Leveraging Seq2seq Language Generation for Multi-level Product Issue Identification

Yang Liu, Varnith Uttam Chordia, Hua Li, Siavash Fazeli,
Yifei Sun, Vincent Gao, Na Zhang
Amazon, Inc.
{yngliun, vchordia, realname, saivash, sunyifei, vincegao, naazhang}@amazon.com

Abstract

In a leading e-commerce business, we receive hundreds of millions of customer feedback from different text communication channels such as product reviews. The feedback can contain rich information regarding customers' dissatisfaction in the quality of goods and services. To harness such information to better serve customers, in this paper, we created a machine learning approach to automatically identify product issues and uncover root causes from the customer feedback text. We identify issues at two levels: coarse grained (L-Coarse) and fine grained (L-Granular). We formulate this multi-level product issue identification problem as a seq2seq language generation problem. Specifically, we utilize transformer-based seq2seq models due to their versatility and strong transfer-learning capability. We demonstrate that our approach is label efficient and outperforms the traditional approach such as multi-class multi-label classification formulation. Based on human evaluation, our fine-tuned model achieves 82.1% and 95.4% human-level performance for L-Coarse and L-Granular issue identification, respectively. Furthermore, our experiments illustrate that the model can generalize to identify unseen L-Granular issues.

1 Introduction

Customer feedback plays a crucial role in continuously improving service quality for e-commerce companies. One important piece of information in customer feedback are the product issues that customers encounter during their order experience, such as product-quality defects and undesired product features. Although negative product experience occurs rarely in mature e-commerce stores, identifying the product issues presented in customer feedback significantly helps sellers understand customers' concerns and facilitates them to further improve customer order experience. As the volume of customer feedback grows rapidly, an automatic and intelligent issue identification system is needed to support the fast-growing e-commerce business. In this paper, we introduce a machine learning solution for multi-level issue discovery by taking advantage of advanced deep language modeling techniques. We show that our approach not only accurately identifies product issues from customer feedback, but also meets the requirements for our use case, which, we believe, is also common to other e-commerce store business.

Identifying product issues from customer feedback has its own unique characteristics as compared to common NLP tasks such as text classification, document summarization, entity extraction, and sentiment analysis. First, we target to extract diverse and dynamic product issues, instead of categorizing them into a fixed set of labels. This is mainly due to three varying factors: the product itself, the customer and the context. Different categories of products naturally have different kinds of issues. Different customers may find different flaws of the same product, depending on their personal taste and preferences; the way they describe issues could vary greatly from one customer to another. In addition, issues are dynamic as different issues emerge and disappear from time to time. For example, a bad local weather can lead to a surge in shipment damage and delivery failure complaints in a short period of time.

Second, to effectively improve customer experience, multi-level issues with different granularity are needed. Higher-level, concise, and consistent issues can be shared with sellers to help them identify trends and hotspots for the flaws in their products; more detailed lower-level issues are needed by business investigators to study root causes of product defects and design treatment accordingly.

The third requirement for product issue identification is about supervised information extraction. Customer feedback often contains contents that are unrelated to product issue. Furthermore, multi-

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 20 - 28
May 26, 2022 ©2022 Association for Computational Linguistics

ple product issues can be tangled or be embedded within one phrase or sentence. Therefore, uncontrolled excerpts from customer feedback can be confusing and overwhelming to downstream users. So our system should focus on product issues and disentangle them to best serve downstream users.

Traditional language processing approaches can not meet all the three requirements discussed above. To start with, traditional classification methods can only produce results from a pre-defined, limited set of labels, and therefore cannot satisfy the first requirement. Unsupervised methods such as topic modeling and clustering, on the other hand, are able to identify diverse patterns varying from a dataset to another and discover new trends. However, it is hard to steer the algorithms to exercise the control over the patterns or type of information they extract from the data. Based on our experience, it is challenging to generate meaningful and highly coherent 'topics' using topic modeling or clustering algorithm based on text embedding techniques. Finally, it is not straightforward to adapt these methods for multi-level issue identification unless we develop multiple models to handle each level individually. For operation efficiency, we prefer to have a single model to complete the whole task so as to minimize model deployment and maintenance cost.

