scientific reports

Check for updates

OPEN

Automatic classification method of e-commerce commodity raw materials through the introduction of self-supervised concepts and the construction of domain ontology

Bing Lei[1,2], Jinghua Wang[3] & Cong Shen[1]✉

The e-commerce platform's function-oriented classification basis will cause items with the same (different) raw materials to be incorrectly classified into different (same) functional categories, posing a challenge to marketing staff who create item sales statistics based on raw materials. Furthermore, it is challenging to promote the present item classification method in engineering applications since it necessitates a high number of manual markings to add labels. As a result, this paper created an item conceptual model to specify the categories and attributes of item raw materials, allowing it to screen item specification samples and automatically add category labels, generate domain-specific lexicon to extract item raw material features, and finally use a machine learning classifier to complete the classification. This research presents a verification of the suggested classification model using flour data from the Chinese e-commerce platform. The experimental results show that the self-supervised learning-based classification method proposed in this article for classifying raw materials of e-commerce items can achieve an accuracy of 91%.

Keywords Self-supervised learning, E-commerce item classification, Item raw materials

Large-scale e-commerce platforms offer a vast array of goods, typically numbering in the millions or more[1]. E-commerce platforms use a range of division techniques to handle items so that customers with varying buying goals may easily find the items they require[2–4]. When an item is released, the merchant determines its category based on the prompts provided by the e-commerce platform. This means that an item may be classified into multiple categories or use different raw materials or ingredients, but it will always be classified at the same level of use. This phenomenon is especially common in the food industry[5]. For instance, cake mixes that contain the same raw material (such wheat core flour) can be categorized as "flour > wheat core flour" or "grain and oil seasoning > Baking ingredients" in China's Jingdong Mall. Whole wheat flour and wheat core flour, for instance, can be used as bread flour. Companies that consider the functionality of their items typically classify these two items as "grain and oil seasoning > Baking materials" when they release them. While this flexible classification clearly helps consumers with their searches, it is exceedingly inconvenient for market workers or merchants who need to track sales of goods based on raw material classification. As a result, it is imperative that items be categorized based on their raw materials in practice rather than their listing in the e-commerce platform's catalogue.

The majority of current studies on the classification of items for e-commerce rely on machine learning techniques[2,4]. On the one hand, the e-commerce platform's item catalogue serves as the primary basis for classifying these methods, and the platform's label serves as the primary basis for feature construction[3]. The feature vocabulary pertaining to item raw materials is rarely used in these processes. As such, there will be a major reduction in the accuracy of transplanting such methodologies into the classification of item raw materials. On the other hand, many of the machine learning-based methods now in use need labor- and time-intensive human labelling, which is frequently impractical in engineering practice[6].

[1]School of Management, Henan University of Technology, Zhengzhou 450001, China. [2]Business Intelligence and Knowledge Engineering Laboratory, Henan University of Technology, Zhengzhou 450001, China. [3]School of Economics and Management, Shangqiu University, Shangqiu 476000, China. ✉email: congshen12345@haut.edu.cn

In order to address the aforementioned issues, this research suggests a self-supervised learning-based approach for classifying raw materials used in e-commerce. First, domain specialists create item domain ontology and establish classification rules for specific items based on raw material characteristics. Second, the samples are separated into normative samples and non-standard samples based on the concept of self-supervised learning. Samples that satisfy domain experts' classification criteria are referred to as normative samples. Regular expressions can be used to derive the item raw material classification labels from normative samples. Machine learning uses these kinds of samples. The machine learning approach presented in this research will predict non-standard samples, which are samples that cannot be retrieved from item raw material classification labels using regular expressions. Thirdly, feature keywords are extracted and a feature matrix for machine learning is constructed using BERT and regular expression based on the item domain ontology. This allows for the automatic classification of item raw materials. The term "self-supervised" referenced in this study represents an application and modification of the self-supervised learning concept, rather than a rigid compliance with established paradigms of self-supervised learning. We have adopted the fundamental concept of self-supervised learning, which involves generating supervisory signals from the intrinsic information within the data. Utilizing a rule system based on domain ontology, we automatically extract the inherent relationships between attributes and categories from commodity text data to produce label information for standardized samples, thereby supplanting the conventional manual labeling process. This approach, while not employing conventional self-supervised techniques like contrastive learning and pre-training tasks, fundamentally aims to diminish dependence on external manual labeling by deriving supervisory signals from intrinsic domain knowledge associations within the data, aligning with the primary objective of self-supervised learning to reduce human intervention.

Our methodology of integrating domain ontology rules with self-supervised concepts is a potential innovation of our work. Current self-supervised learning techniques predominantly depend on the general structural characteristics of data; however, in the context of commodity classification within e-commerce, which possesses pronounced domain attributes, the incorporation of domain ontology knowledge can more precisely identify domain-specific relationships within the data. This approach diminishes the need for manual labeling while improving the domain adaptability of label generation. This distinct implementation strategy seeks to offer an innovative solution for semi-automated categorization jobs in particular fields. The following are the innovations of this paper: first, the ontology of item field is constructed to realize the classification of item raw materials on e-commerce platform; second, the identification rules of normative samples are designed and the labels of normative samples are automatically labelled, based on the concept of self-supervised learning.

The rest of this paper is structured as follows: section "Related work" discusses related work, sect. "Model construction" introduces the proposed self-supervised learning-based model for classifying items based on their raw materials in e-commerce, section "Experimental validation" evaluates our method using real data obtained from Chinese e-commerce platforms, section "Discussion" is devoted to the discussion of this research, and section "Conclusion" concludes with future research directions.

Related work

The primary goal of e-commerce item classification research in academia is to develop automatic techniques for classifying large-scale item texts on e-commerce platforms[7]. Compared to traditional text classification, e-commerce item classification has characteristics such as a huge variety of categories, short and noisy texts, and imbalanced samples in each category. Scholars have suggested improvements to feature vectors, data sources, and classification models in order to increase classification accuracy. Shen et al.[2] addressed the issue of data sparsity using statistical smoothing techniques and created a two-stage learning strategy based on the naive Bayes algorithm to increase classification accuracy. The graded weighted bag-of-words vector (gwBoWV) is a distributed semantic representation technique that Gupta et al.[8] devised to overcome the high-dimensional and sparse problems of BoW or tf-idf feature vectors. In order to tackle the issue of sellers on e-commerce platforms entering item information in an irregular manner, Das et al.[9] introduced a noise detection technique utilizing the Corr-LDA model. Using item titles and descriptions as the classification texts, Cevahir and Murakami[3] extracted item features like model numbers, sizes, and quantities from data from the Japanese e-commerce platform Rakuten. They then developed a deep belief network (DBN) and deep autoencoder (DAE) based classification model. Chen and colleagues[4] introduced a neural item classification model (NPC) designed to address the issues of unstable category vocabulary and idea fuzziness in fine-grained item classification. NPC creates item categories based on item information, including titles, attributes, descriptions, and so on. Deep learning has been used in certain research to classify large-scale e-commerce items, with promising results. In order to address sparsity and scalability concerns, Ha et al.[10] introduced an item classification model based on multiple recurrent neural networks. This model allows many qualities to be incorporated into a common representation. A CNN and Bi-LSTM-based deep learning model for item classification was presented by Kim et al. in 2021. Islam and Alauddin[11] presented a deep convolutional neural network-based categorization technique that relied on item photos rather than item titles and details. An OWL (Open-world Learning) model based on meta-learning was proposed by Xu et al.[12]. It allows the insertion or deletion of new categories without retraining the model, and it only keeps a set of dynamically observable categories. Qiao et al.[13] addressed the distinct requirements of end users and the uniqueness of raw data, assessed the synthetic data set, and enhanced clustering accuracy using the effective paradigm of federated learning (FL), achieving a 9% reduction in convergence time. Tan et al.[14] emphasized the need of recognizing BGP community properties for modeling and developed a graph neural network (GNN) model incorporating convolutional residual networks and fully connected layers, achieving an accuracy rate of 90%.

