Modelling Argumentation for an User Opinion Aggregation Tool

Pablo Weingart, Thiemo Wambsganss, Matthias Söllner
University of Kassel, Bern University of Applied Sciences, University of Kassel
pablo@weingart.co, thiemo.wambsganss@bfh.ch, soellner@uni-kassel.de

Abstract

We introduce an argumentation annotation scheme that models basic argumentative structure and additional contextual details across diverse user opinion domains. Drawing from established argumentation modeling approaches and related theory on user opinions, the scheme integrates the concepts of argumentative components, specificity, sentiment and aspects of the user opinion domain. Our freely available dataset includes 1,016 user opinions with 7,266 sentences, spanning products from 19 e-commerce categories, restaurants, hotels, local services, and mobile applications. Utilizing the dataset, we trained three transformer-based models, demonstrating their efficacy in predicting the annotated classes for identifying argumentative statements and contextual details from user opinion documents. Finally, we evaluate a prototypical dashboard that integrates the model inferences to aggregate information and rank exemplary products based on a vast array of user opinions. Early results from an experimental evaluation with eighteen users include positive user perceptions but also highlight challenges when condensing detailed argumentative information to users.

Keywords: argument mining, user opinion analysis, online review analysis

1. Introduction

Computational methods based on natural language processing have been shown to provide considerable potential to elicit information on phenomena such as user adoption decisions (Kwahk and Kim, 2017) or product design decisions (Yang et al., 2019) from widely available user opinion data. Researchers have used opinion mining, sentiment analysis, topic modelling (e.g., Lee et al., 2023; Cheng and Jin, 2019) or particularly opinion summarization approaches for parsing and aggregating of the data. Opinion-specific document summarization is a commonly used natural language task that employs computational modelling to aggregate (informative) user opinions, generating high-level text that provides user opinion summaries both generally as well as for specific aspects (Bražinskas et al., 2021; Angelidis et al., 2021; Isonuma et al., 2021; Hosking et al., 2023). Yet, the outputs of more recent large language models as well as smaller-sized lower cost models usually remain restricted to high-level text and prone to hallucination, degrading system performances or introduce biases, among others (Ji et al., 2023; Maynez et al., 2020). Notably, the model architectures of generative models do not output numeric data suitable for conventional statistical analyses. In a similar vein, the text remains impractical to parse for users when comparing multiple opinion subjects, such as a set of products within a category. Techniques used by researchers for providing quantifiable metrics, and potentially complement such opinion summaries, commonly derive from aspect-based sentiment analysis (Do et al., 2019; Nazir et al., 2020) and topic modelling to collect insights from user opinions, e.g., pain points (Lee et al., 2023). By default, they operate on all opinion statements equally and do not emphasize more qualitative statements.

One possible solution to expand on these approaches could be argument mining (AM). AM is a research field and method in computational linguistics that expands on sentiment analysis to identify argumentation in natural language text (Lawrence and Reed, 2020). AM can potentially extract more detail on reasoning from user opinions, assess the quality of argumentative statements in user opinions and map these to concepts from argumentation theory that are familiar and traceable to users (Lawrence and Reed, 2020). While there have been promising results in studies from researchers on using AM for the user opinion domain, the literature body is still comparatively underdeveloped. For example, previous AM studies used off-the-shelf models trained for formal debates (Passon et al., 2018) and later fine-tuned models (Chen et al., 2022) to predict the helpfulness of user opinions based on argumentative features in the text. The scope of the studies is relatively narrow. They focus on small subsets of user opinions by only including user opinions specific to certain subjects in their design, e.g., headphones (Chen et al., 2022) or a small set of restaurants, personal computers and hotels (Cattan et al., 2023), indicating a potential lack of general applicability. Finally, the studies do not elaborate on the use of their designs for a user-centered context, e.g., a tool.

To address these shortcomings and provide further understanding on embedding AM approaches, we propose a novel annotation scheme and compile a dataset of 1,016 online reviews with 7,266 sentences consisting of argument components, specificities and sentiments. For the product subset of 566 samples, we provide additional annotations

LREC-COLING 2024, pages 11548–11559

on aspects to relate argumentative statements to common topics in product user opinions, e.g., the delivery process. We evaluate the dataset in a rigorous annotation study, train state-of-the-art models for annotation predictions and evaluate the model inferences on unseen data in a prototypical application with eighteen users.

The annotation scheme was based on the theory of argumentation by (Toulmin, 2003) and upon established argumentation annotation schemes from other domains (Stab and Gurevych, 2014b, 2017; Wambsganss and Niklaus, 2022) as well as user opinions (Wachsmuth et al., 2014b; Chen et al., 2022). The schemes often contain complementary classes on sentiment (Wachsmuth et al., 2014b; Duan et al., 2019), specificity (Lugini and Litman, 2019; Durmus et al., 2019; Wambsganss and Niklaus, 2022) and aspect (Trautmann, 2020). By integrating these additional classes, we aim to integrate insights from complementary concepts used in computational modelling to aggregate and identify relevant user opinions from text. Our scheme and dataset cover multiple user opinion subjects that include a wide variety of products, hotels, restaurants, local services as well as mobile apps.

