Incorporating Customer Reviews in Size and Fit Recommendation systems for Fashion E-Commerce

Oishik Chatterjee
Flipkart Internet Private Limited
Bengaluru, India
oishik.chatterjee@flipkart.com

Jaidam Ram Tej
Flipkart Internet Private Limited
Bengaluru, India
jaidam.ramtej@flipkart.com

Narendra Varma Dasaraju
Flipkart Internet Private Limited
Bengaluru, India
narendra.varma@flipkart.com

ABSTRACT

With the huge growth in e-commerce domain, product recommendations have become an increasing field of interest amongst e-commerce companies. One of the more difficult tasks in product recommendations is size and fit predictions. There are a lot of size related returns and refunds in e-fashion domain which causes inconvenience to the customers as well as costs the company. Thus having a good size and fit recommendation system, which can predict the correct sizes for the customers will not only reduce size related returns and refunds but also improve customer experience. Early works in this field used traditional machine learning approaches to estimate customer and product sizes from purchase history. These methods suffered from cold start problem due to huge sparsity in the customer-product data. More recently, people have used deep learning to address this problem by embedding customer and product features. But none of them incorporates valuable customer feedback present on product pages along with the customer and product features. We propose a novel approach which can use information from customer reviews along with customer and product features for size and fit predictions. We demonstrate the effectiveness of our approach compared to using just product and customer features on 4 datasets. Our method shows an improvement of 1.37% - 4.31% in F1 (macro) score over the baseline across the 4 different datasets.

ACM Reference Format:
Oishik Chatterjee, Jaidam Ram Tej, and Narendra Varma Dasaraju. 2022. Incorporating Customer Reviews in Size and Fit Recommendation systems for Fashion E-Commerce. In *Proceedings of SIGIR eCom'22, SIGIR Workshop on eCommerce (SIGIR eCom'22).* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Online fashion market is expected to grow at 11.4% per year [21]. Returns, where size issue is a considerable piece of the pie, are the bane for the fashion e-commerce companies. Up to 40% of online fashion products are returned [3]. In in-store purchases, consumers prefer to see, touch, and try-on apparel before purchasing. They lack similar engaging experience in online shopping.

In online shopping, consumers rely on symbolic sizes (e.g. 'S', 'M', 'L') to make their purchase decisions. Though symbolic sizes are mentioned in products, they vary between brands. Sometimes there is size variation within a brand [20]. Consumers also use size guides. Size guides provide a mapping from the standard sizes to corresponding physical sizes, in cm or inches. There are multiple symbolic sizing schemes (e.g. 'US', 'UK', 'EU') in a size guide. These size guides are usually at a brand level and do not capture finer fit details of a product. They are cumbersome to enact and require measuring instruments at their disposal. Further, due to vanity sizing, consuming symbolic sizes can be tricky. Thus, sizes mentioned on the products are no longer enough to make a purchase.

Customers when buying a product look for the products with the right fit. They usually return the product if it does not fit. It is important that E-commerce platforms provide accurate recommendation tips to customers. This helps the customers in three ways. Firstly it helps in reducing the customers' returns. Secondly, it eases customers in finding their right fit enriching their shopping experience online and hence might boost conversions. Thirdly it helps in building customer loyalty. Hence we need a good size and fit recommendation system which can help customers in choosing the right fit.

There have been various solutions proposed for size and fit recommendation. Most of the earlier solutions used traditional approaches to embed customer product transactions to determine the right size for the customers. These approaches suffer from the cold start problem because the customer-product transaction data is sparse. Recently deep learning approaches [3, 4, 20] which use customer and product features along with transaction data have been proposed which tries to mitigate this problem. None of the current approaches use the customer reviews present on the product page. These reviews often contain information that can help in predicting size/fit of new customers. When customers return a product they might give some information in the review which indicates the size/fit of the product. Though [1] uses reviews to do fit prediction, they only classify each review as small fit or large. Such a model alone cannot help in recommending size for a new customer. Hence we propose a Deep Learning based approach that uses customer reviews along with product and customer information to predict the right fit.

