Argument Mining for Review Helpfulness Prediction

Zaiqian Chen[1], **Daniel Verdi do Amarante**[2], **Jenna Donaldson**[2], **Yohan Jo**[3], **Joonsuk Park**[2,4]
[1]Columbia University, [2]University of Richmond, [3]Amazon, [4]NAVER AI Lab
zc2666@columbia.edu,
{daniel.verdidoamarante, jenna.donaldson}@richmond.edu,
jyoha@amazon.com, park@joonsuk.org

Abstract

The importance of reliably determining the helpfulness of product reviews is rising as both helpful and unhelpful reviews continue to accumulate on e-commerce websites. And argumentational features—such as the structure of arguments and the types of underlying elementary units—have shown to be promising indicators of product review helpfulness. However, their adoption has been limited due to the lack of sufficient resources and large-scale experiments investigating their utility. To this end, we present the AMazon Argument Mining (AM^2) corpus—a corpus of 878 Amazon reviews on headphones annotated according to a theoretical argumentation model designed to evaluate argument quality. Experiments show that employing argumentational features leads to statistically significant improvements over the state-of-the-art review helpfulness predictors under both text-only and text-and-image settings.[1]

1 Introduction

With the rapid growth of e-commerce, reading product reviews is increasingly becoming a part of online shopping. Going beyond the seller's description of the products, potential customers are considering the firsthand experiences and opinions of those who have already purchased the products. Fortunately, product reviews are quickly accumulating on popular e-commerce websites like *Amazon.com* on a daily basis; however, not all reviews are helpful, necessitating automatic prediction of their helpfulness (Ocampo Diaz and Ng, 2018; Qu et al., 2020).

A wide variety of features for helpfulness prediction has been proposed in the past, including those from the review (Diaz and Ng, 2018), the reviewer (Tang et al., 2013), and the product (Ghose and Ipeirotis, 2011). Among these, argumentational features (AFs) from the review text have shown potential in small-scale experiments (Liu et al., 2017; Passon et al., 2018).

To see how AFs can be useful for predicting the helpfulness of product reviews, consider two reviews on headphones crawled from Amazon.com. Figure 1 is a review that accrued a high number of helpfulness votes, annotated according to our scheme presented in Section 2. The main proposition that the product is a "good item for the price" has been supported by three propositions. One of those is further supported by a firsthand experience (testimony) of the reviewer. In this way, the propositions collectively form a coherent argument. In contrast, Figure 2, a review that had not received any helpfulness votes, has a much more sparse argumentative structure; it lacks support for major points made in the review, nor does it contain a testimony. As demonstrated, AFs can be good indicators of helpfulness.

However, the adoption of AFs has been limited in part by the unavailability of sufficient resources. An-

[Good item for the price.]$_a$ [Easy to charge.]$_b$ [Only one headband to deal with.]$_c$ [A "marriage saver"]$_d$ [because I don't have to listen to my wife's shows, especially when I go to bed before her.]$_e$ [My wife says "I like them".]$_f$

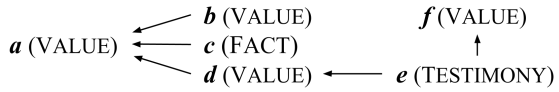


Figure 1: A Review with a High Helpfulness Vote Count. The propositions collectively form a coherent argument. Also, the review does not just contain opinions, but testimony and other objective information.

[Skull Candy's are the best earbuds in their price range that you could buy... until they break.]$_a$ [They include a one year warranty,]$_b$ [but often times, for such a low-priced product, its easier to just purchase a new one.]$_c$ [So definitely, purchase these earbuds,]$_d$ [but don't expect more than a few months out of them until one side burns out.]$_e$

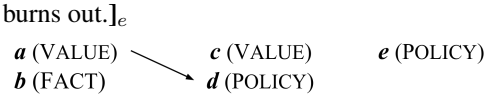


Figure 2: A Review with 0 Helpfulness Votes. Unlike the helpful review example (Figure 1), this review has a poor argumentative structure, e.g., it lacks support for the proposition that the headphones are the best in their price range. Also, it consists mostly of subjective propositions like VALUE and POLICY.