We trained multiple transformer-based classifiers for each of the classes with competitive metrics compared to similar recent studies that model argumentation (Wambsganss and Niklaus, 2022; Weber et al., 2023). We then embedded the classifiers into a prototypical tool for comparing e-commerce products by aggregating user opinions. The tool generates opinion summaries alongside product scores that are easier to parse and compare. We evaluated the tool in an online experiment with 18 potential users to collect user feedback for further development of our annotation approach and tool.

Our work contributes to research by the following. First, we expand the use of AM and argumentation theory in the user opinion domain by creating a novel annotation scheme as well as a freely available dataset[1] and models that are adapted to general user opinions instead of a smaller subsample. Our rigorous annotation study with two independent annotators show that the annotation guidelines lead to satisfying agreements, e.g., for the primary argument component class a Cohen's κ (Cohen, 2013) of 0.739. Second, we provide additional knowledge on classifying aspects within the AM context which was mostly not regarded in contemporary AM designs, but relevant for user opinions. Third, going back to the practical problem we aim to address, we present AM as a linguistic theory-based technique to identify and assess qualitative statements. This approach can be used to enhance existing designs, e.g., when leveraging opinion summarization. Finally, we expand the scope of existing studies by providing a first iteration and early user feedback for an user opinion aggregation tool that is based on our dataset and models. With our work we hope to encourage future research on the use of AM within the user opinion domain as well as on using it as a complementary method for practical use cases.

[1] huggingface.co/mydatasets

2. Related Work

2.1. User Opinion Quality and Argumentation

As user opinions or online reviews on a given subject often occur in significant quantity, a key challenge commonly formulated in the study of aggregating user opinions is the differentiation of (less) relevant user statements and documents (Rietsche et al., 2019; Bražinskas et al., 2021; Chen et al., 2022). According to literature on user opinions, argumentation is a key indicator that influences how they are perceived by users with respect to trustworthiness and persuasiveness, among others (Cheung et al., 2012; Teng et al., 2014). Furthermore, argumentative features have been shown to be an effective predictor of the usefulness of user opinions (Liu et al., 2017; Passon et al., 2018; Chen et al., 2022). As outlined, AM expands on earlier sentiment analysis approaches to explore such reasoning in text for potential use in identifying relevant user statements (Lawrence and Reed, 2020). AM tasks include the identification of argument components and their relations as well as argument quality estimation (Lippi and Torroni, 2016). Since the argumentative structures found in texts are usually domain-specific, AM pipelines commonly necessitate the availability of custom annotated datasets and models (Lawrence and Reed, 2020). Research on cross-domain applicability still remains a challenge and is ongoing (e.g., Fromm et al., 2023).

2.2. Modelling Argumentation in User Opinions

Annotation schemes and datasets for modelling argumentation in texts have been created for a wide variety of use cases such as adaptive argumentation learning (Wambsganss and Niklaus, 2022), detecting argumentative quality (Joshi et al., 2023; Fromm et al., 2023) or for analyzing persuasive essays (Stab and Gurevych, 2014b, 2017). Habernal and Gurevych (2017) model argumentation in user-generated web discourse on political topics and note that the formality of reasoning expressed by users in user-generated content is different compared to other settings. Due to this lack of generalization, adjustments to an annotation scheme for the user opinion domain may be necessary to ensure sufficient annotation reliability. Wachsmuth

et al. (2014a,b) have modelled argumentation in user opinions on hotels with two components: objective facts and subjective opinions, with the latter being positive or negative. Duan et al. (2019) expand on this argumentation structure in hotel user opinions and model major claim, claim, premise, background and recommendation classes as well as their support and attack relations. However, the dataset is relatively small with 85 documents and is filtered by excluding more difficult annotation samples with low annotator agreements. A common use case for AM in an user opinion context is helpfulness prediction (Liu et al., 2017; Passon et al., 2018; Chen et al., 2022). Liu et al. (2017) use a similar modelling approach on hotel user opinions compared to Duan et al. (2019) but omit support and attack relations. Passon et al. (2018) have used an off-the-shelf argumentation model trained on Wikipedia articles for three selected product categories that feature more formal argumentation. Chen et al. (2022) create a dataset of user opinions from Amazon that incorporates rich argumentative information with its own elaborate argumentative structure adapted to user opinions on headphones. The authors then use the resulting model to predict review helpfulness. Cattan et al. (2023) model a hierarchy of key points based on their relation type, such as supporting statements, to structure and quantify the prevalence of key points in a set of user opinion documents. As outlined, the focus on small subsets of user opinions in existing studies highlight the ongoing need to study the modelling of argumentation to a more general user opinion setting. Beyond modelling argumentative components, one potentially promising AM task is argument quality estimation (Fromm et al., 2023; Joshi et al., 2023). This task differs from our aim to model argumentation, but presents opportunities for further study by adapting it to the user opinion domain. Finally, some of the existing approaches neglect additional context such as sentiments or aspects which are otherwise established, especially in complementary non-AM analysis (e.g., Hosking et al., 2023; Wachsmuth et al., 2014b; Trautmann, 2020). Therefore, we aim to model argumentation for a more general set of user opinion categories with additional sentiment and aspect context. Further, the studies do not evaluate their model outputs with user-facing tools on real world usage for which we want to provide additional implementation knowledge for further research.

