


# Assessing the Quality of Online Reviews using Formal Argumentation Theory

Davide Ceolin<sup>1</sup><sup>[0000-0002-3357-9130]</sup>, Giuseppe Primiero<sup>2</sup><sup>[0000-0003-3264-7100]</sup>,  
Jan Wielemaker<sup>1</sup><sup>[0000-0001-5574-5673]</sup>, and Michael Soprano<sup>3</sup><sup>[0000-0002-7337-7592]</sup>

<sup>1</sup> Centrum Wiskunde & Informatica, Amsterdam, The Netherlands  
{davide.ceolin, j.wielemaker}@cwi.nl  
<sup>2</sup> University of Milan, Milan, Italy giuseppe.primiero@unimi.it  
<sup>3</sup> University of Udine, Udine, Italy michael.soprano@uniud.it

**Abstract.** Review scores collect users’ opinions in a simple and intuitive manner. However, review scores are also easily manipulable, hence they are often accompanied by explanations. A substantial amount of research has been devoted to ascertaining the quality of reviews, to identify the most useful and authentic scores through explanation analysis. In this paper, we advance the state of the art in review quality analysis. We introduce a rating system to identify review arguments and to define an appropriate weighted semantics through formal argumentation theory. We introduce an algorithm to construct a corresponding graph, based on a selection of weighted arguments, their semantic similarity, and the supported ratings. We provide an algorithm to identify the model of such an argumentation graph, maximizing the overall weight of the admitted nodes and edges. We evaluate these contributions on the Amazon review dataset by McAuley et al. [15], by comparing the results of our argumentation assessment with the upvotes received by the reviews. Also, we deepen the evaluation by crowdsourcing a multidimensional assessment of reviews and comparing it to the argumentation assessment. Lastly, we perform a user study to evaluate the explainability of our method. Our method achieves two goals: (1) it identifies reviews that are considered useful, comprehensible, truthful by online users and does so in an unsupervised manner, and (2) it provides an explanation of quality assessments.

**Keywords:** Formal Argumentation Theory · Online Reviews · Information Quality

## 1 Introduction

Online reviews can be a valuable source of information, as they allow users to gain from the experience of others who have expressed their opinion about the next product to buy or room to book. Opinions provided by Web users are useful insofar as those of higher quality can be identified. Over the past years, research has characterized reviews’ trustworthiness in several ways: user reputation and quality assessment are among them. However, while reviews are about specific products or services, they represent often express multifaceted views on the target object. To assess the quality and trustworthiness of a review, it is important to understand which arguments it provides, their strength, and on which aspects of a target product they provide positive or negative evidence.

Reviews can be seen in the form of ratings-descriptions pairs. Such form of reviews, common in many e-commerce sites, indicates a rating (often in a 1 – 5 Likert scale) for the quality of a given target product, enriched with textual descriptions motivating the score. We analyze these descriptions to identify arguments that support the corresponding scores. Arguments are identified through natural language processing of such descriptions and ranked according to their importance using the textRank algorithm [16]. The quality of the descriptions is quantified through a readability measure [11]. We formulate, implement, and evaluate a rating system based on formal argumentation theory which collects such sets of pairs when they share a given argument but offer opposing ratings. We study it in depth by addressing the following research questions:

- R1:** Given a set of reviews about the same product, can argumentation reasoning help assessing review quality?
- R2:** Which quality aspects does argumentation reasoning emphasize?
- R3:** Can argumentation reasoning be used to explain review quality?

The rest of this paper is structured as follows. In Section 2 we first provide some informal preliminaries, then develop a preferential argumentation framework. In Section 3 we describe the experimental settings we adopt. In Sections 4, 5, and 6 we present our approaches to RQ1, 2, and 3, and the related results. We discuss the three evaluations in Section 7. In Section 8 we present related work, and in Section 9 we conclude.

## 2 Weight based Preferential Rating Systems

We propose a formal semantics of value-based argumentation that extends the model of Baroni et al. [3] to describe the conflict and support dynamics between topics as arguments within a set of reviews. Let us consider a set of reviews of a given product. We interpret them as nodes of a graph, where edges of the graph express the attack relation between two reviews providing descriptions for at least one common feature of the product, while they assign different scores to it. The semantics of the graph is established by a standard labeling function on vertices:

1. An argument is labeled *in* when all its attackers are *out*;
2. An argument is labeled *out* when at least one of its attackers is *in*;
3. An argument is labeled *undec* if not all its attackers are *out* and no attacker is *in*.

This semantics is aligned with the scores from natural language processing of the reviews and translated in a graph construction algorithm. Topics are grouped using K-means clustering; two reviews with disagreeing ratings attack each other when they share two topics belonging to the same cluster. Attacks follow topic weight ordering and support between arguments is represented indirectly: an argument supports another argument when it attacks its attacker. The weight of the corresponding edge is based on semantic similarity. Grouping reviews per topic allows obtaining a coherent set of reviews identifying the pros and cons of the same item.

**Definition 1 (Review).** A review  $\mathcal{R}_i(t)$  by an agent  $i \in \mathcal{A}$  on a target  $t$  is construed as:

1. A list of topics:  $\mathcal{T} = \{\phi_1; \dots; \phi_n\}$ ;
2. A relevance value  $r(\phi_i) \in [0, 1]$  for each topic  $\phi_i$ ;
3. A semantic similarity value  $sem\_sim(\phi_i, \phi_j) \in [0, 1]$  defined for each pair of topic for each review;
4. A score  $sc(\mathcal{R}_i) = \{1, 2, 3, 4, 5\}$ ;
5. A quality value  $v(\mathcal{R}_i) \in [0, 1]$ .

