Brazilian Portuguese Product Reviews Moderation with AutoML

Lucas Nildaimon dos Santos Silva [1],
Carolina Francisco Gadelha Rodrigues[2], **Ana Claudia Zandavalle** [3], **Tatiana da Silva Gama** [2]
Fernando Rezende Zagatti [1], **Livy Real** [4]
[1] Department of Computing, Federal University of São Carlos, Brazil
[2] Americanas S.A., Rio de Janeiro, Brazil
[3] Federal University of Santa Catarina, Florianópolis, Brazil
[4] Quinto Andar Inc, São Paulo, Brazil
{lucas.silva,fernando.zagatti}@estudante.ufscar.br
{carolfgr25,ana.zandavalle,pro.gamat85,livy.real}@gmail.com

Abstract

Product reviews are valuable resources that assist shoppers in making informed transactions by reducing uncertainty within the purchase process. However, user-generated content is not always secure or adequate. The goal of customer review moderation is to ensure both a secure environment for all parties participating and the integrity of the review information. Content moderation is a difficult task even for human moderators, and in some circumstances, due to the enormous volume of reviews, manual content moderation is not practical. In this paper, we present the experiments carried out using automated machine learning (AutoML) for moderating product reviews on one of Brazil's largest e-commerce platforms. Our machine learning-based solution is faster and more accurate than the previously used content moderation system, performed by a third-party company system dependent on human intervention. Overall, the results showed that our model was 31.12% more accurate than the third-party company system and it had a fast development due to the use of AutoML techniques.

1 Introduction

E-commerce platforms frequently allow customers to provide feedback (reviews) on the products or services they have purchased. Customer reviews are critical mechanisms for reinforcing product and service quality, increasing consumer satisfaction and purchase intent, and identifying areas for business improvement (Geng and Chen, 2021; Askalidis and Malthouse, 2016).

Figure 1 illustrates an example product review from a major online marketplace in Brazil[1].

This type of review is an example of user-generated content (UGC), which is widely considered more trustworthy, authentic, and realistic than firm-generated content. As a result, reviews are critical in assisting other potential customers in their decision-making. However, when dealing with UGC, it is essential to provide a secure environment for users, companies, and brands.

The process of monitoring UGC to ensure that it complies with the platform's rules and guidelines is known as content moderation. This is accomplished by removing or blocking inappropriate content while publishing or approving those that follow the rules. Content can be blocked for a variety of reasons, including violence, nudity, offensiveness, hate speech, and other factors. Therefore, review content moderation is indispensable to provide a safe user experience, and avoid damaging the brand reputation, and loss of revenue.

Content moderation can be manual, automatic, or a combination of the two. In our scenario, manual content moderation is impractical due to the large volume of reviews received by the e-commerce company, as it receives more than 20k reviews weekly. In this work, we describe the process of developing a machine learning-based solution for automated product review moderation. The main goal was to achieve more accurate and efficient results compared to the prior third-party solution adopted by the e-commerce company. We also wanted to internalize the moderation process, which was previously handled by a third-party company. Working on Brazilian Portuguese was one of our major challenges since there was no publicly available content to base our solution on. Indeed there are some reviews corpora available on Portuguese, but those do not count with moderation information.

We organize the rest of the paper as follows: In Section 2, we describe related works. In Section 3, we detail our methodology and experimental design. In Section 4, we present our results. Finally, we conclude in Section 5.

[1]The example translation: The cell phone is very good, the cameras have good quality, and the size is wonderful. Loved it!!! Highly recommended.