[1]This work is not related to Amazon.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8914 - 8922
December 7-11, 2022

notating argument information is challenging due in part to the prevalence of enthymemes, i.e., arguments with premises that have not been stated explicitly, and multiple plausible reconstructions of arguments (Stab and Gurevych, 2017). Still, a review helpfulness corpus with argument information is desirable.

To this end, we present the AMazon Argument Mining (AM^2) corpus, a corpus consisting of 878 Amazon product reviews on headphones annotated with rich argument information. Reviews were annotated according to a complex theoretical argumentation model with two types of support relations prevalent in practical argumentation—REASON and EVIDENCE—as well as five types of elementary units—POLICY, VALUE, FACT, TESTIMONY, and REFERENCE—capturing the appropriate types of support. For instance, VALUE propositions are best supported with REASON, and FACT propositions, EVIDENCE (Park et al., 2015). The annotated reviews have been carefully sampled to control for other factors that affect perceived helpfulness, such as the type of product and how long the given review had been published.

We also demonstrate the efficacy of the corpus by incorporating AFs into the state-of-the-art review helpfulness prediction models. The results demonstrate that the performance improvements are statistically significant for both unimodal (text-only (Dai et al., 2018)) and multi-modal (text-and-image (Liu et al., 2021)) baselines.

The main contributions of this work are twofold: First, we present AM^2—the first argument mining corpus of product reviews, supporting review helpfulness prediction with rich argument information.[2] It is based on a more complex theoretical argumentation model, yet is more than seven times the size of the only existing review helpfulness dataset with argument information (Liu et al., 2017). Second, we build new state-of-the-art review helpfulness predictors under text-only and text-and-image settings by incorporating AFs to existing ones, showcasing the utility of AFs.

2 Annotation Scheme

The annotation scheme is based on a theoretical argumentation model comprised of elementary units and support relations to capture a specific type of argument quality—*evaluability* (Park et al., 2015). In short, an argument is evaluable if every proposition in the argument is accompanied with at least one type of appropriate support, which in turn allows readers to understand the gist of the argument and evaluate its validity and strength. The degree to which an argument in practice meets this requirement determines how evaluable it is.

2.1 Elementary Units

The elementary units in this scheme consist of four types of propositions—FACT, TESTIMONY, POLICY, and VALUE—and REFERENCE:

-**Proposition of Non-Experiential Fact** (FACT): FACT is an objective proposition, meaning it does not leave any room for subjective interpretations or judgements. For example, "and battery life is about 8-10 hours."
-**Proposition of Experiential Fact** (TESTIMONY): TESTIMONY is also an objective proposition. However, it differs from FACT in that it is experiential, i.e., it describes a personal state or experience. For example, "I own Sennheisers, Bose, Ludacris Souls, Beats, etc."
-**Proposition of Policy** (POLICY): POLICY is a subjective proposition that insists on a specific course of action. For example, "They need to take this product off the market until the issue is resolved."
-**Proposition of Value** (VALUE): VALUE is a subjective proposition that is not POLICY. It is a personal opinion or expression of feeling. For example, "They just weren't appealing to me"
-**Reference to a Resource** (REFERENCE): REFERENCE is the only non-proposition elementary unit that refers to a resource containing objective evidence. In product reviews, REFERENCE is usually a URL to another product page, image or video. Also, REFERENCE cannot be supported by other elementary units. For example, "`https://images-na.ssl-images-amazon.com/[...]`"

2.2 Support Relations

Support relations in this scheme are two prevalent ways in which propositions are supported in practical argumentation: REASON and EVIDENCE. The former can support either objective or subjective propositions, whereas the latter can only support objective propositions. That is, you cannot prove that a subjective proposition is true with a piece of evidence:
-**Reason**: For an elementary unit X to be a REASON for a proposition Y, it must provide a reason or a justification for Y. For example, "The only issue I have is that the volume starts to degrade a little bit after about six months."(X) and "and I find I have to buy a new pair every year or so."(Y).
-**Evidence**: For an elementary unit X to be EVIDENCE for a proposition Y, it must prove that Y is true. For example, "`https://images-na.ssl-images-amazon.com/[...]`"(X) and "The product arrived damage[d],"(Y).

