A Transformer-Based Substitute Recommendation Model Incorporating Weakly Supervised Customer Behavior Data

Wenting Ye
wentingye52@gmail.com
Amazon Retails
Seattle, WA, USA

Hongfei Yang
derekyang24@gmail.com
Amazon Retails
Seattle, WA, USA

Shuai Zhao
cheershuaizhao@gmail.com
Amazon Retails
Seattle, WA, USA

Haoyang Fang
kevinfang97@gmail.com
Amazon Retails
Seattle, WA, USA

Xingjian Shi
xshiab@connect.ust.hk
AWS AI
Santa Clara, CA, USA

Naveen Neppalli
naveen.neppalli@gmail.com
Amazon Retails
Seattle, WA, USA

ABSTRACT

The substitute-based recommendation is widely used in E-commerce to provide better alternatives to customers. However, existing research typically uses customer behavior signals like *co-view* and *view-but-purchase-another* to capture the substitute relationship. Despite its intuitive soundness, such an approach might ignore the functionality and characteristics of products. In this paper, we adapt substitute recommendations into language matching problem. It takes the product title description as model input to consider product functionality. We design a new transformation method to de-noise the signals derived from production data. In addition, we consider multilingual support from the engineering point of view. Our proposed end-to-end transformer-based model achieves both successes from offline and online experiments. The proposed model has been deployed in a large-scale E-commerce website for 11 marketplaces in 6 languages. Our proposed model is demonstrated to increase revenue by 19% based on an online A/B experiment.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Learning from implicit feedback**.

KEYWORDS

substitute recommendation; multilingual; weakly supervised learning; natural language processing; selection bias; implicit feedback

ACM Reference Format:
Wenting Ye, Hongfei Yang, Shuai Zhao, Haoyang Fang, Xingjian Shi, and Naveen Neppalli. 2023. A Transformer-Based Substitute Recommendation Model Incorporating Weakly Supervised Customer Behavior Data. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591847

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591847

1 INTRODUCTION

Substitute recommendations are widely adopted in E-commerce by giving customers more options, especially when the reference product is out-of-stock, higher-priced, or lower-rated [7, 10]. It improves the shopping experience and increases customer affinity and loyalty to the E-commerce service provider. It provides alternative products from a reference product, which can be regarded as <reference product, alternative product> pairs. When learning such pairs, existing research usually utilizes customer behavior signals to extract the substitute relationship [4]. Two commonly adopted substitute definitions are *co-view* and *view-but-purchase-another* [12]. These behavior-based heuristics are called buyability signals in this paper since they are correlated with customer purchase behavior.

However, such an approach does not consider product functionality. As shown in Figure 1, although vitamin D and C are usually coupled together based on co-view, they are not substitutable in functionality or product characteristics. Such bad cases harm customer trust and might incur legal regulation issues when claiming them substitute with each other. Therefore, in this paper, we define substitute recommendations based on both *buyability* and *functionality* and then further consider product functionality in the proposed model. Ideally, such functionality relationships can be learned from human-annotated labels. However, in practice, acquiring such information is highly time-consuming and expensive. Hence, we still use the signals from production (i.e., impressions, clicks, purchases, and revenue) to train our models with the awareness of their weak supervision nature. The corresponding technical challenges are described as follows:

(1) Inaccurate supervision: Customer behavior might be confounded by factors other than functional substitutabilities, such as complementary products and customer preference.
(2) Selection bias: It occurs when data samples are not representative of the underlying data distribution [13].
(3) Domain-specific text understanding: E-commerce text has unique characteristics compared to the other public corpus.

To address the above challenges, we first build a dedicated classification dataset to incorporate functionality in evaluation. We use negative sampling augmentation to address selection bias. We adopt a regression setting with log transformation to de-noise the weakly signals and consider product functionality and buyability together. We convert our substitute recommendation problem into a natural

Reference Product:
Vitafusion Extra Strength Vitamin D3 Gummy Vitamins, 120 ct

Alternative Product:
Natvites Vitamin C with Echinacea and Rose Hips Pectin Gummies for Adults 120Count

Figure 1: Popular substitute pair based on co-view substitute definition, but the alternative product has a different function with the reference product.

language matching problem, in which the reference product title is regarded as the "query" and the alternative product title is the "answer" or "document". We use the XLMR [5] into our use case, a state-of-the-art transformer-based multilingual model. It enables us to exploit contextual information efficiently, provides multilingual scalability, and reduces development and maintenance efforts.

