User Aesthetics Identification for Fashion Recommendations

LIWEI LIU, Farfetch
IVO SILVA, Farfetch
PEDRO NOGUEIRA, Farfetch
ANA MAGALHÃES, Farfetch
EDER MARTINS, Farfetch

One of the challenges in fashion recommendations is how to incorporate the concepts of fashion and style to provide a more tailored personalized experience for fashion lovers. Despite that these concepts are subjective, our fashion experts at Farfetch have defined a few key sets of aesthetics which attempt to capture the essence of users' styles into groups. This categorization will help us to understand the customers' fashion preferences and hence guide our recommendations through the subjectivity. In this paper, we will demonstrate that such concepts can be predicted from users' behaviors and the products they have interacted with. We not only compared a popular machine learning algorithm - Random Forest with a more recent deep learning algorithm - Convolutional Neural Network (CNN), but also looked at 3 different sets of features: text, image, and inferred user statistics, together with their various combinations in building such models. Our results show that it is possible to identify a customer's aesthetic based on this data. Moreover, we found that the use of the textual descriptions of products interacted by the customer led to better classification results.

CCS Concepts: • **Information systems → Recommender systems**.

Additional Key Words and Phrases: Recommender Systems, Luxury Fashion, User Segmentation, Aesthetics

ACM Reference Format:
Liwei Liu, Ivo Silva, Pedro Nogueira, Ana Magalhães, and Eder Martins. 2020. User Aesthetics Identification for Fashion Recommendations. In *Proceedings of 2nd Workshop on Recommender Systems in Fashion, 14th ACM Conference on Recommender Systems (recsysXfashion'20)*. ACM, New York, NY, USA, 12 pages.

1 INTRODUCTION

Recommender systems have been an increasingly important part in e-commerce. Some websites are designed to follow a personalized recommendation experience flow [4], such as Netflix (which reported that 75% of the views originate from their recommendations) and Amazon (with 35% of the revenue coming from personalized recommendations). In fact, recommender systems have played such a role in a customers' shopping journey, that today they expect to see recommendations during their interaction with e-commerce websites.

Performing fashion recommendations poses a challenge for those who need to reflect a customer's unique sense of style and preferences in order to enhance a personalized experience for a fashion lover. When a customer visits the website, we should aim to recommend products considering their style preferences and current demand. Recent studies [8] showed that incorporating a user's style into recommendations had mitigated the popular item bias problem in some recommendation domains.

Manuscript submitted to ACM

Moreover, in fashion we need to inspire the customer to discover pieces that resonate with their own style and preferences. This has been a mission at Farfetch, an online luxury fashion retail platform that sells products from thousands of boutiques and brands across the world.

Our fashion experts have defined a set of key aesthetic concepts aiming to reflect our current and target customers' style trends. For females, there are six aesthetics: Arty, Classic, Edgy, Feminine, Minimal and Streetwear. While for males, four aesthetics: Edge, Minimal, Smart and Streetwear (examples in Figure 1). Each of these themes have a product listing page and when customers navigate the page they can start to browse a list of products associated with each of the aesthetics.

Fig. 1. Street aesthetic and feminine aesthetic wear examples

This is a great way to capture a customer's style interests. Based on this data, we build models to identify customer's aesthetics and extend the predictions to the rest of our customer base. Once we know which aesthetics a customer belongs to, we not only can identify our customer "neighbors" more accurately, but also can suggest products more tailored to their aesthetic(s). Similarly to [8], we could add this aesthetic as a feature for our recommendation engine.

In this paper, we present how we can identify customers' aesthetics from their online shopping behaviors and the products they have shown interests in. We compare the performance of our models when using 3 completely different sets of features, namely: text, image, and inferred user statistics. We, also, looked at different combinations of those features. On top of this, we explore the performance difference in using a popular classification algorithm – Random Forest (RF), and a more recent deep learning algorithm – Convolutional Neural Network.

We defined this classification problem as a multi-label classification, since customers can show their interests to multiple aesthetics. For example, a customer can associate thyself with both being Classic and Minimal. When a customer is purchasing for others, the aesthetic interests can be very different, they can even be aesthetics from a different gender as well.