In this paper, we propose an innovative solution by formulating our problem as seq2seq language generation tasks and leveraging deep learning models with multi-task capability to meet all our requirements. In particular, we choose transformer-based seq2seq models as backbone and fine tune them on our data. A single model is trained to generate both the low-level and high-level product issues. Even though the model is trained in a supervised manner with human-annotated data to have better content generation control, as a large-scale language model, the model shows generalization power to discover fine-grained issues from customer feedback that were not seen during training time. To the best of our knowledge, so far it is the most flexible and effective approach for extracting multi-level product issues from e-commerce customer feedback.

2 Related Work

There are various approaches to extract key topics and identify issues from customer feedback. Prior work in this area can be categorized into four major groups: (1) document classification, (2) topic modeling, (3) clustering, and (4) text summarization. Document classification is one of the most well studied NLP tasks, and many deep learning methods (Kim, 2014; Devlin et al., 2018a) have been introduced and excelled at it in recent years. Tong et al. (2018) developed a convolutional neural network to extract reason codes from customer complaints. More recently, Liu et al. (2021a) leveraged BERT-based model to classify different types of fraud elements from Internet fraud complaints. Instead of assigning labels to the entire texts, they map the labels to each paragraph. These methods are not very applicable in the current context since they require a fixed and pre-defined taxonomy.

For topic modeling, latent Dirichlet allocation (LDA) (Blei et al., 2003) is a widely used statistical method for analyzing information in customer reviews and feedback (Mou et al., 2019; Debortoli et al., 2016; Jeong et al., 2019). Zhai et al. (2011) added pre-existing constraints to LDA in order to improve product feature extraction from customer reviews. Bagheri et al. (2014) proposed a Twofold-LDA model to produce topics focused on desired aspects. Srivastava and Sutton (2017) proposed autoencoded variational inference for topic model (AVITM) which yielded much more interpretable topics than LDA. While most traditional topic modeling methods use simple document representations such as the bag-of-words, an alternative approach is to cluster similar documents using document embeddings followed by extracting common topics within each cluster. Du et al. (2016) used GloVe embeddings and K-means clustering for analyzing different aspects in electronics and restaurants reviews. Grootendorst (2020) used BERT embeddings and DBSCAN to create interpretable topics. Although topic modeling and clustering are not constrained to a fixed label set, the results can be inconsistent and incoherent due to unsupervised topic extraction.

One can also pose this problem as a text summarization problem, where a model can generate a summary to present the key information from the input text. Liu et al. (2021b) improved abstractive summarization models for generating summaries about product issues from customer feedback. For our use case, however, it's not clear how to use text summarization to produce multi-level issues and how to disentangle different issues mixed within a summary.

Customer Feedback	L-Granular Issue	L-Coarse Issue
When I received the product, there was no power cable included. The box has been opened previously and the item looks used.	no power cable included, box has been opened, item looks used	missing parts issue, opened box, used condition
Instructions were hard to read, some of it was written in another language. The product also looks different from the picture.	instructions hard to read, written in another language, looks different from picture	instruction issue, not as pictured

Table 1: Example customer feedback texts and L-Coarse & L-Granular issues.