3 The AM^2 Corpus

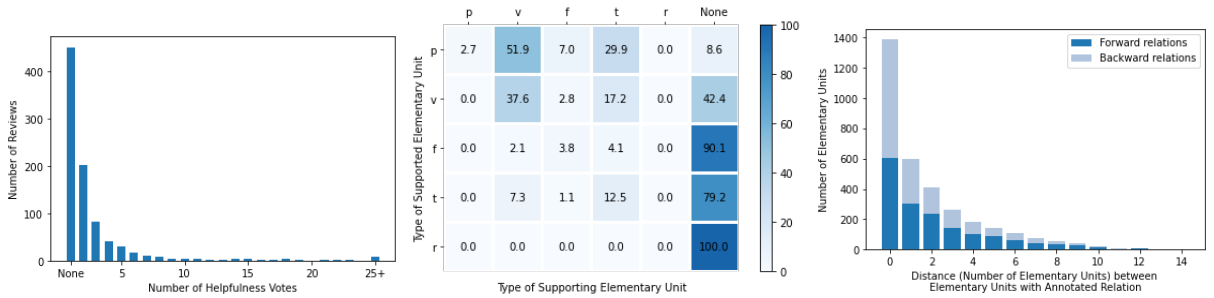
3.1 Preparation

To construct the AMazon Argument Mining (AM^2) corpus, we first sampled product reviews from the UCSD Amazon Review Data (Ni et al., 2019)—a dataset of 233.1M Amazon product reviews. To construct a corpus suitable for analyzing the effects of AFs on helpfulness prediction, the following criteria were imposed: First, the reviews are on headphones only. This is a product category where the variation in reviewed aspects, e.g. sound quality, level of comfort,

[2]The corpus is available at `www.joonsuk.org`.

POLICY	VALUE	FACT	TESTIMONY	REFERENCE	Elementary Unit	Reason	Evidence	Support Relation
123	3,751	280	1,964	3	6,126	3,278	26	3,304

Table 1: Number of Elementary Units and Support Relations in the AM^2 Corpus (878 reviews)



(a) (b) (c)

Figure 3: Statistics of the AM^2 Corpus. (a) Number of reviews by helpfulness vote count. The x-labels are "None, 2, 3, ..., 25+". (Reviews in the UCSD Amazon Data contain the "vote" key only when a review has received two or more votes.) (b) Percentage of types of supporting and supported elementary units. (c) Number of annotated relations between elementary units by distance (measured in # of elementary units in between). "Forward relations" (dark blue) refers to the supported elementary unit appearing *later* in the review than the supporting elementary unit. "Backward relations" (light blue) refers the supported elementary unit appearing *before*. The bars are stacked, meaning the total height represents the count of both cases, as opposed to backward relations consistently outnumbering forward ones by large margins.

stability of connection, etc., across the products is expected to be small. This would prevent spurious associations between product specific keywords and review helpfulness. Second, the reviews have been written within a selected span of 7 days[3]. This ensures that the reviews accrued helpfulness votes for about the same amount of time, off by 6 days at most. Lastly, the reviews consist of 2 to 10 sentences (inclusive) and 10 to 200 words (inclusive). This is to control for the differences in helpfulness resulting from the sheer amount of information.

Of the reviews automatically found to meet these criteria, 475 had accrued helpfulness votes. To match this number, we randomly down-sampled reviews without helpfulness votes from 5.5k to 475, resulting in a total of 950 reviews. The number was further reduced to 878 after a manual reinforcement of the criteria during the annotation process.

3.2 Annotation

The annotation project was completed in three stages—training, annotation and adjudication—using the OVA tool (Lawrence and Reed, 2014). In the training stage, four undergraduate students annotated practice reviews based on an annotation manual and met twice a week to get feedback and discuss issues. These students performed the roles of annotator and adjudicator in the following stages. They were new to argumentation theory, but fluent in English.

In the annotation stage, each review was annotated by two of the four annotators independently. A given review was first split into sentences, each of which was split further if it contains multiple independent clauses or a dependent clause in an argumentative relation, e.g. a "because"-clause. Then, each span other than non-argumentative ones, like greetings and onomatopoeias, were annotated with an elementary unit type. Lastly, support relations between the elementary units were added. The inter-annotator agreement (IAA) between annotators was measured using Krippendorf's α (Krippendorff, 1980): 0.66 for elementary unit types and 0.47 for support relations[4]. Similar to existing argument mining corpora (Park and Cardie, 2018; Egawa et al., 2019, 2020), the level of agreement tends to be moderate due to the difficult and subjective nature of determining underlying argumentative structures.