Provided a list of reviews  $\{\mathcal{R}\}$ , we collect all those with the same target object  $t$  and denote them as  $\{\mathcal{R}(t)\}$ . The list of topics  $\mathcal{T}$  collects the elements characterizing the review content on the target product: for example, on the target "shoes", topics could be "sole", "upper", but also "comfortable for long walks". The relevance value  $r(\phi_i)$  quantifies the importance of topic  $\phi_i$  within the review itself. This is a *de facto* value function from topics to real positive numbers. Likewise,  $sem\_sim(\phi_i, \phi_j)$  represents any of the available functions (e.g. based on thesaurus, symbolical representations of word semantics, term probabilistic co-occurrence) for semantic similarity defined for every couple of topics. Given  $\{\mathcal{R}(t)\}$  and  $\mathcal{T}(t) = \{\phi_1; \dots; \phi_n\}$ , we cluster any two topics according to a function  $sem\_dist(\phi_i, \phi_j)$  defined in the interval  $[0, \infty]$ . Several implementations are possible for this measure. In our case, we use the Word Mover distance [13] to measure the similarity between topics that are represented through text tokens, that are, in fact, groups of words. To identify the ideal number of clusters, we compute the cluster silhouette for a diverse number of clusters and we select the cluster configuration that maximizes this value, i.e. such that the intra-cluster average distance is minimized. The function  $sem\_sim(\phi_i, \phi_j)$  is then computed simply as  $\frac{1}{sem\_dist(\phi_i, \phi_j)}$ . The score  $\mathcal{R}_i$  is the value attributed to the object. We represent the score as an integer from 1 to 5, as it is done in many review systems. We currently consider different values as opposing, and do not consider the absolute difference between them (e.g. treating the difference between  $|sc(\mathcal{R}_i) - sc(\mathcal{R}_j)| > |sc(\mathcal{R}_i) - sc(\mathcal{R}_k)|$ ). Finally, a quality value is used to rank reviews. We especially consider the readability of the review to be an important aspect because: (1) it quantifies how easily a reader might consume it and (2) it might provide a proxy for the quality of the information it contains. In particular, in our experiments, we use the Flesch Kincaid Reading Ease measure [11]. This formula provides reliable scores between 100 (text understandable by 5<sup>th</sup> graders) and 0 (texts understandable by professionals). Other readability measures will be tested in the future.

Topics within the same cluster are those on which reviews' attacks are defined, as reviews showing semantically distant topics might be considered incomparable:

**Definition 2 (Attack).** Review  $\mathcal{R}_i$  attacks review  $\mathcal{R}_j$  with weight  $w$  ( $\mathcal{R}_i \rightarrow_w \mathcal{R}_j$ ) iff

1.  $\mathcal{R}_i(t) = \mathcal{R}_j(t)$ ;
2.  $\{\mathcal{T}(t) \in \mathcal{R}_i \cap \mathcal{T}(t) \in \mathcal{R}_j\} \neq \emptyset$ ;
3.  $sc(\mathcal{R}_i) \neq sc(\mathcal{R}_j)$ ;
4.  $\sum_{\phi_i \in \mathcal{R}_i} (r(\phi_i) \cdot v(\mathcal{R}_i)) > \sum_{\phi_j \in \mathcal{R}_j} (r(\phi_j) \cdot v(\mathcal{R}_j))$ ;
5.  $w = 1 / \sum_{\phi_i \in \mathcal{R}_i, \phi_j \in \mathcal{R}_j} (sem\_dist(\phi_i, \phi_j))$

According to the definition above a review attacks another one if and only if (1) they are about the same target object; (2) they have at least one of the topics of the

target object considered in common; (3) their score is different (as mentioned above, we make at this point no granular distinction between differences in scores); (4) the (sum of the) relevance value(s) of the topic(s) of the attacking review weighted by its quality is higher than that of the attacked review; and finally, (5) attacks are weighted on their importance: the weight of an attack is defined as the inverse of the (sum of the) semantic distance(s) of the topic(s) of the reviews involved, hence expressing the fact that an attack on more closely related topics weights more than one involving distant topics. A rating system is now built as a set of reviews and attacks between them, ordered according to a preference relation based on their weights:

**Definition 3 (Rating System).** A rating system is a tuple  $RS := \langle \{R(t)\}, R^-, \leq \rangle$  where

1.  $\{R(t)\}$  is a list of reviews on target  $t$ ;
2.  $R^- \subseteq \{R(t)\} \times \{R(t)\}$  is a binary relation of attack between reviews, such that  $(\mathcal{R}_i, \mathcal{R}_j) \in R^-$  iff  $\mathcal{R}_i \rightarrow_w \mathcal{R}_j$ ;
3.  $\leq \subseteq R^- \times R^-$  is a preference relation such that  $R^- \leq R'^-$  if and only if  $R^- : \mathcal{R}_i \rightarrow_w \mathcal{R}_j, R'^- : \mathcal{R}_k \rightarrow_{w'} \mathcal{R}_l, w > w'$  with possibly  $j = k$ .