To summarize, our contributions are as follows:

(1) To our best knowledge, it is the first work to consider product functionality in the substitute recommendation.
(2) We employ the state-of-the-art transformer-based model to learn textual information from the product title and fine-tune it in our E-commerce specific domain.
(3) We design new transformation methods and loss function objectives to de-noise label signals from production data and adopt the corresponding negative sampling strategy to improve robustness.
(4) We further consider multilingual support for better scalability. The proposed model is deployed into production and demonstrates success in online and offline experiments.

2 PROBLEM FORMULATION

The primary input to our substitute model is a pair of the reference product and alternative product titles. The model is tasked with predicting customer feedback that is correlated with functionality and buyability. Let u and v be the product titles from the reference and alternative products, and y be the label extracted from the raw customer signal (the count of received impressions, clicks, purchases, etc.). Assuming D is the set of n pairs of reference and substitute products, Div is the loss function measuring the divergence between the model output and the ground truth label, f is the model that outputs substitute score, and θ is model learnable parameters, we can define the learning objective as:

$$l(\theta; D) = \frac{1}{n} \sum_{(u,v,y) \in D} Div(f(u, v|\theta), y) \quad (1)$$

In this paper, we study different y, including click-through rate (CTR, # click / # impression), conversion rate (CVR, # purchase / # click), purchase rate (PR, equal to CTR x CVR), and gross merchandise value (GMV, revenue), as well as different Div and f.

3 METHODOLOGY

In this section, we first discuss feature selection to mitigate weak supervision. Then, we demonstrate the best label as the proxy for buyability and substitutability. Afterward, we present the negative sampling to address selection bias. Finally, we present the model design for multilingual understanding with domain adaptation.

3.1 Feature selection

We use the title as the model's feature to avoid overfitting the noise in customer preference. We discard size and color because their coverage is low, and information is usually duplicated in the title. When the mappings get ingested from retrieval sources, they come with confidence values measuring the quality of the mappings. We also drop these values and source information to avoid the cold start problem. Otherwise, the model cannot process the pairs from a new resource, and it needs to be re-calibrated every time once the upstream modules update.

We used the price information in our early iteration since they improved the offline metric. However, we found that the trained model filtered more high-priced products even if they were substitutable. It was because customer engagement is worse on average for the expensive product. For example, cheap products ($\leq$ \$20) have three times higher average purchase rates than expensive products ($\geq$ \$100). So, as a short-term solution, we dropped these features and left them for future investigation.

3.2 Label engineering

Label engineering aims to find the best proxy labels correlated with functionality and buyability. For functionality, purchases are stronger signals because customers need to pay, while clicks can occur on the non-substitutable product out of the customer's curiosity. Hence, we use purchase rate (PR), the ratio between purchase and impression count, as the training label. Besides, we found that CTR and CVR have a Pearson correlation lower than 0.20, suggesting that using any one alone will lead to a suboptimal buyability ranking. Another commonly adopted approach is to view the problem as a binary classification. Following [11], we can define positive samples as the recommendations purchased by customers at least once and negative samples as the ones that are not.

We also discover the long-tail distribution of the purchase rate. An extremely high purchase rate is likely to be noisy due to insufficient impressions or data collection errors. Hence, we log-transform the label and find the resulting distribution smoother while maintaining the same relative order and achieving better performance.

3.3 Negative sample augmentation

To mitigate the selection bias, we randomly sample pairs of products as negative training samples. Given the broad spectrum of our products, the chance that random mappings are relevant is negligible. In the regression setting, we need to assign a numerical "purchase rate" for the random mappings. Since random mapping

is expected to have lower quality than serving data, we assign a negative value for random negative samples. We perform random sampling before the training instead of for each batch separately, which is computationally efficient and has similar performance as mentioned in [9, 20]. We visualize the score distribution on validation data with and without the negative sampling in the extended version paper on arXiv. With the correct negative values and ratio setting, the model becomes more robust to random mappings. Aside from random negative samples, we also have around 60% of training data with zero purchase rate as the hard negative samples.

3.4 Models

Given the popularity and flexibility of the gradient boosting decision tree (GBDT) model in industry [3, 23], we first build GBDT as the baseline. Then, we propose a transformer-based deep model for better contextual understanding and multilingual scalability. Transformer has achieved great success in the industry recently [19, 22]. In this section, we give a high-level overview of those two models and how to adapt them to our use case.

