"Did you buy it already?", Detecting Users Purchase-State From Their Product-Related Questions

Lital Kuchy
Amazon
Israel
litalku@amazon.com

David Carmel
Amazon
Israel
dacarmel@amazon.com

Thomas Huet
Amazon
Israel
research@thomash.fr

Elad Kravi
Amazon
Israel
ekravi@amazon.com

Abstract

In this study we address the problem of identifying the purchase-state of users, based on product-related questions they ask on an eCommerce website. We differentiate between questions asked before buying a product (pre-purchase) and after (post-purchase). At first, we study the ambiguity that exists in purchase-states' definition, and then investigate the linguistic characteristics of the questions in each state. We analyze the discrepancy between the language models of pre- and post-purchase questions, and offer two classification schemes for this task, both outperform human judgments. We additionally show the effectiveness of our classification models in improving real world applications for both consumers and sellers.

CCS Concepts

• **Information systems → Question answering**.

Keywords

Purchase state classification, Product question answering

ACM Reference Format:
Lital Kuchy, David Carmel, Thomas Huet, and Elad Kravi. 2021. "Did you buy it already?", Detecting Users Purchase-State From Their Product-Related Questions. In *Proceedings of the 44th International ACM SIGIR Conference onResearch and Development in Information Retrieval (SIGIR '21), July 11–15, 2021,Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462940

1 Introduction

eCommerce websites like Amazon and eBay manage question-answering communities around products, enabling users to ask and answer product-related questions. Product questions spread over many areas including product's quality, value-for-money, shipment and service, compatibility, capabilities, and many more.

One of the main factors that characterizes product questions is the asker's purchase state during her purchase journey [5]. According to many studies (e.g., [11, 21]), identifying the customer's purchase state, and in particular the differentiation between pre- and post-purchase stages, is beneficial for providing a better customer service. Consider, for example, a customer who already purchased a product, and now is browsing the product's page. She might be interested in information about installation or usage, while comparison with alternatives is less useful at this stage and may end up with a bad experience.

We categorize a question posted before purchase as a pre-purchase question, while a question posted after as a post-purchase question. Table 1 provides some examples for questions posted before and after purchase. Albeit, for most of the questions (74% in our dataset) we do not have any indication whether the question's related product has already been purchased or not by the user. Some of these questions might be posted by users who haven't bought the product yet (hence no purchase record is available). Others might come from users who bought the product outside of the eCommerce website, or that the product was bought by someone else (e.g. by another family member). This emphasizes the need for automatic detection of the asker's purchase state based on the characteristics of the posted question, even when no purchase information is given.

Clearly, when purchase information is given, the question purchase state is well defined. However, even in that case, what about questions being asked after purchase time and before delivery time? In this twilight zone the purchase state is unclear, since the user who has ordered the item has not been exposed to it yet. Of course, a new purchase state can be defined, however, in most cases delivery time is unknown and is not recorded in our data. Moreover, even when delivery time is given, the actual post-purchase mode begins only after the package is opened and the product is being used. Learning how to correctly label the purchase state of these questions is a research challenge that might have a critical impact on classification accuracy. Therefore, we examine several labeling policies for questions being posted at the critical time window between order time and the (unknown) delivery time.

In addition, classifying the question's purchase state is difficult due to many ambiguous questions. Some questions have a clear pre-purchase state ("Does this have a charger that come with it?"), while others have a clear post-purchase state ("It came with a broken handle, what should I do?"). In contrast, for many questions the purchase state is unclear as they can be asked in both pre- and post-purchase mode ("Is there a warranty?", "what is the return

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462940

Table 1: Examples of pre- and post-purchase questions.

Product	Image	Pre-purchase questions	Post-purchase questions
Ring Video Doorbell (1st Gen)		• How long does the battery last? • Does this work with Google Home? • What is the difference between ring video door and ring video doorbell 2	• How do I enable live view? • When charging, how do you turn it off? • Why is my Ring doorbell battery losing its charge while connected to wires?
eufy by Anker, Robot Vacuum Cleaner		• Does it work on carpets? • Does it work on pet hair? • Is this 220V?	• Can I buy a replacement remote control? • It says not to use on dark floors. Why? How dark? • How long does it take to get a full charge?
Oculus Quest VR Gaming Headset		• When does shipping start? • Can you stream to a roku? • Can you play Minecraft	• Why does it take so long to ship? • Can u upgrade the memory? • Is there a way to cast what you're seeing on a TV or monitor?"

policy?"). Such ambiguous questions may mislead both humans and automatic classifiers in detecting the question purchase state.

Previous studies [5, 11, 21] defined four stages of consumer decision journey, including *awareness, evaluation, purchase* and *post-purchase*. Our study is focused on identifying users' state based on classifying their product related questions to pre- and post-purchase state. It is based on a dataset containing 230K product questions associated with their posting time, purchase time, and with the product related information[1]. We provide a thorough data analysis that highlights the differences between questions at the two purchase states, showing that pre-purchase questions are much more objective in comparison to post-purchase questions which are more subjective and emotional. Additionally, we experiment with two classification approaches, trained over this dataset, that detect the question's purchase state and compare their performance with human performance. Our classifiers, with accuracy above 80%, outperform humans (76% accuracy) at this task.

Finally, we demonstrate the applicability of our classifier, both for buyers and sellers, in two real world tasks. For buyers, we show an example of a question retrieval application. In this task, a user, while browsing a product's page, searches for previously resolved questions (Q&As) that may answer her own needs. Using our classifier we re-rank the retrieved list of Q&As, according to their match with the query's purchase state, showing improvement in user satisfaction. For sellers, we discuss a potential application that utilizes question purchase-state classification for better understanding of customer problems, identifying issues to be solved by different selling divisions like marketing, shipment, and customer service.

