A Multi-model Fusion Approach for Product Classification and Product Substitute Identification on Shopping Queries Data

Yanbo J. Wang
LYZD-FinTech Co., LTD
Beijing, China

Yuhang Guan
LYZD-FinTech Co., LTD
Beijing, China

Yuming Li
University of Auckland
Auckland, New Zealand

Hui Qin
LYZD-FinTech Co., LTD
Beijing, China

Sheng Chen
LYZD-FinTech Co., LTD
Beijing, China

Shilei Shan
LYZD-FinTech Co., LTD
Beijing, China

Xuan Yang
LYZD-FinTech Co., LTD
Beijing, China

Jie Shi
Utrecht University
Netherlands

Siyi Wang
University of California, Los Angeles
USA

ABSTRACT

E-commerce platforms are widely accepted by the public in the era of big data with the rapid development of information technology. With the increasing proportion of e-commerce platforms such as Amazon in international commerce, user search-based analysis and optimization of retrieval results have gradually attracted the attention of the industry, since the effects directly or indirectly result in user experience and transaction rates. Although the application of deep learning in various industries has become more and more mature in recent years, the research on the correlation between user search intent and results is still scarce. Therefore, in this paper, we proposed a multi-model fusion approach for the ESCI challenge to improve the product search in in Amazon KDD Cup 22, and finally achieved the private score of 0.8177 and the public score of 0.8688 on task 2, and the private F1 score of 0.8708 and the public score of 0.8688 in task 3. We finally ranked tenth and fifth respectively.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Product Classification, Product Substitute Identification, Shopping Queries Data, Pre-trained Language Model

ACM Reference Format:
Yanbo J. Wang, Yuhang Guan, Yuming Li, Hui Qin, Sheng Chen, Shilei Shan, Xuan Yang, Jie Shi, and Siyi Wang. 2022. A Multi-model Fusion Approach for Product Classification and Product Substitute Identification on Shopping Queries Data. In *KDD Cup 2022 Workshop: ESCI Challenge for Improving Product Search, August 17, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 4 pages.

KDDCup '22, August 17, 2022, Washington, DC, USA

1 INTRODUCTION

As a platform for online transactions and negotiations for enterprises or individuals, e-commerce platforms have gradually become popular around the world with the development of information technology [Taher 2021]. On the one hand, e-commerce platforms automate and digitize traditional business processes, which can reduce labor and costs. On the other hand, e-commerce breaks through the constraints of time and space, so that transaction activities can be carried out at any time and anywhere, thus greatly improving efficiency. In addition, the ubiquity, global reach, interactivity and information density of e-commerce, along with the advent of large-scale international e-commerce platforms such as Amazon, e-commerce has created more trade opportunities for global enterprises. Relevance matching is the basis for matching user intent with products in e-commerce search. Therefore, improving the relevance of search results plays a significant and positive role in improving customers' purchasing experience and transaction rate.

Although the application of machine learning technology in various industries has generally entered a mature stage in recent years [Rath 2022], the matching problem of retrieval results for users in e-commerce platforms is still a challenge. The notion of binary relevance in existing applications always exists and constrains the searching experience of customers. For example, when a user searches for *"iPhone"* on the Amazon platform, it may be looking for the *"iPhone charger"*. In this case, the search engine needs to understand the correlation between *"iPhone"* and the *"iPhone charger"* so as to ensure the searching experience of users.

Therefore, in the Amazon KDD Cup 22, ESCI challenge for improving product search, based on the Shopping Queries Data Set [Reddy et al. 2022], we proposed a multi-model method for task 2 "MULTICLASS PRODUCT CLASSIFICATION" and task 3 "PRODUCT SUBSTITUTE IDENTIFICATION", and finally achieved the private F1 score of 0.8183 and the public F1 score 0.8177 on task 2, and achieved the private F1 score of 0.8708 and the public F1 score 0.8688 on task 3, ranking 10^{th} and 5^{th} respectively.