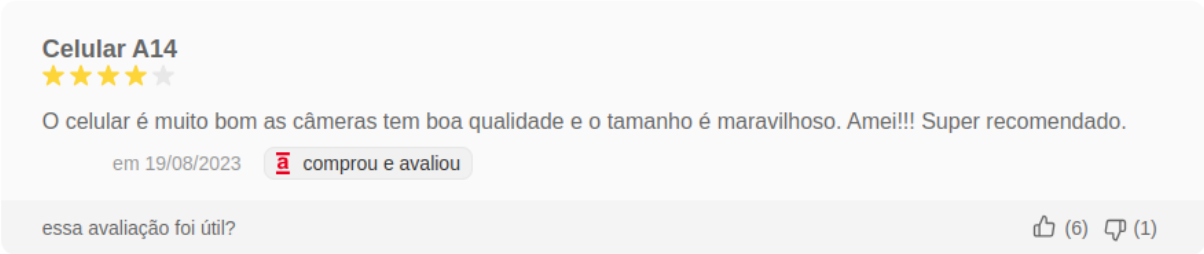

Figure 1: Example of a product review.

2 Related works

There are numerous works in the literature that are related to product reviews. Many of them concern the sentiment analysis of product reviews (Mukherjee and Bhattacharyya, 2012; Yang et al., 2020; Shivaprasad and Shetty, 2017; Haque et al., 2018) and focus on English, which is not the focus of the present work.

Automated content moderation is frequently viewed as a binary classification task that determines whether user-generated content should be published or removed from a platform. (Pavlopoulos et al., 2017; Risch and Krestel, 2018; Shekhar et al., 2020; Ueta et al., 2020; Korencic et al., 2021).

Some of the research literature on automated content moderation addresses issues like algorithmic biases and ethical considerations. (Binns et al., 2017) offers various exploratory methodologies to quantify the biases of algorithmic content filtering systems. (Gillespie, 2020) debates whether or not content moderation should be automated.

The type of data involved in content moderation can vary depending on the application and may involve multimodal tasks such as video moderation (Tang et al., 2021). In our work, we only tackle the textual moderation problem.

The work by (Shido et al., 2022) describes an automated content moderation system for text data that is based on machine learning (ML) models. The system is used to moderate interactions between platform users while transactions are in progress. The system employs a rule-based system, an ML model, and human moderators to detect messages with abusive intent or that may violate platform rules.

The work conducted by (Doan et al., 2021) delves into the application of machine learning for automated content moderation, particularly focusing on user-submitted content related to cosmetic procedures, on the RealSelf.com platform. The study utilized a dataset comprising 523,564 user-submitted reviews on RealSelf.com, each previously categorized as either "published" or "unpublished" by the RealSelf content moderation team. Employing an ensemble approach, the study considered both textual features of the reviews and meta-features associated with the reviewers for effective moderation. Here, we approach this problem similarly, determining whether or not to publish a product review in the product web-page via an ML-based moderation system.

3 Methodology

In this section, we present the business rules that guide our solution as well as the methods and datasets used to build it.

To perform the automatic content moderation, we chose to create a binary Supervised Machine Learning (ML) Model, which requires previously human-labeled examples to learn to classify automatically. It is important to highlight that we aim to have an economic and totally 'inside house' solution, so we did not explore on-demand large language model providers.

Since there was no public available content that could be used to train a classifier, our first step was to build a dataset that represents the business challenge.

3.1 Annotation guidelines

To have a trustful dataset, the labeling instructions must be precise; otherwise, annotators will rely on their subjective judgment, resulting in incorrectly labeled data that harm model learning (Markov et al., 2022). Therefore, an Annotation Guideline, that is, an instructive guide that serves as a guiding document for those involved in the annotation task, with as little personal bias as possible and in a consistent manner, is essential.

First of all, we explored the business rules established by the company to deeply understand all

the issues involved in reviewing content moderation. Reviews that must be made public on the platform must focus specifically on product characteristics such as advantages/disadvantages, quality, size, strengths/weaknesses, etc. This is necessary so that the reviews can assist other consumers in making a purchase decision based on the general aspects of the products themselves, rather than other individual factors in the purchase journey. It is common for the shopper use the review form as an easy way to communicate with the e-commerce platform, e.g., using it to complain about delivery fees or ask for help. Since this information is not helpful in the decision-making process of potential buyers, it is considered inadequate to compose the review information of a given product.

Then, to develop our Guideline, we started with data exploration: we needed to understand the content generated by users, independently of business rules. There is no set methodology for this task, and it can be performed in a variety of ways, such as clustering the data, generating graphics or word clouds. In this project, the exploration was carried out by manually analyzing small batches of aleatory data. The primary goal was to identify recurring issues and to categorize them.