Our contributions are :

- We propose a novel approach of leveraging size and fit information in customer reviews present on E-Commerce platforms for size and fit predictions.
- We demonstrate how user reviews given by customer on product pages of E-Commerce platforms can be embedded using state-of-the-art pre-trained language models (such as BERT [2]) and used along with product and customer features for improving size and fit prediction models. To our

SIGIR eCom'22, July 15 2022, Madrid, Spain

knowledge, this is the first work which utilizes both product-customer features along with customer reviews for doing size and fit predictions.
- We empirically show on 4 datasets curated using the data of one of the largest e-commerce platforms that using size and fit information present in customer reviews does indeed help in predicting correct fit for new customers.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 describes the problem formulation for size and fit recommendation system. Section 4 describes our experiments and datasets used. Experimental results are reported in Section 5 and finally we conclude in section 6.

2 RELATED WORK

In literature, of late there has been a lot of focus on the size and fit problem [18, 19]. Abdulla et al. [17] embed both users and products using skip-gram based Word2Vec model [15] and employ GBM classifier [5] to predict the fit. A latent factor model was proposed by Sembium et al. [18], which was later follow-up by a Bayesian formulation [19] to predict the size of a product (small, fit, large). In [19] Bayesian logistic regression with ordinal categories was used. They proposed an efficient algorithm for posterior inference based on mean-field variational inference and Polya-Gamma augmentation. Guigourès et al. employed a hierarchical Bayesian model [6] for personalized size recommendation. Misra et.al. [16] learn the fit semantics by modeling it as an ordinal regression problem. Then, they employ metric learning techniques to address the class imbalance issues.

Recently, deep learning approaches have been used to solve the size recommendation problem with encouraging results [3, 12, 20]. Deep Learning approaches unlike the traditional approaches are able to scale well with large amounts of data. SFNET [20] provides recommendations at the user cross product level using a deep learning based content collaborative approach. The approach can learn from cross-correlations that exist across fashion categories. They use both purchase and returns data as well as customer and article features for personalized size and fit prediction. Dogani et al. [3] addressed the sparsity problem by learning latent representation at a brand level using neural collaborative filtering [8]. Then, fine-tuning the product representation by transfer learning from brand representation. Lasserre et al. [12] use a deep learning based meta learning approach. Their approach is based on the premise that, given the purchase history of a customer i, products x_j and their corresponding size estimates y_{ij}) share a strong linear relationship. Baier et al. [1] derive product fit feedback from customer reviews using natural language processing techniques which is then used to infer the right fit.

A few approaches explore the use of product images or 3D scan of products to predict the right fit for a customer. SizeNet [10] uses product images to infer whether the product will have fit issues for a customer. ViBE uses a computer vision approach to develop a body-aware embedding that captures garment's affinity with different body shapes [9]. The approach learns the embedding from images of models of various shapes and sizes wearing the product, displayed on catalog. [4] proposes PreSizE – a size prediction framework that utilizes Transformers to capture the relationship between various item attributes (e.g., brand, category, etc.) and its purchased size by encoding items and user's purchase history.

In industrial applications, customers are looped in for size recommendation. These applications ask targeted questions to acquire biometric measurements, e.g. height, weight, age, waist of the customer. They also ask queries related to previous purchases, e.g. "Which brand and size of shoe do you find most comfortable?". These data points are further used to predict for cold start customers. However, the downside is that it may add friction in the consumer path.

In this work, we focus on using customer reviews, purchases and returns data of the customers, product information in predicting the right fit for a customer. Customer reviews might contain crucial size and fit information which might help in predicting the right fit to a customer over using product and customer information alone. To extract the information from customer reviews we have embedded the reviews using a pre-trained language model [2]. A more detailed description of the model is discussed in Section 3.

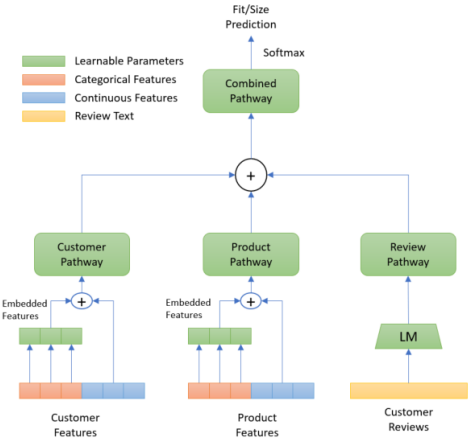

Figure 1: Architecture

3 SIZE AND FIT RECOMMENDATION

We model the size and fit problem as a classification problem where given a product and a customer we want to predict if the product will 'fit' the customer, or it will be 'small' or 'large' for the customer. The following subsections explain the problem formulation and the model architecture.