A common goal of the above research is to facilitate consumers in quickly searching for the desired items. Nevertheless, this might result in the misclassification of things that possess comparable functions or characteristics but are made from distinct raw materials, or items that share the same raw materials but serve

different consumer purposes, leading to erroneous categorization. It is essential for merchants or market workers to categorize and evaluate sales according to the raw materials used. However, the current methods are not suitable for classifying item raw materials, resulting in a low accuracy in classification. Hence, it is imperative to develop an item categorization model that relies on the classification of raw materials. This entails developing an item domain ontology to ascertain the classification labels for the raw materials of items. Subsequently, pertinent characteristics pertaining to the raw materials of the goods are derived from the textual data. Machine learning techniques are utilized to accomplish the categorization of e-commerce items based on their raw materials.

On the other hand, the training sets of the above e-commerce item classification models rely on a multitude of annotated samples or require time-consuming and labor-intensive manual annotation. For example, Shen et al.[2] utilized a substantial number of samples that were categorized by sellers on the eBay site as the training set. Cevahir and Murakami[3] employed a labeled dataset obtained from the Japanese Rakuten platform. Several scholars have employed various techniques for extracting labels. For instance, Das et al.[15] introduced a methodology that utilizes topic models and a simplified manual labeling process to acquire item category labels. Similarly, Chen et al.[4] developed a system that generates detailed item labels by analyzing user search logs.

Self-supervised learning leverages the inherent characteristics of data to extract meaningful features, using auxiliary tasks instead of relying on manual labels or external supervision signals for categorization[16]. Self-supervised learning has been utilized in both the medical domain and the field of sentiment analysis. Chaves et al.[17] employed self-supervised pre-training models to train and derive pseudo labels for the purpose of classifying skin lesions. In their study, Su et al.[18] introduced a progressive self-supervised attention technique for identifying and extracting the most influential language in sentiment prediction. They iteratively extracted feature words using this method and included them into neural network models to enhance the accuracy of sentiment classification. This paper presents a novel approach for classifying raw materials in e-commerce items using self-supervised learning. The method involves dividing samples into standardized and non-standardized categories by constructing an item domain ontology. Labels are then extracted from standardized samples to facilitate machine learning.

Model construction

E-commerce item text information mainly consists of two parts, namely the item title and details, represented as $I = (T, D)$, where T is the item title, typically a group of keywords related to the item's name, attributes, quality, usage, etc., and D is the item details, usually in the form of "attribute: attribute value". Table 1 displays an illustration of the textual information for an e-commerce item. Two samples of flour item texts from Chinese e-commerce platforms are shown in Table 1. Example 1 and Example 2 include both a title and details in Chinese. The Chinese title is a short text that describes the raw material, use, weight, and other attributes of flour. Details give particular information in the form of key-value pairs, such as flour brand, item name, item number, store, item weight, stringiness, package type, and category. For instance, "樱安娜" is the brand name and "金沙河自发粉家用小麦粉包子馒头油条中筋面粉免酵母通用面粉1kg 自发小麦粉1kg*10/20斤" is the item name in example 1. "10050083362787" is the item number for the flour item. In example 2, "华海春" is the brand name and "低筋面粉5斤蛋糕面包家用粉烘焙蒸蛋挞月饼糕点饼干低筋小麦粉1斤 1斤低筋烘培面粉+熟白芝麻100g " is the item name. "10048236400265" is the item number for the flour item. Detailed information is displayed in Table 1 below.

The guiding principles and approach for the model proposed in this paper are as follows: to minimize human intervention and enhance process transparency, to leverage the concept of self-supervised learning for automatic recognition of standard item samples, to extract and label item category tags from standard item samples, ultimately achieving automatic classification based on item raw materials.

The research presents a classification model for categorizing raw materials in e-commerce items. The model consists of four basic components, illustrated in Fig. 1.

Specifically, firstly, construct the item domain ontology. The domain ontology consists of two parts: the conceptual model[19] and the domain-specific lexicon[20]. The conceptual model is primarily composed of item raw material categories and their attributes, while the domain-specific lexicon contains attribute feature terms for each category in the conceptual model. Secondly, identify standardized item samples. Based on the conceptual model and domain-specific lexicon, formulate the correspondence rules between item raw materials categories and attributes. Using the idea of self-supervised learning, check whether the category information provided by the merchant matches the attribute information, and include samples that meet the matching rules as standardized item samples, while the rest are non-standard item samples. Subsequently, automatically add category labels to

Example	Data point	Text content
Example 1	Title	金沙河自发粉家用小麦粉包子馒头油条中筋面粉免酵母通用面粉1kg 自发小麦粉1kg*10/20斤
	Detail	[{"品牌(Brand)":"樱安娜"},{"商品名称(Item Name)":"金沙河自发粉家用小麦粉包子馒头油条中筋面粉免酵母通用面粉1kg 自发小麦粉1kg*10/20斤"},{"商品编号(Item Number)":"10050083362787"},{"店铺(Store)":" 季馨晨旗舰店"},{"商品毛重(Gross Weight)":"100.00 g"},{"筋度(Gluten)":"中筋面粉"},{"包装形式(Packing Form)":"桶装"},{"类别(Item Category)":"自发粉"},{"净含量(Net Content)":"0-500 g"}]
Example 2	Title	低筋面粉5斤蛋糕面包家用粉烘焙蒸蛋挞月饼糕点饼干低筋小麦粉1斤 1斤低筋烘培面粉+熟白芝麻100g
	Detail	[{"品牌":"华海春"},{"商品名称":"低筋面粉5斤蛋糕面包家用粉烘焙蒸蛋挞月饼糕点饼干低筋小麦粉1斤 1斤低筋烘培面粉+熟白芝麻100g"},{"商品编号":"10048236400265"},{"店铺":" 赛豪汇明休闲食品专营店"},{"商品毛重":"1.0kg"},{"商品产地":"中国大陆"},{"类别":"麦芯粉"},{"包装形式":"袋装"},{"筋度":"低筋面粉"},{"净含量":"0-500 g"}]

Table 1. An example of information from an online item description.

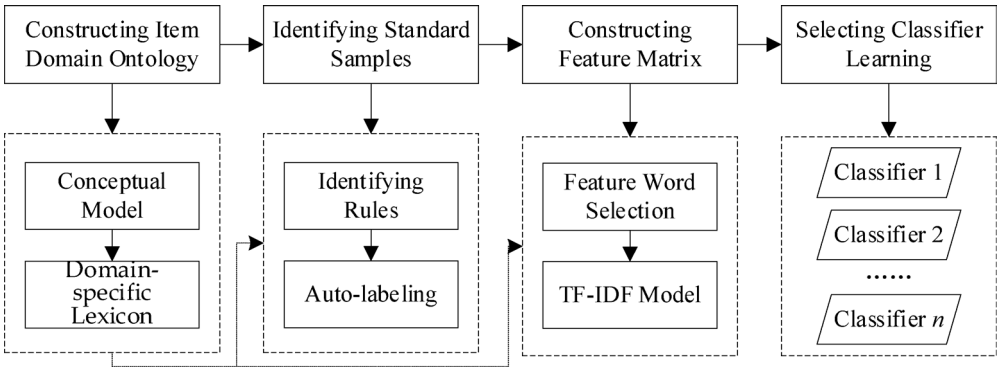

Fig. 1. The classification model constructed in this paper.