During our investigation, we discovered several reviews that included the following themes: Stock; Invoice; Tracking Code; Customer Service; Exchange; Charge-back; Return Delivery; Assembly of Products; Warranty; Coupon; Doubts of Procedures. Because the aforementioned topics are all related and deemed inappropriate for the site, the Guideline unified them all as subthemes of a single category called `Service`.

This process was repeated until all user-generated contexts were fully understood, exemplified, and grouped. A new batch of data was annotated at the end of the Guideline's development to confirm the possible existence of subjects not considered and to clear annotators' doubts. Currently, the Guideline considers nine distinct categories: `Product`, `Advertisement`, `Service`, `Delivery`, `Institutional`, `Inadequate`, `Pre-purchase`, `QnA`[2], and `Vague`. Each theme has a predetermined number of subthemes that are grouped together. Table 1 shows the required action for each of the Guideline's nine categories.

It is important to note that, since we deal with real-world data, there are correlations and dependencies among the classification labels. Reviews that mention both `Product` and `Delivery` aspects are a common example. So, a review as *Produto de ótima qualidade. Comprei no domingo, na quarta feira já recebi em casa*[3], labeled as `Product` and `Delivery`, should be rejected. In this particular case of the `Delivery` label, the company can not guarantee the same delivery conditions to all the customers independently of the shipping address, therefore this information is not considered 'useful to all customers'. The Annotation Guideline is also relevant because it addresses the many interrelationships among labels and how to annotate each sample.

3.2 Dataset annotation

Following the validation of the Guideline, we began the official dataset annotation. The data for this project were extracted randomly over a period of six months. The reviews were annotated binary-style, with ACCEPTED for those that should be published on the site and REJECTED for those that should not. Thus the machine readable dataset was annotated with ACCEPTED/REJECTED labels, being the more complex labels, explained in the previous section, clues to the annotators to consistently arrive in the binary labels in any context. The annotated dataset comprises 3,965 reviews, randomly distributed into 2,379 samples for training and 1,586 samples for testing. Both the training and testing sets consist of 73% positive class samples and 27% negative class samples.

The annotation process involved three annotators, all native speakers of Brazilian Portuguese, with two annotators responsible for the same official batches and a third curating the noisy annotation. As a result, at the end of the task, any disagreements between the two main annotators, as well as any inconsistencies discovered, were resolved before the data was provided to the model. This entire process is critical for solving human error and personal biases and ensuring that the model receives annotated data in the best possible way. In the next subsection, we present our proposed ML pipeline for this task.

3.3 Machine Learning-based moderation

Figure 2 displays the common ML model development pipeline. The first step is data prepara-

[2]QnA, here, stands for Question Answering, a common feature of e-commerce platforms that makes possible sellers answer questions of customers.

[3]Product of great quality. Bought Sunday and received it next Wednesday in my place.

Category	Example	Action
Product	Better than I expected, great cable!	Accept
Advertisement	Product advertisement is different from what was received! The size is too large!	Accept
Delivery	Thank you very much, the product arrived before the expected date. Thank you!	Reject
Institutional	Very efficient and practical to buy on the website, highly recommend it.	Reject
Service	I need the tracking code.	Reject
Inadequate	This challenge is only for those who want to lose weight in a healthy way! [Hyperlink removed]	Reject
Pre-purchase	I haven't purchased it yet, but I hope it's good and doesn't have any defects.	Reject
QnA	I would like to know if this range hood is available for an island?	Reject
Vague	Gospel music 'Diante do Trono'.	Reject

Table 1: Categories and procedures of the annotation guideline.

tion, which involves cleansing and standardizing the data. The second step, feature engineering, involves creating and selecting the features required to train the model. In the third step, algorithm selection and configuration, we test various ML algorithms and hyperparameter values to find those that provide a satisfactory solution. Finally, in the last two steps, we train and evaluate the developed model.