According to this Definition, a rating system contains (1) a set of reviews on the same target, (2) equipped with a set of attack relations, (3) ordered based on their weights. We now define several strategies to establish the attack relations actually included in any given rating system:

**Definition 4 (Full Attack Strategy).**  $\forall R^-, R'^- \in RS$ .

The Full Attack Strategy includes every well-defined attack relation in the graph, i.e. any review attacks any other review with a different score with which it shares a topic within the same semantic similarity cluster and which has a lower weight computed as the relevance of the topic and quality value of the review. From this general case, we offer several pruning strategies on the number of attack relations.

**Definition 5 (Heavy Weight Pruning).**  $R^- \in RS$  iff  $\exists R'^-. R^- \leq R'^-$ .

In the Heavy Weight Pruning, we remove from the rating system the (set of) attack(s) with the lowest weight. By the definition of weight, this reflects the intuition that one removes those attacks based on the reviews having a different score on topics of low relevance, or on semantically distant topics (i.e. the reviews express different views on possibly incomparable aspects of the product). Note that a significant variant of this pruning method consists in removing the attacks with the weight under a certain value, e.g. falling within the last percentile. A more selective pruning strategy is expressed through clustering by semantic similarity:

**Definition 6 (Clustering Pruning).**  $R^- \in RS$  iff  $R^- < R'^-$  for some  $R'^- : \mathcal{R}_i \rightarrow_{w'} \mathcal{R}_j$  and  $w' < n$  such that  $n$  is the chosen value for a given clustering algorithm.

According to this method, an attack relation is considered in the graph of the rating system if and only if its weight is above the clustering threshold for the semantic similarity of the topics involved by the attack. Once the clustering is established of what does it mean for two topics to be similar, any attack which picks topics from distinct clusters is removed and no longer considered. In the following, we consider this pruning strategy as our main one. We now define the labeling of a rating system:

**Definition 7 (Labelling).** Given a rating system  $RS$

- $\{\mathcal{S}(t)\} \subseteq \{\mathcal{R}(t)\}$  is conflict-free iff there are no  $\mathcal{R}_i, \mathcal{R}_j \in \{\mathcal{S}(t)\}$  such that  $(\mathcal{R}_i, \mathcal{R}_j) \in R^-$ ;
- A review  $\mathcal{R}_i \in \{\mathcal{R}(t)\}$  is supported by  $\{\mathcal{S}(t)\} \subseteq \{\mathcal{R}(t)\}$  iff for any  $\mathcal{R}_j \in \{\mathcal{R}(t)\}$  such that  $(\mathcal{R}_j, \mathcal{R}_i) \in R^-$ , it exists  $\mathcal{R}_k \in \{\mathcal{R}(t)\}$  such that  $(\mathcal{R}_k, \mathcal{R}_j) \in R^-$ ;
- A review  $\mathcal{R}_i \in \{\mathcal{R}(t)\}$  is defeated by  $\{\mathcal{S}(t)\} \subseteq \{\mathcal{R}(t)\}$  if and only if it  $\exists \mathcal{R}_j \in \{\mathcal{S}(t)\}$  such that  $(\mathcal{R}_j, \mathcal{R}_i) \in R^-$  and  $\mathcal{R}_j$  is supported by  $\{\mathcal{S}(t)\}$ ;
- A review  $\mathcal{R}_i \in \{\mathcal{R}(t)\}$  which is neither supported nor defeated is undecided.

A conflict-free  $RS$  is possible if and only if every review has the same score for every topic  $\phi_i$  within a given cluster of semantic similarity. The notion of support of a review by a rating system expresses the idea that the score of that review for the given (cluster of) topic(s) is endorsed; the defeat of a review by a rating system expresses the dual idea that the score of that review for the given (cluster of) topic(s) is rejected; an undecided review is one which presents high expected variance on its usefulness in establishing the score of the product.

**Definition 8 (Semantics).** Given a rating system  $RS$

- A conflict free set  $\{\mathcal{S}(t)\} \subseteq \{\mathcal{R}(t)\}$  is admissible iff each  $\mathcal{R}_i \in \{\mathcal{S}(t)\}$  is supported by  $\{\mathcal{S}(t)\}$ ;
- A preferred extension is an admissible subset of  $\{\mathcal{R}(t)\}$  maximal w.r.t. set-inclusion and preference;
- An admissible  $\{\mathcal{S}(t)\} \subseteq \{\mathcal{R}(t)\}$  is a complete extension iff each review supported by  $\{\mathcal{S}(t)\}$  is in  $\{\mathcal{S}(t)\}$ ;
- The most (with respect to the weight of supported reviews) complete extension is the weighted complete extension.

We look for the model which maximizes the number of *in*-nodes with higher weight.

### 3 Experimental Setting

We describe the implementation of the above framework and the dataset adopted.

#### 3.1 Implementation

Figure 1 provides an overview of our implementation,<sup>4</sup> described as follows.

*Feature Extraction.* Given a set of reviews for product target  $t$ , we extract:

1. The set of textual tokens in such reviews to use as the set of topics  $\mathcal{T}$  and their importance in the text  $r(\phi_i)$  for each topic  $\phi_i \in \mathcal{T}$ . Textual tokens are estimated using the Spacy library, their importance is estimated through the pyTextRank library implementing the TextRank algorithm, i.e., computing the PageRank of the tokens in the review based on their textual dependency.
2. The readability scores of the review to use as a proxy for  $sc(\mathcal{R}_i)$ ; again we use the Spacy library and, in particular, the Spacy-readability extension.

