Predicting the Helpfulness of Customer Reviews with Respect to Product Mentions

2410015 Kazuki Urushihara

With the widespread of e-commerce (EC) platforms, product reviews have become an important source of information that strongly influences consumers' purchasing decisions. However, EC sites receive an enormous number of reviews every day, and many reviews do not contain concrete descriptions of product quality or usability. For example, some reviews only include complaints about packaging and delivery service, such as "the delivery was slow," or only describe personal circumstances or shipping experiences, such as "I bought this for my daughter." While such reviews may be meaningful as information about purchasing background or delivery experiences, they do not provide any information about the product itself.

In previous studies on identifying helpfulness of reviews, the dominant approaches estimated helpfulness using surface-level features such as review length, lexical diversity, and sentiment polarity, as well as external indicators such as the number of "helpful" votes. However, these studies did not explicitly evaluate the helpfulness of reviews from the perspective of whether the review mentions the product itself.

This study aims to classify the helpfulness of reviews by focusing on two key aspects: whether "a review mentions the product itself," and whether "a review contains a subjective evaluation of the product." Specifically, this thesis addresses a three-class classification task. The first class is "NotMentioned," which includes reviews that describe only delivery conditions or purchase motivations without mentioning the product itself. The second class is "Mentioned/NotEvaluated," which includes reviews that refer to product features or specifications but do not express a positive or negative evaluation. The third class is "Mentioned/Evaluated," which includes reviews that clearly evaluate product quality or usability.

A distinctive characteristic of the proposed approach is that it uses not only the review text itself but also auxiliary product information, namely the product description and the product name, as model inputs. This allows the model to consider the contextual relationship between the review and the product. In addition, because large-scale manual annotation for helpfulness classification is costly and difficult, pseudo-labeled training data is automatically constructed using a rule-based method and a large language model (LLM).

Two methods for constructing pseudo-labeled datasets are proposed. The first is a rule-based method. Important terms (keywords) are extracted from

a corpus of product descriptions by computing TF-IDF scores, and then a review is classified whether it contains mention to a product by detecting these terms in a review. On the other hand, the presence of evaluation for a product in a review is identified by checking whether evaluation words in an evaluation lexicon appear in the review. Based on these results, each review is assigned to one of the aforementioned three classes. The second method uses an LLM. By providing the review and product-related information in a prompt, the LLM determines which of the three classes the review belongs to. Preliminary experiments showed that the LLM-based method achieved higher labeling accuracy than the rule-based method. Therefore, pseudo-labeled data generated by the LLM were used as training data in the final experiments.

To build a classification model for helpfulness classification, pre-trained language models are fine-tuned. Two model architectures are developed: a one-stage approach that directly classifies reviews into three classes, and a two-stage approach that first determines whether a product is mentioned and then determines whether an evaluation is present. Preliminary experiments revealed that the two-stage approach suffered from error propagation, as misclassifications in the first stage could not be corrected in the second stage, resulting in lower accuracy. In contrast, the one-stage approach achieved more stable performance by learning features across all three classes simultaneously. Accordingly, the one-stage approach is applied in the succeeding experiments. In addition, two pre-trained language models are compared as the base model of the helpfulness classifier: BERT and ModernBERT. ModernBERT is an extension of the standard BERT model, which is better suited for handling longer contexts. It was found that ModernBERT achieved higher overall classification performance by preliminary experiments.

Several experiments were conducted to evaluate the effectiveness of the proposed method using a dataset from Rakuten Ichiba. Pseudo labels were assigned to create training data, while test data were constructed by manually annotating sampled reviews. Accuracy and macro F1 score were used as evaluation metrics, and training and inference were repeated three times to compute average scores. The experimental results showed that our proposed method achieved an accuracy of 74.2% and a macro F1 score of 69.5% at maximum. Next, several combinations of a review, product name, and product description are fed into the helpfulness classifier and the influence of the difference of these inputs was investigated. The results indicated that adding the product name to the model input improved classification performance under many conditions. This may be because product names often contain important features or keywords that help the model capture the relationship between the review and the product. In contrast, adding the product

description did not always improve performance and sometimes degraded performance due to noise. Increasing the maximum input token length also did not necessarily lead to better performance.

Additionally, the proposed lightweight model based on ModernBERT was compared with zero-shot inference using GPT-4o. Although GPT-4o achieved a high accuracy of 79.6% when using "review + product description" as input, the proposed method also achieved approximately 74% accuracy, resulting in a relatively small performance gap. Because LLMs such as GPT-4o require substantial computational resources due to their large number of parameters, their deployment in real-world services can be challenging. In contrast, the proposed method achieves comparable performance with significantly fewer parameters, making it more suitable for practical services that handle large volumes of reviews.

In summary, this study demonstrates that it is possible to construct a labeled dataset without manual annotation by leveraging LLMs, and to build a lightweight and accurate model for helpfulness classification based on product mentions and evaluations. By applying this model, EC platforms can efficiently present reviews that are more useful to consumers, thereby supporting users' purchasing decisions. Future work includes improving the quality of pseudo labels, further optimizing input design, and expanding evaluation datasets.