Our results show that it is possible to identify a customers' aesthetic based on their navigational patterns. Moreover, we found that the use of the textual descriptions of products interacted by the customer led to better classification results.

The main contributions of our work are twofold: (1) a comprehensive characterization of fashion customers based on their behavior on our platform; (2) a comparison of various models over a rich set of features, capable of classifying a customer into a predefined aesthetic.

The rest of this paper is as follows. Section 2 discusses some related work, while Section 3 defines our methodology. Our experimental evaluation is discussed in Section 4. Section 5 summarizes the paper and outlines our future work.

2 RELATED WORK

Multi-label classification is the task of assigning a subset of predefined categories to a given item. Classical approaches are based on binary relevance learning (i. e., construct a binary classifier for each category) [3] or a label powerset, by transforming the problem into a multi-class problem with one multi-class classifier trained on all unique label combinations found in the training data [10]. There is more effort focusing on using deep neural networks in recent works. Nam et al [17] show that a simple NN model trained using cross entropy loss performs, as well as, or even outperforms, state-of-the-art approaches on various textual datasets. Liu et al [11] present a deep learning approach, based on a Convolutional Neural Network (CNN) model tailored for multi-label classification to tackle the problem of Extreme multi-label text classification (when the number of labels is very high). They show that the proposed CNN approach is scalable to large datasets, and produce competitive to superior results with other state-of-the-art in literature. Here, we compare both classical methods and NN ones in our fashion domain, discussing the pros and cons of each one.

Incorporating user's style into recommendations has been delivering promising results on mitigating popular item bias, for example, Iqbal et al [8] incorporates user style into a Variational Autoencoder recommendations framework, and found that this addition allowed more diverse recommendations while maintaining relevance in e-commerce context. There are, also, some studies which tackle the problem of assigning a style to a cloth image [5, 7, 12, 20]. For example, Hadi et al [12] created a crowd sourced dataset to classify clothes according to five different styles. Hsiao et al [7] propose an unsupervised approach to learn a style-coherent representation for items. The method leverages probabilistic polylingual topic models based on visual attributes to discover a set of latent style factors. Unlike them, in this work, we aim to assign a style to a user, not an image. Other works [9, 13, 23] focused on building representations for items that capture somehow the style of the clothes. We focused on understanding the user and their aesthetic preferences.

3 METHODOLOGY

Out of many algorithms that work on multi-label classification, we have selected Random Forest which has been a popular choice and we trusted it to give us a good baseline, and a deep learning model with CNN as in Liu et al [11], which has shown promising results.

Since aesthetics are gender-specific, and also a customer could be purchasing products for others, especially from another gender, we have decided to model customers based on the product gender they have interacted with regardless of the customers' gender. As a result, a customer can be modelled for both Female and Male aesthetics. Unfortunately without purchase context or intention data, we are not able to divide the modelling into a more refined manner. Those kinds of context data are very hard to obtain in an e-commerce environment without disturbing customers' shopping experience.

Having those two main algorithms in mind and a pool of customers with indication of their aesthetic preferences, the rest is an open question on how to choose the training features for the prediction and what variations to explore to obtain the best offline model possible. In this work, we tried 3 different sets of features:

- Users statistics, the categories and brands they have interacted with

- Image embeddings
- Word embeddings

The rest of this section will explain in detail how we used each feature with each of our two classification algorithms.