3.1 Problem Formulation

Given a customer $\mathbf{C}$, a product $\mathbf{P}$, and a set of reviews for the product $\mathbf{R}$, we want to predict if the product will be small, fit or large for the customer. Both product and customer are defined by their respective features $\mathbf{P} = \{p_i\}$ and $\mathbf{C} = \{c_i\}$ where each feature can be either continuous or categorical. $\mathbf{R} = \{r_i\}$ consists of the reviews left by customers on the product page. We define the output space as $\mathbf{F} = \{small, fit, large\}$. We want to learn the probability distribution of $\mathbf{F}$ given $\mathbf{C}$, $\mathbf{P}$, and $\mathbf{R}$ i.e. $p_\theta(f|\mathbf{C}, \mathbf{P}, \mathbf{R})$ where θ denotes the model parameters.

Table 1: Dataset Statistics

Description	Women Jacket	Women Kurta and Kurti	Mens Jean	Mens Polo Tshirt
No. of Customers	14,94,713	61,53,598	31,72,637	21,18,089
No. of Products	47,040	1,20,425	47,140	47,040
% "Small" Instances	4.52%	2.40%	2.92%	3.14%
% "Large" Instances	1.68%	1.30%	2.35%	2.76%

3.2 Neural Network architecture

The network architecture consists of 3 input pathways for customer, product and review inputs which are then followed by a combined pathway that outputs the final prediction. Each pathway consists of a series of residual blocks [7].

Firstly for the product input features, the numerical features like size are normalized and the categorical features like brand, fabric etc are converted to vectors using embedding layers. These are then concatenated together ($\boldsymbol{h_p}$) and passed through a residual block (product input pathway) to generate product embedding ($\boldsymbol{e_p}$). The same is done for the customer features to generate customer embedding($\boldsymbol{e_c}$). For the reviews, each review is encoded using a pre-trained language model (more details in section 3.3). The embeddings are then averaged and passed through another residual block [7] similar to the ones used for customer and product input pathways.

$$h_{pi} = \begin{cases} NormalizationLayer(p_i) & \text{if } p_i \text{ is numerical} \\ EmbeddingLayer(p_i) & \text{if } p_i \text{ is categorical} \end{cases} \tag{1}$$

$$\boldsymbol{h_p} = \bigoplus h_{pi} \tag{2}$$

$$\boldsymbol{e_p} = \text{Product Input Pathway}(\boldsymbol{h_p}) \tag{3}$$

$$h_{ri} = \text{LanguageModel}(r_i) \tag{4}$$

$$\boldsymbol{h_r} = \text{Mean}(h_{ri}) \tag{5}$$

$$\boldsymbol{e_r} = \text{Review Input Pathway}(\boldsymbol{h_r}) \tag{6}$$

$$h_{ci} = \begin{cases} NormalizationLayer(c_i) & \text{if } c_i \text{ is numerical} \\ EmbeddingLayer(c_i) & \text{if } c_i \text{ is categorical} \end{cases} \tag{7}$$

$$\boldsymbol{h_c} = \bigoplus h_{ci} \tag{8}$$

$$\boldsymbol{e_c} = \text{Customer Input Pathway}(\boldsymbol{h_c}) \tag{9}$$

The embedding from the pathways are combined as $[\boldsymbol{e_c}, \boldsymbol{e_p}, \boldsymbol{e_r}]$ and passed through the combined pathway. The combined pathway again consists of a series of residual blocks which is followed by a softmax layer.

$$\boldsymbol{e} = \boldsymbol{e_c} \oplus \boldsymbol{e_p} \oplus \boldsymbol{e_r} \tag{10}$$

$$\boldsymbol{o} = \text{Combined Pathway}(\boldsymbol{e}) \tag{11}$$

$$p_\theta(f|\mathbf{C}, \mathbf{P}, \mathbf{R}) = \text{softmax}(\boldsymbol{o}) \quad f \in F \tag{12}$$

$$y = \underset{f \in F}{argmax}\ p_\theta(f|\mathbf{C}, \mathbf{P}, \mathbf{R}) \tag{13}$$

Table 2: Model Learning Parameters

Parameter	Details
Batchsize	2048
Learning rate	0.01
Optimizer	Adam
Review embedding dimension	768
Input Pathways	#(emb + cont) feat x 25 x 15 x 10
Final Pathway	50 x 100 x 200 x 500 x 3

The number of output labels is related to the reason codes provided by the customer. Customers return products when there are size and fit issues. They select the following reason codes: 'Size smaller' or 'Size larger'. Based on this premise our size recommendations are of the form 'buy one size small' or 'buy one size large'. When the product is true to size no recommendation is shared with the customer. In total there are three classes we wish to determine; small, large and fit corresponding to 'buy one size large', 'buy one size small' and true to size, respectively.