<sup>4</sup> Source code available at: <https://github.com/davideceolin/FARreviews>.

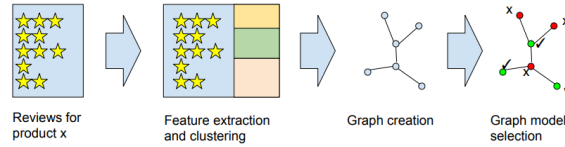


Fig. 1: Overview of the graph creation pipeline

*Argumentation Graph Building.* We proceed as follows:

1. Build the semantic distance matrix of all the tokens in the reviews of that product from each  $sem\_dist(\phi_i, \phi_j)$ . We use the Word Mover distance [13] implemented in Gensim [19] to this aim;
2. cluster tokens according to their semantic similarity. We use K-means and we identify the optimal number of clusters using the silhouette method;
3. represent an argumentation graph as a NetworkX Directed graph where: (1) nodes represent reviews; and (2) links represent attacks. Reviews attack all other reviews with a lighter score that share the same topic and disagree on the rating.

*Graph Solution.* In order to identify the models of the graph, we implement a SWISH Prolog-based solver also available as a standalone service accessed via a customized extension of the Python Prolog Penguins library.<sup>5</sup>

### 3.2 Dataset

We evaluate the above model on the Amazon Review Dataset [15], in particular on the Amazon Fashion 5-core dataset, which consists of 3,562 reviews (3,009 after duplicate removal) provided by 406 users about 31 products. For each review, the dataset reports:

- the id of the review author;
- the timestamp and the text of the review;
- the id of the product reviewed;
- the rating given to the product (on a 1-5 Likert scale);
- the number of upvotes a given review received. Note that users can only indicate whether they found a given review useful, not the contrary.

### 3.3 Argumentation Graph Building Example

Reviews and their argumentation graph are compared for explainability. Details of the graph construction process are given below in Section 4. Here we provide an example:

Review 1: ‘We have used these inserts for years. They provide great support.’  
(5 stars)

<sup>5</sup> <https://swish.swi-prolog.org/p/argue.swinb>

Review 2: ‘This is my 6th pair and they are the best thing ever for my plantar fasciitis and resultant neuromas. Unfortunately, the ones I ordered from Smart-Destination must be seconds as they kill my feet. The hard plastic insert rubs on the outside edges of my feet. I am unable to exchange them as I waited one day too late to use them in my walking shoes.’ (2 stars).

The two reviews have no textual token in common, however, some of their tokens are semantically related. For example, ‘these inserts’ (Review 1) and ‘hard plastic insert’ are semantically close enough to belong to the same cluster. This means that we capture an attack between the two, from Review 1 (readability 102.5) to Review 2 (readability 73.44). The weight of the attack is given by the sum of the importance of all the tokens of the two reviews which co-occur in a cluster, weighted on the semantic similarity between each pair of tokens and on the readability of the review itself. The semantic similarity is computed after stop words removal (the above tokens are ‘hard plastic insert’ and ‘inserts’). This process is repeated with all the tokens shared between two reviews and with all the review pair combinations for a given product.

## 4 RQ1 - Review Quality Assessment Evaluation

We consider here the ability of our system to discriminate reviews’ quality.

### 4.1 Baselines and Evaluation Settings

We created two baselines:

**Unsupervised (K-Means).** We extract a set of basic textual features from the reviews (e.g., text length) and we cluster them using the K-Means algorithm with  $K = 3$ .

**Supervised (SVC).** Using the same features as above, we split the dataset and use the first 30% of reviews to train a Support Vector Classifier to classify the remaining 70%. To allow a fair comparison between the three methods, we convert the number of upvotes into two buckets, to mimic the classification obtained with our method. We provide three variations on this, with thresholds at 1, 5, and 10 upvotes.

We evaluate our framework under three different settings:

**Argumentation Framework** We adopt the dataset described in Section 3.1.

**Argumentation Framework Weighted** We adopt the dataset described in Section 3.1, but we apply a decaying function to the number of upvotes based on their age. The decaying function we use is  $w(x) = \frac{t_{max}-t_x}{t_{max}-t_{min}}$  where  $t_{max}$  and  $t_{min}$  are the highest and lowest timestamps in the dataset;  $t_x$  is the timestamp of review  $x$ . Since the argumentation framework result is compared with a snapshot of the upvotes collected at a given time, this decaying function compensates for the fact that the older reviews had a higher chance to get upvotes than the younger ones.

**Argumentation Framework Weighted (Upvotes>0)** Since votes can only be up and not down, we cannot tell whether zero-votes reviews deserve zero votes or negative votes. We focus here on reviews that received at least one upvote.

Table 1: Average number (left) and sum (right) of upvotes received by the reviews in each class. The average of upvotes in the class should be maximized, the average in the *out* class minimized. In Arg. Framework Weighted, a temporal decaying function (see Section 4.1) is used. In Arg. Framework Weighted (> 0 upvotes), we consider reviews with at least 1 upvote (see Table 6).