The rest of the paper is organized as follows: Section. 2 presents previous studies. Section. 3 models the problem and discusses how data was labeled. In Section. 4 we present our classification approaches. Section. 5, presents our dataset, and the evaluation process. In Section. 6 we present analysis of potential applications. We conclude our work in Section 7.

[1]We thank Amazon for certifying this data for our study. To the best of our knowledge, 1) this is the first dataset of pre- and post-purchase questions to become available to the public; 2) this dataset cannot be considered as a representative sample of the entire set of Q&As on Amazon website

2 Related Work

Consumer purchase journey. Detecting the different stages during the customer purchase journey has been extensively studied in economics and related areas [6, 18]. Court et al. [5] proposed a new approach for modeling the consumers purchase journeys. By examining the purchase decisions of almost 20K consumers, they offered the following four stages model: (1) *awareness* – the moment when a customer is aware of the existence of the product, (2) *evaluation* – customer pro-actively investigates or compares it to other options, (3) *purchase* – acquires it and (4) *post-purchase* – actually uses and reviews it. Ozer and Gultekin [18] surveyed 207 buyers, and found a connection between impulse tendency and pre-purchase mood which might cause impulse buying behavior. Another application detected purchase states, and specifically post-purchase state, for marking verified purchase content such as reviews [6]. Anderson et al. [1] analyzed the differences between verified and non-verified purchase reviews, showing that non-verified reviews are significantly more negative and more likely to contain deception cues.

Purchase state classification. Several studies [11, 21, 23] focused on automatic detection of purchase states. Vázquez et al. [21] classified user-generated content, extracted from different social media sources, into the four purchase states. Their rule-based classification method was based on extracted features such as lexical items and linguistic patterns (e.g., future tense). The work of Hasan, et al. [11] who focused on two main stages of the shopping journey, pre- and post- purchase states, is close in spirit to ours. They studied the task of classifying consumer review fragments into the two states. They trained classifiers using two sets of features: (1) textual features, and (2) manually crafted features based on part-of-speech (POS) tags. We also experimented with these features in our classification model. In contrast to our work, both previous works used a relatively small manually annotated dataset—600 labeled items in [21], and 4K in [11].

Zhou et al. [23] developed a deep learning framework to assign consumer journey states to web activities such as search, clicks and emails. They used a dataset of 2B activities, using clicks as a positive

signal, and utilized their framework to optimize advertisement-matching for the different purchase states. Their results showed an improvement over baseline approaches of up to 3.6% in AUC.

Following previous efforts to classify users into their purchase states, we set our goal to develop a purchase state classification method for product-related questions. In contrast to previous studies that manually annotated data by experts or by crowd-sourcing, we obtain the purchase state labels by processing purchase information of the question's related product, showing classification accuracy that exceeds human judgments.

3 Pre-Post Purchase Classification

Our dataset consists of a set of tuples $r =< q, t_{pos}, t_{pur}, t_{ans}, p >$ where q is a product question, posted at time t_{pos} on product p which was purchased by the asker at time t_{pur}. $t_{ans}(> t_{pos})$ is the question answering time (or null if the question has not been answered). p encapsulates the purchased product data, including product title, product image, and other relevant information.

Our goal is to learn a classification function, $F : (q \times p \rightarrow \{pre, post\})$ that gets a question and the corresponding product as input, and outputs question's purchase state. The labels used for the training phase are determined by

$$l(r) = \begin{cases} t_{pos} - t_{pur} < \Delta & pre \\ otherwise & post \end{cases} \quad (1)$$

where $\Delta \geq 0$ is the critical time window between purchase time and the expected delivery time. Since we do not have the exact delivery time, we set $\Delta = 24$ hours based on our data analysis (See Subsection 3.1), and considering Amazon's one-day shipping policy[2] which claims that in many locations (in the US) delivery takes place within 24 hours.

For the rest of the paper we denote pre-purchase questions as *pre questions* and post-purchase questions as *post questions*

3.1 Analysis

Next, we learn the characteristics of pre- and post questions by analyzing the questions in our dataset. The dataset is further described in Section 5.1.

3.1.1 Posting and purchase time relationship. An analysis of the difference between t_{pos} and t_{pur} is presented in Figure 1. We can see that most questions are being asked in proximity to purchase time. Moreover, the distribution of pre questions over the time difference is more dense around purchase time, as most are posted 24 hours prior to purchase time, while the distribution of post questions is more flatten. Additionally, analyzing the question answering time reveals that questions are answered quickly: 30% of the questions are answered within 2 hours and over 50% within one day. 30% of the questions in our dataset were not answered at all.

3.1.2 Textual discrepancy. We analyze the textual discrepancy between pre- and post questions by training two language models (LM)[3]. One is trained on questions posted before the purchase time and the second is trained on questions posted 48 hours and more after the purchase time, focusing mainly on highly confident post questions. We analyze the distance of both LMs to a background

[2]https://www.amazon.com/gp/help/customer/display.html?nodeId=202053420

[3]We implemented a LM based on NLTK package (https://www.nltk.org/api/nltk.lm.html), which includes Laplasian smoothing.