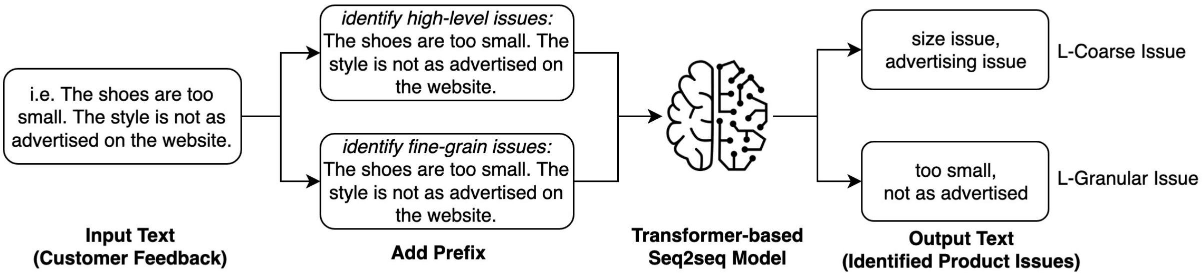
i.e. The shoes are too small. The style is not as advertised on the website.
Input Text (Customer Feedback)
identify high-level issues: The shoes are too small. The style is not as advertised on the website.
identify fine-grain issues: The shoes are too small. The style is not as advertised on the website.
Add Prefix
Transformer-based Seq2seq Model
size issue, advertising issue
L-Coarse Issue
too small, not as advertised
L-Granular Issue
Output Text (Identified Product Issues)

Figure 1: Illustration of multi-level issue generations enabled by multi-tasking models.

3 Proposed Approach

3.1 Multi-level Issue Identification

We aim to have comprehensive issue representations due to the large variety of intents and issues customers can express through customer feedback. One solution is to build a hierarchical architecture. According to the abstraction levels, more detailed and trivial issues will sit on the leaves whereas more general and conclusive issue categories will sit on the branches and major chunks. In this work, we decide to approach the problem starting with a two-level issue representation: L-Coarse and L-Granular.

An L-Granular issue is a faithful representation of the original customer communication and it captures concrete and fine-grain issues in a free form text. We deliberately leave L-Granular issues unaggregated to preserve original customer expression so that they will serve as a solid foundation for subsequent bottom-up aggregation. Each L-Coarse issue is an aggregation of multiple L-Granular issues, which describes an abstract issue shared across multiple customer communications. To fully capture customer concerns and differentiate the nuances among issues, we allow multiple issues, at both L-Coarse and L-Granular levels, for each customer feedback. Table 1 provides three examples of customer feedback together with their corresponding L-Coarse and L-Granular issues[1].

3.2 Seq2seq Learning for Issue Identification

We tackle the issue identification problem using a seq2seq learning approach. In this approach, we format our problem as text-to-text tasks, where the input text is customer feedback and the output text is the literal text representing the identified issues. We fine tune a seq2seq model to capture issues that are relevant to product defects. This approach is illustrated in Figure 1.

Leveraging the versatility of the text-to-text format, Raffel et al. (2019a) and Aribandi et al. (2021) demonstrated the capability and advantage of handling multiple tasks within a single model. They inspired us to take the advantage of these multi-task learning techniques to train a single model to generate both L-Coarse and L-Granular issues simultaneously. As illustrated in Figure 1, we append different customized prefixes to the same input text to distinguish different tasks and pair the prefixed sample with the corresponding target text (L-Coarse or L-Granular issues). The model is trained to produce different levels of issues according to the prefixes in the input. In particular, the prefixes we use are *"identify high-level issues:"* and *"identify fine-grain issues:"* for L-Coarse- and L-Granular-issue generations, respectively. For the target text, we use comma to separate multiple issues for both L-Coarse and L-Granular issues. In our training, we use the natural mix (1:1) of the two task data.

[1] Due to confidentiality, all customer feedback examples in this paper are composed by the authors and are used for demonstration only.

4 Experiments

4.1 Dataset

We collected 9200 samples of negative customer feedback across different post-order communication channels from online e-commerce stores. We asked subject matter experts [2] to identify L-Coarse and L-Granular issues for each customer feedback text with emphasis on extracting the information related to product issues. The annotated dataset contains 70 unique L-Coarse issues and 6906 unique L-Granular issues. The large difference in the numbers of unique issues shows the different characteristics of L-Coarse and L-Granular issues: L-Coarse issues are organized and abstract whereas L-Granular issues are detailed, diverse, and are often in free form. As illustrated in Figure 2, for the distribution of the top 50 L-Coarse issues, it is highly skewed and long tailed. In our experiments, we used human-annotated issues as the target text during model training. The train/test split ratio is 80:20.