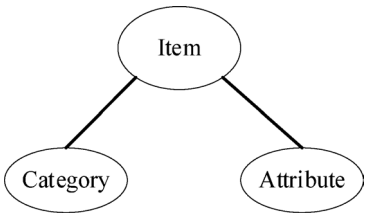

Fig. 2. Item ontology conceptual model.

standardized item samples. Thirdly, enrich the domain-specific lexicon item entity names using BERT[21] and use the domain-specific lexicon as an auxiliary segmentation tool to tokenize item texts. Then, build the feature matrix based on the TF-IDF model[22]. Finally, select a classifier for learning. After obtaining the automatically labeled standardized item labels and feature matrix, select a machine learning classifier for learning and training, and finally achieve category prediction based on item raw materials.

Constructing the item domain ontology

In order to determine the categories of e-commerce items based on their raw materials, it is necessary to have domain experts who can establish categorization criteria and develop a domain ontology. This encompasses a theoretical framework and specific examples. The conceptual model depicts the arrangement of item categories within this e-commerce sector and the distinct characteristics associated with each category. In this case, instances refer to the vocabulary used inside a specific domain, and they serve as a concrete representation of the conceptual model. Their primary functions are twofold: firstly, to serve as a benchmark for evaluating the adherence of item samples to categorization criteria, and secondly, to provide a reliable point of reference for extracting precise characteristics of item raw materials.

Constructing the conceptual model

The item ontology conceptual model constructed in this paper is shown in Fig. 2. The item ontology is represented by the binary tuple $O = (C, A)$, where C represents the categories of item raw materials formulated by domain experts, and A represents the expert-defined item attributes. Taking flour items sold on e-commerce platforms as an example, the item category defined by flour domain experts is wheat flour, and the item attributes include flour raw materials, flour gluten strength, flour origin, etc.

Item attributes consist of primary attributes and supplementary attributes, represented as $A = (K, S)$, where primary attributes K represent the essence of the item, such as raw materials, and supplementary attributes S are additional explanations of the item, such as origin, brand, function, etc. Taking flour as an example, the primary attribute is the raw material of the flour, and the supplementary attributes include flour gluten strength, flour quality, flour brand, flour origin, etc.

Generating the domain-specific lexicon

During the instantiation of the conceptual model, domain experts supply the categories and characteristics of the item, while the attribute values are acquired by comparing the item details information provided by e-commerce platforms. Table 1 displays a significant amount of terminology associated with attribute characteristics in the item details information. The vocabulary is presented in the form of key-value pairs, represented as $D = [\{\, attribute :\ attribute\ value \,\}]$. For example, the details of a certain flour item are:

[{Brand: Xinliang}, {Gluten Strength: Medium-Gluten Flour}, {Accuracy: Standard Flour}, {Origin: Xinxiang, Henan}]

Considering that regular expression[23] is a powerful text matching tool that allows users to search and extract text according to custom rules, regular expressions can be used to obtain attribute feature vocabulary from item details.

Given the item attributes $A = (K, S)$ to obtain the corresponding content of A, it is necessary to construct matching rules for both K and S. It's important to note that the presentation of domain ontology content on different platforms may not be entirely consistent, so it is necessary to construct appropriate matching patterns based on the presentation form of the content to obtain the content of the domain ontology. For instance, e-commerce platforms like Tmall and JD in China present item details in the form $D = \{attribute: attribute\ value\}$. However, there are differences in the matching rules between the two platforms due to the presence of spaces in the item details on the Tmall platform. The matching rules are as follows:

r'{"Kt": "(.+?)"}
r'{"St": "(.+?)"}
r'{"Kj":"(.+?)"}'.
r'{"Sj":"(.+?)"}'.

The details information on Suning platform is presented in the form of $D1 = [attribute \quad attribute\ value]$, and the construction rules are as follows:

r'Ks (\w+){1}', r'Ss (\w+){1}'.

The attribute vocabulary extracted by regular expressions is relatively rough, so further processing is needed. The precise steps are as follows: Initially, analyze the entity lexicon. The second objective is to augment the domain vocabulary through the application of the Levenshtein distance[24].

Regarding processing, certain words can be too lengthy and unwieldy, failing to accurately and straightforwardly convey the issue. In this instance, the lexicon can be divided to derive fresh terms, so reducing the material to improve readability and recognizability. Taking flour as an example, the extracted origin information is: {China mainland xx province xx city xx county, China xx autonomous region xx autonomous prefecture, China xx city xx district, USA, Russia, …}

As seen, the origin information is quite long, and it's necessary to process it by dividing the origin information into the pattern of "(country - province - city)". When the country is a non-Chinese country, the country is the name of that country, and the province and city are ignored. Otherwise, the country is set to China, and the province and city are set to the corresponding information.

Additionally, we employed the Levenshtein distance to quantify the similarity between strings, facilitating the matching of attribute values within a specified edit distance. The specific process is as follows. First, import the Levenshtein distance package, then set the standard word list according to the opinions given by domain experts, next set the threshold for measuring text distance using the Levenshtein distance (generally set to 2 characters), and finally remove duplicates from the matched keywords and add them to the original domain word library. For example, in this study, for the keywords of the flour accuracy attribute, we introduced the Levenshtein distance for expansion. The original keywords were ["特一粉", "特二粉", "普通粉", "标准粉"]. After fuzzy matching using the Levenshtein distance, we obtained ["标准面粉", "特制一等面粉", "特制二等面粉", "普通面粉"] as the keywords. The specific code is shown in Fig. 3. This can effectively expand the scope of the domain word library, help to expand the coverage of the standard samples, and reduce the gap in manual annotation. Similarly, we can also match more domain words based on other attributes of flour, such as category, brand, and elasticity.

Meanwhile, by constructing a synonym dictionary for domain attribute words to expand synonyms, different expressions with the same semantic meaning are included in the matching scope (for example, "wheat germ powder" and "wheat core powder"). This adjustment significantly expands the coverage of the standard sample, reduces the missed detections caused by strict matching, and thereby reduces the gap in manual annotation.

Identification of standardized item name samples

Upon examining the item description, it is evident that only a small number of merchants include item category information based on the raw ingredients in the item specifics. Other merchants either lack the provision of it or offer information that deviates from the standard. This study presents a method for recognizing standardized

```
standard_words_jinli=["高筋", "中筋", "低筋"]
def find_fuzzy_matches_jinli(candidates, standards, max_distance=None):
    if max_distance is None:
        max_distance = min(2, len(candidates) // 2)  # 动态阈值

    related_words = set()
    for word in candidates:
        for std in standards:
            if levenshtein_distance(word, std) <= max_distance:
                related_words.add(word)
    return list(related_words)

fuzzy_words = find_fuzzy_matches_jindu(data['筋力'].value_counts().index.tolist(),standard_words_jinli)
print("模糊匹配相关词:", fuzzy_words)
```

```
模糊匹配相关词: ['高筋面粉', '低筋面粉', '中筋', '高筋', '中筋面粉']
```

Fig. 3. Match fuzzy keywords based on the Lewinstein distance.

item samples in order to automatically acquire item raw material categories and minimize excessive resource use. The method is illustrated in Fig. 4.