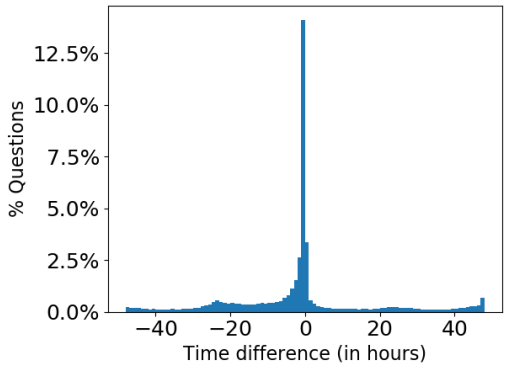

Figure 1: Question distribution over the time difference between posting and purchase time, in 1-hour bins, for questions posted within 96-hours window around purchase time.

LM consisting of all the questions, comparing the sets of prominent words that mostly contribute to the KL divergence [14] between the two models, results are presented in Table 2.

It can be seen that pre-purchase prominent terms are in present or future tense ('will', 'would', 'is', 'are', 'be'), and consisting compatibility terms ('fit', 'work', 'come'), while the language of post-purchase questions is more personal and possessive ('I', 'my', 'mine'), and in past tense ('received', 'bought', 'did'). This analysis, emphasizes the discrepancy between languages of the two types of questions, as reflected by their different language models.

Table 2: Prominent words of pre- and post-purchase LMs. Words are those with the largest contribution to the KL-divergence from a general LM trained over all questions.

Pre-purchase			Post-purchase (≥ 48 hours)		
this	will	fit	i	to	do
does	is	a	not	how	get
size	work	or	and	my	why
with	what	are	but	replacement	recieved
these	the	would	was	there	bought
be	many	model	no	mine	when
come	of		anyone	did	

In order to deep dive into the differences between pre and post questions, we trained a sequence of language models, $\{LM_i\}$, $i \in [-96 \ldots, +96]$, where LM_i is a model trained on the subset of all questions with i-hours difference between t_{pos} and t_{pur}. We calculate the KL divergence score between these LMs and the pre and post LMs, trained over all pre and post questions, receptively. Figure 2 shows the KL divergence scores of the one-hour models with respect to the pre and the post LMs. We can see the following phenomena: At first, in a 24-hour time window after purchase time, the divergence of the one-hour LMs from both the pre and the post LMs is very similar, suggesting that pre and post questions, at this time window, are very similar in terms of their language hence they can be either classified as pre- or post-purchase. Furthermore, prior to purchase time, the 1-hour LMs are clearly more similar (diverge

less) to the pre LM, while 24 hours after the purchase time, they are clearly more similar to the post LM.

The second interesting phenomenon is the significant drop in divergence of the 1-hour LMs of questions posted closely to purchase time. We assume this drop follows the large number of such questions (see Figure 1). In this case, the language models of these bins are similar to both pre- and post-purchase LMs. Finally, we see a 24-hour periodical drop in the KL score. These drops also correlate with the 24-hour periodical picks in question distribution, as can be seen in Figure 1. We hypothesize that customers are more active (ask and purchase) during (a periodical) one-hour slot of the day, therefore, since more questions are posted, and more of their associated products are purchased at this time slot, there are more questions with zero time difference, modulo 24, hence the periodical drops in KL divergence.

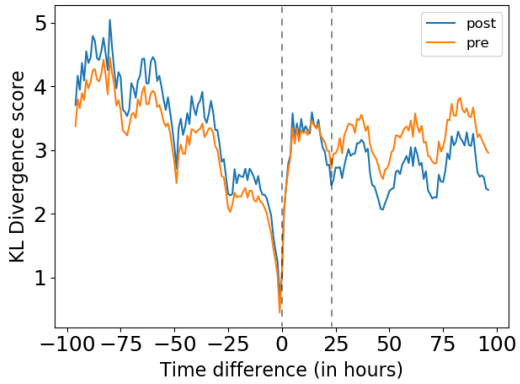

Figure 2: KL divergence score of one-hour LMs, from the pre and the post LMs; for questions posted within 96-hour window around purchase time.

4 Classification Approaches

Next we present our classification approaches. We begin with a logistic regression-based classifier that sheds some light on the different properties of pre and post questions. We then present a deep neural network classifier based on a pre-trained RoBERTa model [16]. We use $\Delta = 24$ hours as our labeling threshold (See Equation 1).

4.1 Logistic Regression

In this section we describe several linguistic features of product-related questions and analyze their capability to discern between the purchase state classes [11, 21]. These features are utilized to train a Logistic Regression classifier[4], jointly with question's unigram and bi-gram textual features; for each n-gram we calculate the tf-idf value.

4.1.1 Sentiment Polarity Analysis. Assuming people tend to be more emotional after purchase than before purchase, we hypothesize that this may lead to an express of negative sentiment in the question language through complains about malfunctions and other defects. In order to examine this hypothesis, we train a classifier for measuring sentiment polarity in pre and post questions.

Sentiment analysis was vastly learned in recent years (e.g. [2]) with the goal of correctly classifying positive and negative sentiment in textual content. In order to identify the question's sentiment we train a classifier using BERT model [7], over 'Amazon Review Data' [17] which consists of 200K Amazon product reviews. The sentiment of each review is labeled by associated human star rating (1-2 stars - negative label, 4-5 - positive label, 3 star reviews were omitted). In order to adjust the review-based sentiment classification model into the questions domain, we apply an unsupervised domain adaptation technique, as suggested by Ganin and Lempitsky [10]; we add a domain classifier, predicting whether the given input is a question or a review. The model is optimized to increase the loss of the domain classifier, while decreasing the loss of the sentiment classifier, for learning domain invariant features. We use the network's output as the sentiment score.

We present the distribution of the predicted sentiment for pre and post questions in Table 3. It can be seen that the majority of post questions are negative in comparison with pre questions. For illustration, look at the pre questions "I love these shorts. Does Lee offer this style in pants?" compared to the post questions "Where are the other 3 packs that I paid for? I purchased a pack of 4 but received only one."As expected, people are more critical and sentimental after purchase.