3. Dataset Construction

3.1. Data Source

The dataset covers 1016 online reviews from a set of five domains: e-commerce products, local services, hotels, restaurants and mobile apps. We combine these domains because they have been of most interest, as indicated by the AM studies that analyzed them individually. To expand this representative sample, we also included opinions on apps which have strongly been featured in non-AM studies (e.g., Maalej et al., 2016). The 566 reviews for e-commerce products were taken from the Amazon review dataset (Ni et al., 2019). 19 product categories of the dataset were selected for our sample. Product categories on human-produced media such as books were omitted, since the review style and structure was found to be significantly more complex compared to the general perspective which we wanted to address. To balance the amount of low-effort reviews we set the minimum length of the sampled reviews to 200 characters, as done in similar studies (e.g., Bražinskas et al., 2021). We sampled 30 reviews for each category and checked every review for comprehensibility and coherence. In total, four incoherent reviews were identified and removed. Beyond product reviews, we also chose to sample other review types to increase the applicability of the annotation scheme, the dataset and its models within the user opinion domain. As such, 150 reviews were taken from the Yelp dataset (Yelp) to cover the local services, hotel and restaurant domains, with 50 reviews each. To sample the reviews, we added a bias to include more helpful opinions that were upvoted by other users. According to literature on user opinions (Liu et al., 2017; Passon et al., 2018; Chen et al., 2022), helpful reviews contain better argumentation which we wanted to reflect in the dataset as well. Without the bias, the sample distribution leans heavily towards unvoted reviews. The 300 reviews for mobile apps were scraped from the Google Play Store based on recency. Unlike the previous subsamples, these reviews were constrained to six brokerage apps to enable researchers to test use cases and metrics that are specific to a comparable set of opinion subjects. Otherwise, the reviews were randomly sampled to cover a broad range of items reviewed on the respective platform.

3.2. Annotation Scheme

We created the annotation scheme and dataset with two objectives in mind. First, we aimed to identify argumentative statements that provide a base to aggregate and quantify relevant statements from user opinions. Second, we wanted to provide more knowledge on analyzing the level of argumentation in user opinions for a broad set of domains. Building upon the large body of related work, the annotation scheme was based on the theory of argumentation by (Toulmin, 2003) which is established in AM studies to model argumentation (Lawrence and Reed, 2020). Further, we based the scheme on knowl-

edge from existing annotation schemes from other domains (Stab and Gurevych, 2014b, 2017; Wambsganss et al., 2020) as well as user opinions (e.g., Wachsmuth et al., 2014b; Chen et al., 2022). The studies also contain complementary classes on sentiment (Wachsmuth et al., 2014b; Duan et al., 2019), specificity (Lugini and Litman, 2019; Durmus et al., 2019; Wambsganss and Niklaus, 2022) and aspect (Trautmann, 2020) to provide additional argumentation context which is studied in non-AM studies as well. Therefore, we created four annotation classes for a sentence classification task: argument component, argument specificity, argument sentiment and argument aspect. As highlighted earlier, argumentation is a key differentiator for the perceived quality and helpfulness of a review (e.g., Cheung et al., 2012; Liu et al., 2017) which we aim to model with the the argument component class. Argument specificity is derived from sentence specificity and describes the amount of detail provided in a sentence and is a more general differentiator to the quality of text (Ko et al., 2019). For example, user opinions contain generic statements such as "I love this product" which we wanted to identify to rank these statements lower in potential later designs. The sentiment class is used to differentiate the polarity of the user statements. Finally, the aspect class references the topic that the argument references. Examples for the annotation classes are visualized in Table 1 and 2.

Argument component As outlined, we based our annotation scheme on the use of argumentation in user opinions on the Toulmin model Toulmin (2003) and similar studies that model argumentation successfully for the user opinion (Wachsmuth et al., 2014a; Duan et al., 2019) and other domains (Stab and Gurevych, 2014b, 2017; Duan et al., 2019). These studies propose that an argument consists of multiple components: claim and premise being the most fundamental and common. They also address relations between the individual argument components (Duan et al., 2019; Chen et al., 2022). As previously suggested, the role of argumentation in user opinions is less formal (Habernal and Gurevych, 2017), hence researchers use simpler argumentation structures to annotate argumentative components (Wachsmuth et al., 2014a; Liu et al., 2017) or only operate on small subsets of a larger sample (Cattan et al., 2023; Chen et al., 2022). We found that the randomly sampled reviews for our more general perspective share this sentiment and usually don't use elaborate argumentative backing. The documents often consist of sequences of statements (Wachsmuth et al., 2015) and seldom feature clear support or oppose relations between argument components which may result in unsatisfying annotator agreements due to blurry annotation boundaries. Potential reasons include the personal motivations users have that do not result in an incentive to persuade the reader in favor or in opposition to the product (Yoo and Gretzel, 2008). We therefore simplified our argument component class to classify claim, premise and background statements. The claim class contains any statements that reference an assessment or claim about the reviewed item. A premise was defined as (potential) reasoning for a claim. Premises thus contain less subjective statements and provide greater insight compared to claims. The background class refers to all other statements which are less relevant to the stated annotation objective by neither providing reasoning or claims on the reviewed item.