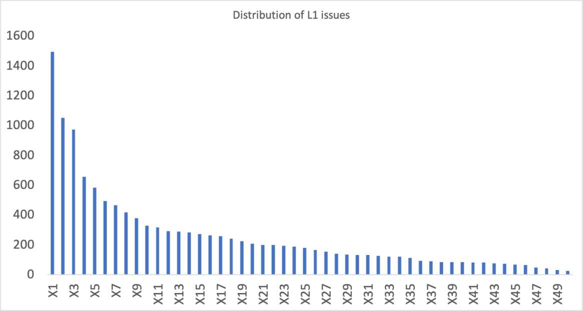

Figure 2: Distribution of anonymized L-Coarse issues.

4.2 Models & Training Settings

We compare three transformer-based seq2seq models: BART (Lewis et al., 2019a), T5 (Raffel et al., 2019a), and Pegasus (Zhang et al., 2020). We adopt the pretrained models from the HuggingFace[3] implementation and fine tune the models on our training dataset.The details regarding the training parameters we used can be found in the appendix C. We implement several training techniques to improve model performance and help the models better adapt to our application:

Auxiliary Tasks Recent studies have shown that multi-tasking helps improve model performance (Raffel et al., 2019a; Aribandi et al., 2021). In our case, in addition to L-Coarse and L-Granular issue generations, we include other NLP tasks into model training. The auxiliary tasks include summarization, token infilling, and classification. Details about the auxiliary tasks can be found in Appendix A.

Sentence & Issue Shuffling A unique characteristic of issue identification is that the order of individual issues does not matter, as long as all relevant issues are correctly identified from customer feedback. This is contrary to conventional language generation tasks, such as summarization and translation, where the order of generated words must follow natural language syntax. Hence, we shuffle both input sentences and target issues during model training to induce the model to learn to ignore the ordering information and achieves better performance.

4.3 Evaluation Metrics

Similarity Measure To evaluate our models' performance, we first measure the similarity between model-generated and human-annotated issues. We employ two sets of metrics to measure both lexical and semantic similarity.

- **Lexical Metrics**: We compute the ROUGE-1 & ROUGE-2 (Lin, 2004) metrics between the model-generated text and human-annotated issues. They measure the overlapping unigrams and bigrams between the texts.

- **Semantic Metrics**: We define two new metrics - **SimCSE Precision** & **SimCSE Recall** - which evaluate the precision and recall at the customer feedback level, based on the pairwise similarity of the sentence embeddings for model-generated and human-annotated issues. Let the model-generated issues be represented by $X_1, X_2, .., X_I$ and the human-annotated issues by $Y_1, Y_2, .., Y_J$. We measure the pairwise similarity between each issue pair X_i and Y_j ($0 \leq i \leq I$ and $0 \leq j \leq J$) as the cosine similarity of their corresponding SimCSE embeddings (Gao et al., 2021). A similarity higher than a given threshold is considered a match between a pair of issues as seen in equation 1. Based on the number of matches, we are able to compute precision and recall. The threshold (0.7) is chosen based on our

[2] All annotators followed the confidentiality policy when conducting the labeling jobs.