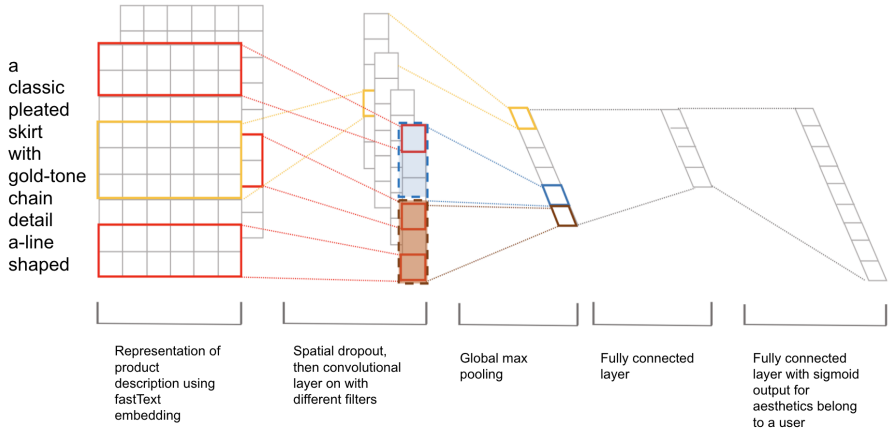

Fig. 2. Architecture used in this paper, graph adapted from [11]

3.1 General Users Statistics Model

In the users' statistics type of features, we have aggregated features per customer and applied the Random Forest algorithm to build a classifier. The features used include: the number of sessions, number of clicks (within six months), number of orders (within two years), number of returned items, the average discount a customer's purchases, the total gross margin values, etc. We also included the categories and the brands a user has interacted within the selected date range. Bearing in mind the curse of dimensionality problem, as the categories and brands data suffer a typical long-tail problem (i.e., the most popular categories cover the majority of the clicks). In this work, we only used the top 100 most popular categories, and the top 100 most popular brands as features.

When using categories or brands as features, their values are normalized weighted actions counts, here we use a category as an example:

$$V(u, cat) = \frac{\sum_{a \in allActions} w_a * Count(a, u)}{\sum_{a \in allActions} w_a} \quad (1)$$

where, u is representing the customer, a is a particular action (i.e. click, order, add to wish list, add to bag, return) and $Count(a, u)$ the count of how many times the customer performed this action. Each of the actions has a weight w_a based on its importance on our platform. Generally, the weight for click action is the smallest, and order action is the largest.

3.2 Image Embedding Model

CNN have been successful in solving computer vision problems in recent years [6] [21] [22]. There are a few well-known network architectures such as VGG16, VGG19 [21], ResNet50 [6], Inception V3 [22] pre-trained on the ImageNet dataset.

In this work, we used the last convolutional layer of ResNet50 for the image feature extraction due to our previous success in other projects in practice. All the product image embeddings are represented as a 1x2048 dimensional vector.

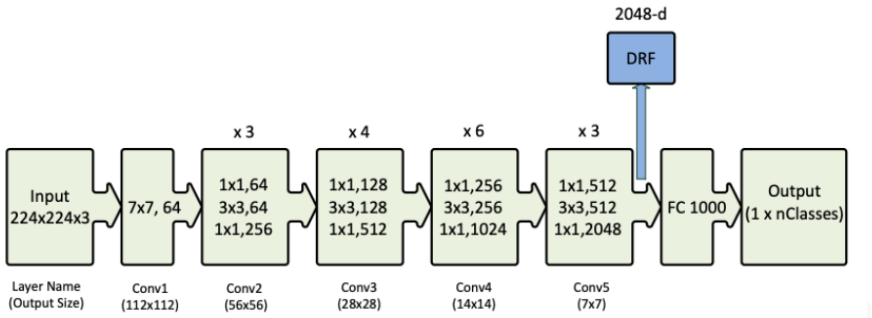

Fig. 3. ResNet50 architecture with residual units, the size of filters and the outputs of each convolutional layer [6] [15] .

Here, image embeddings were treated in different ways in order to extract relevant information so that they can be used as features. First, we tried to aggregate all the products image embeddings together per user, i.e., calculating the average, minimum, maximum or quantile of the embeddings of all the products a user has interacted with. This will finally provide a 2048 vector per user, which is then used as a feature.

