% for text and image contribution, respectively. As a follow-up of this work, we plan to conduct a comparative study of fusion engines (early versus late) on a VA recommendation task.

Finally, we note that our application asked participants to rate one painting randomly selected from each of the nine categories of our dataset. This resulted in a 9-dimensional preference elicitation vector with associated weights, which is perhaps small, considering that previous work asked participants to rate up to 80 paintings [73]. However, we have not observed substantial overlaps in the rankings produced by each RecSys engine, which indicates that each participant received truly personalized recommendations. Further, unlike our experiments, previous work was conducted in a very controlled setting. In general, preference elicitation is a longstanding challenge in designing real-world RecSys applications. Ideally, VA RecSys needs to interact with new visitors to gather as much information as possible, however people are not always willing to provide information or answer lengthy questionnaires [56]. This makes the task of providing personalized VA contents rather challenging. Hence, instead of relying on explicit user profiling, future work should investigate efficient strategies to extract maximal information with minimal user engagement. Nevertheless, we should note that our study reflects a high level of realism, in terms of ecological validity: anybody can access the application with any device and receive VA recommendations from any of our RecSys engines in real-time.

9 CONCLUSION

Understanding how users' perceive and interact with highly subjective content such as artwork is an extremely challenging task due to the complexity of the concepts embedded within artworks and the emotional and cognitive reflections they may trigger on users. We have studied the elements of visual art recommendation, i.e. techniques to uncover latent semantic relationships embedded within paintings, leveraging textual and visual information, as well

as their combination. To evaluate the performance of each approach, we adopted user-centric evaluation measures.

Our findings open an interesting perspective to understand how users perceive and interact with artwork. Overall, we can conclude that the semantics of paintings cannot be represented only by visual features nor textual descriptions, since the emotional and cognitive reflections they may trigger on users are quire diverse and often unpredictable. Although hybrid approaches of fusing visual and textual features showed clear performance improvements, more research remains to explore how to improve further the quality of recommendations.

Ultimately, this paper may benefit the HCI community by offering a systematic examination of how to uncover semantic information from different data sources in a way that users will perceive as high-quality personalized content. Our work has potential applications well beyond the scope of this paper, such as user modeling, intelligent user interfaces, and adaptive user interfaces, among others. Our dataset, software, and models are publicly available at https://github.com/Bekyilma/VA_RecSys.

ACKNOWLEDGMENTS

This work was supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001 and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

REFERENCES

[1] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory — ICDT 2001*, Jan Van den Bussche and Victor Vianu (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 420–434.

[2] Maha Amami, Gabriella Pasi, Fabio Stella, and Rim Faiz. 2016. An lda-based approach to scientific paper recommendation. In *International conference on applications of natural language to information systems*. Springer, 200–210.

[3] Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe, and José Ochoa Luna. 2014. Online Courses Recommendation based on LDA.. In *SIMBig*. Citeseer, 42–48.

[4] LM Aroyo, Y Wang, R Brussee, Peter Gorgels, LW Rutledge, and N Stash. 2007. Personalized museum experience: The Rijksmuseum use case. In *Museums and the Web 2007 (San Francisco CA, USA, April 11-14, 2007. Proceedings)*. Archives & Museum Informatics.

[5] Monika Bansal, Munish Kumar, Manish Kumar, and Krishan Kumar. 2021. An efficient technique for object recognition using Shi-Tomasi corner detection algorithm. *Soft Computing* 25, 6 (2021), 4423–4432.

[6] Catarina Barata, M Emre Celebi, and Jorge S Marques. 2018. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE journal of biomedical and health informatics* 23, 3 (2018), 1096–1109.

[7] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[8] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. 1999. When Is "Nearest Neighbor" Meaningful?. In *Database Theory — ICDT'99*, Catriel Beeri and Peter Buneman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 217–235.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[10] Fidel Cacheda, Víctor Carneiro, Diego Fernández, and Vreixo Formoso. 2011. Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB)* 5, 1 (2011), 1–33.

[11] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. Brief: Binary robust independent elementary features. In *European conference on computer vision*. Springer, 778–792.

[12] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 160–172.