Argument specificity is a concept we adapt from multiple related AM studies (Lugini and Litman, 2019; Durmus et al., 2019) to assess the persuasiveness and quality of argumentative components (Carlile et al., 2018; Wambsganss et al., 2020). Specificity is used to differentiate between generic and more thoughtful argumentative statements (Ko et al., 2019). Since the user opinion domain features more informal argumentation, which results in less clear annotation boundaries, we initially derived the two binary annotation classes "general" and "specific". We define them as follows: general statements are generic and provide marginal insight for potential readers, while specific statements should include more nuanced insights. To enable a clear differentiation when annotating, we defined the minimum criteria a statement had to pass to be annotated as specific: they had to reference an aspect, instead of the item generally, as well as provide a descriptive adjective for the aspect. For example, the statement "the yoga mat is great" references the item generally and uses a non-descriptive adjective. Finally, in agreement with literature on user opinions (Wachsmuth et al., 2014b; Rietsche et al., 2019), we noticed that a large part of the statements refer to individual experiences and subjective usage of the reviewed item, an example being "Ive charged my Note4 5 times on this already and it still has room for more". These statements don't provide the same level of insight as more descriptive and factual statements about the product, but they allude to potential usage and can still be relevant indicators when a large amount of reviewers mentions similar experiences. As a consequence, we split the specific class and added the experience class for annotation.

Argument sentiment relates to sentiment analysis and is present in multiple AM studies from which our scheme derives as well. The concept is relatively simple and usually differentiates positive and negative sentiment in argumentative components (Wachsmuth et al., 2014b; Duan et al., 2019). The concept is also used successfully in many other

Example	Component	Specificity	Sentiment
I am absolutely thrilled with the dividers.	Claim	General	Positive
Damaged in 2 places when arrived in 3 boxes, [...].	Premise	Experience	Negative
I purchased 4 of these to use [...] in our kitchen.	Background	General	None
This smoke detector uses 2 AA batteries instead of a 9V.	Premise	Specific	Positive
Cute but tiny!	Claim	General	Balanced

Table 1: Shortened example annotations for argument component, specificity and sentiment

sentiment-based designs for extracting information from general text and user opinions (Nazir et al., 2020; Do et al., 2019). Hence, our initial goal was to differentiate statements that raise positive, neutral or negative points about the reviewed item. By annotating this class, argumentative components can be filtered and presented as positive or negative in a potential tool to the user. Off-the-shelf sentiment classifiers are ubiquitous. While testing these classifiers, we noticed inaccuracies, including: some statements reference both positive and negative points or they reference competitors and do not actually apply to the reviewed item. To address the former, we added an additional "balanced" sentiment to the traditional set of "positive", "neutral" and "negative" sentiments. For the latter issue, the annotation guideline specifies that the classified sentiment should refer to the opinion subject. Hence if a statement concerns a competing product and is either positive or negative, the opposite label was to be used.

Argument aspect Extracting aspects is a common part of pipelines for extracting information from user opinions in both traditional (Nazir et al., 2020; Do et al., 2019) and recent literature (Angelidis et al., 2021; Isonuma et al., 2021; Hosking et al., 2023). Studies on aspect extraction within the AM context are more recent and operate within the debate setting (Reimers et al., 2019; Trautmann, 2020; Ruckdeschel and Wiedemann, 2022) as well as for user opinions (Dragoni et al., 2018). To build on this work and differentiate what topics are discussed in the identified argumentative statements, we opted to add an aspect class to our annotation scheme. We did not incorporate existing approaches directly because they either did not generalize beyond the debate setting or, more importantly, require the statements to mention the aspect explicitly in the statement (Dragoni et al., 2018). For our sampled user opinions we found that users often do not directly mention the aspect, e.g., "For the time being, these get the job done". Further, explicit approaches analyze each text individually, but do not include the implicit domain knowledge that we wanted to embed for potential applications of our dataset. For annotating the argument aspect, we created and iteratively refined a set of aspects for the product review subsample. This subsample was chosen because potential tool users we interviewed predominantly used user opinions for comparing products on e-commerce websites. We defined an aspect as a generic topic that is mentioned in argumentative statements for a wide array of product categories. We differentiate the following argument aspects: delivery, function, general (sentiment), installation, pricing, (customer) service, style, usage, and none. A brief description and example is visualized in Table 2. Unlike the other classes, a statement can be annotated with one or more aspect labels as some user statements contain multiple aspects. The aspects were modelled for use in the tool evaluation and testing its general viability. Hence, they do not generalize beyond the e-commerce product context.