2 TASK DESCRIPTION

2.1 Data Description

The Shopping Queries Data Set [Reddy et al. 2022] for this task is a large-scale human-annotated dataset derived from the search data of Amazon platform users, including English, Japanese, and Spanish. The task defines the correlation between products in the search results into four classes (ESCI):

- Exact (E): The item is relevant to the query and meets all query specifications
- Substitute (S): The item is somewhat relevant, it does not satisfy all the aspects of the query, but the item can be used as a functional substitute
- Complement (C): The item does not satisfy the query, but can be combined with an exact item
- Irrelevant (I): The item is irrelevant, or it fails to satisfy the central aspect of the query

The requirement of task 2 is to give a query and a list of products retrieved by the query, and classify the correlation between the retrieved products and the query products as one of ***E***, ***S***, ***C***, and ***I***. The requirement of task 3 is to measure the ability of the query system to find substitutes for the retrieved products, which can be regarded as changing the multi-classification task in task 2 into a binary classification task. Given an input example, in which **product_id** represents the id of the product **11 degrees** to be queried, and **query_locale** represents the region to which the query language belongs. **Example_1** and **example_2** are all queries based on the us-English environment.

example_id	**query**	**product_id**	**query_locale**
example_1	11 degrees	product0	us
example_2	11 degrees	product1	us

Table 1: Input example of the Shopping Queries Dataset

After inputting the query information in Table 1, task 2 needs to return the correlation label between the query and the product in each example. For example, the correlation between **11 degrees** and **product0** in **example_1** is **exact**.

example_id	**esci_label**
example_1	exact
example_2	complement

Table 2: The output of Task 2

Task 3 needs to identify whether the query product and the given product are substitutes. For example, the query **11 degrees** and **product0** in **example_1** in Table 1 are substitutes.

example_id	**substitute_label**
example_1	substitute
example_2	no_substitute

Table 3: The output of Task 3

2.2 Evaluation

In this task, we choose to use the F1 score as the evaluation criteria for evaluating the results, which is a classic metric used in statistics to measure the accuracy of classification models. The F1 score can be regarded as a harmonic average of the precision and recall of the model, with a maximum value of 1 and a minimum value of 0. In view of the fact that the distribution of categories in the dataset is not balanced, in task 2, four categories account for 65.17%, 21.91%, 2.89% and 10.04% respectively, while in task 3, the two categories account for 33% and 67% respectively. Therefore, the micro averaging F1 score was chosen as the specific evaluation metric in these two tasks. The calculation process is as follows:

$$F_1 = 2 \times \frac{precisionrecall}{precision + recall} = \frac{TN}{TP + \frac{1}{2}(FP + FN)}$$

Where Precision can be regarded as the measure of quality, and recall can be regarded as the measure of quantity.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

In which the explanation of TP, FP, TN, FN is as follows:

- TP: True Positive, classified as a positive sample, which is actually a positive sample.
- FP: False Positive, classified as a positive sample, but is actually a negative sample.
- TN: True Negative, classified as a negative sample, which is actually a negative sample.
- FN: False Negative, is classified as a negative sample, but is actually a positive sample.

3 METHODOLOGY

Given that Task 2 and Task 3 are identical in requirements and data form, the only difference is the number of categories to be classified. Therefore, we use the same model fusion methods and tricks in these 2 tasks. Our methodology is based on the fusion of three pre-trained language models to achieve a better result, where three models are xlm-roberta-large [Conneau et al. 2019], infoxlm-large [Chi et al. 2020] and rembert-large [Chung et al. 2020].

3.1 Data Preprocessing

Since the product catalogue in the Shopping Queries Data Set [Reddy et al. 2022] has multiple attributes (product_title, product_description, etc.), we use the **[SEP]** token to segment the text of each field, connect it to the model input text, and use the **[CLS]** token vector as the potential feature of the data.

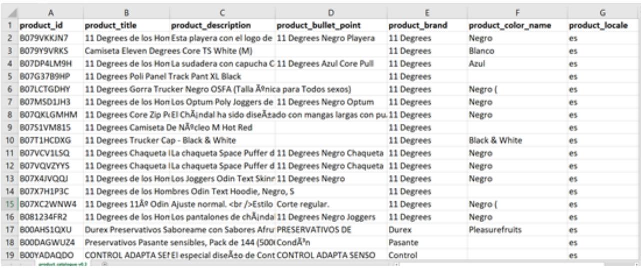
The original data in the Shopping Queries Data Set is:

After preprocessing, the input token is connected as:

[CLS] <query_content> [SEP] <title_content> [SEP] <bullet_point> [SEP] <brand> [SEP] <color_name> [SEP] <locale> [SEP] <description>

3.2 Pre-trained model selection

Since our approach is to fuse the multiple model to achive a better result, in order to choose the right pre-trained model to be fused, we