The main point of this work was to create an ML model for the binary text classification task. As a result, the techniques used in each step of the ML pipeline had to be suitable for dealing with text data. It was also relevant to the project to pursue the lowest possible costs and necessary time for inferences and (re)training the models; therefore we focus exclusively on shallow learning methods (Zhang and Ling, 2018; Janiesch et al., 2021; Silva et al., 2021).

The first step in data preparation was to concatenate the review title and body so that we could treat it as a single input. Subsequently, we convert all text to lowercase before removing punctuation, accents, and special characters with regular expressions. Subsequently, we delete duplicated reviews. In feature engineering, we use the term frequency–inverse document frequency (TF-IDF) method (Aizawa, 2003) to generate our features, and SelectKBest, a feature selection technique based on univariate statistical tests, to select only the K highest scoring features. For algorithm selection and configuration, we use AutoViML and Auto-sklearn (Feurer et al., 2015) automated machine learning systems (AutoML) to help us accelerate experimentation. AutoML systems automatically configure, train, and compare multiple ML algorithms, reducing the need for human intervention to test different ML algorithms and hyperparameter values (Hutter et al., 2019). Auto-sklearn serves as a versatile end-to-end AutoML system with multiple machine learning algorithms. However, as of this experiment, AutoViML offers only two algorithm options: the random forest (RF) and the naive Bayes algorithms. It's noteworthy that while AutoViML automatically generates and optimizes text vectorization, Auto-sklearn necessitates prior text vectorization. These automated solutions assisted in selecting hyperparameter values for feature generation with TF-IDF and feature selection with SelectKBest, ultimately guiding us to employ an RF algorithm for training our final model. To evaluate the proposed model, we employ common machine learning evaluation metrics, including precision, accuracy, recall, and the F1-score. Table 2 displays the results of the first version of the model, which we call in-House V1, in the test dataset.

	Precision	Recall	F1-score	Number of samples
REJECTED	0.79	0.73	0.76	433
ACCEPTED	0.90	0.93	0.92	1153
Mean Value	0.85	0.83	0.84	

Table 2: Results of the first version of the model in the test dataset.

3.4 Qualitative error analysis and model improvements

The qualitative analysis of the errors is one method for obtaining valuable information about the model's behavior. It was possible to obtain inputs for the creation of new training sets more focused on the problem by analyzing the incorrect predictions in the test dataset of the first version of the model.

		Predicted Class	
		REJECTED	ACCEPTED
True Class	REJECTED	314	122
	ACCEPTED	62	1088

Table 3: Confusion matrix for the first version of the model in the test dataset.

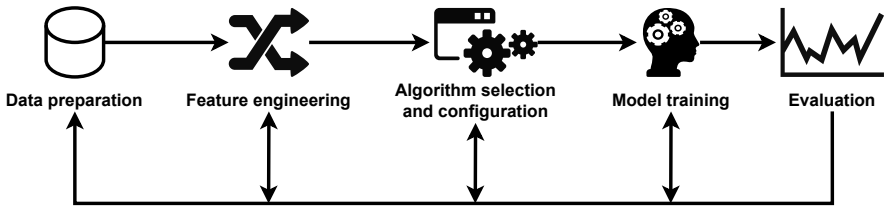

Figure 2: Common machine learning model development pipeline.

Table 3 displays the confusion matrix for the first version of the model in the test dataset. False negatives (FN) are assessments that should have been classified as ACCEPTED by the model, but were predicted as REJECTED in the context of this project. False positives (FP) are instances where the model should have predicted REJECTED but instead was classified as ACCEPTED. The errors patterns discovered were grouped through a qualitative analysis of the 184 misclassified reviews (FN + FP).

Regarding the reviews that were not accurately blocked by the model (FP), the main area for improvement should be centered on reviews that pertain specifically to the `Delivery` and `Service` contents. Other relevant contexts to be improved were disclosure of sensitive information, as well as contexts related to legal matters and pre-purchase evaluations.