3.4.1 GBDT model. To adapt GBDT in our use case, we need to featurize the text into a fixed-length vector. Firstly, we use word embedding to embed each word into a low-dimensional vector, which has proved effective in many areas [1, 14]. Specifically, we first remove the stop words using the NLTK library and encode the words with the FastText word embedding [1]. Then, we sum over each word embedding to get fixed-size embedding. Our early experiments show that the sum performs better than average. Lastly, the embeddings from both products are concatenated and fed into the GBDT model for learning.

3.4.2 Deep learning model. One disadvantage of the GBDT model is that it completely disregards the word order in the sentence. Besides, since the word embedding is pre-trained on the public corpus and cannot be fine-tuned in an end-to-end manner [18], making it difficult to learn with our E-commerce domain-specific data [17]. Recently, deep learning based method has achieved great success in different application [2, 6, 16, 21]. To solve this issue, we utilize the transformer-based neural network [6] and fine-tune it on our dataset for domain adaptation. In our case, we adopt the interaction-based model, which achieved higher accuracy than the representation-based model in our preliminary experiments. Specifically, we use XLMR [5] as the model backbone, which achieves state-of-the-art performance on cross-lingual benchmarks such as GLUE [15].

4 EXPERIMENT

In this section, we first describe the dataset and evaluation methods. Then we compare our proposed method with baselines and conduct an ablation study in the offline experiment. Lastly, we present the online impact on a large-scale worldwide E-commerce website.

4.1 Training dataset

We use the historical aggregated traffic feedback data, which record the count of customer behavior for a specific mapping pair since inception, including impressions, clicks, purchases, and GMV. We only keep the recommendation with over 250 impressions to balance the signal quality and size of the dataset. Since we only use the aggregated count of customer behavior, no customer identification information is touched. Furthermore, we exclude the mappings whose query products occur in the validation data to avoid data leakage [8]. There are 460k mappings in the training set. It consists of data from 11 countries: US (English), UK (English), DE (German), FR (French), JP (Japanese), CA (English), IT (Italian), ES (Spanish), IN (English), AU (English), and MX (Spanish).

4.2 Evaluation dataset and metrics

We prepare the two datasets to evaluate the two aspects of our substitute recommendation: functionality and buyability.

Functionality classification dataset contains 215k mappings from product managers' audits on traffic data, random negative samples, and good/bad mappings from traffic data based on customer signal. It is a binary classification dataset where the mappings are classified as substitutable or non-substitutable. Substitutable products are defined as the products which the customer can choose without critical compromises. The ratio between positive (substitutable) and negative (non-substitutable) samples is kept to 6:4, which is similar to the production distribution. The area under the precision-recall curve (AuPRC) is used as the metric.

Buyability ranking dataset contains 222k mapping in the traffic dataset with more than 500 impressions. A higher impression threshold is used for a more reliable purchase rate estimation. Normalized discounted cumulative gain (NDCG) over the PR is used to evaluate the buyability ranking performance. We first calculate the NDCG for each query product based on the model score and ground truth purchase rate and then take the average over all query products to get the final metric.

4.3 Offline experiment

In this section, we validate the proposed design choices by checking the offline metrics. To reduce the search space, we sequentially search for the best setting of the individual components in our design and keep it in the following experiments. We conduct the offline evaluation on the US fold of the data for the first two experiments and all marketplace data for the last multilingual experiment. For data safety, the performance was reported as the delta over the baseline, which is marked by an underline and dash. "NA" means the model cannot handle the corresponding data/language.

4.3.1 The choice of objective loss function. We compare functionality AuPRC and buyability NDCG under different objectives in Table 1. The baseline model is GBDT.

First, we observe that using PR as supervision achieves the best performance in AuPRC, NDCG@CVR, and NDCG@PR. It is reasonable that the CTR model has slightly better performance in CTR ranking. The CVR model behaves poorly because of its sample size and high variance (click count is around 100 times smaller than impression count). The GMV is strongly correlated with seasonal trends and product prices. Hence, it is noisy and leads to little learning. Log transformation can further improve the PR model in functionality classification (AuPRC from +1.76% to +5.62%) and buyability ranking (NDCG@PR from +1.58% to +2.15%) because it avoids overfitting the noisy label and focuses more on ranking.