4.1.2 Subjectivity. Previous studies (e.g. [24]) have detected language subjectivity by training a classification network using product reviews as subjective examples and product descriptions as objective examples. We take a similar approach to train a BERT-based classifier using a balanced dataset of $200K$ product review and description pairs, uniformly sampled from 'Amazon Review Data' [17]. We use similar domain adaptation technique, as described for sentiment polarity analysis, to adapt to the question domain, verifying that classification decision is domain invariant. We use the network's output as the subjectivity score.

The distribution of subjective/objective labels over pre and post questions is depicted in Table 3. Post questions are more subjective (74% vs 26%), compared to pre questions which are perfectly balanced (50% vs 50%), indicating that people are more emotionally involved with the product after purchase.

Table 3: Sentiment and subjectivity distribution over pre and post questions.

Label	Sentiment		Subjectivity	
	Negative	Positive	Subjective	Objective
Pre	44%	56%	50%	50%
Post	60%	40%	74%	26%

4.1.3 Specificity. Following the insights from sentiment and subjectivity analysis, showing that post questions are more negative and more subjective than pre questions, we hypothesize that they are also more specific since they focus more on product usage and specific details. In order to examine this hypothesis, we implement a set of linguistic features, as proposed by Li and Nenkova [15]

[4]Using Sklearn package: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

for specificity classification , including text's length, number of occurring numbers, capital letters, non-alphanumeric symbols, and more. We call them specificity features hereafter.

We measure the distribution of these feature values over the purchase states. Question length was found to be the most separable feature between classes: post questions tend to be longer than pre questions, with an average of 14.67 words per question (std: 7.98, median: 13) compared to an average of 11.29 words in pre questions (std: 6.71, median: 9).

4.1.4 Part-of-speech (POS) tags patterns. Hasan et al. [11] implemented four groups of POS tags patterns to capture the relations between review posting time and product-purchase time. The pattern groups represent 'ownership', 'quality', 'experience', and 'modality' relationships. We utilize the NLTK framework[5] to annotate the question with POS tags, and create 4 pattern groups, similarly to [11]. Figure 3 presents the portion of questions per label that applies to each of the pattern groups. It can be seen that 'ownership' and 'experience' groups are associated with a significantly higher portion of the post questions, as one would expect.

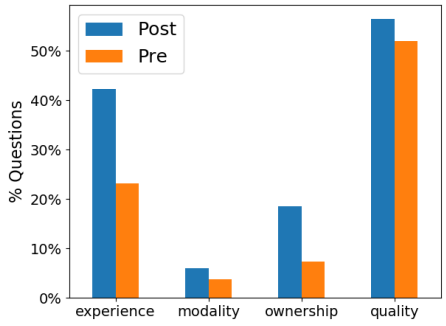

Figure 3: Percentage of pre and post questions applying to the different POS tag pattern groups.

4.1.5 Question Type and Expected Answer Type. Next we trained a classifier that predicts the question type and the expected answer type. Question's type is categorized into one of the values: 'product details', 'products relation', 'usage', 'product search', 'general knowledge' and 'subjective opinion'. Expected answer type is categorized into one of the values: 'yes/no', 'numeric value', 'date and time' and 'free text'.

We found that many usage questions are post questions (30% vs 10%) (e.g. "How long does it take to charge?"), while more questions about product details are pre questions (45% vs 36%). For expected answer type – 68% of post questions associated with free text (only 43% for pre questions) and 12% of pre questions expecting a numeric answer ("How much trim is there?", "What is the size of the doll?") compared to only 5% of the post questions.

4.2 DNN-Based Classification

Deep neural networks are known to work well for text classification [13] without the need of extensive domain-dependent feature extraction. Recently, Transformers models have been shown to outperform other networks over various tasks [22]. We utilize[6] a state-of-the-art, pre-trained, RoBERTa model [16] to train a purchase state classifier for product questions. The input for the RoBERTa model is the question's text and its output is an embedding vector, that is fed into a fully connected layer,

In addition, products may hold a useful information for detecting question's purchase state. For example, some products are prone to attract more pre questions while others would attract post ones. In order to capture the product's context we feed the RoBERTa model with product's title in addition to the question's content.

We also experiment with the set of features, described in Subsection 4.1, by applying Z-score Normalization[7] over their values, and then concatenating them to the RoBERTa's deep embedding vector. Finally, we feed the new vector into a fully connected layer for final classification, as presented in Figure 4.

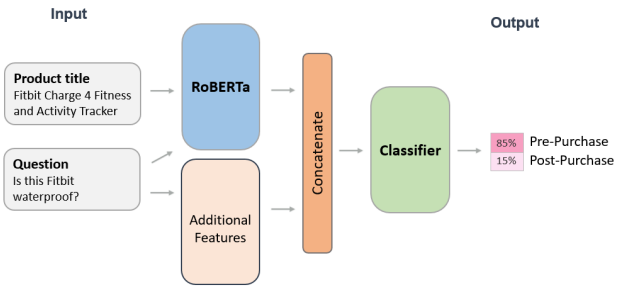

Figure 4: The DNN classification framework, including a question and product's title.

5 Evaluation

Next, we describe the evaluation framework for the question's purchase state classification. We begin with the description of our dataset and the implementation details of the methods we experimented with. We then present the result of our evaluation and analyze classification errors.

5.1 Dataset

Our evaluation framework is based on a novel dataset[8] of product-related questions, including their textual content, and gap, in hours between purchase and posting time. Each question is also associated with the related product details, including its id and title.