Method	Out	In	Method	Out	In
Arg. Framework	2.3	0.5	Arg. Framework	35	<b>1553</b>
Arg. Framework Weighted	<b>0.0</b>	0.4	Arg. Framework Weighted	<b>0</b>	<b>1210</b>
Arg. Fram. Weigh. (>0 upvotes)	<b>0.0</b>	4.2	Arg. Fram. Weigh. (>0 upvotes)	<b>0</b>	<b>1210</b>
Unsupervised (K-Means)	2.5	0.3	Unsupervised (K-Means)	662	926
Supervised (SVC) @1	<b>0.0</b>	5.7	Supervised (SVC) @1	26	1165
Supervised (SVC) @5	0.1	10.2	Supervised (SVC) @5	304	765
Supervised (SVC) @10	0.3	<b>17.7</b>	Supervised (SVC) @10	610	459

## 4.2 Results

We run the above algorithm and we obtain a classification of product reviews as *in* or *out*. No review is classified as *undecided*. Table 1 shows the average number and sum of upvotes that the review in a given class received. For example, the reviews that are labeled as *out* (i.e., rejected) by the weighted version of our framework got, on average, 4.2 upvotes, and reviews classified as *out* got on average 0.0. We considered the possibility of computing precision and recall of our method. However, precision and recall imply the existence of negative samples, while upvotes are only positive values. Artificially introducing a threshold to split reviews into positive and negative items would be possibly misleading. A “one-size-fits-all” would hardly work in this case: such a threshold could have to vary per product or product type and could have to take into account also temporal aspects. For instance, less popular products could receive fewer reviews and have a smaller chance to get upvotes. Thus, their threshold should be lower than that of popular products. At the same time, the rareness of reviews alone cannot be considered a sufficient reason to set the bar low: those few reviews could get few upvotes because of their poor quality. Therefore, we limit the comparison with the baseline approaches to Table 1. With these considerations in mind, to allow a comparison between our method and SVC, we still introduce the use of thresholds to convert the multivalued classification of SVC into binary values. For this, we use three thresholds, 1, 5, and 10 (the mean number of upvotes received by a review in the ground truth is 0.55, median 0). We deepen these considerations in Section 7.

## 5 RQ2 - Multidimensional Review Quality Assessment

The evaluation of the argumentation theory-based review assessment by correlation with upvotes uses the latter as the only ground truth provided in the dataset at our disposal, but they also show important limitations. First, upvotes collect only positive votes: if a review did not get a high number of upvotes, it could be either of low- or average-quality. Second, the semantics of upvotes is rather vague and broad: since they are the only means for readers to express their endorsement, they can capture appreciation in a too broad sense. Third, upvotes might depend on the order with which reviews



are exposed to users and their age. We extend our analysis on the quality of reviews to obtain a more thorough and detailed gold standard. We crowdsource answers to questions regarding quality aspects of a significant number of reviews, as detailed below.

## 5.1 Crowdsourcing Setting

We collect 380 reviews by first randomly selecting one of the products reviewed, then one of its reviews. This ensures that the products are fairly represented since the dataset is rather skewed. Considering that, in total, the dataset is composed of 3,009 reviews, our sample has a confidence interval of 6.19 with a confidence level of 99%. We ask each worker to evaluate the quality of 10 reviews, and each review is evaluated by 5 workers. Workers are located in the US, and the tasks (which are rewarded 0.9\$) are performed through Amazon Mechanical Turk.<sup>6</sup>

*Task Description.* We present the worker with a product description as provided in the Amazon dataset. Then, we present the review, and we ask the worker to assess the review on a 5-level Likert scale (from -2, completely disagree, to +2, completely agree), across the following quality dimensions:

**Truthfulness:** measures the overall truthfulness and trustworthiness of the review.

**Reliability:** the review is considered reliable, as opposed to reporting unreliable information. *Example (label: +2 Completely agree): "They fit great, look great, are quite comfortable and are just what I was looking for!"*

**Neutrality:** the review is expressed objective terms, as opposed to resulting subjective or biased. *Example (label: -2 Completely disagree): "Love them!!"*

**Comprehensibility:** the review is comprehensible/understandable/readable as opposed to difficult to understand. *Example (label: +2 Completely agree): "They run big. Order a full size smaller"*

**Precision:** the review is precise/specific, as opposed to vague. *Example (label: +2 Completely agree): They run big. Order a full size smaller.*

**Completeness:** the review is complete as opposed to partial. *Example (label: +2 Completely agree): "I actually have 3 pairs of these trainers. They are very comfortable, there is a neoprene sleeve that goes around your ankle that makes them the most comfortable for me compared to normal athletic shoes. They run a little narrow - for me this is perfect, but you may want to round up on the size or try on in the store first if your feet are on the wider side."*

**Informativeness:** The review allows deriving useful information as opposed to well-known facts and/or tautologies. *Example (label: +1 Agree): "Love these shoes! Needed new running shoes and these are perfect. Light weight and fit great!"*

The above dimensions are based on previous work on multidimensional quality assessment [6]. However, with reviews, it is very hard for the workers to determine the truthfulness of information because they need to assess the authenticity of the review itself, which is often subjective. So, we adapt the quality dimensions from the literature to represent more subjective aspects like reliability.