3.3 Review Embeddings

We embed reviews using a pre-trained language model before passing on to review input pathway. We use BERT [2] as the pre-trained language model here (we have experimented with different language models and found that BERT [2] gives the best results). We pre-train BERT [2] in a variety of ways and report numbers on each pre-trained model. We train some of the BERT [2] model on review fit classification task where each review is labeled {small, fit, large} which helps the model learn about the fit information present in the review text that can help in subsequent size and fit prediction task. More details is given in section 4.2. Each product may have zero, one or many reviews. Since we want a single review feature for each product, therefore in case of many reviews we aggregate the review embedding vectors of all the reviews for a given product to get a single vector. To do this we average over all the embedding vectors. In case of zero reviews we use a default vector of all zeros.

4 EXPERIMENTS

We demonstrate that size and fit prediction models can leverage information present in customer reviews to improve their performance by comparing our approach with SFNet [20] which uses only customer and product features on 4 different datasets (each a separate category) - **Women Jacket**, **Women Kurta and Kurti**, **Mens Jean** and **Mens Polo Tshirt**.

Table 3: F1-Macro score of all the models (OR: Only Reviews, BM: BERT-Modcloth, BFC: BERT-FC, FB: FKBERT, FFC: FKBERT-FC). The numbers denote the improvement in % over the baseline(SFNet).

Model	Women Jacket	Women Kurta and Kurti	Mens Jean	Mens Polo Tshirt
SFNet (Baseline)	38.26	35.6	37.19	37.2
OR	-15.68%	-8.13%	-12.77%	-13.12%
BM	+0.68%	+1.94%	+0.83%	+0.43%
BFC	+3.27%	+2.81%	+1.00%	+2.15%
FB	+1.93%	+1.57%	**+1.37**%	+2.34%
FFC	**+4.31**%	**+3.09**%	**+1.37**%	**+2.42**%

4.1 Dataset

We have collected product details, purchase and returns data, and customer reviews for the following categories: **Women Jacket**, **Women Kurta and Kurti**, **Mens Jean** and **Mens Polo Tshirt**. We label each purchase with small or large based on if the customer has returned the product stating small size or large size as the reason respectively. If the customer has not returned the product, then we label it as fit. Table 1 shows the data statistics of all the categories we have trained on. We have split each dataset in 80:10:10 train validation and test split. We also create a reviews dataset for each category which is used to train the BERT model [2] for embedding the customer reviews. Here each review is labeled as small, fit or large based on the criteria mentioned above. We also use a public dataset - Modcloth [16], for training the BERT model [2] to show the effectiveness of transfer learning when labeled reviews dataset is not available.

4.2 Models

We have trained the language model in the following ways:

- **BERT-Mocloth**: We have trained a BERT model [2] on the fit classification task with the Modcloth reviews dataset.
- **BERT-FC**: We have trained a BERT model [2] on the fit classification task with the reviews of the given dataset.
- **FKBERT**: We have trained a ROBERTa model [13] on various text e-commerce data such as product description, reviews, question-answers, etc. on the masked language modeling task.
- **FKBERT-FC**: We have fine-tuned the FKBERT model on the fit classification task with reviews of the given dataset.

We also report numbers on **Only Reviews** which uses the predictions by the bert model directly and determines the class of the product using majority voting of the review predictions. For the baseline we compare against **SFNet** model [20] which uses just customer and product features for prediction.

4.3 Hyperparameters, Training and Evaluation

We have used the same hyperparameters for all the datasets. Details of the hyperparameters are given in Table 2.

We have first trained the BERT [2] model on the reviews dataset. We have used AdamW [14] optimizer with a learning rate set to 2e-5 for training the BERT model.

For training the rest of the model, we have used Adam [11] optimizer with learning rate 1e-2. Since the dataset is highly skewed towards the fit class, hence we have re-sampled the fit class instances such that the number of fit instances in the training data is equal to the number of small + large instances. This prevents the model from over-fitting to the fit class. The model is trained for a maximum of 100 epochs and the epoch with the best validation F1 is chosen. For evaluation, we report the F1-macro scores.