3.3. Annotation Process

The annotation process was split into two parts. The annotation of the main 746 online reviews was performed by two annotators. The annotators were selected based on their extensive experience on interacting with user opinions on e-commerce platforms for a wide variety of products, as evident in the dataset. Furthermore, familiarity with basic argumentation theory was developed and evaluated through additional workshops and trainings. The annotations were performed independently to analyze inter-rater agreements, refine the annotation guideline as well as discussing edge cases. Before annotation, all reviews were sentence split with the Natural Language Toolkit (NLTK) python library. All annotation classes were defined as multi-class problems, with the exception of argument aspect as multi-label. As exclaimed, the annotation for aspects was specific to the 566 samples of the Amazon online review dataset. The guideline described definitions, rules and examples for the annotation for each class. It was iteratively refined and discussed through training sessions and validated by two independent senior researchers concerning the criteria of robustness, conciseness, extensibility and comprehensibility. Finally, all remaining disagreements were analyzed and resolved through a final workshop with both annotators. In case of conflict, a third senior researcher was consulted. To expand the dataset for more product categories, one annotator annotated the 270 remaining online reviews. This extension part of the dataset was

Aspect	Description	Example
Delivery	Delivery process	Damaged in 2 places when arrived [...].
Function	Quality and functionality	The flashlight has a range of 100m.
General	General sentiments	I love this product.
Installation	Ease of first time usage	I installed this in 20 min.
None	Unrelated statements	I bought this for my wife.
Price	Pricing assessment	The toy is expensive for what it does.
Service	Customer service experience	I had to pay the return fee myself.
Style	Appearance and design	The blue color is beautiful.
Usage	Fun and usage for user purposes	I played the game for 2 hours with my son.

Table 2: Descriptions and examples for the aspect class

thus not subject to the aforementioned supervision and correction and is subsequently of lower quality. All sentences were annotated with the *Argilla* annotation tool in a randomized order. The classes were labelled in this order: argument component, specificity, sentiment and aspect.

4. Dataset Analysis

4.1. Inter-annotator Agreement

To evaluate the annotation agreement with respect to the chosen classes as well as assess the adaption of our guideline, we used two inter-annotator agreement measures: Cohen's κ (Cohen, 2013) and Krippendorff's α (Krippendorff, 2004). As shown in Table 3, substantial Cohen's κ agreements have been obtained for all classes according to the scale provided by Landis and Koch (1977). The agreement scores according to Krippendorff's metric are moderate for the argument component and specificity class which indicates potentials for further improvement. The scores are competitive compared to other AM datasets which also present challenges with annotator agreement for argumentative components (e.g., Wambsganss and Niklaus, 2022; Weber et al., 2023). Notably, the argumentation review dataset on hotel user opinions by Duan et al. (2019) presents better annotation agreements, but filters documents with a lower agreement score than 0.5, excluding a large majority of all documents, and is thus not comparable. We concluded that our annotation guidelines for argument component, specificity, sentiment and aspect class lead human annotators to a satisfying agreement for the domain of general user opinions.

Class	Cohen's κ	Krippendorff's α
Component	0.739	0.572
Specificity	0.681	0.478
Sentiment	0.853	0.739
Aspect	0.797	0.646

Table 3: Inter-annotator agreement scores for the argument component, specificity, sentiment and aspect classes

4.2. Dataset Statistics

As described in the previous section on the annotation process, the dataset consists of two parts. The main part of the corpus consists of 5,166 sentences with 99,915 tokens that were extracted from 746 user opinions. On average, each review has 6.92 sentences and 133.93 tokens. The extension part of the corpus consists of 2,100 sentences with 43,771 tokens that were extracted from 270 online reviews. On average, each review has 7.78 sentences and 162.11 tokens. Counts for each of the classes can be found in Tables 5 and 6.

5. Models

Dataset preparation and modelling After reaching a satisfying agreement in our annotation study and annotating the full dataset, our next aim was to predict the annotated labels. For this purpose, we set up a multi-class classification task for each of the classes to classify user opinion sentences. For the aspect class, we've mapped all samples with multiple associated classes were consolidated into a category named *multi*. In our experiments the multi-label setup lead to worse results due to the relative low occurrence of sentences with multiple aspects in the samples. The main part of the dataset was chosen for fine-tuning since it features higher quality annotations, as exclaimed. We evaluated the pretrained models of the *transformers* Python library and chose a RoBERTa-based model (*xlm-roberta-base*) for fine-tuning on the dataset as it has been used when evaluating recent AM datasets (e.g., Wambsganss and Niklaus, 2022; Weber et al., 2023) and can be used for comparison to evaluations of other datasets. The tokenization was automatically handled by the associated tokenizer by the *transformers* library. We used the larger DeBERTa (*deberta-v3-large*) and sentence transformer models for additional improvement of the metrics (*mpnet-paraphrase-mpnet-base-v2*). A

Class	F1	Accuracy	Precision	Recall
		RoBERTa		
Component	0.764	0.766	0.764	0.766
Sentiment	0.785	0.787	0.784	0.787
Specificity	0.729	0.728	0.733	0.728
Aspect	0.627	0.683	0.609	0.683
		DeBERTa		
Component	0.803	0.804	0.803	0.804
Sentiment	0.852	0.853	0.851	0.853
Specificity	0.795	0.793	0.801	0.793
Aspect	0.705	0.721	0.720	0.721
		MPNet sentence transformer		
Component	0.789	0.789	0.793	0.789
Sentiment	0.825	0.827	0.826	0.827
Specificity	0.756	0.754	0.761	0.754
Aspect	0.768	0.765	0.774	0.765

Table 4: Metrics for the component, sentiment, specificity, and aspect classes by model architecture

learning rate of $2e^{-5}$ was chosen for the fine-tuning process. For evaluation and validation purposes, we partitioned the dataset into an 80-20 split, where 80% was utilized for training and 20% was kept as test data. All models were fine-tuned for three epochs. Links to the DeBERTa models for testing inputs on the *Huggingface* platform are available on the dataset page.