<sup>6</sup> <http://mturk.com>

## 5.2 Results

Assessments were collected and we checked whether the scores in any of the evaluated dimensions showed a correlation with the *in-out* evaluation of the review by our algorithm. Since our classification consists of two labels only, while the crowdsourced data are multidimensional and finer-grained, we performed a set of analyses at diverse levels of aggregation, starting from splitting the reviews into *in* and *out*, obtaining:

- a  $\chi^2$  on the two sets review scores: no significant difference is identified;
- a Mann-Whitney test on the average score per dimension: no significant difference between the two sets of reviews is identified;
- t-test or Mann-Whitney test when comparing the raw scores on each dimension: no significant difference is identified.

Then, we aggregate the scores in two ( $[-2,0],[1,2]$ ) and three ( $[-2,-1], [0],[1,2]$ ):

- a  $\chi^2$  test on the two sets reviews still does not identify any significant difference in the distribution of scores;
- a Mann-Whitney test on the average score per dimension identifies a significant difference in the distribution of scores;
- at 90% confidence, a significant difference is identified in the distribution of the comprehensibility and the overall truthfulness scores of the two distributions.

In other words, when the crowdsourced scores are expressed on a coarse scale (two- or three-valued), our classification identifies two sets of reviews, where those labeled as *out* have higher comprehensibility and a higher overall-truthfulness than those labeled as *in*. Since readability score plays a role in the argumentation framework, those results might just be linked to the use of those scores. However, the readability score has a correlation of 0.24 with the crowdsourced comprehensibility, and of only 0.02 with the overall truthfulness. Thus, the identification of the review with higher truthfulness can be attributed to the whole framework.

## 6 RQ3 - Explainability Evaluation

We run an explorative questionnaire<sup>7</sup> to evaluate whether our approach provides informative explanations on the decision taken about the reviews (*in/out* outcome). We select two reviews about the same product, one accepted, and one rejected by our system. We show the argumentation graph on which the judgment is based and we ask the respondent whether the graph helps in understanding the underlying reasoning using a 1-5 Likert scale. Users can provide additional feedback. Table 2 shows the distribution of the 31 anonymous responses received, while Figure 2 shows an example question.

According to these results, the argumentation graph does indeed help in explaining the outcome. Since the outcomes vary from ‘poorly informative’ (1) ‘to very informative’ (5), the results are explanatory on both reviews (although for review 2 the signal is stronger). An important aspect of consideration as a possible limitation is that in argumentation-based reasoning arguments are valid until attacked and this translates into reviews accepted because not attacked.

<sup>7</sup> The questionnaire is available at <https://forms.gle/srGJpGyYBzWd9RTaA>.

Table 2: Distribution of the answers regarding the helpfulness.

Informativeness	1 (Poorly Informative)	2	3	4	5 (Very Informative)
Review 1 (accepted)	0	6	11	12	2
Review 2 (rejected)	0	1	7	15	8

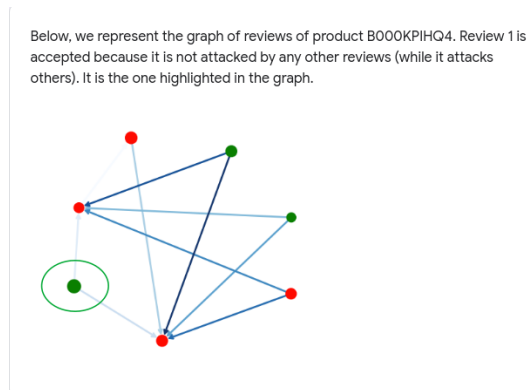


Fig. 2: Example question. Nodes represent reviews, (green *in*, red *out*) for the argument-based review classification, arrows represent attacks, their shade expresses semantic similarity.

## 7 Discussion

We now discuss the results related to each research question.

### 7.1 RQ1 - Given a set of reviews about the same product, can argumentation reasoning help assess review quality?

*Our method (especially in the improved versions, see rows 2 and 3 of Table 1) identifies two clusters of reviews where those in have a higher chance of having more upvotes than those out. Also, the method identifies the majority of the reviews that received upvotes.*

The first difference between the unsupervised approach and the proposed argumentation framework concerns labeling. The results reported in Table 1 assume arbitrarily that one of the two classes predicted by the K-means method equals the *out* class, the other the *in* class. However, we do not have any means to label the classes in this respect. So, while the performance of the two methods looks similar when considering the averages in Table 1, this may not be the case. For most of the remaining performance reported in Table 1, our method outperforms K-means. The supervised approaches are those showing the best performance in terms of the distribution of the average number of upvotes. Supervised approaches focus on identifying the peculiarities of reviews that hint at their upvotes. They do so at the dataset level, they make use of labeled data (number of upvotes per review) and can identify those reviews that meet these criteria. These methods achieve high accuracy of the number of upvotes estimated for a given review. However, they do so at the expense of a significant amount of upvotes missed, as

the right table of Table 1 shows. Measuring performance as precision and recall would have meant comparing our method on the mere ability to identify reviews having at least  $n$  upvotes for an arbitrary threshold  $n$  (this step is necessary to transform the number of upvotes in the ground truth into binary values comparable with our classification). This goes beyond our goals and just amplifies the results reported in the left table of Table 1. The correct threshold should depend on the number of reviews received by a given product, etc. We use thresholds to transform SVC in binary outcomes, though, because of the quantitative nature of SVC. SVC predicts the number of upvotes received by a review. Setting a threshold introduces the mentioned limitations but, in this case, performance would have been measured in terms of error of the number of upvotes predicted. Thresholds mainly reduce the granularity of such metrics but necessarily introduce some error: reviews which got  $n$  upvotes for  $0 < n < \text{threshold}$  are labeled as out, thus affecting the performance reported in the right table of Table 1. Also, the good performance of the supervised methods comes with limitations:

**Need for training data.** Being supervised, SVC craves for labeled data; in production, the system might be affected by the cold start problem.