Specifically, first, formulate the correspondence rules between item categories and attributes based on the item ontology conceptual model. Taking C and A as the representation of category and attribute, there is a correspondence relationship $C \rightarrow A$. Secondly, extract the category and attribute values of the item samples using regular expressions. Third, use the idea of self-supervised learning to check whether the extracted information matches the corresponding rules. If it matches, it is a standardized sample.

Building corresponding rules

According to the conceptual model, formulate the corresponding rules between item categories and attributes. Use A to represent the item category information and A to represent the item attribute information. There is a correspondence rule $C \rightarrow \{A_1, A_2, \ldots, An\}$, where $A_1, A_2, \ldots, An$ are specific attributes. It is worth noting that after the correspondence between the standardized category C and attribute A, when checking the item samples, the category A can only be determined if the main attribute is satisfied. If the main attribute cannot determine the category of the sample, it needs to be determined based on the main attribute and supplementary attributes. For example, for corn flour, the corresponding rule is:

Corn flour → {(corn, maize), (medium gluten, low gluten), (Xinliang, Ying'anna, Yesanpo…), (Liaoning, Shandong, Henan…)}

If the sample attribute values include corn, the category of the sample is directly determined as corn flour. If the sample attributes include both corn and sorghum, it is necessary to count the word frequency of corn and sorghum, as well as the value range of the supplementary attributes, to determine the item category.

Extracting item sample information using regular expressions

First, construct regular expression matching rules to extract item category and type information. Considering that different platforms provide inconsistent item category information, there are differences in constructing regular expressions. Specifically:

r'{"Flour Type": "(.+?)"} .findall('Sample [Details]') and r'{"Flour Attribute": "(.+?)"} .findall('Sample [Details]').

Secondly, after running the code, the category and attribute information of each item sample can be obtained, and then they are constructed in the form of A → A, and then waiting for subsequent verification.

Verifying standardized item name samples

After obtaining the relationship between item sample categories and attributes, it is necessary to verify the samples using the principle of self-supervised learning[16]. This approach achieves two main goals. Firstly, it efficiently obtains labels by reducing the need for manual labeling, thereby reducing time and cost in engineering applications. Secondly, it ensures high matching accuracy of standardized samples by learning the golden rules provided by the item ontology. This allows for more precise verification of the compatibility between sample attributes and categories, resulting in standardized samples based on item raw materials as the classification basis.

The implementation procedure comprises two distinct steps. Firstly, verify the existence of the item category and its compliance with the predefined concept model provided by the merchants. If either criteria is not satisfied, it signifies that the sample is non-standard and should be included in the test set. Secondly, if the item category satisfies the criteria of the domain ontology, additional verification is necessary to ensure that the

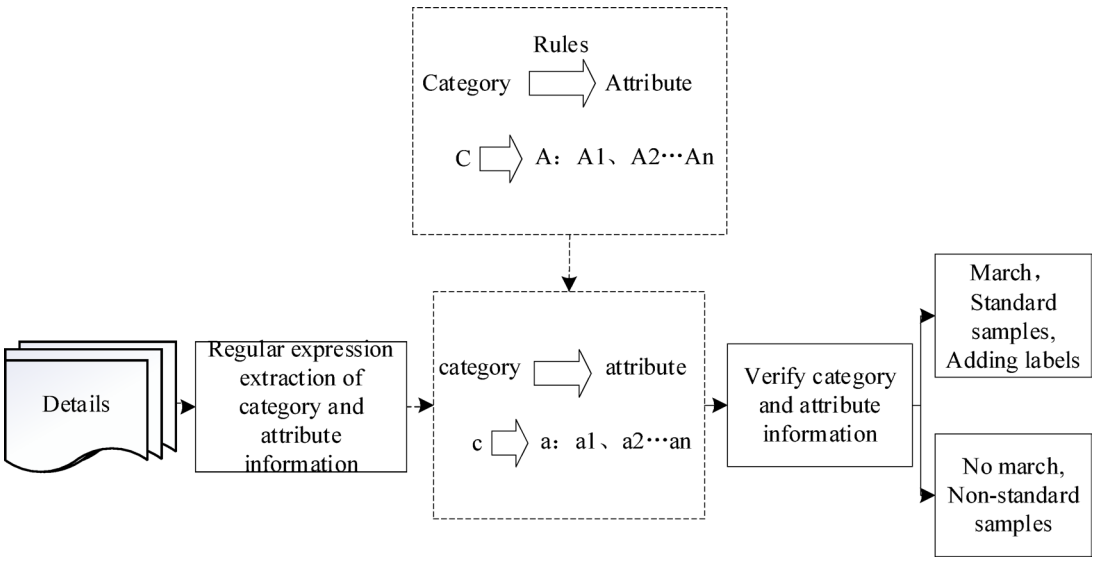

Fig. 4. Method for identifying standard item samples.

Fig. 5. Process of constructing item feature keywords.

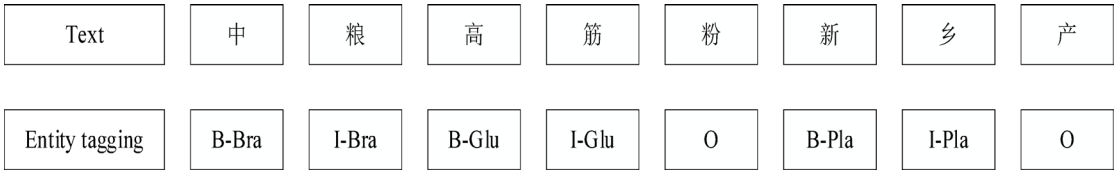

Fig. 6. BERT entity tagging example.

attribute values conform to the established standards. If there is no match, the sample should be included in the test set as well. Simultaneously, the corresponding samples should be included in the training set and assigned category labels automatically.

Constructing feature matrix

The specific process of constructing the feature keyword for the item is shown in Fig. 5. Firstly, BERT (Bidirectional Encoder Representations from Transformers) is used to extract entity keywords from the item. These extracted keywords are then processed to enrich the domain-specific lexicon. Next, a segmentation tool is used to tokenize the text and create a feature matrix based on the TF-IDF (Term Frequency-Inverse Document Frequency) model.

Enriching the domain-specific lexicon

The domain-specific lexicon plays a crucial role in text segmentation and feature extraction for item texts. Considering that item titles often contain entity keywords that represent item characteristics, extracting and adding them to the domain-specific lexicon significantly enhances its vocabulary size and improves the accuracy of text segmentation and feature extraction. To extract entity keywords from item titles, this study utilizes a pre-trained language model called BERT (Bidirectional Encoder Representations from Transformers). BERT is a bi-directional training language model based on deep Transformer architecture, which was introduced in late 2018. It captures rich semantic information from large-scale unlabeled text and continuously adjusts model parameters for prediction, making it highly effective in extracting contextual features[21].