In the adjudication stage, the annotations for each review was revisited by an independent adjudicator who chose what to include in the final annotation.

3.3 Resulting Corpus

Our final corpus consists of 878 reviews, 6,126 elementary units and 3,304 support relations as summarized in Table 1. We designate a random split of the corpus into a training set (614 reviews) and a test set (264 reviews) to support intrinsic evaluation of argument mining systems in future work.

The corpus exhibits several interesting characteristics. First, about half of the reviews have received fewer than two helpfulness votes, with the count halved as the vote count increases (Figure 3(a)).

Also, there are noticeable patterns concerning which types of elementary units are supported by which types

[3]The actual span chosen was 2/27/2016 to 3/4/2016, which is a 7-day span surrounding 3/1/2016, the day when the most number of reviews were written for headphones.

[4]The IAA was measured by treating IDs of supporting elementary units as labels for the supported elementary unit.

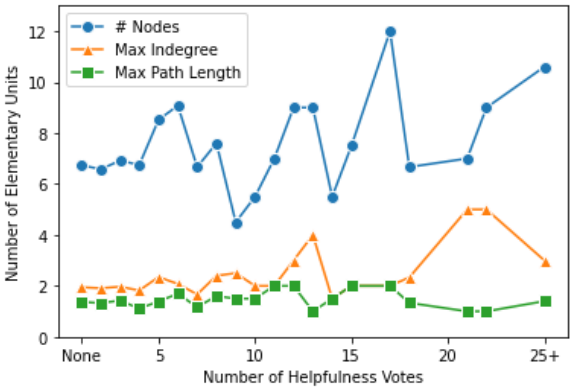


Figure 4: Comparison of Reviews by Helpfulness Votes. The graph plots the number of elementary units averaged across reviews in the training set with the same helpfulness vote count. (Note, graph theory terminology is used in the legend for brevity—consider arguments as graphs as shown in Figure 1.) Reviews with more helpfulness votes tend to be longer (# Nodes) and provide more support for the main proposition (Max Indegree). However, the longest chains of reasoning are of similar lengths as those in reviews with fewer helpfulness votes (Max Path Length).

of elementary units (Figure 3(b)): Most of FACT and TESTIMONY have not been supported, meaning reviewers neglected to provide reason or evidence for objective propositions. Regarding the two types of subjective propositions, only 8.6% of POLICY has not been supported, unlike VALUE, 42.4% of which has not been supported. This significant difference may be from POLICY being a stronger form of opinion. That is, reviewers may have felt more obligated to provide support when they tell others to do something, rather than simply state how they feel.

Lastly, Figure 3(c) shows that about 40% of support relations are between two adjacent elementary units (x=0). Also, there is an interesting difference in support relations between adjacent elementary units (x=0) and those between elementary units that are farther apart (x≥2): For the former, there are more backward relations—the supported elementary unit appears *before* the supporting elementary unit; for the latter, the opposite is the case. One explanation for this phenomenon is that there are common sentence structures in which a reason follows immediately after a proposition, e.g. "I like these," / "because they are comfy."

A closer analysis of the training set reveals interesting similarities and differences in the argumentative characteristics of reviews with and without many helpfulness vote counts (Figure 4). Reviews with more helpfulness votes tend to be longer and provide more support for the main proposition. However, the longest chains of reasoning are of similar lengths as those in reviews with few helpfulness votes. In other words, the additional elementary units found in more helpful reviews tend to serve the role of providing additional support for the main proposition, rather than make the chain of reasoning longer as additional intermediate premises.

4 Review Helpfulness Prediction

4.1 Task Formulation

Following Liu et al. (2021), we formulate review helpfulness prediction as a ranking task. More specifically, let p_i be the description (and images in the multi-modal scenario) of the ith product, with the associated reviews denoted as $R_i = \{r_{i,1}, \ldots, r_{i,M}\}$. Here, each $r_{i,j}$ is assigned a helpfulness score of $s_{i,j} \in \{0, \ldots, 4\}$ derived from grouping actual helpfulness vote counts into five bins with powers of 2 as boundaries, i.e., [0,1], [2,3], [4,7], [8,15], [16,∞). The goal is to find f such that the ranking of the reviews in each R_i based on $\hat{s}_{i,j} = f(p_i, r_{i,j})$ best match that based on $s_{i,j}$.

