Proactive and Automatic Detection of Product Misclassifications at Massive Scale

Ling Jiang[*]
Amazon
Sunnyvale, CA, USA
jiangll@amazon.com

Xiaoyu Chu[*]
Amazon
Sunnyvale, CA, USA
xiaoychu@amazon.com

Saaransh Gulati
Amazon
Sunnyvale, CA, USA
saaransg@amazon.com

Pulkit Garg
Amazon
Sunnyvale, CA, USA
pulkitg@amazon.com

Andrew Borthwick[†]
Amazon
Seattle, WA, USA
andborth@amazon.com

Gang Luo[†,‡]
Amazon
University of Washington
Seattle, WA, USA
luougang@amazon.com

ABSTRACT

In e-commerce, product classification is widely used for various purposes. Misclassifying products can cause compliance issues and hurt the company's reputation. To address this problem, we propose an automated system to proactively detect product misclassifications by overcoming several challenges. A large e-commerce retailer can sell billions of distinct products, on which many thousands of classification tasks are performed. At this massive scale, we need to quickly detect misclassifications under a limited budget. In this talk, we point out these challenges and show how we design our system to handle them. When evaluated on a set of Amazon's product classification data, at an overhead of <10% of the classification cost, our system automatically identified and corrected many misclassifications, which would take a human many thousand years to manually find and 14.6 years to manually review and correct if our system were not used.

CCS CONCEPTS

• Applied computing → Online shopping; • Computing methodologies → Machine learning.

KEYWORDS

E-commerce, classification, misclassification, scalability

ACM Reference format:

Ling Jiang, Xiaoyu Chu, Saaransh Gulati, Pulkit Garg, Andrew Borthwick, and Gang Luo. 2023. Proactive and Automatic Detection of Product Misclassifications at Massive Scale. In *Proceedings of the 32st ACM International Conference on Information and Knowledge Management (CIKM'23), October 21-25, 2023, Birmingham, UK.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583780.3615510

1 INTRODUCTION

In e-commerce, product classification is widely used for various purposes such as computing sales taxes and regulatory and legal compliance (e.g., tagging products illegal for sale to children and products that are non-air-transportable). Misclassifying products can give customers bad experiences, make the company face legal consequences, and hurt its reputation. For example, if a product is misclassified as non-air-transportable, customers may encounter delays in receiving it. As another example, some products are banned from sale in certain regions. If misclassifications enable customers living there to buy such products, the company could be fined. To reduce the damage caused by product misclassifications, we propose an automated system to proactively detect them. This requires us to address several challenges:

1) **Massive scale**: A large e-commerce retailer can sell billions of distinct products, on which many thousands of classification tasks covering various use cases are done. Thus, quadrillions (10^{15}) of product classifications need to be handled.
2) **Limited budget**: The company's budget is finite. We need to detect product misclassifications under a limited budget.
3) **Low latency**: Once products are classified, we need to quickly detect misclassifications to minimize their caused damage.
4) **Mixed types of misclassifications:** Both machine learning classifiers and manually formed, keyword rule-based classifiers are used to classify products. We need to detect misclassifications made by both types of classifiers. Yet, prior literature on this topic mainly focuses on detecting misclassifications made by machine learning classifiers [1].
5) **Lack of labeled data**: Due to the huge human labeling cost, we cannot afford to ask humans to give enough labels of

[*] Denotes co-first authors who made equal contributions to the paper
[†] Denotes co-senior authors
[‡] Gang Luo is an Amazon Scholar as well as a Professor at the University of Washington. This paper describes work done at Amazon.

CIKM '23, October 21–25, 2023, Birmingham, UK
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615510

misclassified products for each classification task. For a typical task, we have few or no labels of misclassified products.

6) **High precision**: We cannot afford to ask humans to manually review all of the misclassifications detected by our system. Instead, we want our system to automatically correct as many of the detected misclassifications as possible. To reach this goal, our system needs to accurately detect misclassifications, preferably with a precision ≥ 95%.