[3] https://huggingface.co/

Model	L-Coarse				L-Granular			
	ROUGE		SimCSE		ROUGE		SimCSE	
	R-1	R-2	P	R	R-1	R-2	P	R
MCML-BERT	-24.1%	-41.8%	-27.5%	-35.7%	-	-	-	-
BART	+1.6%	+0.3%	-3.4%	+4.9%	-2.9%	-4.0%	-7.4%	0.0%
Pegasus	+1.2%	+1.0%	+0.5%	-0.7%	-3.9%	-5.9%	-1.8%	-9.3%

Table 2: Performance comparison. The performance numbers are shown as the difference from the T5 performance. Here R-1 & R-2 refers to ROUGE-1 and ROUGE-2 scores respectively, whereas P refers to precision and R refers to recall.

empirical study on the inter-label similarity distribution.

$$match(X_i, Y_j) = \begin{cases} 1, & \text{if sim}(X_i,Y_j) > 0.7 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where sim(X_i,Y_j) is the similarity score between a model generated issue and human-annotated issue. More details with examples on the calculation of this metric can be found in the Appendix B.

Human Evaluation Both ROUGE and SimCSE evaluations have their limitations, i.e. ROUGE score fails to capture semantic similarity while SimCSE evaluation relies on the embedding quality. Both methods are influenced by annotation noise. Thus, we also conduct human evaluation to assess the quality of the model- and human-generated issues. To have an unbiased evaluation, we adopt a double blind approach, where we first shuffle the model outputs and human annotations and then ask human auditors to adjudicate the issues (shuffled) according to the following criteria:

- Rating 1: All the issues are correctly identified.
- Rating 2: At least one issue is missing.
- Rating 3: At least one issue is incorrectly assigned.

5 Results

In this section, we report the results and analysis from our study. To our knowledge there are no publicly-available datasets in e-commerce or performance benchmarks for the problem of multi-level product issue identification. Thus, we conduct the study using our private customer-feedback dataset. Due to confidentiality, we can not report absolute model performance numbers but only relative ones compared to the baseline models.

5.1 Model Performance

First, we compare the performance of different transformer-based seq2seq models, including BART (Lewis et al., 2019b), T5 (Raffel et al., 2019b), and Pegasus (Zhang et al., 2019). We attempted to provide a quantitative comparison between our seq2seq approach and other approaches such as topic modeling and clustering. However, the outputs from these alternative approaches are not directly comparable to ours. Instead, we train a multi-class multi-label classification model for L-Coarse issues using BERT (Devlin et al., 2018b) (MCML-BERT) as a baseline for comparison. We do not apply the classification modeling approach to the L-Granular-issue prediction task due to the large number of L-Granular issues.

Results on the test dataset are provided in Table 2, where we choose T5 as the base model. We observe that for L-Coarse issues, MCML-BERT significantly underperforms in comparison to all the seq2seq models. Among the seq2seq models, for L-Coarse-issue identification, BART and Pegasus perform marginally better than T5. For L-Granular-issue identification, however, T5 shows consistent better performance. Due to the best overall performance of T5, in the following sections, we will focus on the results produced by T5.

5.2 Zero- and Few-shot Learning

As we aim to identify diverse and dynamic issues, it's important for the model to be able to discover novel issues with zero or only few training samples. Here, we examine the zero-shot and few-shot learning capability by varying the amount of samples containing specific issues in the training dataset. Specifically, we select two frequent issues from L-Coarse and L-Granular respectively. We fine-tune T5 using the training dateset containing a fraction of samples with those selected issues, then evaluate the model on the same test dataset.

Table 3 shows the relative model performance (F1-scores) as a function of the fractions of samples

Level - Selected Issue	100% samples	75% samples	50% samples	25% samples	0% samples
L-Coarse - X1	100.0%	98.5%	98.3%	92.2%	0%
L-Coarse - X2	100.0%	77.2%	69.9%	56.1%	0%
L-Granular - X3	100.0%	97.1%	96.0%	90.7%	75.8%
L-Granular - X4	100.0%	95.3%	92.7%	81.5%	68.6%

Table 3: Relative SimCSE F1-score performance with different proportions of training examples containing the selected issues (anonymized).

exposed during model training. In this table, we take the model performance when 100% samples with the selected issues are included in training as baseline. In general, the performance decreases as fewer training samples are included. For L-Granular issues, even when there is no sample of selected issues present during model training, the model is still able to identify the correct issues with a high F1-scores (75.8% and 68.6%). It indicates that the model generalizes well on identifying unseen fine-grained issues from customer feedback. On the other hand, though the model fails to recognize unseen L-Coarse issue (no exposure during training), its detection performance improves quickly: with 25% samples, model's F1-scores rise to 92.2% and 56.1%. This shows that our approach, fine-tuning a pre-trained seq2seq model such as T5, is label efficient.