We also tried to cluster all the products within each category when using image embeddings, on the assumption that the products of a category belonging to a certain aesthetic may share some visual similarity. For example, 3 clusters were formed under the Tops category, and named as cluster_tops_1, cluster_tops_2, and cluster_tops_3 , and these 3 names will be features for the Tops category. Hence, the final features will be all the clusters out of all the categories, and the values of these features per user are whether that user has interacted with a product from that cluster or not. In this work, K-means clustering and also variations of the image embedding dimensions using PCA are also experimented. Both sets of image features are used as input for Random Forest algorithm.

3.3 Word Embedding Model

Product descriptions are used as another alternative to generate features on the assumption that there would be some indication of the style of products from the way they are described. We first tried to use term frequency - inverse document frequency (TF-IDF) to prepare the values of each token. The tokens are generated through a series of NLP processes such as converting words to lower case, removing punctuation, stemming and finally transforming to tokens.

In the next iteration, FastText [1] was used to generate the embeddings for each word. Since most brands in our text are non frequent occurrences, FastText is a better choice in comparison with word2vec [16] or GloVe [19]. FastText represents words as the sum of a bag of characters of n-gram.

FastText vectors are then used as input to Random Forest and we also trained a CNN model using them. The architecture for training this neural network can be observed in figure 2.

The model consumes the embedding of the words/tokens from each product description out of all the interacted products per user using a pre-trained FastText model, which is followed by a spatial dropout and a convolutional layer with different filters. Then a global max pooling layer, which was flattened out and passed to a fully connected layer.

Finally, we reach the output layer corresponding to the number of aesthetics. The multi-label loss function following the equation 2 [11]:

$$\min_{\Theta} -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L} y_{ij}\log\left(\hat{p}_{ij}\right) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in \boldsymbol{y}_i^+}\frac{1}{\left|\boldsymbol{y}_i^+\right|}\log\left(\hat{p}_{ij}\right) \tag{2}$$

where Θ represents model parameters, $\boldsymbol{y}_i^+$ represents the set of relevant labels of instance i and $\left(\hat{p}_{ij}\right)$ is the model prediction for user i on label j.

4 EXPERIMENTS AND RESULTS

4.1 Dataset and Evaluation

Each of the aesthetic concepts has its own URL link which will lead to a list of products associated with that aesthetic on Farfetch website. We defined that a customer is interested in some aesthetic when they navigated to an aesthetic listing page and then proceeded to click in at least one of the listed products, considering the last six months period. For simplicity, only information about Female Aesthetics are shown here. As you can see in Figure 4, the total number of users that interacted with each of the aesthetic concepts is quite balanced, with Streetwear having the highest number of users (more than 7000) and Artistic the lowest numbers of users (a few more than 5000).

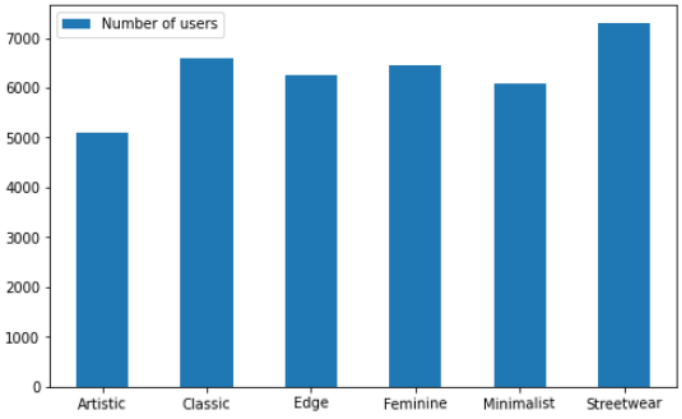

Fig. 4. Total number of users that interacted with each female aesthetic concept

There are 31,900 users that clicked in aesthetic listing pages from female gender and interacted with the products presented in those pages. Typically, 50% of those users have done more than 230 actions that can be just clicking in the product, adding it to the wishlist/bag or purchasing it.

Nonetheless, most of those actions seem to be related to products belonging to the same aesthetic concept as, on average, each user has interacted with 1.18 aesthetic concepts, as shown in Figure 5. In fact, more than 85% of the users only interacted with one aesthetic and respective products.