Objective	Δ AuPRC	ΔNDCG@CTR	ΔNDCG@CVR	ΔNDCG@PR
Regression				
CTR+MSE	-	-	-	-
CVR+MSE	-23.3%	-4.60%	-3.42%	-5.88%
PR+MSE	+1.76%	-1.60%	+2.74%	+1.58%
GMV+MSE	-23.59%	-4.60%	-3.42%	-6.00%
PR+Log+MSE	**+5.62%**	-1.26%	**+3.19%**	**+2.15%**
Classification				
purchase > 0 + logistic	**+5.62%**	-1.78%	+2.17%	+0.79%
purchase > 0 + hinge loss	-3.42%	-3.35%	-0.68%	-2.83%

Table 1: Model performance with different objectives. The baseline model is underlined, and its score is marked by a dash.

Model	US		All	
	AuPRC	NDCG	AuPRC	NDCG
Monolingual model				
GBDT (US)	**+3.8%**	-1.1%	NA	NA
RoBERTa (US)	+0.3%	-0.1%	NA	NA
Multilingual model				
XLMR (US)	-	-	-	-
XLMR (EN)	+0.2%	+0.2%	+1.1%	+0.7%
XLMR (All)	+0.6%	**+0.2%**	**+2.0%**	**+1.3%**

Table 2: Monolingual vs. Multilingual model. The baseline is underlined, and its score is marked by a dash.

Name	Δ Revenue	Δ PR
No model (V0)	-	-
Naive GBDT (V1)	+10%	-12.6%
Robust GBDT (V2)	+19%	+24.1%
Robust XLMR (V3)	+19%	-2.5%

Table 3: Online model performance gain for each launch.

Second, it shows that logistic regression (binary cross-entropy) performs worse than the log-transformed PR model in ranking. The reason is that PR can preserve more information than a binary label for the model to identify the high-performing pairs.

4.3.2 Monolingual vs. Multilingual. In this section, we compare GBDT and Transformer model in different dataset settings. We focus on the cross-lingual ability of the Transformer. To save space, only the US and all 11 marketplaces performance are reported in Table 2.

GBDT model remains a strong baseline in the US but can only support English. RoBERTa can achieve comparable performance in the US with a lower AuPRC but higher NDCG score. If only provided with the monolingual corpus, the multilingual model behaves similarly to its monolingual counterpart (XLMR (US) vs. RoBERTa). However, it performs surprisingly well in the non-English marketplace, even without supervision. For example, XLMR (US) achieves higher AuPRC for JP and DE than the US with only English train data. It demonstrates its capacity for cross-lingual inference in the E-commerce domain. We can further improve US performance by 0.6% in AuPRC with data from other marketplaces (both English and non-English). It validates that the multilingual model can generate universal embedding for different languages and achieve better performance by utilizing more data from other languages.

4.4 Online customer impact

We have experimented with three different model variations. We present the customer impact during the experiments in Table 3, in which each variation is compared against its predecessor, and we only compare the marketplace that the new model supports (US for GBDT model, and 11 marketplaces for Robust XLMR model). We set the predict score threshold based on the performance on historical data, and then order the unfiltered substitutes by their respective scores. Note that "No model" means we only filter and rank the mappings with the upstream score. For the GBDT model, we concatenate the other numerical features with sentence embedding as input. The experiment details can be found in our extended paper on arXiv.

The naive GBDT model increased the revenue by 10% because of larger coverage. However, it suffered from selection bias and lowered the purchase rate by 12.6%. Besides, the Naive GBDT model is only evaluated on the serving data. Hence the model selection is suboptimal. With negative sampling and replacing CTR with PR, the Robust GBDT model significantly outperformed Naive GBDT with 19% incremental revenue and 24.1% purchase rate increase. Robust XLMR further drove 19% additional revenue with only 2.5% purchase rate decrease. The improvement is mainly driven by the non-US marketplaces. They had no dedicated model support before, and it obtained 22.3% higher PR and 0.6% higher revenue.

5 CONCLUSION AND FUTURE WORK

The paper explores the substitute recommendation's goal as optimizing buyability and functionality. The issues of inaccurate supervision, selection bias, and domain gap in the E-commerce corpus are identified and provided with the techniques to solve them. The proposed method is demonstrated to be effective in both offline and online experiments. In future work, we will build a clean human-labeled dataset with functionality, buyability, and complementary data and learn a multi-task model jointly.