5.3 Human Evaluation

Human evaluation is performed on 850 records randomly sampled from the test dataset. We ask auditors to evaluate both model-generated and human-annotated issues following double blind auditing procedures (see Section 4.3 for details). We obtain the ratings assigned by auditors for all the samples. We then compute the performance metric that is defined as the (percentage of Rating-1 samples + 0.5 $*$ percentage of Rating-2 samples). Table 4 shows the model performance relative to human annotator. As can be observed from this table, our model achieves 82.1% and 95.4% of human-level performance for L-Coarse and L-Granular issue identification, respectively. Given the complexity and challenges of the tasks, such results indicate the effectiveness of our approach. On the other hand, the better L-Granular performance aligns with our previous observations that model generalizes better on L-Granular issues than L-Coarse issues. We hypothesize that this is due to that fact that L-Granular issues are concrete ones while L-Coarse issues are more abstract, which is more challenging for the model to learn, as also observed in Zeyu Liu (2021).

Issue Level	Relative to Human Performance
L-Coarse	82.1%
L-Granular	95.4%

Table 4: Model performance based on human auditing.

5.4 Ablation Study

We examine the effect of adding auxiliary tasks and sentences & issues shuffling to model training. Here, the baseline model is a T5 model that's fined tuned using only L-Coarse & L-Granular issue identification tasks where the issues are presented in a fixed order.

Additional Training Techniques	L-Coarse	L-Granular
Summarization, Token-infilling, classification	-0.16%	+1.44%
Sentences & issues shuffling	+3.48 %	-0.89%
All tasks above	+4.40 %	+1.33%

Table 5: Effect of auxiliary tasks and sentences & issues shuffling (SimCSE F1-scores).

Table 5 shows the change in SimCSE F1-scores when including the additional training techniques. As can be seen from this table, auxiliary tasks improve L-Granular performance while sentences & issues shuffling gives better performance at the L-Coarse level. We achieve the best overall model performance by combining both set of techniques.

6 Conclusion

In conclusion, we analyzed the challenges in identifying diverse and multi-level product issues from customer feedback. To overcome these challenges, we creatively formulated our problem as a seq2seq modeling problem and leveraged text-to-text transfer learning framework. We utilized multi-tasking to generate product issues at multiple levels using a single model, which minimizes operational cost. Results show that our model performs closely to human on issue identification tasks. We also observed that our approach is label efficient and our model generalizes well to identify unseen L-Granular issues. Our next step is to explore ways to improve our model's performance and generalizability on L-Coarse issue prediction.

References

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Ayoub Bagheri, Mohamad Saraee, and Franciska De Jong. 2014. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40(5):621–636.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Stefan Debortoli, Oliver Müller, Iris Junglas, and Jan Vom Brocke. 2016. Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, 39(1):7.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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

Hui Du, Xueke Xu, Xueqi Cheng, Dayong Wu, Yue Liu, and Zhihua Yu. 2016. Aspect-specific sentimental word embedding for sentiment analysis of online reviews. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 29–30.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Maarten Grootendorst. 2020. Bertopic: leveraging bert and c-tf-idf to create easily interpretable topics (2020). *URL https://doi. org/10.5281/zenodo*, 4381785.

Byeongki Jeong, Janghyeok Yoon, and Jae-Min Lee. 2019. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48:280–290.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019b. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tong Liu, Siyuan Wang, Jingchao Fu, Lei Chen, Zhongyu Wei, Yaqi Liu, Heng Ye, Liaosa Xu, Weiqiang Wang, and Xuanjing Huang. 2021a. Fine-grained element identification in complaint text of internet fraud. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3268–3272.