The dataset contains 940K product-related questions, provided by Amazon customers, spanned over 740K unique products, that were sampled from Amazon Question & Answers dataset, and were publicly available on Amazon Website. Data was filtered to include only English questions with more than two words. Only 26% of the questions (232K) include purchase information; this subset is used for training and evaluating our examined classifiers.

Ground truth labels were inferred using a comparison of question posting and purchase times (see Section 3). Questions posted at least $\Delta = 24$ hours after purchase time were labeled as post questions,

[5] http://www.nltk.org/

[6] We use the Transformers framework: https://huggingface.co/transformers

[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[8] The data is available on https://registry.opendata.aws/pre-post-purchase-questions/.

while all the rest labeled as pre questions. We denote this labeling as 24*H-labels*. The portion of pre and post questions, using 24*H*-labels, is 58% and 42% respectively. Questions were split to train, validation, and test sets, consisting of 160K, 40K and 32K questions respectively.

Duplicate questions We examine to which extent duplicate questions exist in our dataset. Such questions with contradicting labels may lead to classification errors. In order to examine duplication prevalence we cluster similar questions, using the DBSCAN algorithm [4]. DBSCAN is a density-based algorithm that applies two parameters: *eps*, denoting a similarity-distance threshold, and n, denoting a threshold on the cluster's maximum number of data-points . The semantic similarity between two questions is measured by the cosine between the questions' embedding, where the question text is embedded using Google's Universal Sentence Encoder [3].

We apply the clustering algorithm with strict parameters ($eps = 0.1, n = 2$), for clustering near-duplicate questions. 6.5% of the questions were found in clusters with contradicting labels.

5.2 Methods

We experiment with the following classification methods:

- **Majority vote.** Persistently return the majority class, pre-purchase state in our case (for 24*H*-labels).
- **Logistic regression (LR_Q).** A logistic regression classifier[9], trained using the question's uni-gram and bi-gram textual features only.
- **Logistic regression with manual features (LR_{Q+F}).** We add the manual features described in Section 4.1 to the logistic regression classifier. Features include sentiment and subjectivity scores, specificity, POS patterns, question type, and expected answer type.
- **RoBERTa based methods.** Our NN-based classifier as presented in Section. 4. We consider three variations – one utilizes only the question text (*RoBERTa*$_Q$), the second utilizes the question and the product title (*RoBERTa*$_{QT}$), and the third adds the manual features, denoted as *RoBERTa*$_{QT+F}$.
- **Human judgments.** In order to measure the task difficulty, we asked human annotators to classify product-related questions to their purchase state, based on the question text and the associated product title and image. Annotations were done using the Appen framework[10], where each question was judged by five annotators. Figure 5 illustrates the annotation task page as presented to annotators. There was high agreement among annotators (Fleiss' kappa = 0.58). It turned out that people tend to agree whether a question was asked before or after purchase, however, as we will see, their accuracy is inferior to the automatic approaches.

5.3 Results

Table 4 reveals the accuracy of the different classifiers. Accuracy and F1 are measured with 0.5 threshold over classifier's output probability. It can be seen that maximal accuracy is around 80%. Overall, at least 20% of the questions are misclassified in all methods, indicating the difficulty of this task, as we further discuss in Section. 5.4.

[9]We also examined SVM and Naïve Bayes, though they had weaker performance we only report results for the logistic regression classifier.
[10]www.appen.com

Figure 5: Annotation task page as presented to annotators.

Table 4: Accuracy and F1 results of the different classification methods. Statistically significance with respect to Human and to RoBERTa$_Q$ methods is marked by '*' and '' respectively, using McNemar test with p-value <0.05 [8]**

Model	Acc.	F1
Majority vote	0.577	0.366
LR$_Q$	0.741	0.729
LR$_{Q+F}$	0.756	0.744
RoBERTa$_Q$	0.801*	0.790*
RoBERTa$_{QT}$	**0.805****	**0.795****
RoBERTa$_{QT+F}$	0.804	0.793
Human	0.756	0.725

The features described in Section 4.1 assist LR$_{Q+F}$ to outperform LR$_Q$ and humans. However, best results are achieved by RoBERTa$_Q$ and RoBERTa$_{QT}$ models that were found to be statistically significant better compared to humans. RoBERTa$_{QT}$ significantly outperforms RoBERTa$_Q$, denoting the product contribution to the question classification task.

Table 5: Classification results separated for pre- and post-purchase classes for 24*H*-labels labeling scheme.

Method	Pre			Post		
	Prec.	Recall	F1	Prec.	Recall	F1
Human	0.720	**0.943**	0.817	**0.866**	0.500	0.634
RoBERTa$_Q$	0.789	0.895	0.839	0.825	0.673	0.7419
RoBERTa$_{QT}$	**0.797**	0.887	**0.840**	0.819	**0.692**	**0.750**

Precision-recall trade-off. Diving into the performance over individual classes, as presented in Table 5, we observe that all the methods has better precision in classifying post questions. Albeit, this is achieved in the price of low recall for this class. Meaning, the examined methods, and especially Human (pre-purchase recall of 0.943), tend to classify questions as pre questions.