Given the reviews that were erroneously blocked by the model (FN), efforts to address this issue should focus on contexts related to product reviews containing negative sentiments. Based on this analysis, a new batch of 1,000 reviews focused on the identified contexts underwent annotation and curation. Subsequently, incorporating this fresh batch of data, we augmented the dataset's size to 4,965 samples, comprising 2,979 samples in the training dataset and 1,986 samples in the test set. This expansion furthered the balance in class distribution, with the positive class accounting for 62% of samples and the negative class for 37% in both the training and test sets. Next, we used the new datasets to create a second version of the model (in-House V2) using the same ML pipeline as before.

4 Results

The qualitative error analysis enabled the development of a second version of the model with the goal of improving on the first version's misclassifications. Table 4 shows the results of the model's second version in the test dataset, and Table 5 displays the confusion matrix.

	Precision	**Recall**	**F1-score**	**Number of samples**
REJECTED	0.85	0.81	0.83	742
ACCEPTED	0.89	0.91	0.90	1244
Mean Value	0.87	0.86	0.87	

Table 4: Results of the second version of the model in the test dataset.

		Predicted Class	
		REJECTED	ACCEPTED
True Class	REJECTED	597	145
	ACCEPTED	108	1136

Table 5: Confusion matrix for the second version of the model in the test dataset.

Since the two models were developed using different training and test datasets, it is difficult to make a fair comparison between them. However, we can still evaluate and compare their generalization capacities by looking at the results achieved in both versions of the test datasets. By comparing the results in Table 2 and Table 4, we observed that in-House V2 shows a more balanced performance between classes, with improved results in relation to the negative class.

Since our primary goal was to develop a ML model that could replace the third-company moderation system, we needed to compare their performances to determine if the proposed model was adequate for the task. Table 6 compares the results achieved by the in-House V2 model and the third-party company in the test dataset. Overall, our proposed model surpasses the baseline results set by the third-party company.

4.1 Model evaluation in production

Even with satisfactory metrics from offline model evaluation, it was necessary to assess the model's performance in production. To accomplish this, we used the Shadow Deployment strategy, in which the proposed system is deployed in parallel with the official system in production. The proposed

	REJECTED			ACCEPTED			
Approach	Precision	Recall	F1- Score	Precision	Recall	F1- Score	Number of test samples
in-House V2	**0.85**	0.81	**0.83**	0.89	**0.91**	**0.90**	1986
Third-party company	0.56	**0.91**	0.69	**0.91**	0.58	0.71	1986

Table 6: Comparison of the results for the two approaches of moderation.

system receives and moderates the same content as the official system, but its predictions are not used. Instead, the responses of the proposed model were saved for future comparisons of the two systems.

Following a two-week testing period, the two models moderated approximately forty thousand reviews. To conduct a manual analysis, a sample of 5,000 data points was chosen at random and annotated by humans in accordance with the guidelines. The data points in the sample were divided into two groups: those in which both the in-House system and the third-party company's system agreed on the classification, and those in which the two systems gave different answers for the same review content. Regarding the agreements, both systems achieved an accuracy value of 0.89. In relation to the disagreements, the in-House V2 model correctly predicted 81.1% of the 2,500 analyzed samples, against 18.9% achieved by the third-party company system. Overall, the results of this analysis showed that our model was 31.12% more accurate than the third-party company system. The average moderation time of the in-House V2 model was 771 milliseconds per review, compared to 68 minutes for the third-party company's system, which most likely included human moderators, indicating that the in-house solution has a much faster ability to provide quality information to the customer.

5 Conclusion

In this paper, we outline our methodology for constructing a machine learning-driven moderation system, aimed at curtailing the dissemination of unsolicited content within the customer reviews section of one of the largest Brazilian e-commerce website. Our solution, founded primarily on the implementation of TF-IDF features, a Random Forest model, and AutoML, demonstrated robust performance in terms of time efficiency and precision in this task. Although we had the possibility of utilizing more sophisticated techniques, such as transformer-based models, we opted for a straightforward, yet effective solution, especially considering inference and (re)training costs. (Silva et al., 2021) showed that for downstream tasks, classical machine learning techniques can achieve the same results as deep learning techniques, being the inference time of transformer-based models up to 9 times more than classical approaches.