In this study, a BERT entity tagging process based on BIO (Begin, Inside, Outside) labeling is employed. BIO labeling is an abbreviation for the beginning, inside, and outside of an entity. The specific tagging process is as follows: Firstly, collect text data with entity annotations. Each entity is labeled with specific entity tags, such as B-Bra for the beginning of a brand entity, I-Bra for the inside of a brand entity, and O for content outside of an entity. Secondly, choose a suitable BERT model. In this study, the "bert-base-chinese" pre-trained model[25] for Chinese entities is utilized. Thirdly, create a fine-tuning dataset for Chinese entity recognition and fine-tune the pre-trained model using this dataset. Finally, predict the entities in the item text.

For example, in the text "中粮高筋粉新乡产" (as shown in Fig. 6), "中粮" is labeled as a brand entity, "高筋" is labeled as a gluten strength entity, and "新乡"is labeled as a location entity. Initially, the text is separated into "中 粮 高 筋 粉 新 乡 产" as a single text; The text "中" is marked as the brand entity to start B-Bra, the text "粮" as the brand internal I-Bra, the text "高" as the strength entity to start B-Glu, the text "筋" as the strength internal I-Glu, the "新" as the place name entity to start, and the "乡" as the place name entity. These markings are used to identify and designate the corresponding entities for individual characters. Once the non-physical components "粉" and "产"are identified, the flour item entity name is obtained.

To enhance the domain lexicon, we chose the "RaNER Named Entity Recognition - Chinese - E-commerce Domain - Fine-grained" entity extraction model from Alibaba ModelScope[26]. This model is predicated on the Chinese e-commerce sector and is capable of executing entity recognition and extraction on the titles of the training samples presented in this article. The model is specifically loaded in the ModelScope online notebook, where sample title text is input for recognition and entity word output. Subsequently, words with high recognition probabilities (typically set above 0.5, and in this article, set to 0.7) are extracted based on attributes (category, brand, place of origin) within the domain ontology. Finally, single characters and nonsensical characters are removed and recorded in the domain vocabulary. The particular code is illustrated in Fig. 7. In summary, employing this entity extraction model enhances the accuracy of domain-specific terminology, hence improving text segmentation of the test set and yielding more precise feature extraction.

Tokenization and feature matrix construction

Initially, using a tokenization tool to import the stop word dictionary and domain-specific word library. The stop word dictionary facilitates the elimination of frequently occurring words, symbols, or idioms that lack

```python
from collections import defaultdict
def get_target_entities_batch(data_list, prob_threshold=None):

    result = defaultdict(set)  # 使用集合自动去重

    for ner_dict in data_list:
        for entity in ner_dict["output"]:
            # 概率过滤
            if prob_threshold is not None and entity["prob"] <= prob_threshold:
                continue

            if entity["type"] == "品牌":
                result["品牌"].add(entity["span"])
            elif entity["type"] == "地点地域_产地":
                result["产地"].add(entity["span"])
            elif entity["type"] == "产品_核心产品":
                result["核心产品"].add(entity["span"])
            elif entity["type"] == "适用范围_适用对象":  # 修正字段名
                result["适用对象"].add(entity["span"])

    return {k: list(v) for k, v in result.items()}
```

Fig. 7. Settings for entity extraction codes in the e-commerce field.

substantive significance in the text. The domain-specific lexicon improves the tokenization process by precisely retaining terms that depict the attributes of the items. For example, in the text segment T:

5 kg of low-gluten flour for cakes and bread, along with 100 g of roasted white sesame seeds.

After tokenization, its content is Ts: 5/kg/of/low/gluten/flour/for/cakes/and/bread/along/with/100/g/of/roasted/white/sesame/seeds.

After removing stop words, it is Tss: low/gluten/flour/cakes/bread/roasted/white/sesame/seeds after loading the domain-specific lexicon, it is $Tssd$: low gluten/flour/cakes/bread/white sesame seeds.

After obtaining the feature words, it is necessary to vectorize them to form a word vector matrix. In this paper, the TF-IDF word vector model is used to process the feature words. TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used technique in text information retrieval and text mining, which measures the importance of a word in a document collection[27]. TF-IDF consists of two parts: Term Frequency (TF), which refers to the frequency of a word appearing in a document. It is commonly used to represent the importance of a word in a document, where higher frequency indicates higher importance. Inverse Document Frequency (IDF) refers to the frequency of a word appearing in the entire document collection. IDF is used to measure the rarity of a word in the entire document collection, where rarer words have higher weights.

The formulas for calculating TF, IDF, and TF-IDF are as follows (1–3):

$$TF_{i,j} = \frac{n_{i,j}}{\sum_{k=1}^{k} n_{k,j}} \tag{1}$$

$$IDF_i = \log\left(\frac{n_d}{df\,(d, w_i) + 1}\right) \tag{2}$$

$$TFIDF = TF * IDF \tag{3}$$

Translation: In the formulas, $TF_{i,j}$ represents the frequency of the term w_i appearing in the document d_j, $n_{i,j}$ is the number of times w_i appears in the document d_j, the denominator is the sum of the frequencies of all feature words in the document d_j, k is the number of different words in the document d_j. IDF_i represents the inverse document frequency of the feature word w_i in the document d_j, n_d is the total number of documents in the text corpus, $df(d, w_i)$ is the number of documents in which the feature word w_i appears, and adding 1 is to prevent the denominator from being 0.

Using TF-IDF as the feature vectorization tool, the feature words of standardized and non-standardized item samples are transformed into a feature matrix for subsequent machine learning training and prediction.

Classifier selection

Once the item category labels and feature matrix have been acquired, it is crucial to choose an appropriate classifier for the purpose of training and testing. Classifiers encompass conventional machine learning techniques like logistic regression, naive Bayes, random forest, and XGBoost[28], as well as advanced deep learning algorithms such as CNN, RNN, and LSTM[29]. The focus of this work is to classify brief texts. Conventional machine learning

Classification	Attribute				
	Raw material	Gluten strength	Quality	Brand	Origin
Wheat flour	Wheat	High-gluten flour	Special grade A	Xinliang	Henan
Whole wheat flour	Whole wheat	Low-gluten flour	Regular	Huidao	Shanxi
Buckwheat flour	Buckwheat	Medium-gluten flour	Standard	Tianye	Neimeng
Corn flour	corn	Low-gluten flour	Standard	Heyuan	Liaoning
Glutinous rice flour	Glutinous rice	Low-gluten flour	Regular	Xinliang	Henan
Rice flour	rice	Low-gluten flour	Regular	Xingyi	Yunnan
Others	Sorghum	Low-gluten flour	Special grade B	Xiaji	Jiangsu

Table 2. Example of flour domain lexicon.

Sample examples	Sample classification	Sample raw materials	Within classification range	Meets recognition rules	Is standardized
Example 1	Self-rising flour	Wheat	No	No	No
Example 2	Glutinous rice flour	Wheat	Yes	No	No
Example 3	Corn flour	Corn	Yes	Yes	Yes
Example 4	Buckwheat flour	Buckwheat(frequency = 3),whole Wheat(frequency = 1)	Yes	Yes	Yes

Table 3. Example of identification of standardized item samples.

methods can attain high accuracy in this scenario, however deep learning algorithms necessitate substantial processing resources and a large quantity of labeled data, rendering them inappropriate for this methodology.

Experimental validation

To validate the feasibility and effectiveness of the classification model mentioned above, this study conducted a classification of flour items on e-commerce platforms obtained through web crawling. Comparative experiments were designed to further verify the accuracy of the classification model.

Data source

The study's data was obtained in May 2022 from e-commerce platforms (Tmall, Suning, and JD) utilizing web crawling techniques. The collection comprises 21,590 entries of flour items. Following the elimination of duplicates, noise, and the treatment of missing values, a grand total of 18,045 samples were ultimately selected for analysis.