Moreover, as shown in Figure 6, it is clear that the aesthetic concepts do not have any correlation between them, i.e. the labels are independent of each other. This proves to be very important when it comes to the modelling task because it supports the idea of using multiple binary classification models (one for each label), rather than training a single multi-label model.

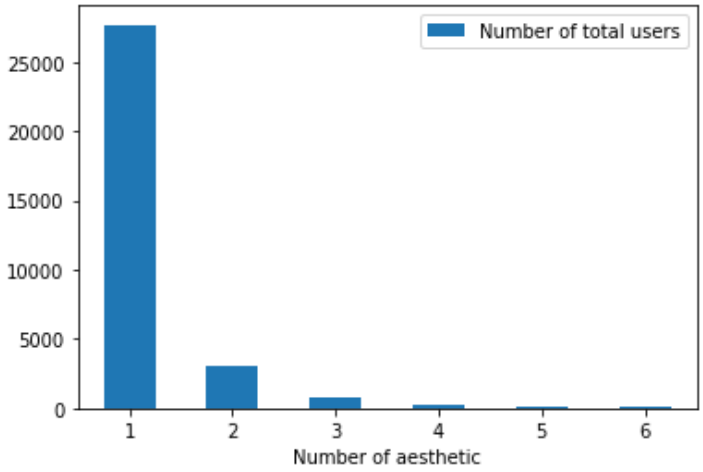

Fig. 5. Total number of users by number of aesthetic concepts they interacted with

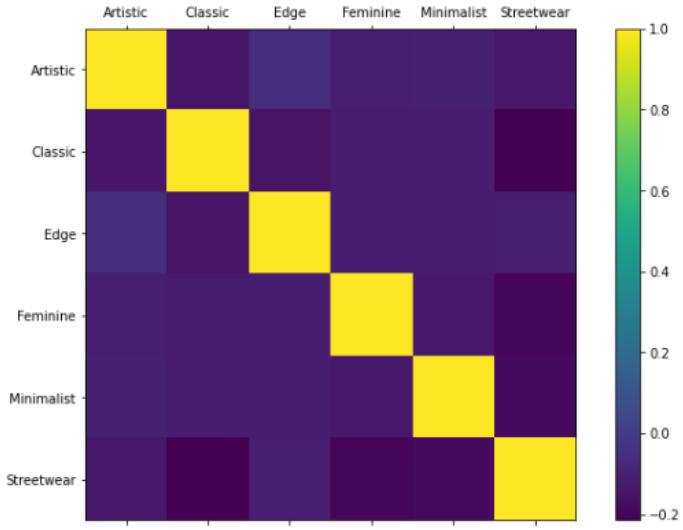

Fig. 6. Correlation between aesthetic concepts (labels)

We randomly splitted our data in 75% for training and 25% for testing, in a 5-fold cross-validation setting. Precision, Recall and F1 were used as the evaluation metrics. In the case of binary models, we used micro aggregation on those three metrics.

4.2 Discussion

Experiments were carried out testing on a few main variations, aiming to help us to understand what could be the best model for productization. Only the results from modelling Female Aesthetics are shown here, for the interest of clarity. The Male Aesthetics results achieved similar outcomes when it comes to modelling purposes.

4.2.1 Multi label configuration. We conducted preliminary experiments to determine what is the best technique to approach our multi label problem. We tested with both classical approaches (binary relevance and label powerset) and, also, with the default *scikit-learn* implementation of random forests (that supports multi label classification[1]).

Overall, models trained using binary relevance strategy performed slightly better than models that employed the power set technique. To our surprise, we noted that in some cases, the *scikit-learn* multi-label model had very odd performance. For example, when using TF-IDF features, the recall is almost 1, i.e. almost all the predictions are 1 for all classes. This indicates that just averaging impurity reduction across all the outputs could not be enough to build a reliable multi label model for our scenario. So, we decided to use binary relevance for training all the Random Forest models.

4.2.2 Dealing with data imbalance. When creating a classifier for each class, our dataset may become imbalanced, as all other classes will represent negative samples. As shown in Figure 5, the percentage of users having more than one aesthetic is small, less than 15% of the whole population. Therefore, 85% of the samples are considered the negative class. This could cause imbalance problems models, so we tested with some strategies to deal with this problem.