5 RESULTS

We perform a set of experiments on the 4 different datasets mentioned in section 4.1. Our baseline model is the SFNET [20] model which does not use customer reviews. We report the improvement of F1 (Macro) score over the baseline (SFNet [20]) on the 4 models discussed in 4.2 in table 3. Results show that there is a significant improvement when reviews are used with other features. Even without labeled reviews data, we get **0.43% - 1.94%** improvement on the baseline model using transfer learning from public datasets. Also, we observe that pre-training with e-commerce textual data also improves performance as can be seen from **FKBERT** and **FKBERT-FC** numbers. The biggest improvement is shown when the language model is first pre-trained on generic e-commerce data and then fine-tuned on the fit classification task. **FKBERT-FC** shows **1.37% - 4.31%** improvement in macro F1 over the baseline. We also observe that reviews by themselves cannot be used alone for fit prediction as shown by the low numbers of the **Only Reviews** model.

6 CONCLUSION AND FUTURE WORK

Size and Fit recommendation is an important problem in e-commerce because it helps customers in choosing the right fit and thereby reduces size and fit related returns. Most of the earlier works embedded user and product information to predict the right fit. In this paper, we propose a deep learning based approach to predict the fit based on customer reviews along with customer and product information. Through extensive experimentation on different datasets curated from data of one of the largest e-commerce platforms, we show the effectiveness of our approach.

We plan to extend this work by including other customer feedback like customer uploaded images and customer QnA to predict the right fit. Customers sometimes ask questions related to size and fit which get answered by other customers who have already bought the product. Also using customer uploaded images along with product information might help the model to learn the interactions between user and product to predict the right fit.

REFERENCES

[1] Stephan Baier. 2019. Analyzing Customer Feedback for Product Fit Prediction. *CoRR* abs/1908.10896 (2019). arXiv:1908.10896 http://arxiv.org/abs/1908.10896

[2] 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).

[3] Kallirroi Dogani, Matteo Tomassetti, Saúl Vargas, Benjamin Paul Chamberlain, and Sofie De Cnudde. 2019. Learning Embeddings for Product Size Recommendations.. In *eCOM@ SIGIR*.

[4] Yotam Eshel, Or Levi, Haggai Roitman, and Alexander Nus. 2021. PreSizE: Predicting Size in E-Commerce using Transformers. *arXiv preprint arXiv:2105.01564* (2021).

[5] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[6] Romain Guigourès, Yuen King Ho, Evgenii Koriagin, Abdul-Saboor Sheikh, Urs Bergmann, and Reza Shirvany. 2018. A hierarchical bayesian model for size recommendation in fashion. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 392–396.

[7] 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.

[8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182.

[9] Wei-Lin Hsiao and Kristen Grauman. 2020. ViBE: Dressing for diverse body shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11059–11069.

[10] Nour Karessli, Romain Guigourès, and Reza Shirvany. 2019. Sizenet: Weakly supervised learning of visual size and fit in fashion images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[11] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Julia Lasserre, Abdul-Saboor Sheikh, Evgenii Koriagin, Urs Bergman, Roland Vollgraf, and Reza Shirvany. 2020. Meta-learning for size and fit recommendation in fashion. In *Proceedings of the 2020 SIAM international conference on data mining*. SIAM, 55–63.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[14] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[16] Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 422–426.

[17] G Mohammed Abdulla, Shreya Singh, and Sumit Borar. 2019. Shop your Right Size: A System for Recommending Sizes for Fashion products. In *Companion Proceedings of The 2019 World Wide Web Conference*. 327–334.

[18] Vivek Sembium, Rajeev Rastogi, Atul Saroop, and Srujana Merugu. 2017. Recommending product sizes to customers. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 243–250.

[19] Vivek Sembium, Rajeev Rastogi, Lavanya Tekumalla, and Atul Saroop. 2018. Bayesian models for product size recommendations. In *Proceedings of the 2018 World Wide Web Conference*. 679–687.

[20] Abdul-Saboor Sheikh, Romain Guigourès, Evgenii Koriagin, Yuen King Ho, Reza Shirvany, Roland Vollgraf, and Urs Bergmann. 2019. A deep learning system for predicting size and fit in fashion e-commerce. In *Proceedings of the 13th ACM conference on recommender systems*. 110–118.

[21] Statista. 2020. Fashion eCommerce report 2020. Retrieved Jan 21, 2021 from. *https://www.statista.com/study/38340/ecommerce-report-fashion/* (2020).