Identification of standardized item samples

First, the domain lexicon is generated based on the flour ontology concept model. The ontology conceptual model is composed of different types and attributes of flour. The flour categories are provided by domain experts and include wheat flour, whole wheat flour, buckwheat flour, corn flour, glutinous rice flour, rice flour, and others. The flour attributes are obtained from the flour descriptions, using regular expressions to extract attribute values for flour raw materials, gluten strength, quality, brand, and place of origin from the item details. The generated flour domain-specific lexicon is shown in Table 2.

Next, the identification rules for standardized item samples are constructed based on the domain-specific lexicon. The implementation process of these rules is as follows: Extract the category and attribute values of the flour samples. Check whether the category is within the range of categories provided by the domain experts. If it falls outside of this range, the sample is considered a non-standardized sample. If the category is within the range, examine if the flour raw material information in the attributes matches the requirements of the domain-specific lexicon for a single category. If it matches, the sample is considered a standardized item sample and the corresponding category label is added. If the flour raw material information matches the requirements of multiple categories, further verification is needed. This is done by comparing the frequency[30] of the raw material information. The category is determined based on the raw material with the highest frequency. If the frequencies are equal, the requirements of the supplementary attributes are used to determine the category. An example of the identification of standardized item samples is shown in Table 3.

Ultimately, the standardized item samples are allocated to the training set, and the non-standardized item samples are allocated to the testing set. The unstandardized data are labeled manually based on the related criteria for assessing the machine learning model. A total of 8,351 standardized item samples were obtained during the identification process, representing 46.3% of the total samples. The samples included 5,814 samples of wheat flour, 863 samples of whole wheat flour, 699 samples of buckwheat flour, 161 samples of maize flour, 101 samples of glutinous rice, 79 samples of rice flour, and 656 samples of other flour varieties.

Experimental analysis

Evaluation criteria

The evaluation criteria for classification are usually precision (P), recall (R), and F1 score[31]. In this study, the method used is to treat the target labels as positive class and other categories as negative class. A confusion matrix is constructed to calculate the metrics for each category label. The number of samples correctly classified as a certain category label is denoted as TP (true positive), the number of samples incorrectly classified as that label is denoted as FP (false positive), the number of samples correctly classified as other category labels is denoted as TN (true negative), and the number of samples incorrectly classified as other category labels is denoted as FN (false negative). The formulas for calculating P, R, and F1 are as follows (4–6):

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

Selection of classifier

By comparing the performance of different classifiers for the classification of e-commerce item raw materials, a specific classifier is selected. The compared classifiers include Logistic Regression (LR) classifier[32], Multinomial Naive Bayes (MNB) classifier[33], Decision Tree (DT) classifier[34], Random Forest (RF) classifier[35], Radial basis function SVM classifier[36], and Shallow neural network classifier. In the experiment, except for the different models, other features and conditions are the same, and the specific results are shown in Table 4. According to Table 4, it is evident that the logistic regression classifier performs the best, with precision (P), recall (R), and F1 score all reaching 0.91, indicating the best performance. According to the classification results in Table 4, the p-values, r-values and f1-values of the radial basis function SVM and the shallow neural network classifier are all higher than 0.85, indicating a good classification effect. However, their p, r and f1 values are lower than those of the logistic regression classifier, and thus cannot classify the flour data more precisely. Therefore, this study will use the logistic regression classifier as the classification tool for training and prediction.

Comparative analysis

To assess the efficacy of the e-commerce flour classification method presented in this paper, two controlled experiments were conducted. The first aimed to compare the performance of feature extraction and classification with and without a domain lexicon to evaluate the lexicon's contribution. The second sought to compare the accuracy of a Chinese zero-shot large language model with the classification model based on domain ontology proposed herein to validate the latter's advantages. This study's initial controlled experiment is to evaluate the efficacy of the domain-specific lexicon in extracting properties of raw materials for e-commerce products. The comparison of results is based on the utilization of the domain-specific lexicon. The objective is to evaluate the benefits and drawbacks of feature extraction for e-commerce item raw materials, comparing the existing classification methods with the suggested method in this research. Table 5 presents the categorization outcomes of e-commerce flour items utilizing the domain-specific vocabulary suggested in this research. In contrast, Table 6 displays the categorization outcomes of e-commerce flour items using conventional approaches, specifically without employing the domain-specific vocabulary.

Table 5 demonstrates that employing a domain-specific lexicon for extracting characteristics of flour raw materials yields a classification accuracy of 0.91. Out of all the options, wheat bran flour and glutinous rice flour demonstrate the highest level of classification accuracy, surpassing 90%. Both whole wheat flour and glutinous rice flour also exhibit a level of precision ranging from 85% to 90%. Buckwheat flour, corn flour, and other categories exhibit a somewhat lower level of accuracy, however all surpass the threshold of 75%. The accuracy achieved using a weighted average is 90%. Additional examination indicates that the decreased precision in distinguishing buckwheat flour, maize flour, and other flour types is a result of sporadic similarities between buckwheat flour and whole wheat flour, causing some uncertainty. The presence of corn flour is common in mixed grain items, and the incorporation of mixed grains in other item categories may have a little impact on the classification of corn flour. The category of "other" flour is extensive, and during the prediction process, there

Model	Precision ratio (*P*)	Recall rate (*R*)	F1
Logistic regression classifier	0.91	0.91	0.91
Multivariate Bayesian classifier	0.86	0.46	0.57
decision tree classifier	0.80	0.65	0.69
Random forest classifier	0.79	0.71	0.73
Radial basis function SVM classifier	0.88	0.86	0.86
Shallow neural network classifier	0.87	0.85	0.86

Table 4. Classification results of different machine learning models.

Classification	Precision	Recall	F1	Number of Categories
Wheat flour	0.95	0.96	0.95	6922
Whole wheat flour	0.87	0.77	0.82	773
Corn flour	0.71	0.90	0.80	354
Buckwheat flour	0.76	0.86	0.81	284
Glutinous rice flour	0.94	0.73	0.82	44
Rice flour	0.90	0.90	0.90	67
Other	0.79	0.74	0.76	1250
Avg/total	0.91	0.91	0.91	9694

Table 5. Classification results using the domain-specific lexicon.

Classification	Precision	Recall	F1	Number of Categories
Wheat flour	0.95	0.95	0.95	6922
Whole wheat flour	0.86	0.77	0.81	773
Corn flour	0.70	0.88	0.78	354
Buckwheat flour	0.76	0.86	0.81	284
Glutinous rice flour	0.94	0.73	0.82	44
Rice flour	0.86	0.90	0.88	67
Other	0.77	0.73	0.75	1250
Avg/total	0.90	0.90	0.90	9694

Table 6. Classification results without using the domain-specific lexicon.

Category	Precision ratio (*P*)	Recall rate (*R*)	F1	Category correspondence number
Wheat flour	0.90	0.85	0.88	6922
Whole wheat flour	0.55	0.79	0.65	773
Corn flour	0.72	0.83	0.77	354
Buckwheat flour	0.52	0.85	0.65	284
Glutinous rice flour	0.10	0.50	0.16	44
Rice flour	0.04	0.40	0.07	67
Others	0.41	0.07	0.13	1250
avg/total	0.78	0.74	0.74	9694

Table 7. StructBERT zero-sample classification - Chinese-base classification results.

may be instances of things that the classifiers have not completely "learned," resulting in reduced classification accuracy.

