

Multiclass Product Classification Based On Multilingual Model and LightGBM (Team:Uni)

Peng Zhang*
Zhejiang University
China

Linghan Zheng*
Ant Group
China

Ruiqing Yan*
CNIC, CAS; BIPT
China

Changyu Li*
University of Electronic Science and
Technology of China
China

Rui Hu
Ant Group, China
China

Sheng Zhou
Zhejiang University
China

Jinrong Jiang; Lian Zhao
CNIC, CAS; University of Chinese
Academy of Sciences
China

Qianjin Guo; Qiang Liu
AAI, BIPT
China

BACKGROUND

- The primary objective of ESCI competition is to build new ranking strategies and identify interesting categories of results (i.e., substitutes) that improves customer product searching experience.
- The dataset is multilingual that includes English, Japanese, and Spanish.
- Task 1: Query-Product Ranking
- Task 2: Multiclass Product Classification
- Task 3: Product Substitute Identification

BACKGROUND

- Task 2 & Task 3 share the same training dataset with different classification targets
- Task 2: Multiclass Product Classification
 - Given <Query, product> pairs, predict 4 labels. (**exact, complement, irrelevant, substitute**)
- Task 3: Product Substitute Identification
 - Given <Query, product> pairs, predict 2 labels. (**no_substitute, substitute**)
- **We participated in Task 2 and Task 3. We won the 3rd place in both tasks.**

BACKGROUND

Product Information

- product_locale
- product_id
- product_title
- product_color_name
- product_brand
- product_description
- product_bullet_point (Nan)

Amazon.es/vidaXL-diván-madera-200x90-blanco/dp/B06X3ZWQ72/ref=cm_cr_arp_d_product_top?ie=UTF8

vidaXL Sofá Cama Extensible de Doble Altura Madera Pino Blanco Sillón Dormir

Marca: vidaXL 32 valoraciones

No disponible.

Color	Blanco
Marca	VidaXL
Estilo	Moderno
Capacidad de asientos	2
Forma	Rectangular
Material del marco	Madera
Material	Madera de pino
Requiere montaje	Sí

~ Ver menos

Advertencia: El producto puede requerir montaje.

Detalles del producto

Is Discontinued By Manufacturer : No
Producto en Amazon.es desde : 30 abril 2014
Fabricante : vidaXL
ASIN : B06X3ZWQ72
Referencia del fabricante : 242953
Opiniones de los clientes: 32 valoraciones