4.2 Experimental Setup

The experiments were conducted on Amazon product reviews for the Electronics category used in Liu et al. (2021), which consists of 13,205 product descriptions and 324,907 reviews in the training and development set, and 3,327 and 79,570 in the test set, respectively.

We tested the efficacy of AFs by incorporating them into two state-of-the-art baselines: a unimodal model that takes as input the product description and review text—Convolutional Kernel-based Neural Ranking Model (Conv-KNRM) (Dai et al., 2018), and a multi-modal model that takes as input product description and review text, as well as the images by the seller and reviewer—the Multi-perspective Coherent Reasoning (MCR) (Liu et al., 2021)

First, we trained the argument mining model by Morio et al. (2020) on AM^2. We then used it to extract AFs for the Amazon reviews from the Electronics category. While the output of this model is far from perfect, this approach allowed us to run experiments in a large scale and make fair comparisons against the baselines. Given the challenging nature of acquiring gold-standard argument annotations, we believe this setup reflects a realistic usage of our corpus.

Then, we incorporated AFs into the baseline models by encoding them in a vector and concatenating it to the review vector passed to the final classification layer of the respective models. Here, we considered constructing argumentational feature vectors using a multilayer perceptron (MLP) and a Graph Convolutional Network (GCN) (Kipf and Welling, 2017). In the MLP approach, we constructed a numeric vector storing the counts of each elementary unit by type, as well as the number of support relations present in the review. It was then passed to an MLP to encode the information in a dense vector. In the GCN method, the elementary units and their types, as well as support relations linking them, were represented as a graph, which was encoded in a dense vector by a GCN. In essence, elementary units were first represented by node embeddings,

Model	mAP	N@3	N@5
Conv-KNRM	52.6	40.5	44.2
Conv-KNRM + AFs (MLP)	**53.4**	**42.7**	**46.0**
Conv-KNRM + AFs (GCN)	52.9	41.4	45.0
MCR	56.0	46.5	49.7
MCR + AFs (MLP)	56.1	47.4	50.3
MCR + AFs (GCN)	**56.6**	**48.1**	**51.0**

Table 2: Helpfulness Prediction Results. Experiments on Amazon product reviews (Electronics category) exhibit the benefit of argumentational features (AFs) on unimodal (text-only; Conv-KNRM) and multi-modal (text-and-image; MCR) state-of-the-art models. The improvements over the respective baselines are statistically significant ($p<0.001$).

Model	mAP	N@3	N@5
Conv-KNRM + AFs' (MLP)	51.9	40.7	43.8
Conv-KNRM + AFs' (GCN)	52.3	41.0	44.1
MCR + AFs' (MLP)	56.3	47.7	50.6
MCR + AFs' (GCN)	56.5	47.7	50.7

Table 3: Helpfulness Prediction Results (with argument mining component trained on the UKP persuasive essays dataset, instead of AM^2). The performance is consistently worse than that of the best models in Table 2.

then combined based on the graph structure through rounds of message passing to form a global embedding for the entire review.

4.3 Results and Analyses

Table 2 shows that incorporating AFs improves the performance—measured in mean average precision (mAP) and normalized discounted cumulative gain (N@3 and N@5)—in both text-only and text-and-image scenarios. The improvements over the respective baselines are statistically significant ($p<0.001$), as measured with paired bootstrapping with the bootstrap size of 1,000 (Berg-Kirkpatrick et al., 2012). Note that the increase in performance is noticeably greater in the text-only case. This is expected, since AFs capture additional information about the review text, and the overall impact of textual information is greater in the text-only scenario.

A closer observation into individual reviews reveals the value of AFs. Figure 5 is a review that contains well-structured arguments, providing explicit reasons for propositions. For instance, the subjective proposition that the product is "one of the best keyboards" they have used is accompanied with two objective reasons describing the clean layout and chiclet style keys. Providing appropriate types of support for propositions in this way is precisely what our AFs are designed to capture; indeed, the ranking of this review increased after incorporating AFs: For Conv-KNRM, the ranking went from 11th to 8th (MLP) and 9th (GCN), and for MCR, from 11th to 4th (MLP) and 10th (GCN).