Table 4 shows the performance metrics we evaluated for the three model architectures. We opted for a weighted average for evaluating these metrics to take into account the uneven class distribution. Compared to the smaller RoBERTa model, the fine-tuned DeBERTa and the MPNet sentence transformer model provided superior results. However, this advantage came with a higher computational cost during training and diminished efficiency during inference. Notably, the sentence transformer model yielded improvements for the aspect class, possibly due to less obvious classification boundaries within the text that the MPNet architecture was able to capture better. The metrics are competitive or better compared to similar recent AM dataset evaluations (e.g., Wambsganss and Niklaus, 2022; Weber et al., 2023). For the argument sentiment and aspect classes we observe lower performance compared to traditional studies. This was expected since we added additional labels and encapsulate additional domain knowledge specific to user opinions.

Application of models To better illustrate the practical applications of our dataset, we developed a dashboard that consolidates product data from Amazon reviews. The design of this platform was informed by insights drawn from eleven interviews with individuals who use user opinions for comparing products online. For the user interface we leveraged common design patterns from literature on dashboards (Bach et al., 2022). Our principal objective in developing this system was to offer a comprehensive and balanced overview of e-commerce products based on user opinions while enabling time-efficient comparison of the reviewed products through ranking and scoring the products. To test the efficacy and functionality of our system, we selected yoga mats as a representative product class for the prototype because it is neutral and relatable to a large part of potential users. Further, user opinions for this product class are available in large quantity in the underlying Amazon online review dataset (Ni et al., 2019). A visualization of the system is provided in Figure 1.

Beyond providing openly accessible information such as product rating and description, the system has two primary functions. First, the system provides one general and six aspect-specific scores that rank the analyzed products and provide insight into a product's relative strengths and weaknesses. Second, the system displays text summaries generated with the *ChatGPT-3.5-turbo* model via the OpenAI API to present more detailed insight, if needed by the user. The tool uses the argument component and specificity predictions to filter the user statements for use in the scoring calculation as well as for the generated summaries. The rationale for the scoring is that more general claims should be weighted less than more substantial statements. It derives from the literature on user opinion helpfulness, argumentation theory and specificity as presented in earlier sections. For the summaries, the filtering was used to reduce to fill the available context size with more relevant text as well as to decrease usage cost and generation time of the used API. The argument aspect predictions are used to filter user statements for the specified aspects and create six aspect-specific scores. Notably, the dataset provides nine aspect classes. The *general* and *none* aspects were omitted because they

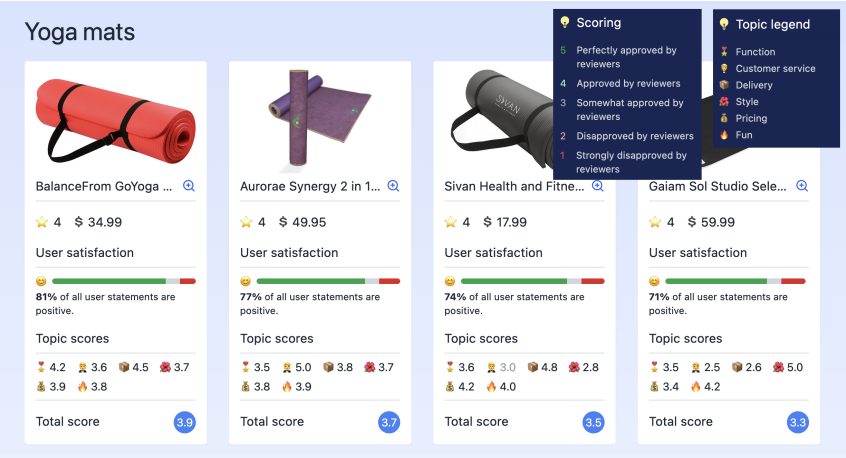

Figure 1: Overview and ranking of the products in the system

provided no insights specific to an aspect. The *installation* aspect was omitted because no inferences were made in this exemplary case, possibly because yoga mats are ready for use immediately. Finally, the argument sentiment class backs the polarity of the scoring algorithm. The calculation sets a score of three as a baseline for an average product. The percentage share of positive statements is then normalized to a range of -3 to 0 for products with a positive share of less than 65 and a range of 0 to 2 for products with a share of more than 70. This approach was chosen to increase the impact of negative reviews. A naive approach resulted in highly uniform scores due to a significant data bias towards positive sentiments. Finally, we used a "relative impact" multiplier to reduce the impact of small, imbalanced samples. For an extreme example, a product that has only five negative statements on delivery in a sample of 700 reviews should not be subjected to a score that would otherwise be -3. The final value is then added to the baseline score. While the scoring has been iteratively refined as part of our study, it has not been a focus of the study. Hence, more tests and improvements are needed to reduce biases specific to user opinions.