To the best of our knowledge, no existing solution for detecting misclassifications can address all of these challenges and fulfill all of our requirements simultaneously.

2 MISCLASSIFICATION DETECTION SYSTEM

In this section, we outline our automated system for detecting product misclassifications. Our system contains three modules (see Figure 1) and uses extra information beyond that used by the product classifiers. To improve efficiency, in the candidate generation module, we first use low-cost models to quickly reduce the search space and form an initial set of potential candidates of misclassifications. Then in the candidate evaluation module, we use a more expensive model to identify actual misclassifications from this set. Finally in the feedback and correction module, we automatically correct the identified misclassifications.

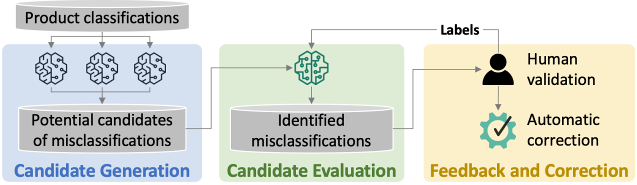

Figure 1: Our automated system for detecting product misclassifications.

Candidate generation module. This module uses five low-cost models, each leveraging distinct information to quickly capture misclassifications from a different angle. The potential candidates of misclassifications found by each model are merged to form our initial set of potential candidates. This module requires no labels of misclassified products. Instead, it uses information on:

1) **Product family.** A product family is a group of products that are almost identical to each other except along certain dimensions such as size and color. Typically, the products in the same family should have the same classification results. We enlist inconsistent classification results among these products as potential candidates of misclassifications.

2) **Product similarity.** We use features such as embeddings of product names and images to find similar products. Many of these features are unused in the product classifiers, as they need to be lean enough to be able to do classifications in real time. Intuitively, similar products tend to have the same classification results. We use this insight to identify some potential candidates of misclassifications.

3) **Product-task relationship.** For binary classification tasks, we formulate product misclassification detection as a recommender system problem, by treating tasks, products, and classification results as users, items, and users' ratings of items, respectively. By comparing the product classifications and the products recommended to the tasks, we identify some potential candidates of misclassifications.

4) **Task similarity.** For the classification tasks, we use their textual descriptions to compute their similarities. Intuitively, if two tasks are very similar to each other, correspondence can be established between their classes, with the same product tending to belong to the matching classes. We use this insight to identify some potential candidates of misclassifications.

5) **Task correlation.** The amount of labelled data available for training the classifiers varies by classification tasks. We use the PECOS (Prediction for Enormous and Correlated Output Spaces) algorithm [2] to capture task correlations and improve the classification results for the tasks with limited labels. This gives us some potential candidates of misclassifications.

Candidate evaluation module. Using some labels of misclassified products that humans give across all classification tasks, we train a machine learning model to find among the initial set of potential candidates of misclassifications, the items that are indeed misclassifications. To reach a high precision, we use a more expensive model than those used in the candidate generation module plus some extra product features unused in that module.

Feedback and correction module. For each classification task, we ask humans to label a sample of misclassifications identified by the candidate evaluation module and compute a separate precision of detecting misclassifications. If the precision is $\geq$ a given threshold t like 95%, our system automatically corrects the misclassifications that the candidate evaluation module identified for the task. Otherwise, if the precision is $< t$, we incorporate the human-provided labels to re-train the model used in the candidate evaluation module. We keep iterating until the precision is $\geq t$.

3 EVALUATIONS

We evaluated our system on a subset of Amazon's product classification data. At an overhead of <10% of the product classification cost, our system automatically identified and corrected many misclassifications, which would take a human many thousand years to manually find and 14.6 years to manually review and correct if our system were not used.

4 RELEVANCIES TO CIKM

Our talk covers several topics of interest to the CIKM community: practical industry challenges, efficient massive-scale data mining, machine learning algorithms, and multi-modal data.

REFERENCES

[1] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS'18)*. 5546-5557.

[2] Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S. Dhillon. 2022. PECOS: Prediction for enormous and correlated output spaces. *Journal of Machine Learning Research*, 23(98):1-32.