We first tried to use the "class_weight" parameter in Random Forest implementation to automatically adjust the class weights inversely proportional to class frequencies in the input data [18]. We also tried to use SMOTE [2] which aims to create synthetic data to help to reduce the data imbalance problem. Our results, summarized in Table 1, confirm that there is a significant improvement when we balanced the dataset, either using class weight or SMOTE. In fact, the first approach seems to produce better results. We did try different options under the SMOTE domain, all of which produced very close results. This set of experiments were all carried out using general stats features in Random Forest binary classifiers, although very similar results could be obtained in each of the feature sets.

Table 1. Evaluation results when applying different treatments for data imbalance with general stats in binary Random Forest classification models

Treatment	F1	Precision	Recall
No treatment	0.043	0.771	0.022
Class_Weight Balanced	0.442	0.360	0.572
SMOTE	0.400	0.345	0.476
SMOTE combined with Class_Weight Balanced	0.402	0.347	0.478

4.2.3 Impact of the choice of training features. This set of experiments looks into how the same model performs under different feature sets. We trained one model for each feature set combination for female aesthetics using Random Forest. We also trained an additional Random Forest model over a random generated feature set. This could be seen as a lower bound for our metrics (in fact, as expected, all the models outperformed the random one). Table 2 shows the approximate size of each feature set.

[1]It builds a single generalized model capable of processing output correlations. To build a tree, it uses a multi-output splitting criteria computing average impurity reduction across all the outputs. To the best of our knowledge, this could be viewed as a kind of greedy label powerset technique.

Table 5 shows our results. In general, word embedding (TF-IDF) performs better than other models using a single feature set, without any dimensionality reduction. From the reasoning of features choices, it makes sense that features generated from product descriptions perform well, since some words could be a strong indication of aesthetics. Interestingly, when using FastText the results decrease, this could indicate that there are some words with a special meaning in the fashion domain that have not been captured by an embedding model trained over a more general text dataset. An in depth analysis of this fact is out of scope of our objective and is defined as a future work.

The general stats feature model results seem promising, particularly on recall, indicating that users in a particular customers' aesthetics share some navigation patterns.

On the other hand, models based on image features are performing the worst. The main reason might be because in the same aesthetic concept those features can be too generic since the products can look very different, i.e. with different patterns, different shapes, etc.

Finally, some variations we tried seemed to show no sensitivity in the outcome, as we can see in the image embedding and general stats combination (Table 5). This could indicate that those features do not have complementary information that could be exploited by the Random Forest algorithm.

4.2.4 CNN results. Deep learning models seem to be another promising way of modelling, especially with the recent multi-label classification development [11]. With that in mind, we compared Random Forest with CNN (Section 3.2) in a multi-label classification setting. To do so, we chose the best Random Forest model trained over a single feature set – Word embedding (TF-IDF) – and tested it against the CNN model trained on the same feature set.

As you can see on Table 5, the deep learning models tend to need more data, we tested using the last 3 and 6 months of products that a user has interacted with. This increase in data had little impact on Random Forest, but helped CNN to significantly improve recall. In general, the results seem promising. CNN models performed better on precision at the cost of recall leading to a worst F1 when compared with Random Forest. It is possible that with experiments using a bigger dataset the CNN model can be further improved. Also, our CNN model is not fully optimized, in particular, we did not conduct a deep study on the impact of the class imbalance on the CNN model which may affect its performance. In future iterations, we plan on exploring different techniques to tackle this problem.

4.2.5 Different loss functions. We have hypothesised over using binary cross entropy or categorical cross entropy as the loss function in a multi-label classification setting. The authors in [11] used binary cross entropy, on the other hand [14] mentioned that categorical cross entropy seems to perform better. We tried both in our CNN model and, in our case, the binary cross entropy loss function had a better F1 score.