	A	B	C	D	E	F	G
1	product_id	product_title	product_description	product_bullet_point	product_brand	product_color_name	product_locale
2	B079VKKJN7	11 Degrees de los Hon	Esta playera con el logo de	11 Degrees Negro Playera	11 Degrees	Negro	es
3	B079Y9VRKS	Camiseta Eleven Degrees Core TS White (M)			11 Degrees	Blanco	es
4	B07DP4LM9H	11 Degrees de los Hon	La sudadera con capucha C	11 Degrees Azul Core Pull	11 Degrees	Azul	es
5	B07G37B9HP	11 Degrees Poli Panel Track Pant XL Black			11 Degrees		es
6	B07LCTGDHY	11 Degrees Gorra Trucker Negro OSFA (Talla Ãºnica para Todos sexos)			11 Degrees	Negro (	es
7	B07MSD1JH3	11 Degrees de los Hon	Los Optum Poly Joggers de	11 Degrees Negro Optum	11 Degrees	Negro	es
8	B07QKLGMHM	11 Degrees Core Zip P	El ChÃ¡ndal ha sido diseÃ±ado con mangas largas con pu		11 Degrees	Negro	es
9	B07S1VM815	11 Degrees Camiseta De NÃºcleo M Hot Red			11 Degrees		es
10	B07T1HCDXG	11 Degrees Trucker Cap - Black & White			11 Degrees	Black & White	es
11	B07VCV1LSQ	11 Degrees Chaqueta I	La chaqueta Space Puffer d	11 Degrees Negro Chaqueta	11 Degrees	Negro	es
12	B07VQVZYYS	11 Degrees Chaqueta I	La chaqueta Space Puffer d	11 Degrees Negro Chaqueta	11 Degrees	Negro	es
13	B07X4JVQQJ	11 Degrees de los Hon	Los Joggers Odin Text Skinr	11 Degrees Negro	11 Degrees	Negro	es
14	B07X7H1P3C	11 Degrees de los Hombres Odin Text Hoodie, Negro, S			11 Degrees		es
15	B07XC2WNW4	11 Degrees 11Âº Odin	Ajuste normal. Estilo	Corte regular.	11 Degrees	Negro (	es
16	B081234FR2	11 Degrees de los Hon	Los pantalones de chÃ¡nda	11 Degrees Negro Joggers	11 Degrees	Negro	es
17	B00AHS1QXU	Durex Preservativos Saboreame con Sabores Afru		PRESERVATIVOS DE	Durex	Pleasurefruits	es
18	B00DAGWUZ4	Preservativos Pasante sensibles, Pack de 144 (500		CondÃ³n	Pasante		es
19	B00YADAQDO	CONTROL ADAPTA SE	El especial diseÃ±o de Cont	CONTROL ADAPTA SENSO	Control		es

Figure 1: The data samples in the product catalogue in the Shopping Queries Data Set

first conduct comparative experiments on some of the commonly used pre-trained language models. We compare the out-of-fold score (oof_score) and subscore (sub_score) of various pre-trained language models, and we finally choose 3 pre-trained models, xlm-roberta-large [Conneau et al. 2019], infoxlm-large [Chi et al. 2020] and rembert-large [Chung et al. 2020] for further fusion according to the results shown in Table 4.

Model Name	oof_score	sub_score
Multilingual-MiniLM-L12-H384	0.72018	0.723
bert-base-multilingual-cased	0.72862	0.734
infoxlm-base	0.73432	0.742
xlm-roberta-large	**0.7566**	0.76
rembert	**0.7561**	0.759
twitter-xlm-roberta-base	0.7312	0.738
infoxlm-large	**0.7554**	0.759
roberta-large-us	0.7686	

Table 4: The comparation of the pre-trained language model

3.3 Model Structure

We use the features of the **[CLS]** token of all hidden layers output by the pre-trained model, a total of 24 feature vectors are connected into a 24 * hidden_size feature matrix, and then use three convolution kernels with the sizes of 5* 24, 7* 24 and 9* 24 respectively to extract features. After maximum pooling, a 1*hidden_size feature is obtained for classification. The model results of this part are shown in the following figure:

During training, we use 5 Dropout layers with different parameters for the features obtained from the above structure, perform parallel processing on the features and calculate the average loss.

3.4 Model Fusion

Based on the results of the comparative experiments in Table 4, we choose xlm-roberta-large [Conneau et al. 2019], infoxlm-large [Chi et al. 2020] and rembert-large [Chung et al. 2020] for the model fusion. Although the score of roberta-large-us is also high, considering that 3 languages are included in the task data, we temporarily exclude pre-trained models based on monolingual environments. In this task, we adopt weight fusion; the weights of the three models are 0.31, 0.31 and 0.38 respectively.