In order to further examine the trade-off between precision and recall for the two classes, we analyze the precision-recall graph as presented in Figure 6. It can be seen that RoBERTa$_{QT}$ is superior to Human and to LR$_{Q+F}$ algorithms for the complete range. We also present the results of the RoBERTa$_{QT+F}$ algorithm (RoBERTa

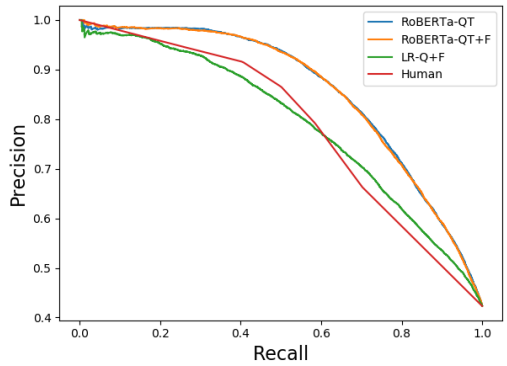

Figure 6: Recall-Precision graph of the different classification methods.

enhanced with the manual features), showing that the extracted features do not contribute to the network performance.

Feature importance. We conduct the following ablation test to quantify the contribution of the different features to purchase state classification. We train a logistic regression classifier that uses (1) a single feature out of the features discussed in Section. 4.1, and (2) all the features but a single feature. Accuracy and F1 score are depicted in Table 6. In both configurations (single and all-but-one) the question text is found as the most important feature. Among the crafted features, the question type and expected answer type contribute significantly to overall performance. Sentiment score, on the other hand, has a marginal contribution, in agreement with Table. 3. We note that this set of features improve the performance of the logistic regression classifier by 1.5% (see Table 4), however, they did not contribute to the performance of RoBERTa$_{QT}$, as shown in Figure 6. We assume no improvement is observed by adding these features since their functionality is already been captured and implicitly covered by the network.

Table 6: Ablation analysis of a logistic regression classifier, while trained with only one of the features (two left columns), and with all but one (two right columns).

Feature	This feature only		All but one	
	Accuracy	F1	Accuracy	F1
All	0.756	0.744	0.756	0.744
Question text	0.741	0.729	0.703	0.687
Subjectivity score	0.637	0.632	0.754	0.741
Sentiment score	0.585	0.572	0.755	0.743
Specificity features	0.627	0.582	0.755	0.743
POS patterns groups	0.619	0.594	0.756	0.744
Question and-expected answer types	0.646	0.586	0.703	0.687

5.4 Error Analysis

Next, we dive into the classification results to better understand the cases which are harder for the RoBERTa$_{QT}$ classifier.

Error-rate. Figure 7 presents the classifier's error rate as a function of the difference between posting time and purchase time. It can be seen that much more errors are made on post questions. The error rate grows at the [0,24] interval due to purchase state ambiguity at this critical time zone (See Section 3), and drops afterward for larger positive time differences, as the purchase state becomes more apparent.

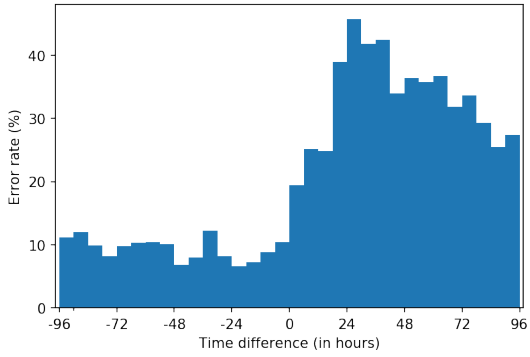

Figure 7: Error-rate of the RoBERTa$_{QT}$ classifier, measured over all questions posted within 96-hour window around their purchase time, in 1-hour bins.

Human vs. the machine. We next present a qualitative analysis of classification errors, comparing between Human and RoBERTa$_{QT}$ errors. Examples are given in Table 7.

- *Ambiguity.* Questions about return policy and warranty can be posted before or after purchase, hence are wrongly classified many times by humans. The classifiers outperform humans on this type of ambiguous questions since they can learn, for example, that the majority of warranty questions are posted after purchase.
- *Before delivery.* All questions posted during the critical time zone, the 24H time window after purchase time, are considered pre-purchase, according to the 24H labeling scheme. A significant portion of these questions are actually post questions.
- *Cynicism.* Cynical and sarcastic questions (e.g., "Do the knives actually cut?") are confusing as they may present pre-purchase intent while they are actually subjective post questions, typically with a negative sentiment.
- *Verification.* Sometimes, post questions can be easily answered by inspecting the product at hand (e.g., "What is included in the box?"). These questions usually confuse humans as well as classifiers. We denote them as verification questions.

Number of training examples. Previous studies on purchase state classification used relatively small datasets (600 in [21] and 4K in [11]). In the following we analyze the impact of the dataset size on classification accuracy.

Transformer-based classifiers, using a pre-trained model, can be fine-tuned with a relatively small number of examples [7]. Usually, few thousands of examples are enough to fine tune the model, and additional examples contributes marginally to the classification accuracy. Since our dataset is relatively large, we examine the effect

Table 7: Examples for questions wrongly classified by Human and by RoBERTa$_{QT}$.

Question	Label	Human	RoBERTa$_{QT}$
"What type of songs can be preloaded?"	Pre	✓	✗
"How do I know how much space is left on the internal storage?"	Post	✓	✗
"What is the return policy for this item?"	Pre	✗	✓
"Is there a warranty on this watch?"	Post	✗	✓
"How can I get a replacement key?"	Pre	✗	✗
"Do the knives actually cut?"	Post	✗	✗

of the sample size on performance, by training the classifier with an increasing number of examples. Results are presented in Table. 8. It can be seen that the number of examples has a significant impact on classifier's performance, 6% accuracy improvement while the number of examples is increased from 1K to 160K. Furthermore, it seems that adding more training data can potentially improve performance even further.

Table 8: Accuracy and F1 as a function of number of training examples (in thousands).