4.2.6 Aesthetic result breakdown. In order to evaluate how well the best model – Random Forest with word embeddings (TF-IDF) – perform in each Aesthetic label, we present the F1, Precision, and Recall on Table 6. We can see that there

Table 2. Approximate size of each feature set.

Feature set	**Approximate size**
General stats	300
Image embedding (clusters)	2048
Word embedding (TF-IDF)	455K
Word embeddings (FastText)	300

Table 3. Evaluation results when using different sets of training features with binary Random Forest classification models.

Features	F1	Precision	Recall
Word embedding (TF-IDF)	0.525	0.586	0.476
Word embedding (TF-IDF) + Image embedding	0.524	0.585	0.474
General Stats + Word embedding (TF-IDF) + Image embedding	0.507	0.513	0.500
General Stats + Word embedding (TF-IDF)	0.503	0.521	0.486
General Stats	0.442	0.360	0.572
General Stats + Image embedding	0.420	0.382	0.467
Word embeddings (FastText)	0.418	0.335	0.554
Image embedding (clusters)	0.348	0.295	0.424
Random	0.257	0.197	0.370

Table 4. Evaluation results when using Random Forest comparing to using a CNN deep learning model with different time length in multi-label classification.

Data time range	Features	Algorithm	F1	Precision	Recall
6 months	Word embedding (TF-IDF)	RF	0.525	0.586	0.476
3 months	Word embedding (TF-IDF)	RF	0.505	0.555	0.463
6 months	Word embedding - FastText	CNN	0.404	0.680	0.288
3 months	Word embedding - FastText	CNN	0.307	0.687	0.199

Table 5. Evaluation results when using different loss functions for CNN deep learning model trained using word embedding - FastText.

Loss function	F1	Precision	Recall
Binary cross-entropy	0.404	0.680	0.288
Categorical cross-entropy	0.284	0.702	0.178

is a positive correlation between class frequency and the evaluation metrics which might be an indicator that more popular aesthetics are easier to classify.

Table 6. Evaluation results breakdown by Female Aesthetic for Random Forest with word embeddings (TF-IDF).

Aesthetic	Frequency	F1	Precision	Recall
Streetwear	0.230	0.635	0.739	0.557
Classic	0.207	0.589	0.671	0.524
Feminine	0.203	0.483	0.554	0.428
Edge	0.196	0.484	0.551	0.431
Minimalist	0.192	0.499	0.583	0.437
Artistic	0.158	0.442	0.525	0.383

As you can see from the example of Figure 7, a user that was interested in the products presented was correctly labeled as belonging to the Feminine Aesthetic. Looking at the products, we can say that they look like very feminine products, however, the images are very diverse between each other. This observation, might sustain the fact that image embeddings do not provide useful information to the models when compared with other features.

On the other hand, if we look at the product descriptions, we can see that there are some words in common between products like "midi" and "dress" that were crucial for the model to classify the Aesthetic correctly.

Fig. 7. List of products that a user, correctly labeled as Feminine Aesthetic, interacted with.

5 CONCLUSION AND FUTURE WORK

We have explored identifying customers' style preference through aesthetic concepts in various ways. We demonstrated that those aesthetics could be inferred from the customer's online shopping behaviors and the products they have shown interests in. In the end, using Random Forest with binary relevance to tackle this multi-label problem seems like the best option for the dataset we currently have available. Also, using text features generated from product description seems to have a better performance when compared with other feature sets (image embedding, user general stats).

As a future work, we will carry out a customer survey aiming to collect more data on our customers perception about their aesthetics. With this type of data, we can evaluate our models with customer survey data to see whether the results will align with what we have found in this paper. Moreover, we will further explore our CNN model, in particular, we want to study the class imbalance problem and the reason why FastText embedding performs worse than the TF-IDF feature set.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606* (2016).

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2020. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv preprint arXiv:1106.1813* (2020).