Upon comparing Tables 5 and 6, it is apparent that there has been an enhancement in the overall accuracy of the e-commerce flour item classification. Regarding accuracy, whole wheat flour, maize flour, other categories, and glutinous rice flour have each experienced an increase of 1%, 1%, 2%, and 4% respectively. Regarding recall, wheat bran flour, the "other" category, and maize flour have each experienced a respective increase of 1%, 1%, and 2%. Regarding the macro F1 metric, there has been a 1% improvement in whole wheat flour, a 1% improvement in the "other" category, a 2% improvement in maize flour, and a 2% improvement in glutinous rice flour. The e-commerce item classification model described in this research, which utilizes a domain-specific vocabulary, is capable of accurately extracting raw material properties of items. Furthermore, its classification performance surpasses that of existing e-commerce item classification approaches.[37]

For the second controlled experiment, in order to verify the effectiveness of the classification accuracy of the method proposed in this research, we selected a currently popular large language model for the controlled experiment. Specifically, a typical large model that supports Chinese zero-shot classification in Alibaba's ModelScope community was selected, namely StructBERT zero-sample classification - Chinese -base. e designed a unified prompt template for the test set samples, requiring the large language model to classify based on the given flour category. Finally, we compared the effectiveness of the Chinese zero-shot classification large model and the classification model proposed in this paper based on the f1 value of the classification results[38]. The classification results of the large prediction model are shown in Table 7. According to Table 7, the classification effects of glutinous rice flour, sticky rice flour, and other types of flour are poor, with an overall f1 value of 0.74, which is lower than the f1 value of the classification model designed in this paper (the f1 value of this paper is 0.91).

The experimental results show that, on the same test set, the automatic label assignment method based on ontology guidance outperforms the selected zero-sample LLM model in terms of classification accuracy. Analyzing the reasons, the main factor lies in that StructBERT zero-sample classification - Chinese-base relies more on general semantic understanding and context reasoning. For samples with specific concept systems and semantic constraints within a certain domain, it is prone to generate ambiguity due to the lack of precise guidance from domain knowledge. However, in our study, our method can more accurately capture the domain-specific associations between samples and labels. Thus, it improves the classification accuracy. Through the above supplementation and optimization, we further highlight the innovation points of the research and verified the novelty and superiority of the proposed process through experiments.

Discussion

This study utilizes flour data from Chinese e-commerce platforms to validate the efficacy of the suggested product categorization algorithm. The experimental findings indicate that the logistic regression approach attains a classification accuracy of 91%. Nonetheless, as the experiment was limited to flour data within a singular category, we will further examine the model's universality in this part.

We assert that the overarching framework of this research methodology extends beyond the flour sector and possesses applicability in other domains. This method's fundamental process involves creating a product domain ontology, standardizing product sample identification, developing a feature matrix, and selecting a classifier for learning purposes. This approach is constant across several sectors of application.

In the development of the product domain ontology, applicable to flour or other items, both the conceptual model and the domain lexicon are essential. The conceptual model comprises categories of product raw materials and their characteristics, whilst the domain lexicon serves as the attribute feature vocabulary for each category inside the conceptual model. In standardizing product sample identification, rules are established based on the conceptual model and domain lexicon, utilizing self-supervised learning to verify the congruence of category and attribute information provided by merchants, thereby filtering standardized samples and automatically assigning labels. In generating the feature matrix, BERT is employed to extract product entity names to enhance the domain lexicon, which is subsequently utilized as an auxiliary word segmentation method in conjunction with the TF-IDF model to develop the feature matrix. Subsequently, upon acquiring the automatically annotated labels and the feature matrix, a suitable machine learning classifier is chosen for training and learning to facilitate category prediction.

The critical aspect necessitating modification according to certain domains resides in the formulation of the domain ontology, which is contingent upon the expertise within the specific field. Various goods across distinct fields possess diverse categories of raw materials and attribute characteristics; therefore, in developing the conceptual model and the domain lexicon, as well as in defining corresponding regular expression matching rules, adjustments must be made in accordance with the specific knowledge of that field.

The term "self-supervised" in this study represents an adaptation and application of the fundamental concept of self-supervised learning, rather than adhering rigidly to established paradigms of self-supervised learning. This discrepancy has been explicitly articulated through modifications in the paper's language. The primary objective of self-supervised learning is to diminish dependence on externally labeled data by deriving supervisory signals from the intrinsic content or structure of the data, thus facilitating model training and problem resolution. This study's approach design is explicitly focused on this fundamental objective. The huge quantity of e-commerce product data renders the expense of manually labeling each sample exceedingly high. Consequently, we incorporate domain ontology knowledge to produce initial supervisory signals, specifically the labels of normative samples, so substituting the conventional manual labeling process. Consequently, with these supervisory signals, the model enhances its understanding of implicit features within the data, thereby augmenting its capacity to process unlabeled samples. The procedure continually follows the fundamental principle of minimizing dependence on external manual labeling and attaining automated processing via data-related information, aligning with the primary objective of self-supervised learning. Our decision to forgo popular self-supervised methods, such as contrastive learning, is predicated on the specific research context and task demands. The product classification task in e-commerce possesses significant domain attributes, with the features and logical relationships of product data being heavily reliant on domain expertise. Conversely, conventional techniques such as contrastive learning depend predominantly on overarching structural characteristics of the data (for instance, semantic similarities in text or visual attributes in images), hence complicating the precise identification of domain-specific correlations in product data. Consequently, we opted to integrate domain ontology rules with the concept of self-supervised learning, so producing supervisory signals that align more closely with domain specifications through these rules, and subsequently extracting the implicit features within the domain. This strategy not only minimizes manual intervention but also improves the adaptability of label creation to other domains, rendering it a suitable alternative for particular task scenarios.

Conclusion

This work successfully implemented a self-supervised learning model to classify materials of e-commerce items. As a result, e-commerce platforms can now classify items based on their original materials. The resulting conclusions were as follows. (1). The study creates an item domain ontology and defines item categories based on raw materials, which serves as a robust framework for classifying objects that share the same raw materials. Additionally, it provides useful insights for firms and workers engaged in sales statistics and management pertaining to the raw ingredients of the items. (2) Employing self-supervised learning for the identification of standard samples minimizes the need for human processing and preserves resources, rendering it exceedingly practical for engineering applications. (3) Classifying e-commerce items involves categorizing short texts,

and classical machine learning approaches can achieve accurate results. Developing intricate deep learning algorithms for classification is unnecessary, leading to a substantial reduction in burden.

The suggested classification model does possess certain constraints. Initially, there are difficulties in formulating regulations for determining standard item samples, as the attribute data offered by various platforms may differ. It is crucial to carefully evaluate both the number of attributes and the diversity of attribute values. An attribute design that is more permissive could lead to a higher number of identified standard item samples, whereas a setup that is more stringent could result in a drop in the number of identified samples. Furthermore, the logistic regression technique can be optimized to better handle the increased number of item attributes in order to raise classification accuracy, given its intrinsic linearity. This research, while novel, possesses certain limitations. Future research will involve a comprehensive comparative case study on integrating domain ontology rules with self-supervised reasoning and the rule-based methodology. By means of intuitive case comparisons, readers will discern the distinctions in processing capacity and efficacy between the two methodologies more clearly.