	1	10	50	100	160
Accuracy	0.742	0.777	0.794	0.801	0.805
F1	0.724	0.767	0.783	0.789	0.795

6 Discussion

We now discuss how purchase state classification can be utilized to improve potential services and applications. We offer two potential applications – for buyers and for sellers. Then, we analyze the sensitivity of our classifier to different labeling schemes and different labels distribution.

6.1 Potential Applications

6.1.1 Question Retrieval. Customers browsing a product page might be interested in resolved questions, asked by other users, to answer their own question. Since many products attract many Q&As, a search service over the product's Q&As is required for supporting users who seek for a resolved question. For example, Figure 8 presents the Amazon product details page, where users can pose a query and retrieve relevant Q&As.

Let's assume a user who issues a query while in a post-purchase mode; which type of resolved questions, pre or post, are more of interest for her? To answer this question we conduct the following experiment: given a product, p with associated Q&As set, Q_p, we experiment with two retrieval methods: 1) the BM25 algorithm [19]; 2) boosting the textual score with purchase label matching score. Boosting was done by a linear combination of both scores, with equal weights.

Our evaluation benchmark contains a sample of 350 products, each is associated with two randomly selected questions $q_{pre}, q_{post} \in Q_p$ (denoted as queries), one per each purchase state. The purchase state for all the questions in Q_p was determined by the RoBERTa$_{QT}$ classifier. We applied the two retrieval methods to rank the Q&As in $Q_p \setminus \{q_{pre}, q_{post}\}$.

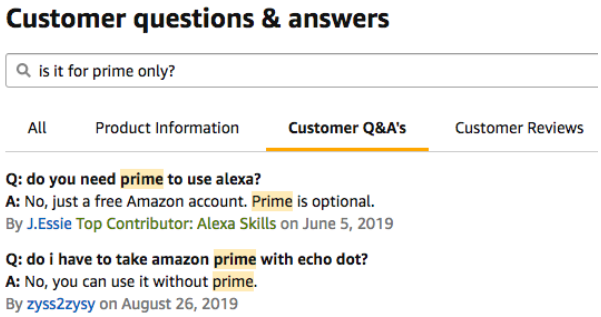

Figure 8: Amazon's question search page for Echo Dot (3rd Gen).

We evaluated the rankings by manual judgments of the search results. Annotators were exposed to the product title and figure (as context), the query, and the question, and were asked to vote for question usefulness to the query, out of the options 'very relevant', 'somewhat relevant' and 'irrelevant'. Each <query, question> pair was judged by 5 annotators. Overall inner-agreement was moderate (Fleiss' kappa of 0.31).

We measured the performance of the two retrieval methods by NDCG@k ($k \in \{1, 3, 5\}$). Results are presented in Table 9, for pre and post queries, and for products with at least 100 questions. It can be seen that for post questions, when boosted with the purchase state match, performance is improved. For NDCG@3 and NDCG@5 improvement is statistically significant. On the other hand, for pre questions, boosting brings no improvement. We hypothesize that when users are in post-purchase state, boosting is more crucial, as pre questions are of less interest. On the other hand, in a pre-purchase mode, all questions, including post questions, are of interest for purchase decision, therefore boosting pre-purchase questions has no effect on performance.

Table 9: NDCG results for the question-retrieval task, using vanilla BM25 and BM25 with label's boost. Statistically significance with respect to 'BM25' method is marked by '*' , using McNemar test with p-value <0.05.

Purchase state	Retrieval method	NDCG@1	NDCG@3	NDCG@5
Pre	BM25	0.685	0.751	0.849
	w/ label boost	0.685	0.750	0.850
Post	BM25	0.583	0.652	0.756
	w/ label boost	**0.598**	**0.673***	**0.777***

6.1.2 Understanding the Customer needs. Sellers can benefit from understanding customer needs, along their decision journey [5, 9], by analyzing the customer questions being asked on their product. For example, when many users ask about product's compatibility, this is an important signal for a missing or unclear information in the product's presentation on the site. Classifying the user questions into purchase states, and identifying the main topics being asked in each state, is beneficial for improving customer support.

We present an example of this analysis on a sample of popular products (with > 1000 questions). For each product, questions are first classified into the two purchase states (using RoBERTa$_{QT}$) and then clustered to identify the main topics being asked. Clustering was done by first embedding the questions using universal-sentence-encoder[4], then applying the DBSCAN [4] algorithm with eps = 0.1 for pre questions, eps = 0.2 for post questions and n = 2. Finally, we label each of the top clusters by selecting a random question from the cluster.

Table 1 presents a few pre and post questions, representing typical topics being asked on several products. It can be seen that pre-questions are indeed related more to product's presentation issues (compatibility, functionality, included items), to be handled by the marketing division, while post-questions mostly refer to shipment, operation and maintenance, to be handled by the customer service department.

6.2 Sensitivity Analysis

Question labeling. In all experiments discussed so far we set the labeling border to 24H, labeling any question posted up to 24 hours after purchase as a pre-question, and all the rest as post questions. How varying the labeling threshold, Δ, will affect the classification task? We examine this question by training a sequence of RoBERTa$_{QT}$ models using different threshold values $\Delta \in \{0, 12, 24, 36, 48\}$.

We evaluate the various trained models using two different testing schemes. The first one is based a high-confident examples where only questions with high certainty are sampled – pre-questions with purchase time < 0 and post-questions with purchase time > 48. We test all trained models using this test-set. Results are shown in the left two columns of Table 10.

For the second testing scheme, each model is tested on its own unique test set, where test questions are labeled exactly the same as training questions – according to the given Δ. Results are shown in the two right columns of Table 10.

Table 10: Accuracy and F1 results of RoBERTa$_{QT}$, using two different labeling schemes.

