Categorizing Review Helpfulness Using Abstract Dialectical Frameworks

Atefeh Keshavarzi Zafarghandi[1,2(✉)] and Davide Ceolin[1]

[1] Human-Centered Data Analytics, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{atefeh.keshavarzi.zafarghandi,davide.ceolin}@cwi.nl
[2] Vrije University Amsterdam,Amsterdam, The Netherlands

Abstract. Consumer reviews are a vital aspect of the decision-making process for both buyers and companies in the era of e-commerce and online shopping. However, the helpfulness of reviews varies widely, and the abundance of available information can make it difficult to identify the most informative ones. Therefore, categorizing product reviews based on their helpfulness is a critical task. Review helpfulness can be determined by considering several features, such as readability, sentiment, word count, and coherence between the sentiment and score of a review. This article proposes a method for categorizing review helpfulness based on readability and coherence. Our approach employs abstract dialectical frameworks (ADFs), which use interpretation-based semantics to evaluate the acceptability of arguments. We tailor a specific ADF to each review to assess its helpfulness and provide clear explanations for our labeling decisions. We use the grounded semantics of ADFs, which provides information that no one can argue against, to justify our labels and enhance the value of our process. Our method can also be used as a system to give feedback to the review authors on why their reviews may not be helpful and how they can improve them in the future by considering readability and coherence factors. Moreover, our method can work on both small and large data-sets, which may not be feasible with machine learning methods that require a lot of training data.

Keywords: Online reviews · Abstract dialectical frameworks · Explainable artificial intelligence

1 Introduction

In today's world, we are inundated with an enormous amount of online reviews that play a critical role in shaping the opinions of new readers. Product reviews are invaluable sources of information for both customers and companies. Companies can use reviews to gain insight into what their customers want and need in a product, while customers can use reviews to make informed purchasing decisions.

Published by Springer Nature Switzerland AG 2024
T. Muller et al. (Eds.): IFIPTM 2023, IFIP AICT 694, pp. 89–104, 2024.
https://doi.org/10.1007/978-3-031-76714-2_6

However, there are significant challenges in utilizing online reviews, including inconsistency and difficulty in determining credibility. In order to extract valuable insights, structuring online reviews is essential. These challenges have motivated our approach, which aims to evaluate the helpfulness of reviews. While metrics such as verified purchases, reviewer reputation, and platform reliability are commonly used to evaluate the source of the reviews, there is still a large area of unexplored territory in categorizing reviews based on their helpfulness.

To evaluate the helpfulness of reviews, we employ formal argumentation, specifically Abstract Dialectical Frameworks (ADFs) [3–5]. ADFs provide a framework for modeling different relationships among arguments, including conflicts. The semantics of ADFs offers a systematic and rigorous approach to assessing the quality and credibility of arguments. ADFs can also be used to elucidate the reasons behind the acceptance of arguments.

In this study, we examine reviews that include both text and a score. In different works, different factors, such as readability, word count, sentiment, and number of sentences, are considered to indicate the helpfulness of the reviews. However, the importance of these factors may vary in different data-sets. We choose the Amazon Fashion 5 reviews data-set by McAuley et al. (2015), which contains some reviews that are voted by other users. We consider the voted reviews as helpful ones, because they have attracted attention from the readers. After running different classifiers on this data-set, we find that readability is the most important feature for classifying the reviews, and the second most important feature is the sentiment of the review. However, the correlation of these factors and voted reviews is not very high. Therefore, we consider readability as a crucial factor for indicating the helpfulness of a review, not only because it is the most influential feature based on the classifiers, but also because a good review should be understandable and easy to read. Additionally, we expect the text of a review to support its score. For example, a 5-star review should provide sufficient evidence of the positive aspects of a product that justify the high score. While it is acceptable to mention a few negative points, the focus should be on highlighting the strengths of the product that lead to the 5-star rating. We introduce the concept of coherence in a review, which means that the text should match the review's score. A 4 or 5-star review is coherent if the text emphasizes the positive aspects of the product, while a 1, 2 or 3-star review is coherent if the text highlights the negative aspects of the product.

We have three levels to determine the helpfulness of a review: very helpful, helpful, or unhelpful. Our goal is to assign a review to one of these categories based on its features. To achieve this, we use formal argumentation to establish a connection between the features of a review and its associated label. We consider arguments for each feature and label, as well as auxiliary arguments to link the review features to its label. Although it is possible to categorize reviews in these three categories with the use of a table, we choose to do so by formal argumentation for several reasons. By constructing an ADF associated with a given review, we can visualize the connections between a review's features and its label. The semantics of ADFs are then used to assess the acceptance of

arguments, specifically label arguments, i.e., label of the review. The ADF is not only used to label reviews but also to provide an explanation for the model's decision. Furthermore, in this work we only consider two factors for labeling reviews, but if we have n factors and want to use a table for the labeling, we would need a table with 2^n rows to indicate the label of a given review.