User evaluation To evaluate our tool and collect further change requests, we conducted a task-based online experiment on Prolific with 18 random participants. We chose Prolific due to its response quality and large variety of potential samples (Peer et al., 2017). The sampling was restricted to native English speakers. Ten participants identified as female, eight as male. Their mean age was 39.16. In the task scenario, the participants were asked to recommend a yoga mat to a hypothetical wellness online community and provide balanced reasoning based on the (optional) information provided in the dashboard. This scenario was chosen to replicate the role of website creators that aggregate product information from user opinions. Finally, the participants were asked to describe their experience with the tool, task scenario and potential improvements. The response has been generally positive, praising the additional data that is provided. Yet, some participants voiced concerns about the complexity of the user interface. Further, they expressed lower intents of usage by noting that rating, pricing and product description are enough to guide their recommendation decisions. Potential reasons include that more detailed information is not as useful to the highly general sample that was randomly selected for the online experiment, unlike to the one we interviewed previously. Taking this feedback into account, we simplified the user interface and hide more complex information by default. For the next iteration of the experiment, we are making adjustments to the participant sample and task scenario to reflect a user group that has more interest in detailed information about the subject.

6. Conclusion

With our work, we propose a novel annotation scheme for modelling argumentation in diverse user opinion domains, in contrast to previous argument mining (AM) studies which focused on a single domain or smaller subsamples. Based on the scheme, we introduce a dataset that contains 1,016 online reviews with 7,266 sentences annotated for argumentative components, argument specificity, sentiment and aspect. With these classes, we adapt and integrate research on specificity, aspect and sentiment from related studies to extract additional context. The annotation study shows that annotating is reliably possible. Additionally, the performance metrics of the models that were trained on the dataset are competitive with comparable AM settings for other domains. Finally, we embed the models into a first iteration of an user-facing tool that aggregates user opinion texts and scores that

are easy to parse and compare for multiple products. With the study, we want to encourage further research on two research directions. First, the use of AM for the general user opinion context requires more development, especially when used for user-facing tools. Second, we want to showcase AM as a potential complementary technique to supplement designs that aggregate information from user opinions with additional argumentation-derived metrics, specifically concerning opinion summarizations.

7. Limitations

Our work provides several areas to iterate upon with further research. First, some classes only featured moderate aggrements due to annotation conflicts (see Table 3). To improve annotation quality, all disagreements were discussed and resolved together, if needed with a third senior researcher. Notably, the dataset consists of a main and extension part (26.57%). The latter was only annotated by one annotator and hence not subject to the same level of quality control as the main part. The modeling of argumentation remains a challenge that is also present in similar studies that feature comparable or worse scores on annotation aggreement (Wambsganss and Niklaus, 2022; Duan et al., 2019; Weber et al., 2023; Park and Cardie, 2018). Subsequently, the same also applies for the evaluation metrics for our models (see Table 4) which provide further base for improvement. More generally, comparing metrics to other studies in this domain is challenging because the annotation schemes vary highly in complexity (e.g., Wachsmuth et al., 2014a; Stab and Gurevych, 2014a; Chen et al., 2022) or no models have been evaluated (Duan et al., 2019). Still, our models feature comparable or better evaluation metrics with respect to similar approaches for other domains, e.g., persuasive writing (Wambsganss et al., 2020; Weber et al., 2023). With respect to the annotation scheme, the argument component, specificity and sentiment classes are generally applicable to user opinions, but aspects are tailored to e-commerce product user opinions. We chose this tradeoff to incorporate implicit domain knowledge for our practical use case. Beyond the use case, the categories may not be granular enough and practitioners may need to adapt the aspect categories to their own setting. The user evaluation only presents a first iteration to guide further development of a tool that uses AM and opinion summarization for aggregating user opinions. The responses highlight challenges to use AM for the user opinion context which need to be addressed with further experiments that use larger sample sizes and incorporate rigorous, theory-based constructs to represent valid and comparable results. Finally, this work is built on AM and comparable studies that modeled argumentation with a supervised way, featuring potential disadvantages that unsupervised approaches from other research streams on user opinion not have.

8. Bibliographical References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293.

Benjamin Bach, Euan Freeman, Alfie Abdul-Rahman, Cagatay Turkay, Saiful Khan, Yulei Fan, and Min Chen. 2022. Dashboard design patterns. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):342–352.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. *arXiv preprint arXiv:2109.04325*.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. *arXiv preprint arXiv:2306.03853*.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922.

Mingming Cheng and Xin Jin. 2019. What do airbnb users care about? an analysis of online review comments. *International Journal of Hospitality Management*, 76:58–70.

Cindy Man-Yee Cheung, Choon-Ling Sia, and Kevin KY Kuan. 2012. Is this review believable? a study of factors affecting the credibility of online consumer reviews from an elm perspective. *Journal of the Association for Information Systems*, 13(8):2.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Hai Ha Do, Penatiyana WC Prasad, Angelika Maag, and Abeer Alsadoon. 2019. Deep learning for aspect-based sentiment analysis: a comparative review. *Expert systems with applications*, 118:272–299.