	High-confidence		Δ-based	
Label method	Acc.	F1	Acc.	F1
0H-labels	0.811	0.806	0.771	0.769
12H-labels	0.817	0.809	0.797	0.788
24H-labels	0.817	0.809	0.805	0.795
36H-labels	0.816	0.805	0.803	0.789
48H-labels	0.811	0.799	0.800	0.780

We can see that in general, accuracy is higher over the High-confidence test set as the classification task is easier. However, the difference in performance of the various models is marginal. For the Δ-based testing scheme, we can see a significant higher accuracy of the 24H-model, and a drop in model accuracy as the threshold is moving away for both sides. This supports our selection of the 24H threshold as a reasonable border between pre and post questions.

Classification accuracy over unlabeled data. The majority of questions (74%) do not have purchase information. Will our classifiers which were trained over labeled data, maintain their performance over this unlabeled data? In order to answer this question we first examine the distribution of pre and post questions on a large sample of unlabeled data by weak labeling these question through invocation of the RoBERTa$_{QT}$ classifier. It turns out that 86% of the questions in this sample are pre questions, while only 14% are post questions (in comparison to 58% versus 42% of pre- and post questions in the labeled dataset).

We then sampled a new test set with a corresponding ratio of pre and post questions (86/14% respectively) from the labeled dataset. The performance of the RoBERTa$_{QT}$ classifier is reduced on this test set, with an accuracy of 0.866, and macro f1 of 0.718 (in comparison to 0.805 and 0.795, respectively, on the full labeled test set). The reason for performance reduction is an improved precision over pre questions (0.957 vs 0.789) and a reduced precision over post questions (0.419 vs 0.825). We hypothesize that the reason for performance reduction is that the classifier predicts (and makes mistakes on) post questions in a much higher portion of their actual prevalence. Utilizing a classifier trained over data with a class distribution that is different from the target data attracts much attention recently [12, 20] and we leave this interesting direction for a future work.

7 Conclusion

In this work we studied the problem of classifying the purchase state of a user who asks a product-related question. We analyzed a dataset, collected from Amazon catalog, consisting of questions asked by customers who purchased the associated product. We studied the correct approach for labeling this data, addressing the dilemma of how to label questions asked after purchase time and before delivery time. We showed that pre-purchase questions are more positive, objective and broad than post-purchase questions.

We then developed two classification frameworks that classify a given question into its purchase state; a logistic regression classifier based on domain specific features of the questions, and a DNN model. The accuracy of the DNN model is about 80%, significantly outperforming the logistic regression classifier, as well as human judgments. Our error-analysis shows that while on one hand classifiers struggle in detecting post-purchase questions, on the other hand, enlarging the size of the dataset can help to improve their performance. We demonstrated how our classification approach can assist question retrieval on the product page while boosting the ranks of post-purchase Q&As for post queries, as well in identifying main topics of concerns that should attract the seller attention.

In a future work, we plan to study whether our classification approach, designed for questions posted on the product page, can be also applied for queries asked outside the product page (i.e., in a general product search scenario). Such classification can assist, for example, by offering alternative products for pre-purchase queries, and complementary products for post-purchase queries. Another research direction is to keep improving the classifier performance, for example, by modeling the user intention of asking questions while being in a pre-or in post- purchase mode.

References

[1] Eric T Anderson and Duncan I Simester. 2014. Reviews without a purchase: Low ratings, loyal customers, and deception. *Journal of Marketing Research* 51, 3

(2014), 249–269.

[2] John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic*.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *CoRR* abs/1803.11175 (2018). arXiv:1803.11175 http://arxiv.org/abs/1803.11175

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[5] David Court, Dave Elzinga, Susan Mulder, and Ole Jorgen Vetvik. [n.d.]. The consumer decision journey. *McKinsey Quarterly* ([n. d.]), 1–11. http://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/the-consumer-decision-journey/

[6] Michael Crawford, Taghi M Khoshgoftaar, Joseph D Prusa, Aaron N Richter, and Hamzah Al Najada. 2015. Survey of review spam detection using machine learning techniques. *Journal of Big Data* 2, 1 (2015), 23.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1383–1392.

[9] David C Edelman and Marc Singer. 2015. Competing on customer journeys. *Harvard Business Review* 93, 11 (2015), 88–100.

[10] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

[11] Mehedi Hasan, Alexander Kotov, Aravind Mohan, Shiyong Lu, and Paul M Stieg. 2016. Feedback or research: separating pre-purchase from post-purchase consumer reviews. In *European Conference on Information Retrieval*. Springer, 682–688.

[12] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439* (2020).

[13] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[14] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[15] Junyi Jessy Li and Ani Nenkova. 2015. Fast and accurate prediction of sentence specificity. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 188–197.

[18] Leyla Ozer and Beyza Gultekin. 2015. Pre-and post-purchase stage in impulse buying: The role of mood and satisfaction. *Journal of retailing and consumer services* 22 (2015), 71–76.

[19] Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

[20] Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548* (2018).

[21] Silvia Vázquez, Óscar Muñoz-García, Inés Campanella, Marc Poch, Beatriz Fisas, Nuria Bel, and Gloria Andreu. 2014. A classification of user-generated content into consumer decision journey stages. *Neural Networks* 58 (2014), 68–81.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[23] Yichao Zhou, Shaunak Mishra, Jelena Gligorijevic, Tarun Bhatia, and Narayan Bhamidipati. 2019. Understanding consumer journey using attention based recurrent neural networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3102–3111.

[24] Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor Detection in Product Question Answering Systems. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.). ACM, 519–528. https://doi.org/10.1145/3397271.3401077