[3] Krzysztof Dembczynski, Weiwei Cheng, and Eyke Hüllermeier. 2010. Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains.. In *ICML*, Johannes Fürnkranz and Thorsten Joachims (Eds.). Omnipress, 279–286. http://dblp.uni-trier.de/db/conf/icml/icml2010.html#DembczynskiCH10

[4] Daniel Faggella. 2017. The ROI of recommendation engines for marketing. https://martechtoday.com/roi-recommendation-engines-marketing-205787. Accessed: 2020-07-29.

[5] Diogo Gonçalves, Liweu Liu, and Ana Magalhães. 2019. How big can style be? Addressing high dimensionality for recommending with style. *CoRR* abs/1908.10642 (2019). http://dblp.uni-trier.de/db/journals/corr/corr1908.html#abs-1908-10642

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[7] Wei-Lin Hsiao and Kristen Grauman. 2017. Learning the Latent "Look": Unsupervised Discovery of a Style-Coherent Embedding from Fashion Images.. In *ICCV*. IEEE Computer Society, 4213–4222. http://dblp.uni-trier.de/db/conf/iccv/iccv2017.html#HsiaoG17

[8] Murium Iqbal, Kamelia Aryafar, and Timothy Anderton. 2019. Style Conditioned Recommendations. *CoRR* abs/1907.12388 (2019). http://dblp.uni-trier.de/db/journals/corr/corr1907.html#abs-1907-12388

[9] Hanbit Lee, Jinseok Seol, and Sang-goo Lee. 2017. Style2Vec: Representation Learning for Fashion Items from Style Sets. *CoRR* abs/1708.04014 (2017). arXiv:1708.04014 http://arxiv.org/abs/1708.04014

[10] Feng Liu, Xiaofeng Zhang, Yunming Ye, Yahong Zhao, and Yan Li. 2015. MLRF: Multi-label Classification Through Random Forest with Label-Set Partition. In *Advanced Intelligent Computing Theories and Applications*, De-Shuang Huang and Kyungsook Han (Eds.). Springer International Publishing, Cham, 407–418.

[11] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep Learning for Extreme Multi-label Text Classification. 115–124. https://doi.org/10.1145/3077136.3080834

[12] Alexander C. Berg Tamara L. Berg M. Hadi Kiapour, Kota Yamaguchi. 2014. Hipster Wars: Discovering Elements of Fashion Styles. In *European Conference on Computer Vision*.

[13] Ana Rita Magalhães. 2019. The trinity of luxury fashion recommendations: data, experts and experimentation.. In *RecSys*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 522. http://dblp.uni-trier.de/db/conf/recsys/recsys2019.html#Magalhaes19

[14] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the Limits of Weakly Supervised Pretraining. *arXiv preprint arXiv:1805.00932* (2018).

[15] Ammar Mahmood, Ana Giraldo Ospina, Mohammed Bennamoun, Senjian An, Ferdous Sohel, Farid Boussaid, Renae Hovey, Robert B. Fisher, and Gary Kendrick. 2020. Automatic Hierarchical Classification of Kelps using Deep Residual Features. *arXiv preprint arXiv:1906.10881* (2020).

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. (2013), 1–12. arXiv:1301.3781 http://arxiv.org/abs/1301.3781

[17] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-Scale Multi-label Text Classification — Revisiting Neural Networks. In *Machine Learning and Knowledge Discovery in Databases*, Toon Calders, Floriana Esposito, Eyke Hüllermeier, and Rosa Meo (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 437–452.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[20] Edgar Simo-Serra and Hiroshi Ishikawa. 2016. Fashion Style in 128 Floats: Joint Ranking and Classification Using Weak Data for Feature Extraction.. In *CVPR*. IEEE Computer Society, 298–307. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2016.html#Simo-SerraI16

[21] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[22] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv preprint arXiv:1512.00567* (2015).

[23] Andreas Veit, Balazs Kovacs, Sean Bell, Julian J. McAuley, Kavita Bala, and Serge J. Belongie. 2015. Learning Visual Clothing Style with Heterogeneous Dyadic Co-Occurrences.. In *ICCV*. IEEE Computer Society, 4642–4650. http://dblp.uni-trier.de/db/conf/iccv/iccv2015.html#VeitKBMBB15