**Arbitrary parameters.** When comparing the two methods, we had to convert the estimated number of upvotes into two classes. This is arbitrary because it corresponds to answering a question like “how many upvotes does a review need to receive to be accepted?”. This has led to testing the three different parameters.

**Lack of explanations.** The method is meant to estimate the number of upvotes received by each review. However, when deciding whether to consider a given review or not based on such estimates, it is important to understand how such reasoning was performed. Inspection on the importance would require additional efforts.

These limitations are not shown by our method, which is unsupervised and explainable. Also, from the diverse evaluation settings, we learned the following lessons.

*Lesson learned 1: Time matters* When inspecting the reviews in the *out* class, the high average is due to just one review labeled as out, despite having received 35 upvotes. This is the oldest review of that product; 6 more reviews, received about 6 years later, had 0 upvotes. Given that these newer reviews got a lower chance to get an upvote because they are more recent, we discounted the number of upvotes based on the age of the review. This improves the system performance (see Table 1).

*Lesson learned 2: Non-attacked reviews should not necessarily be accepted* In formal argumentation theory, arguments are accepted until they are defeated. However, not yet attacked reviews could get zero upvotes for a variety of reasons (e.g., they are out of topic). On a long-tail distributed dataset, this affects the results obtained. This is the reason why the reviews classified *in* have a low average number of upvotes. As shown in the third row of Table 1, the performance on the reviews with at least one upvote is higher. Table 3 provides an overview of the number of reviews per class.

## 7.2 RQ2 - Which quality aspects does argumentation reasoning emphasize?

*The labeling obtained by our argumentation framework is correlated with the comprehensibility and with the overall truthfulness of the reviews. As already pointed out in*

Table 3: Number of reviews classified as in and out, split on the number of upvotes.

Class	in	out
Reviews with 0 upvotes	2,706	14
Reviews with at least 1 upvote	288	1

Section 5, the readability scores alone would not be able to point out the reviews having higher overall truthfulness. This result has a twofold consequence. First, it supports the argumentation-based approach and the need for logical reasoning to be performed on top of the ranked arguments to obtain labeling that correlates with overall truthfulness and comprehensibility. Second, it points out other quality aspects that we might consider in future extensions of our framework. E.g., completeness might be correlated to the number of *in* arguments in a review. Here we learned an important lesson.

*Lesson learned 3: Granularity and Semantics matter* While quality is subjective and contextual, it is also possible to define which aspects of quality we are interested in. This is important to allow a more precise understanding of the argumentation outcome. Also, the current implementation of the framework provides a three-valued assessment and, as expected, correlation with crowdsourced ratings emerges only when these are aggregated in buckets. Future extensions of the framework might consider a fine-grained representation of acceptance/rejection of arguments.

### 7.3 RQ3 - Can argumentation reasoning be used to explain review quality?

*According to the exploratory study described in Section 6, argumentation graphs are useful to explain review assessment.* The study was meant to provide a first indication about the hypothesis that argumentation graphs are useful to explain review assessment. The responders agreed with this idea: 45,2% of them rated informativeness at level 4 or 5 (very informative) for the first question, 73,6% for the second. This will be further explored in the future. “How to better represent attack weights?” and “which level of complexity users can handle?” are examples of questions we will tackle.

## 8 Related Work

This work falls within the growing family of weighted argumentation frameworks extending standard Dung’s setting, including Preferential Argumentation Frameworks [1,17,2] and Value-based Argumentation Frameworks [4,5]. A specific approach is represented by systems defining preferences based on weighted attacks, see [9], establishing that some inconsistencies are tolerated in the set of arguments, provided that the sum of the weights of attacks does not exceed a given value. Weights can be used to provide a total order of attacks, see [14]. This approach can be generalized in several ways: in [8] a different way of relaxing the admissibility condition and strengthening the notion of defense is presented; in [7] different selections on extensions based on the order of weights are proposed. Our work also relies on an ordering on weighted attacks, essential differences being that:

1. the definition of weights is given by the semantic distance between topics;
2. the clustering of attacks is based on weights;
3. the pruning of the graph is based on the order, as distinct from the selection of the model based on the maximization of the weight of accepted arguments.

Research on the assessment of quality and credibility of product reviews has focused mostly on linguistic aspects, e.g. based on readability and linguistic errors [10,22,12,18]. While such approaches can be a source of inspiration for future extensions, the main difference with our approach is the combination of such linguistic aspects with argumentation reasoning. A similar extension can be obtained by looking into credibility factors, as in [21]. Lastly, [23] looks for a junction between natural language processing and argumentation reasoning. While it classifies more thoroughly the diverse tokens as different kinds of arguments, it does so semi-automatically, while we take an automatic unsupervised approach. Refining the characterization of arguments is one aspect we intend to improve in the future. Regarding the crowdsourced assessment of online information, we refer the reader to [20], although their focus is on political statements, and their assessment is mono-dimensional. A multidimensional approach is adopted in [6], where Web documents are assessed by experts (nichesourcing).