In addition, the consistent improvement in performance suggests that AFs are helpful even if the underlying argument mining system is not perfect. Also,

[This is one of the best keyboards I've ever used]$_a$ [The layout is simple and clean]$_b$ [I love how the chiclet style keys are evenly spaced out and how they have just enough resistance to pop back out quickly after you press them]$_c$ ... [It's light, yet feels well made and solid]$_g$ [The rubber feet at the bottom prevent it from slipping, and the light indicators in the top [...]]$_h$ ...

a (VALUE) ← ***b*** (FACT) ***g*** (VALUE) ← ***h*** (FACT)
c (FACT)

Figure 5: A Helpful Review Whose Ranking Increased with an Incorporation of AFs. The ranking increased from the 11th place to the 8th (MLP) and the 9th (GCN) for Conv-KNRM; and to the 4th (MLP) and the 10th (GCN) for MCR. Note, only the elementary units in support relations are shown.

annotations for reviews on headphones seem to be useful for training argument mining systems for reviews on the broader Electronics category. This is promising given the perhaps inevitable use of argument mining systems to extrapolate from the limited data with argument information annotated.

We also experimented with a popular argument mining dataset—UKP persuasive essays dataset (Stab and Gurevych, 2014)—in place of AM^2 to investigate the feasibility of using existing resources for mining AFs. This dataset is annotated with three types of elementary units (major claim, claim, and premise) and two types of relations (support and attack). Their theoretical argumentation model builds on perhaps the simplest model consisting of claim and premise, with an additional elementary unit and relation suitable for persuasive essays. However, it still does not capture appropriateness of support as the argumentation model adopted in this work does. This, along with the difference in domain, likely caused the consistently worse performance reported in Table 3. This in turn empirically shows the utility of AM^2.

5 Conclusion

AFs are highly relevant to determining the helpfulness of product reviews, as they capture several dimensions of quality, such as informativeness (e.g. claims accompanied with premises are more informative than unsubstantiated claims) and clarity (e.g. reviews written in a clear argumentative structure are easier to understand than obscure writing). Yet, there had not been sufficient resources to facilitate their adoption. To this end, we presented AM^2, a corpus of Amazon product reviews with argument information annotated, along with experimental results showing its efficacy. In the future, we hope to develop review helpfulness predictors that better leverage AFs; methods for incorporating AFs have not been studied extensively in this work, and effective approaches may lead to large gains in the overall performance.

6 Limitations

One limitation of this work is that the experiments were conducted on reviews from the Electronics category only. The underlying argument mining system was trained on AM^2 consisting of reviews for Headphones, which is a subcategory of Electronics in Amazon's product hierarchy. The experiment results suggest that it can handle other reviews from the Electronics category. However, its performance on reviews from unrelated categories, e.g. Arts & Crafts, may be poor, which in turn could limit the applicability of this corpus to helpfulness prediction of reviews from other domains.

Also the size of the corpus is smaller than ideal. While it is comparable to that of other popular argument mining corpora, e.g. 402 persuasive essays (Stab and Gurevych, 2017) and 731 e-rulemaking user comments (Park and Cardie, 2018), it is small nonetheless. This is likely to have a negative impact on the argument miner's performance, which in turn would make it difficult to observe the full potential of AFs on review helpfulness prediction. Note, argument mining corpora are generally small, because annotating argument information is difficult: It requires proper training of annotators, and annotation itself is time-consuming due to the complex and subjective nature of identifying underlying argumentative structures.

7 Ethical Considerations

The corpus presented in this work is a subset of the UCSD Amazon Review Dataset (Ni et al., 2019), which is a large collection of reviews publicly available from the US Amazon website. Annotations were completed by undergraduate students who volunteered to participate in the project and were compensated financially, $9.5/hour, or received course credits.