[13] Jie Chen, Li-hui Zou, Juan Zhang, and Li-hua Dou. 2009. The Comparison and Application of Corner Detection Algorithms. *Journal of multimedia* 4, 6 (2009).

[14] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

[15] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 585–592.

[16] Louis Deladiennee and Yannick Naudet. 2017. A graph-based semantic recommender system for a reflective and personalised museum visit: Extended abstract. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. 88–89. https://doi.org/10.1109/SMAP.2017.8022674

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Roman Egger and Joanne Yu. 2021. Identifying hidden semantic structures in Instagram data: a topic modelling comparison. *Tourism Review* (2021).

[19] Abigail R Esman. 2012. The World's Strongest Economy? The Global Art Market.

[20] John H Falk. 2016. *Identity and the museum visitor experience*. Routledge.

[21] Angus Graeme Forbes, Saiph Savage, and Tobias Höllerer. 2012. Visualizing and verifying directed social queries. In *IEEE Workshop on Interactive Visual Text Analytics. Seattle, WA, Citeseer*. Citeseer.

[22] Sarah Frost, Manu Mathew Thomas, and Angus G Forbes. 2019. Art i don't like: An anti-recommender system for visual art. In *Proceedings of Museums and the Web*.

[23] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[24] Willem Robert van Hage, Natalia Stash, Yiwen Wang, and Lora Aroyo. 2010. Finding your way through the Rijksmuseum with an adaptive mobile museum guide. In *Extended semantic web conference*. Springer, 46–59.

[25] Hebatallah A Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, and Joeran Beel. 2019. Bert, elmo, use and infersent sentence encoders: The panacea for research-paper recommendation?. In *RecSys (Late-Breaking Results)*. 6–10.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: a visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 309–316.

[28] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[29] D Helal, H Maxson, and J Ancelet. 2013. Lessons learned: Evaluating the Whitney's multimedia guide. *línea).< http://mw2013. museumsandtheweb. com/paper/lessons-learned-evaluating-the-whitneys-multimedia-guide/>, acceso* 25 (2013).

[30] María Hernández-Rubio, Alejandro Bellogín, and Iván Cantador. 2020. Aspect-based active learning for user preference elicitation in recommender systems. In *Proc. CIRCLE Workshop*.

[31] A Victor Ikechukwu, S Murali, R Deepu, and RC Shivamurthy. 2021. ResNet-50 vs VGG-19 vs training from scratch: a comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Global Transitions Proceedings* 2, 2 (2021), 375–381.

[32] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[33] Budi Juarto and Abba Suganda Girsang. 2021. Neural Collaborative with Sentence BERT for News Recommender System. *JOIV: International Journal on Informatics Visualization* 5, 4 (2021), 448–455.

[34] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2016).

[35] Mehmet Oguz Kelek, Nurullah Calik, and Tulay Yildirim. 2019. Painter classification over the novel art painting data set via the latest deep neural networks. *Procedia Computer Science* 154 (2019), 369–376.

[36] Kalliopi Kontiza, Olga Loboda, Louis Deladiennee, Sylvain Castagnos, and Yannick Naudet. 2018. A museum app to trigger users' reflection. In *International Workshop on Mobile Access to Cultural Heritage (MobileCH2018)*.

[37] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[38] Tsvi Kuflik, Zvi Boger, and Massimo Zancanaro. 2012. *Analysis and Prediction of Museum Visitors' Behavioral Pattern Types*. 161–176.

[39] Tsvi Kuflik, Einat Minkov, and Keren Kahanov. 2014. Graph-based Recommendation in the Museum. In *DMRS*. Citeseer, 46–48.

[40] Dor Lavi, Volodymyr Medentsiy, and David Graus. 2021. consultantbert: Fine-tuned siamese sentence-bert for matching jobs and job seekers. *arXiv preprint arXiv:2109.06501* (2021).

[41] Bin Li and Dimas Lima. 2021. Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering* 2 (2021), 57–64.

[42] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2020. Dynamic attention-based explainable recommendation with textual and visual fusion. *Information Processing & Management* 57, 6 (2020), 102099.