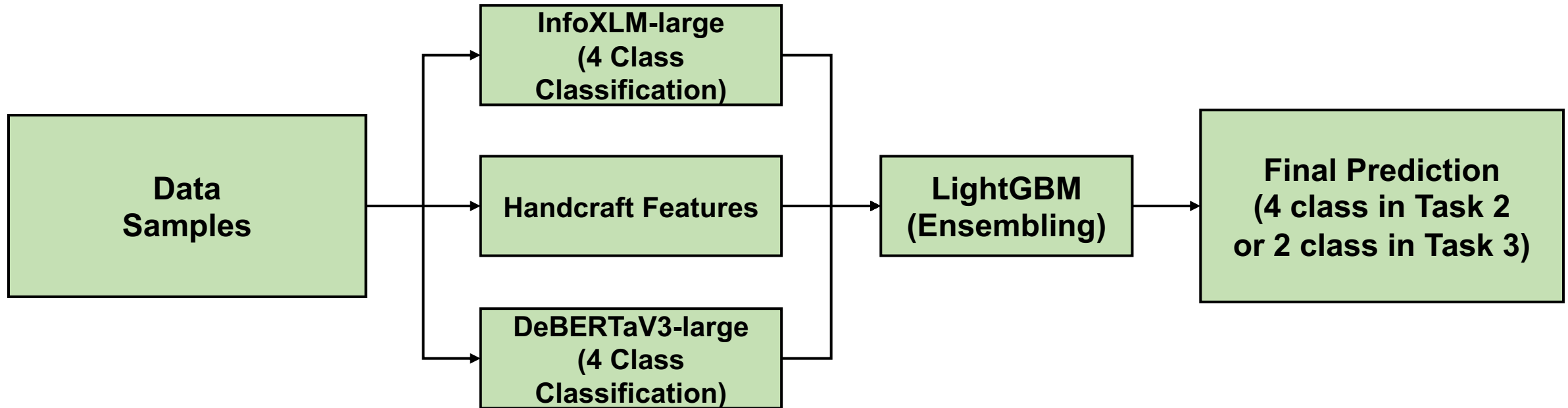
Descripción del producto

Este sofá cama doble es una excelente solución para personas con espacio limitado. Proporciona espacio para dormir para dos personas, al tiempo que maximiza el espacio disponible en el suelo. La cama de abajo se estira suavemente, por lo que se puede convertir fácilmente en un sofá durante el día y en una cama cómoda por la noche. Además, la cama inferior está equipada con una tabla lateral que evitará que el usuario se caiga mientras duerme. Toda la estructura de la cama tiene una construcción de madera robusta, lo que la hace muy duradera y adecuada para el uso diario. El montaje es bastante fácil.Tenga en cuenta que la entrega incluye la estructura de la cama solamente, los colchones no están incluidos.

- Color: Blanco
- Material: Estructura de madera de pino + listones de madera contrachapada
- Dimensiones cama superior: 205 x 97,5 x 66 cm (largo x ancho x alto)
- Dimensiones cama inferior: 200 x 94,5 x 19 cm (largo x ancho x alto)
- Tamaño adecuado del colchón: 200 x 90 cm (ancho x profundo) (colchones no incluidos)
- Cama inferior equipada con 4 ruedas
- Las dos camas se pueden usar por separado
- Fácil montaje

- NONE of the following external data is used:
 - Product info pages and images crawled by web crawlers
 - Data leakage from Task 1
 - Training dataset of Task 1
- Disclaimer: We didn't use any external data, but the top 2 teams all stated in their respective posts that their solutions used the external data and gained significant benefits.

OUR SOLUTION



OUR SOLUTION – Language Models

- InfoXLM
 - One of the most powerful cross-lingual pre-trained models
- DeBERTaV3
 - ~70% of the sample pairs are English corpus in this dataset.
 - One of the most powerful pre-trained English language models

OUR SOLUTION – Training Strategy

- Input of the Language Model :

*Input = [CLS] + Query + [SEP] + product_id + [SEP] + product_brand + [SEP]
+ product_color + [SEP] + product_title_name + [SEP] + product_bullet_point + [SEP]
+ product_description + [SEP]*

- Take Language Model as a feature extractor and then a fully connected layer will transform the feature into the two/four class probability
- Other Useful Train Strategies:
 - FGM
 - Rdrop
 - Hard Samples Mining

OUR SOLUTION - Stacking

- We use LightGBM as the final predictor, which contains three different inputs:
 1. 4-fold cross-validation probability of InfoXLM-large
 2. 4-fold cross-validation probability of DeBERTaV3-large
 3. Handcraft Features
 - Match scores based on query term and product term, e.g. Jaccard similarity between query and product_title
 - Query_locale (tell the model which country this sample comes from)
 - Aggregation features of predicted results of all samples under the same query
 -

Results

Language Models	Input Length	Training Strategy	Model Ensembling (Kfold)	Handcraft Features	Model Ensembling Strategy	Online F1 score
InfoXLM-large	180	-	5	-	Average	0.8142(public)
InfoXLM-large	128	Rdrop, FGM	5	-	Average	0.8163(public)
InfoXLM-large + DeBERTaV3	128	Hard-Sample-Mining, Rdrop, FGM	4	-	Average	0.8184(public)
InfoXLM-large + DeBERTaV3	128	Hard-Sample-Mining, Rdrop, FGM	4	Add	LightGBM Blending	0.8281(public)
InfoXLM-large + DeBERTaV3	128	Hard-Sample-Mining, Rdrop, FGM	4	Add	LightGBM Blending	0.8274(private)

Table 1: Performance comparison with different models

Only use Task 2 dataset and achieve 0.8274 on private dataset

ACKNOWLEDGMENTS

- We thank both KDD organizers as well as Amazon for holding such a great competition.
- This study is supported by the National Natural Science Foundation of China (Grant No.41931183). The numerical calculation in this work were carried out on the SunRising-1 computing platform.
- This study is also supported by National Natural Science Foundation of China (Grant No: 62106221)