One intended use of the corpus is to train review helpfulness predictors. In this regard, models trained with this corpus may further perpetuate biases that are present in the Amazon review platform, if any. Such biases may include a tendency to interpret reviews written by a member of a particular demographic group as more helpful than those written by reviewers outside the group. Note that while not using the user information would mitigate this problem, specific style of structuring arguments may be common within a given demographic group and the model may wrongfully associate it with reviews being helpful.

Acknowledgments

We thank the University of Richmond and the Thomas F. and Kate Miller Jeffress Memorial Trust, Bank of America, Trustee for their generous support for this project. We also thank Dinesh Puranam for the helpful discussions.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL, pages 995 – 1005.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126–134, New York, NY, USA. Association for Computing Machinery.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 422–428, Florence, Italy. Association for Computational Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2020. Corpus for modeling user interactions in online persuasive discussions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1135–1141, Marseille, France. European Language Resources Association.

Anindya Ghose and Panagiotis G. Ipeirotis. 2011. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications.

J. Lawrence and C. Reed. 2014. AIFdb corpora. In *Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014)*, pages 465–466, Pitlochry. IOS Press.

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. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1363, Copenhagen, Denmark. Association for Computational Linguistics.

Junhao Liu, Zhen Hai, Min Yang, and Lidong Bing. 2021. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5927–5936, Online. Association for Computational Linguistics.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2020. Towards better non-tree argument mining: Proposition-level biaffine parsing with task-specific parameterization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3259–3266, Online. Association for Computational Linguistics.

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.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia. Association for Computational Linguistics.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in erulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ICAIL '15, pages 206–210, New York, NY, USA. ACM.

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)*, Paris, France. European Language Resources Association (ELRA).

Marco Passon, Marco Lippi, Giuseppe Serra, and Carlo Tasso. 2018. Predicting the usefulness of Amazon reviews using off-the-shelf argumentation mining. In *Proceedings of the 5th Workshop on Argument Mining*, pages 35–39, Brussels, Belgium. Association for Computational Linguistics.

Xiaoru Qu, Zhao Li, Jialin Wang, Zhipeng Zhang, Pengcheng Zou, Junxiao Jiang, Jiaming Huang, Rong Xiao, Ji Zhang, and Jun Gao. 2020. *Category-Aware Graph Neural Networks for Improving E-Commerce Review Helpfulness Prediction*, page 2693–2700. Association for Computing Machinery, New York, NY, USA.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

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

Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-aware review helpfulness rating prediction. *Proceedings of the 7th ACM conference on Recommender systems*, pages 1–8.

A Appendix

A.1 Annotation Details

Along with the annotations collected in this work (See Table 4), the AM^2 corpus also comes with the metadata available from the UCSD Amazon Review dataset (Table 5).

Field	Description
ID	ID of the elementary unit
Text	Text of the elementary unit
Type	POLICY, VALUE, FACT, TESTIMONY or REFERENCE
Reasons	List of reasons (elementary unit IDs)
Evidence	List of evidence (elementary unit IDs)

Table 4: Annotations Available for Elementary Units.

Field	Description
ReviewID	ID of the review
ElementaryUnits	List of elementary units and annotations (See Table 4)
Vote	# of helpfulness votes ($0 \leq$ n)
Verified	Was the review written by a verified purchaser of the product? (T/F)
ReviewTime	Submission time ("mm dd, yyyy")
ReviewerID	ID of the reviewer
ReviewerName	Name of the reviewer
Summary	Summary of the review
UnixReviewTime	Submission time (Unix time format)
Image	Images posted with the review
ASIN	ID of the product
Overall	Rating of the product ($1 \leq$ n ≤ 5)
Style	Product metadata (e.g. size, color)
Total	# of reviews written for the product
Existing	# of reviews written for the product at the time this review was written

Table 5: Metadata Available for the Reviews.

A.2 Argument Mining

For the argument mining task, as in Morio et al. (2020), the inputs will be the texts of online product reviews, separated into their constituent elementary unit spans. Each span is a section of the text representing an elementary unit of the argument, represented as a node in a graph. The outputs of the task would be the elementary

Prediction Target	F_1	# of instances
Support Relation	18.2	1072
Elementary Unit Type (macro-ave.)	49.6	1898
Fact	21.4	92
Testimony	65.1	570
Value	68.8	1208
Policy	43.1	28

Table 6: Argument Mining Results.

unit type of each argument component as well as the support relations of the argument graph, represented as edges.