Mauro Dragoni, Celia da Costa Pereira, Andrea GB Tettamanzi, and Serena Villata. 2018. Combining argumentation and aspect-based opinion mining: the smack system. *AI Communications*, 31(1):75–95.

Xueyu Duan, Mingxue Liao, Xinwei Zhao, Wenda Wu, and Pin Lv. 2019. A hotel review corpus for argument mining. In *Cognitive Systems and Signal Processing: 4th International Conference, ICCSIP 2018, Beijing, China, November 29-December 1, 2018, Revised Selected Papers, Part I 4*, pages 327–336. Springer.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining relative argument specificity and stance for complex argumentative structures. *arXiv preprint arXiv:1906.11313*.

Marcio Fonseca, Yftah Ziser, and Shay B Cohen. 2022. Factorizing content and budget decisions in abstractive summarization of long documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6341–6364.

Michael Fromm, Max Berrendorf, Evgeniy Faerman, and Thomas Seidl. 2023. Cross-domain argument quality estimation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13435–13448.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. *arXiv preprint arXiv:2305.11603*.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, 9:945–961.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. Arganalysis35k: A large-scale dataset for argument quality analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Domain agnostic real-valued specificity prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6610–6617. Issue: 01.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. 38:787–800. Publisher: Springer.

Kee-Young Kwahk and Byoungsoo Kim. 2017. Effects of social media on consumers' purchase decisions: evidence from taobao. *Service Business*, 11:803–829.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Yukyung Lee, Jaehee Kim, Doyoon Kim, Yookyung Kho, Younsun Kim, and Pilsung Kang. 2023. Painsight: An extendable opinion mining framework for detecting pain points based on online customer reviews. *arXiv preprint arXiv:2306.02043*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using argument-based features to predict and analyse review helpfulness. *arXiv preprint arXiv:1707.07279*.

Luca Lugini and Diane Litman. 2019. Argument component classification for classroom discussions. *arXiv preprint arXiv:1909.03022*.

Walid Maalej, Zijad Kurtanović, Hadeer Nabil, and Christoph Stanik. 2016. On the automatic classification of app reviews. *Requirements Engineering*, 21:311–331.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

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)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of amazon reviews using off-the-shelf argumentation mining. *arXiv preprint arXiv:1809.08145*.

Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of experimental social psychology*, 70:153–163.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

Roman Rietsche, Daniel Frei, Emanuel Stoeckli, and Matthias Söllner. 2019. Not all reviews are equal - a literature review on online review helpfulness.

Mattes Ruckdeschel and Gregor Wiedemann. 2022. Boundary detection and categorization of argument aspects via supervised learning. In *Proceedings of the 9th Workshop on Argument Mining*, pages 126–136.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Shasha Teng, Kok Wei Khong, Wei Wei Goh, and Alain Yee Loong Chong. 2014. Examining the antecedents of persuasive ewom messages in social media. *Online information review*, 38(6):746–768.

Stephen Toulmin. 2003. *The uses of argument*. Cambridge University Press. OCLC: 1198272922.

Dietrich Trautmann. 2020. Aspect-based argument mining. *arXiv preprint arXiv:2011.00633*.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow-a general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014a. Modeling review argumentation for robust sentiment analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 553–564.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014b. A review corpus for argumentation analysis. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 115–127. Springer.

Thiemo Wambsganss and Christina Niklaus. 2022. Modeling persuasive discourse to adaptively support students' argumentative writing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8748–8760.

Thiemo Wambsganss, Christina Niklaus, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. A corpus for argumentative writing support in german. *CoRR*, abs/2010.13674.

Florian Weber, Thiemo Wambsganss, Seyed Parsa Neshaei, and Matthias Soellner. 2023. Structured persuasive writing support in legal education: A model and tool for german legal case

solutions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2296–2313.

Pablo Weingart, Thiemo Wambsganss, and Matthias Soellner. 2023-05-11. A taxonomy for deriving business insights from user-generated content.

Bai Yang, Ying Liu, Yan Liang, and Min Tang. 2019. Exploiting user experience from online customer reviews for product design. *International Journal of Information Management*, 46:173–186.

Yelp. Yelp dataset.

Kyung Hyan Yoo and Ulrike Gretzel. 2008. What motivates consumers to write online travel reviews? *Information Technology & Tourism*, 10(4):283–295.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

A. Appendix

Class	Main count	Extension count
Claim	2,572	832
Premise	1,354	536
Background	1,240	732
General	2,495	1,142
Specific	1,274	583
Experience	1,397	375
Positive	1,797	749
Negative	1,868	548
Neutral	1,194	688
Balanced	307	115

Table 5: Basic class counts for the argument component, specificity and sentiment classes

Class	Main count	Extension count
Service	34	38
Delivery	42	45
Usage	212	242
Function	795	652
General	327	301
Installation	89	38
Price	89	72
Style	129	37
None	1068	732

Table 6: Basic counts for the aspect classes