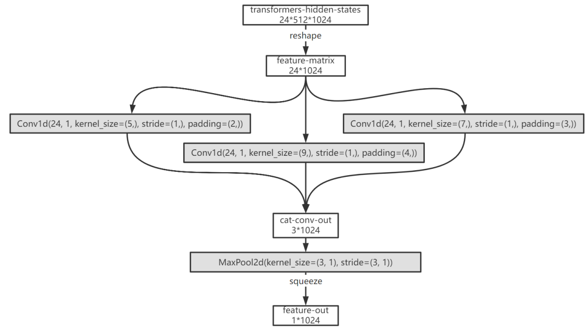


Figure 2: The structure between the pre-trained model and output

$$Logits = 0.31 * RoBERTa + 0.31 * InfoxLm + 0.38 * Rembert$$

3.5 Tricks

In order to optimize the prediction effect of a single model, we tested a variety of training techniques. On the basis of model fusion, we add some widely used tricks, such as Pre-trained MLM Task, Adversarial Training, Pseudo Labeling, Focal Loss and Labelsmooth.

Trick	oof_score	sub_score
base	0.7547	0.8087
Pre-trained MLM Task	0.7594	0.8113
Adversarial Training	0.7589	0.8082
Pseudo Labeling	0.7596	0.8117
Focal Loss	0.7429	0.8032
Labelsmooth	0.7524	0.8083

Table 5: The tricks and corresponding score

3.5.1 Pseudo Labeling. Pseudo Labeling is a concept in semi-supervised learning, which can facilitate models to learn better from unlabeled information. The principle is to use the existing labeled data to train a model, and then use the trained model to predict the unlabeled data, then add the predicted labels and data of the unlabeled data to the training set for training, thereby improving the generalization ability of the model.

We use the xlm-roberta large model that performs best in Table 4 to make predictions on the test set of Shopping Queries Data Set, and then sort according to the maximum value of the probability vector to find a threshold. After that, we select the test set data with high confidence as pseudo labeling data, and add it to the training process of the model.

In order to determine the appropriate threshold, we use the same method to sort the validation set, select 10000 continuous predictions, and slide the calculation accuracy. When the accuracy is close to the overall validation set, we calculate the average value of the maximum probability of the current region as the threshold of the segmentation validation set. Through such segmentation, pseudo labeling data can be selected as much as possible while ensuring the accuracy of pseudo labeling data.

3.5.2 Pre-trained MLM Task. On the basis of the original pre-training model, we use all the data from the Shopping Queries Data Set to process an additional pre-train, so as to better fit the data in this task. The pre-training is optimized only for MLM Task, and the training data is constructed using a 30% Mask proportion. The original xlm-roberta-large and infoxlm-large were pre-trained, and the rembert model was not pre-trained due to time and equipment reasons.

3.5.3 Inference speed up. In order to get the prediction results of the three models within the specified time, we mainly used two methods.

- Since FP16 model performs about twice as fast as FP32 model [Fabien-Ouellet 2020], we use semi-precision FP16 to make the final prediction.
- Sort the input data according to the length of tokens, and dynamically complete the input data according to the maximum length of a single bath in Dataloader, so as to reduce the unnecessary computation generated by large area zero complement pairs.

4 RESULT AND CONCLUSION

The public widely accepts E-commerce platforms in the era of big data with the rapid development of information technology. In this paper, we proposed a multi-model fusion approach for ESCI challenge for improving product search in Amazon KDD Cup 22. We compared multiple pre-trained language models in our experiments and finally selected three multi-language pre-trained models for fusion. Then we used the Pre-trained MLM task, pseudo labeling and other tricks for further tuning in the optimization stage and finally achieved the private F1 score of 0.8183 and the public score of 0.8177 on task 2, and the private F1 score of 0.8708 and the public score of 0.8688 in task 3. We finally ranked 10^{th} and 5^{th} respectively.

REFERENCES

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. *arXiv preprint arXiv:2007.07834* (2020).

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821* (2020).

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).

Gabriel Fabien-Ouellet. 2020. Seismic modeling and inversion using half-precision floating-point numbersSeismic modeling and inversion using FP16. *Geophysics* 85, 3 (2020), F65–F76.

Mamata Rath. 2022. Machine learning and its use in e-commerce and e-business. In *Research Anthology on Machine Learning Techniques, Methods, and Applications*. IGI Global, 1193–1209.

Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. arXiv:2206.06588

Ghada Taher. 2021. E-commerce: advantages and limitations. *International Journal of Academic Research in Accounting Finance and Management Sciences* 11, 1 (2021), 153–165.