Given a product review text consisting of N tokens, with M spans, let (s_j, e_j) be the starting and ending token indices for the jth elementary unit span. Thus, $0 \leq s_j \leq e_j \leq N$ and for each span j, we predict its elementary unit type and outgoing edges.

The results of the best 3-fold cross validated model are shown on Table 6.

A.3 Implementation Details

The first model we used was the argument mining model from Morio et al. (2020). This model is comprised of BiLSTM encoders, task-specific encoding layers, and biaffine attention modules for support relation prediction. For inference on the larger Electronics category of the Amazon product review dataset from Liu et al. (2021), we segment the review text into elementary units by sentence. The hyperparameter settings are show in Table 7, mostly matching those reported by Morio et al. (2020). There were approximately 3 hyperparameter search trials, each consisting of a training run and a evaluation run. Each training run took approximately 2 hours while each evaluation run took approximately 15 minutes.

The next models we used were for the helpfulness prediction task, namely Conv-KNRM and MCR. Each are neural ranking models, with Conv-KNRM only taking in text data while MCR taking in image in addition to text data. The Conv-KNRM model consists of convolutional layers, cross matching layers, kernel pooling and learning-to-rank layers. MCR consists of a text convolutional neural network (CNN) encoder, pretrained Faster-RCNN image features encoded by a self-attention module, intermodal and intramodal coherence modules, and intra-review coherent reasoning modules. For our MLP method of encoding AF features, we encode the counts of each elementary unit by type and the number of support relations present in a review by passing each into a separate MLP with output dimension equivalent to other final representations. For our GCN method of encoding AF features, we stack 3 GCN layers before finally passing the node representations through a global mean pool operation to generate the review embedding, again equivalent in dimension to other final representations. In both the MLP and GCN method, we concatenate the AF embeddings with the final representations before the final scoring layer. The hyperparameter settings of each models are show in Table 8,

There were approximately 5 hyperparameter search trials per model, each consisting of a training run and a evaluation run.

All times are reported as using 1 GPU from an academic server maintained by the University of Richmond. The server contains 4 available GPUs, each with 50 GB of RAM.

Hyperparameter	Value
GloVe dimension	300
GloVe linear projection dimension	100
POS linear projection dimension	100
ELMo type	elmo_2x4096_512_2048cnn_2xhighway_options
Input dropout rate	0.45
BiLSTM encoder dimension	400
BiLSTM encoder stack	1
BiLSTM type dimension	300
BiLSTM type stack	3
Recurrent dropout of all BiLSTMs	0.33
Output dropout of all BiLSTMs	0.25
Dimension of all MLPs	700
Dropout of all MLPs	0.25
Activation of all MLPs	LeakyReLU with negative slope 0.1
$(\lambda_{\text{edge}}, \lambda_{\text{type}})$	(0.5, 0.5)
Learning rate	0.0012
Epoch	100
Mini-batch	16

Table 7: Argument Mining hyperparameter settings attained through manual tuning based on validation F_1 score

Hyperparameter	Value
GloVe dimension	300
All text encoder kernel sizes	[1, 3, 5]
All text encoder hidden dimension	128
Conv-KNRM number of Gaussian kernels	11
Conv-KNRM Gaussian sigma	0.1
Conv-KNRM Gaussian exact sigma	0.0001
Conv-KNRM epochs	15
MCR common space dimension	64
MCR image encoder layers	3
MCR image encoder input dimension	2048
MCR image encoder embedding dimension	128
MCR image encoder dropout	0.5
MCR image encoder ReLU dropout	0.3
MCR image encoder attention heads	4
MCR image encoder attention dropout	0.3
MCR image encoder feed forward network dimension	512
MCR cross modal match dimension	128
MCR coherent encoder dimension	128
MCR coherent encoder layers	2
Batch size	2
Optimizer	Adam
Learning rate	0.0001
Loss	Ranked Hinge Loss
Loss Margin	1
Learning rate scheduler	Reduce on Plateau
Learning rate scheduler factor	0.5
Learning rate scheduler patience	4

Table 8: Conv-KNRM and MCR hyperparameter settings attained through manual tuning based on validation mAP, N@3, and N@5 scores