[43] Ioanna Lykourentzou, Xavier Claude, Yannick Naudet, Eric Tobias, Angeliki Antoniou, George Lepouras, and Costas Vasilakis. 2013. Improving museum visitors' Quality of Experience through intelligent recommendations: A visiting style-based approach. In *Workshop Proceedings of the 9th International Conference on intelligent environments*. IOS Press, 507–518.

[44] Elmar Mair, Gregory D Hager, Darius Burschka, Michael Suppa, and Gerhard Hirzinger. 2010. Adaptive and generic corner detection based on the accelerated segment test. In *European conference on Computer vision*. Springer, 183–196.

[45] Karim Malik, Colin Robertson, Steven A Roberts, Tarmo K Remmel, and Jed A Long. 2022. Computer vision models for comparing spatial patterns: understanding spatial scale. *International Journal of Geographical Information Science* (2022), 1–35.

[46] R. Mayer and S. Sheehan. 1991. *The Artist's Handbook of Materials and Techniques*. Viking. https://books.google.lu/books?id=tQ9nYreyGwEC

[47] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[48] Pablo Messina, Manuel Cartagena, Patricio Cerda, Felipe del Rio, and Denis Parra. 2020. CuratorNet: Visually-aware Recommendation of Art Images. (2020).

[49] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2017. Exploring Content-based Artwork Recommendation with Metadata and Visual Features. *arXiv preprint arXiv:1706.05786* (2017).

[50] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2019. Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction* 29, 2 (2019), 251–290.

[51] F. Nah. 2004. A study on tolerable waiting time: how long are web users willing to wait? *Behaviour & Information Technology* 23 (2004). Issue 3.

[52] Yannick Naudet, Angeliki Antoniou, Ioanna Lykourentzou, Eric Tobias, Jenny Rompa, and George Lepouras. 2015. Museum personalization based on gaming and cognitive styles: the BLUE experiment. *International Journal of Virtual Communities and Social Networking (IJVCSN)* 7, 2 (2015), 1–30.

[53] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics, 100–108.

[54] Jakob Nielsen. 2009. Powers of 10: Time Scales in User Experience. https://www.nngroup.com/articles/powers-of-10-time-scales-in-ux/.

[55] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.

[56] Bilih Priyogi. 2019. Preference elicitation strategy for conversational recommender system. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 824–825.

[57] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[59] Pradeep Rai and Shubha Singh. 2010. A survey of clustering techniques. *International Journal of Computer Applications* 7, 12 (2010), 1–5.

[60] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Process. Mag.* 34, 6 (2017), 96–108.

[61] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019). arXiv:1908.10084 http://arxiv.org/abs/1908.10084

[62] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[63] Deepjyoti Roy and Mala Dutta. 2022. A systematic review and research perspective on recommender systems. *Journal of Big Data* 9, 1 (2022), 1–36.

[64] Ayodeji Olalekan Salau and Shruti Jain. 2019. Feature extraction: a survey of the types, techniques, applications. In *2019 International Conference on Signal Processing and Communication (ICSC)*. IEEE, 158–164.

[65] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* 11, 9 (2020), 1141–1152.

[66] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[67] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[68] Eirini Eleni Tsiropoulou, Athina Thanou, and Symeon Papavassiliou. 2017. Quality of Experience-based museum touring: A human in the loop approach. *Social Network Analysis and Mining* 7, 1 (2017), 1–13.

[69] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* (2008), 2579–2605.

[70] Eliséo Véron and Martine Levasseur. 1989. *Ethnographie de l'exposition: l'espace, le corps et le sens*. Bibliothèque publique d'information du Centre Pompidou.

[71] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. 28, 4, Article 20 (2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[72] Bereket Abera Yilma, Najib Aghenda, Marcelo Romero, Yannick Naudet, and Hervé Panetto. 2020. Personalised visual art recommendation by learning latent semantic representations. In *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA)*. IEEE, 1–6.

[73] Bereket Abera Yilma, Yannick Naudet, and Hervé Panetto. 2021. Personalisation in Cyber-Physical-Social Systems: A Multi-Stakeholder Aware Recommendation and Guidance. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 251–255. https://doi.org/10.1145/3450613.3456847

[74] Feng Zhao, Yajun Zhu, Hai Jin, and Laurence T Yang. 2016. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Generation Computer Systems* 65 (2016), 196–206.

[75] He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498* (2021).