The incorporation of AutoML facilitated the acceleration of the solution prototyping process, thereby affording additional time to create comprehensive annotation guidelines. This, in turn, led to high-quality labeling of the data utilized for the model's training. The approach to model development was centered on data, emphasizing the importance of data quality for robust model creation.

After conducting both offline and online evaluations, we have determined that the in-House V2 model outperforms the third-party moderation previously utilized in terms of both speed and accuracy. Accordingly, our solution has superseded the previous system, and it is now the primary method employed to moderate customer reviews on the e-commerce website.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Georgios Askalidis and Edward C. Malthouse. 2016. The value of online customer reviews. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, page 155–158, New York, NY, USA. Association for Computing Machinery.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics*, pages 405–415, Cham. Springer International Publishing.

Alicia Doan, Nathan England, and Travis Vitello. 2021. Online review content moderation using natural language processing and machine learning methods: 2021 systems and information engineering design symposium (sieds). In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE.

Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter.

2015. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems 28 (2015)*, pages 2962–2970.

Ruoshi Geng and Jun Chen. 2021. The influencing mechanism of interaction quality of ugc on consumers' purchase intention – an empirical analysis. *Frontiers in Psychology*, 12.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. 2018. Sentiment analysis on large scale amazon product reviews. In *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*, pages 1–6.

Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Cham.

Christian Janiesch, Patrick Zschech, and Kai Heinrich. 2021. Machine learning and deep learning. *Electronic Markets*.

Damir Korencic, Ipek Baris, Eugenia Fernandez, Katarina Leuschel, and Eva Sánchez Salido. 2021. To block or not to block: Experiments with machine learning for news comment moderation. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 127–133, Online. Association for Computational Linguistics.

Todor Markov, Chong Zhang, Sandhini Agarwal, Tyna Eloundou, Teddy Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2022. A holistic approach to undesired content detection in the real world. *arXiv preprint arXiv:2208.03274*.

Subhabrata Mukherjee and Pushpak Bhattacharyya. 2012. Feature specific sentiment analysis for product reviews. In *Computational Linguistics and Intelligent Text Processing*, pages 475–487, Berlin, Heidelberg. Springer Berlin Heidelberg.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

Julian Risch and Ralf Krestel. 2018. Delete or not delete? semi-automatic comment moderation for the newsroom. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 166–176, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ravi Shekhar, Marko Pranjić, Senja Pollak, Andraž Pelicon, and Matthew Purver. 2020. Automating news comment moderation with limited resources: benchmarking in croatian and estonian. *Journal for Language Technology and Computational Linguistics*, 34(1):49–79.

Yusuke Shido, Hsien-Chi Liu, and Keisuke Umezawa. 2022. Textual content moderation in C2C marketplace. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 58–62, Dublin, Ireland. Association for Computational Linguistics.

T. K. Shivaprasad and Jyothi Shetty. 2017. Sentiment analysis of product reviews: A review. In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 298–301.

Diego F. Silva, Alcides M. e. Silva, Bianca M. Lopes, Karina M. Johansson, Fernanda M. Assi, Júlia T. C. de Jesus, Reynold N. Mazo, Daniel Lucrédio, Helena M. Caseli, and Livy Real. 2021. Named entity recognition for brazilian portuguese product titles. In *Intelligent Systems*, pages 526–541, Cham. Springer International Publishing.

Tan Tang, Yanhong Wu, Yingcai Wu, Lingyun Yu, and Yuhong Li. 2021. Videomoderator: a risk-aware framework for multimodal video moderation in e-commerce. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):846–856.

Shunya Ueta, Suganprabu Nagarajan, and Mizuki Sango. 2020. Auto content moderation in c2c e-commerce. In *2020 USENIX Conference on Operational Machine Learning (OpML'20)*, page 33.

Li Yang, Ying Li, Jin Wang, and R. Simon Sherratt. 2020. Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE Access*, 8:23522–23530.

Ying Zhang and Chen Ling. 2018. A strategy to apply machine learning to small datasets in materials science. *npj Computational Materials*, 4.