6 BIOGRAPHY

Presenter Wenting Ye was an applied scientist at Amazon. He earned a Master of Computational Data Science degree from Carnegie Mellon University. His research interest includes information retrieval and natural language processing. He has published papers in various conferences and served as committee members in COLING, ACL, NeurIPS, and SIGIR. He is currently a senior machine learning engineer in Bytedance.

Amazon.com, Inc is an American multinational technology company focusing on e-commerce, cloud computing, online advertising, digital streaming, and artificial intelligence.

REFERENCES

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[2] Jingjun Cao, Zhengli Wu, Wenting Ye, and Haohan Wang. 2017. Learning functional embedding of genes governed by pair-wised labels. In *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. 397–401. https://doi.org/10.1109/CIAPP.2017.8167247

[3] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[4] Tong Chen, Hongzhi Yin, Guanhua Ye, Zi Huang, Yang Wang, and Meng Wang. 2020. Try this instead: Personalized and interpretable substitute recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 891–900.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[6] 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).

[7] Hyunwoo Hwangbo, Yang Sok Kim, and Kyung Jin Cha. 2018. Recommendation system development for fashion retail e-commerce. *Electronic Commerce Research and Applications* 28 (2018), 94–101.

[8] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. 2012. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 4 (2012), 1–21.

[9] Sen Li, Fuyu Lv, Taiwei Jin, Guli Lin, Keping Yang, Xiaoyi Zeng, Xiao-Ming Wu, and Qianli Ma. 2021. Embedding-based Product Retrieval in Taobao Search. *arXiv preprint arXiv:2106.09297* (2021).

[10] Weiwen Liu, Yin Zhang, Jianling Wang, Yun He, James Caverlee, Patrick PK Chan, Daniel S Yeung, and Pheng-Ann Heng. 2021. Item relationship graph neural networks for e-commerce. *IEEE Transactions on Neural Networks and Learning Systems* (2021).

[11] Hanqing Lu, Youna Hu, Tong Zhao, Tony Wu, Yiwei Song, and Bing Yin. 2021. Graph-based Multilingual Product Retrieval in E-Commerce Search. *arXiv preprint arXiv:2105.02978* (2021).

[12] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 785–794.

[13] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. 2020. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*. 1863–1873.

[14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[15] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[16] Haohan Wang, Xiang Liu, Yifeng Tao, Wenting Ye, Qiao Jin, William W Cohen, and Eric P Xing. 2018. Automatic human-like mining and constructing reliable genetic association database with deep reinforcement learning. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*. World Scientific, 112–123.

[17] Qiong Wu, Christopher G Brinton, Zheng Zhang, Andrea Pizzoferrato, Zhenming Liu, and Mihai Cucuringu. 2021. Equity2vec: End-to-end deep learning framework for cross-sectional asset pricing. In *Proceedings of the Second ACM International Conference on AI in Finance*. 1–9.

[18] Qiong Wu, Adam Hare, Sirui Wang, Yuwei Tu, Zhenming Liu, Christopher G Brinton, and Yanhua Li. 2021. Bats: a spectral biclustering approach to single document topic modeling and segmentation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 5 (2021), 1–29.

[19] Qiong Wu, Wen-Ling Hsu, Tan Xu, Zhenming Liu, George Ma, Guy Jacobson, and Shuai Zhao. 2019. Speaking with actions-learning customer journey behavior. In *2019 IEEE 13th International conference on semantic computing (ICSC)*. IEEE, 279–286.

[20] Han Zhang, Songlin Wang, Kang Zhang, Zhiling Tang, Yunjiang Jiang, Yun Xiao, Weipeng Yan, and Wen-Yun Yang. 2020. Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2407–2416.

[21] Shijie Zhang, Hongzhi Yin, Qinyong Wang, Tong Chen, Hongxu Chen, and Quoc Viet Hung Nguyen. 2019. Inferring Substitutable Products with Deep Network Embedding.. In *IJCAI*. 4306–4312.

[22] Shuai Zhao, Wen-Ling Hsu, George Ma, Tan Xu, Guy Jacobson, and Raif Rustamov. 2020. Characterizing and Learning Representation on Customer Contact Journeys in Cellular Services. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3252–3260.

[23] Shuai Zhao, Achir Kalra, Chong Wang, Cristian Borcea, and Yi Chen. 2019. Ad Blocking Whitelist Prediction for Online Publishers. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, 1711–1716.