## 9 Conclusion and Future Work

This paper presents a framework for classifying reviews’ quality based on a combination of NLP and argumentation reasoning. We evaluate the framework on a real-world dataset showing that this approach partly outperforms baseline unsupervised and supervised approaches, while also providing explainable results. A deeper analysis of the quality of the reviews based on crowdsourcing highlights that the argumentation framework is actually capable of identifying those reviews that the users perceive as more comprehensible and truthful. Also, a two- or three-level scoring of reviews across multiple quality dimensions reveals to be the ideal level of granularity. We also run a user study that confirms the ability of argumentation graphs of providing useful explanations. This argumentation-based framework represents a first step towards a reliable and transparent assessment of the quality of online opinions.

We foresee several future developments for this work. Firstly, the framework should be extended by discounting the weight of the review and its attacks considering the temporal aspect (e.g., using weight  $w(x)$  of Section 7). Secondly, the model could account for a different semantics of nodes *in* and *out* to prevent that novel reviews be automatically *in*. Thirdly, we will improve the identification of the arguments among the review tokens. Lastly, we plan on analyzing a larger number of datasets and reviews.

**Acknowledgements** This work is partially supported by The Credibility Coalition.

## References

1. Amgoud, L., Cayrol, C.: A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence* **34**, 197—215 (2002)

2. Amgoud, L., Vesic, S.: Two roles of preferences in argumentation frameworks. In: Proceedings of ECSQARU. pp. 86–97. Springer (2011)
3. Baroni, P., Caminada, M., Giacomin, M.: Abstract argumentation frameworks and their semantics. In: Baroni, P., Gabbay, D., Giacomin, M. (eds.) Handbook of Formal Argumentation, chap. 4. College Publications (2018)
4. Bench-Capon, T.J.M.: Value-based argumentation frameworks. In: Proceedings of NMR Workshop. pp. 443–454 (2002)
5. Bench-Capon, T.J.M.: Persuasion in Practical Argument Using Value-based Argumentation Frameworks. *Journal of Logic and Computation* **13**(3), 429–448 (2003)
6. Ceolin, D., Noordegraaf, J., Aroyo, L.: Capturing the ineffable: Collecting, analysing, and automating web document quality assessments. In: Proceedings of EKAW. p. 83–97. Springer (2016)
7. Coste-Marquis, S., Konieczny, S., Marquis, P., Ouali, M.A.: Selecting extensions in weighted argumentation frameworks. In: Proceedings of COMMA. IOS Press (2012)
8. Coste-Marquis, S., Konieczny, S., Marquis, P., Ouali, M.A.: Weighted attacks in argumentation frameworks. In: Proceedings of KR. p. 593–597. AAAI Press (2012)
9. Dunne, P.E., Hunter, A., McBurney, P., Parsons, S., Wooldridge, M.: Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence* **175**(2), 457–486 (2011)
10. Ghose, A., Ipeirotis, P.G.: 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 (2011)
11. Kincaid, J., Fishburne, R., Rogers, R., Chissom, B.: Derivation of new readability formulas for navy enlisted personnel. research branch report 8–75. Tech. rep., Chief of Naval Technical Training: Naval Air Station Memphis (1975)
12. Korfiatis, N., García-Bariocanal, E., Sánchez-Alonso, S.: Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications* **11**(3), 205–217 (2012)
13. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: Proceedings of ICML. p. 957–966. JMLR.org (2015)
14. Martínez, D.C., García, A.J., Simari, G.R.: An abstract argumentation framework with varied-strength attacks. In: Proceedings of KR. pp. 135–144. AAAI Press (2008)
15. McAuley, J.J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of SIGIR. pp. 43–52. ACM (2015)
16. Mihalcea, R., Tarau, P.: TextRank: Bringing order into text. In: Proceedings of EMNLP. pp. 404–411. ACL (2004)
17. Modgil, S.: Reasoning about preferences in argumentation frameworks. *Artificial Intelligence* **173**(9), 901–934 (2009)
18. Ocampo Diaz, G., Ng, V.: Modeling and prediction of online product review helpfulness: A survey. In: Proceedings of ACL. vol. 1, pp. 698–708. ACL (2018)
19. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of NLPFrameworks Workshop. pp. 45–50. ELRA (2010)
20. Roitero, K., Soprano, M., Fan, S., Spina, D., Mizzaro, S., Demartini, G.: Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor’s Background. In: Proceedings of SIGIR. p. 439–448. ACM (2020)
21. Wathen, C.N., Burkell, J.: Believe it or not: Factors influencing credibility on the web. *Journal of the American Society for Information Science and Technology* **53**(2), 134–144 (2002)
22. Wu, P., Van Der Heijden, H., Korfiatis, N.: The influences of negativity and review quality on the helpfulness of online reviews. In: Proceedings of ICIS. pp. 3710–3719 (2011)
23. Wyner, A., Schneider, J., Atkinson, K., Bench-Capon, T.: Semi-automated argumentative analysis of online product reviews. In: Proceedings of COMMA. pp. 43–50. IOS Press (2012)