Data availability

All data generated or analysed during this study are included in this published article and its supplementary information files. The data that support the findings of this study are available on request from the corresponding author.

Received: 28 May 2025; Accepted: 29 January 2026
Published online: 10 February 2026

References

1. Tan, L., Li, M. Y. & Kok, S. E-commerce product categorization via machine translation. *ACM Trans. Manage. Inform. Syst.* **11** (3), 1–14 (2020).
2. Shen, D., Ruvini, J. D., Mukherjee, R. & Sundaresan, N. A study of smoothing algorithms for item categorization on e-commerce sites. *Neurocomputing* **92**, 54–60 (2012).
3. Osaka, J. Large-scale Multi-class and Hierarchical Item Categorization for an E-commerce Giant. In *Proceedings of COLING the 26th International Conference on Computational Linguistics: Technical Papers*. Available: https://aclanthology.org/C16-1051 (2016).
4. Fine-Grained Item Categorization in E-commerce. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. Available: https://dl.acm.org/doi/https://doi.org/10.1145/3357384.3358170
5. Ma, P. H. et al. Ahuja. Deep learning accurately predicts food categories and nutrients based on ingredient statements. *Food Chem.* **391**, 133243 (2022).
6. Kim, Y. & Lee, H. J. and J. Shim. Developing Data-Conscious deep learning models for item classification. *Appl. Sci.* **11**(12), 5694, (2021).
7. Lu, H. et al. AutoD: intelligent blockchain application unpacking based on JNI layer deception call. *IEEE Netw.* **35** (2), 215–221 (2021).
8. Product Classification in E-Commerce using Distributional Semantics. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. Available: https://arxiv.org/abs/1606.06083
9. Washington, D. C. USA. Large-scale taxonomy categorization for noisy item listings. *IEEE International Conference on Big Data (Big Data)*. Available: https://ieeexplore.ieee.org/abstract/document/7841063
10. Large-Scale Item Categorization in e-Commerce Using Multiple Recurrent Neural Networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Available: https://dl.acm.org/doi/abs/https://doi.org/10.1145/2939672.2939678
11. Dhaka, B. A. & Novel idea of classification of e-commerce items using deep convolutional neural network. In *4th International Conference on Electrical Engineering and Information & Communication Technology*. (2018). https://ieeexplore.ieee.org/abstract/document/8628161
12. San Francisco, C. A. USA. Open-world Learning and Application to Item Classification. WWW '19: The World Wide Web Conference. Available: https://dl.acm.org/doi/abs/https://doi.org/10.1145/3308558.3313644
13. Qiao, C., Zeng, Y., Lu, H., Liu, Y. & Tian, Z. An efficient incentive mechanism for federated learning in vehicular networks. *IEEE Netw.* **38** (5), 189–195 (2024).
14. Tan, Y., Huang, W., You, Y., Su, S. & Lu, H. Recognizing BGP communities based on graph neural network. *IEEE Netw.* **38** (6), 282–288 (2024).
15. Valencia, S. Web-scale language-independent cataloging of noisy item listings for e-commerce. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Available: https://aclanthology.org/E17-1091
16. Liu, X. et al. Self-Supervised learning: Generative or contrastive. *IEEE Trans. Knowl. Data Eng.* **35** (1), 857–876 (2023).
17. An evaluation of self-supervised pre-training for skin-lesion analysis. European Conference on Computer Vision. Available: https://link.springer.com/chapter/https://doi.org/10.1007/978-3-031-25069-9_11
18. Su, J. et al. Enhanced aspect-based sentiment analysis models with progressive self-supervised attention learning. *Art. Intell.* **296**, 1–16, (2021).
19. Maass, W. & Storey, V. C. Pairing conceptual modeling with machine learning. *Data Knowl. Eng.* **134**, 101909 (2021).
20. Kalra, V., Kashyap, I. & Kaur, H. Improving document classification using domain-specific vocabulary: Hybridization of deep learning approach with TFIDF. *Int. J. Inform. Technol.* **14** (5), 2451–2457 (2022).
21. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Available: https://arxiv.org/abs/1810.04805
22. Dey, A., Jenamani, M. & Thakkar, J. J. Lexical TF-IDF: An n-gram feature space for cross-domain classification of sentiment reviews. In *International Conference on Pattern Recognition and Machine Intelligence* (Springer), pp. 380–386. (2017).
23. Flores, C. A. & Figueroa, R. L. Pezoa. Active learning for biomedical text classification based on automatically generated regular expressions. *IEEE Access.* **9**, 38767–38777 (2021).
24. Hassan, S. U., Ahamed, J. & Ahmad, K. Analytics of machine learning-based algorithms for text classification. *Sustain. Oper. Computers.* **3**, 238–248 (2022).
25. Yu, Y. et al. Chinese mineral named entity recognition based on BERT model. *Expert Syst. Appl.* **206**, 117727 (2022).
26. Zhang, X., Jiang, Y. & Wang, X. Domain-specific NER via retrieving correlated samples. arXiv preprint arXiv:2208.12995, (2022).
27. Aizawa, A. An information-theoretic perspective of tf–idf measures. *Inf. Process. Manag.* **39** (1), 45–65 (2003).
28. Dhaka, B. Performance analysis of supervised machine learning algorithms for text classification. In *2016 19th International Conference on Computer and Information Technology (ICCIT)*. Available: https://ieeexplore.ieee.org/abstract/document/7860233

29. Shrestha, A. & Mahmood, A. Review of deep learning algorithms and architectures. *IEEE Access.* **7**, 53040–53065 (2019).
30. Kadhim, A. I. Survey on supervised machine learning techniques for automatic text classification. *Artif. Intell. Rev.* **52**, 273–292 (2019).
31. Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR).* **34** (1), 1–47 (2002).
32. Divya, R. Shantha Selva Kumari, R. Genetic algorithm with logistic regression feature selection for alzheimer's disease classification. *Neural Comput. Appl.* **33**, 8435–8444 (2021).
33. Xu, S., Li, Y. & Wang, Z. Bayesian multinomial Naïve Bayes classifier to text classification. *Adv. Multimed. Ubiquitous Eng.*, pp. 347–352, (2017).
34. Yuvaraj, N. et al. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Comput. Electr. Eng.* **92**, 107186 (2021).
35. Speiser, J. L., Miller, M. E., Tooze, J. & Ip, E. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* **134**, 93–101 (2019).
36. Lalwani, P., Mishra, M. K., Chadha, J. S. & Sethi, P. Customer churn prediction system: A machine learning approach. *Computing* **104**, 271–294 (2021).
37. Zhang, P. E-commerce products recognition based on a deep learning architecture: Theory and implementation. *Future Gener. Comput. Syst.* **125**, 672–676 (2021).
38. Yin, W. & Hay, J. D. Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. arXiv:1909.00161, (2019).

Acknowledgements

This research was financially supported by National Social Science Fund of China (Grant No. 24BGL311), Social Science Innovation Fund Project E-commerce and Business Intelligence Research Innovation Team of Henan University of Technology (Grant No. 2025-SKTXTD-07), National Social Science Fund of China (Grant No. 25FGLB037), and Research Project of Henan Provincial Social Sciences Association (Grant No. SKL-2025-1953).

Author contributions

Lei Bing: concept, administration, writing, method, software, fundingWang Jinghua: writing, software, analyzeShen Cong: writing, editing, analyze, funding.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-38214-2.

Correspondence and requests for materials should be addressed to C.S.