Yang Liu, Yifei Sun, and Vincent Gao. 2021b. Improving factual consistency of abstractive summarization on customer feedback. In *Proceedings of The 4th Workshop on e-Commerce and NLP*, pages 158–163.

Jian Mou, Gang Ren, Chunxiu Qin, and Kerry Kurcz. 2019. Understanding the topics of export cross-border e-commerce consumers feedback: an lda approach. *Electronic Commerce Research*, 19(4):749–777.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019a. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019b. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215.

Xuesong Tong, Bin Wu, Shuyang Wang, and Jinna Lv. 2018. A complaint text classification model based on character-level convolutional network. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, pages 507–511. IEEE.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jungo Kasai Hannaneh Hajishirzi Noah A. Smith Zeyu Liu, Yizhong Wang. 2021. Probing across time: What does roberta know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820—-842.

Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. 2011. Constrained lda for grouping product features in opinion mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 448–459. Springer.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

A Auxiliary Tasks

- Summarization: We use a customized dataset containing 14k customer-feedback records with human-annotated summaries. For this task, we train the model to mimic human-generated summaries.

- Token Span infilling: It has been recently shown (Devlin et al., 2018a) that a masked language modeling based objective results in superior performance on various downstream tasks. The objective for the model is to predict the missing tokens during model pre-training. We follow a similar approach by randomly replacing spans of token by single sentinel tokens. The target sentence corresponds to all of the dropped-out spans of tokens, delimited by the same sentinel tokens used in the input sequence. The number of samples for this task is 7k.

- Classification: For this auxiliary task, we use a customer-feedback dataset that has been labelled using a fixed set of product-related categories. There are 10k samples in total. We cast this classification problem into a text-to-text format, where the input is the customer feedback text and the output is a text-based label.

B SimCSE Evaluation

For each sample, we compute pairwise cosine similarity base on the SimCSE (Gao et al., 2021) sentence embedding computed for each of the model-generated and human-annotated issues. If the similarity is greater than a threshold for an issue pair, we increment the number of matches by 1, even if there is more than one match for a given issue. The threshold was selected based on L-Coarse and L-Granular intra-issue similarity distributions, i.e., the distribution of all the pairwise cosine similarity between the unique issues from the data. Based on our emperical study, we choose 0.7 as the threshold. Table 6 illustrates the SimCSE precision and recall calculation process. Note that the issues are comma delimited. We can see for the first example in the table although "minor scratches" and "minor cosmetic defects" vary lexically, the two are semantically similar, which is reflected in their greater than 0.7 SimCSE similarity. On the other hand, "minor scratches" has a SimCSE similarity lower than 0.7 with "counterfeit issue" and hence not a match, which is consistent with our intuition. The number of matches includes all the issues with similarity greater than 0.7. Based on the number of matches, we can calculate the instance level precision and recall as usual. Averaging instance level precision and recall, we can obtain precision and recall for aggregated dataset level.

C Training Parameters

For training transfomer-based seq2seq models, we select maximum input text and target text lengths as 512 and 64, respectively. We use a batch size of 40 distributed equally over 8 GPUs, with a learning

Model Output	Human Annotation	# Match
minor scratches	minor cosmetic defects, fitting issue	1
size issue, not as described	size issue, material issue, not as described	2
quality issue	quality issue	1

Table 6: Calculation of the Precision and Recall, with similarity threshold of 0.7 for L-Coarse

rate of 5e-5. We train this setup over 25 epochs on a p3.16xlarge EC2 instance with distributed model and data parallelism to fine-tune the model. During inference we use beam search (Sutskever et al., 2014) to generate the target text sequence with a beam width of 2 and length penalty $\alpha = 2.5$ (Wu et al., 2016).