Labeling reviews based on their helpfulness can enhance the precision of statistical results by filtering out unhelpful reviews. A transparent labeling process achieved through a dialogue argumentation can explain the rationale behind each label choice. This can also provide valuable feedback to writers of reviews.

The remainder of the paper is organized as follows. Section 2 discusses the related work with our proposal. We give an overview of abstract dialectical frameworks in Sect. 3. In Sect. 4.1, we outline the characteristics of helpful reviews. In Sect. 4.2, we discuss the use of the Abstract Dialectical Framework (ADF) as a model for labeling reviews. Section 5 presents a statistical analysis of our data set. Additionally, this section describes the implementation of the method presented in Sect. 4.2 for labeling the reviews in this data set. Section 6 presents a formal argumentation dialogue aimed at addressing the explanation gap in establishing the relationship between a review's features and its assigned label. We conclude the paper and present outline the future work directions in Sect. 7.

2 Related Work

The quality evaluation of online opinions, specifically analyzing reviews, has received significant attention in recent years. A number of studies have been conducted in this area, including research on detecting review spam [19] and identifying deceptive reviews [10]. Formalisms of argumentation have been proposed as a tool to model the relationship between reviews and to assess their quality [7]. Additionally, argumentation formalisms have been employed to assess the different aspects of a review [15]. In our work, we leverage argumentation formalisms to evaluate the helpfulness of reviews, taking into consideration two critical features, namely readability and coherence between the review text and the assigned score. We believe that labeling reviews based on their helpfulness is a critical step for verifying their trustworthiness and is a necessary prerequisite for identifying a set of consistent reviews.

With the increasing interest in explainable artificial intelligence (XAI), the use of argumentation has become prevalent in generating explanations for automated systems' decisions [17]. Dialectical explanations for argument-based reasoning have been formalized in [14], while [21] has developed an argumentation-driven recommender system that provides explanations extracted from argumentation frameworks. Dialogical argumentation, represented by qualitative argumentation frameworks (e.g., [11,20]) or quantitative argumentation frameworks (e.g., [8,9]), has been explored in various settings, including fake news detection [8] and exchanging explanations with users [9]. In our context, we also utilize dialogical argumentation to provide explanations supporting the review's label.

The quality of a text can be assessed using readability measures [16]. A good product review should be concise, clear, and offer relevant information. We link

review helpfulness to comprehension ease. Readability measures are widely used for text quality assessment. In [6], various readability measures were evaluated on our dataset, and we chose the Dale-Chall measure [12] due to their similarity. We consider the coherence of the reviews as another factor of helpfulness and define coherence as a measure of how well a review supports its score.

3 Preliminary: Abstract Dialectical Frameworks

We summarize key concepts of abstract dialectical frameworks [3,5].

Definition 1. *An abstract dialectical framework (ADF) is a tuple* $D = (A, L, C)$ *where:*

- A *is a finite set of arguments (statements, positions);*
- $L \subseteq A \times A$ *is a set of links among arguments;*
- $C = \{\varphi_a\}_{a \in A}$ *is a collection of propositional formulas over arguments, called acceptance conditions.*

An ADF can be represented by a graph in which nodes indicate arguments and links show the relation among arguments. Each argument a in an ADF is labeled by a propositional formula, called acceptance condition, φ_a over $Par(a)$ such that, $Par(a) = \{b \mid (b, a) \in L\}$. The acceptance condition of each argument clarifies under which condition the argument can be accepted. An argument a is called an *initial argument* if $Par(a) = \{\}$.

Furthermore, acceptance conditions indicate the set of links implicitly. Thus, in a concrete example of ADFs, we oftentimes only define acceptance conditions explicitly and implicitly define links via the variables of the propositional formulas. That is, for the reason of brevity we avoid presenting the set of links of ADFs in our examples.

A *three-valued interpretation* v (for D) is a function $v : A \mapsto \{\mathbf{t}, \mathbf{f}, \mathbf{u}\}$, that maps arguments to one of the three truth values true ($\mathbf{t}$), false ($\mathbf{f}$), or undecided ($\mathbf{u}$). Given an interpretation v (for D), the partial valuation of φ_a by v, is $\varphi_a^v = \varphi_a[b/\top : v(b) = \mathbf{t}][b/\bot : v(b) = \mathbf{f}]$, for $b \in Par(a)$. Semantics for ADFs can be defined via the *characteristic operator* Γ_D, presented in Definition 2.

Definition 2. *Let* D *be an ADF and let* v *be an interpretation of* D*. Applying* Γ_D *on* v *leads to* v' *such that for each* $a \in A$*,* v' *is as follows:*

$$v'(a) = \begin{cases} \mathbf{t} & \text{if } \varphi_a^v \text{ is irrefutable (i.e., } \varphi_a^v \text{ is a tautology)}, \\ \mathbf{f} & \text{if } \varphi_a^v \text{ is unsatisfiable}, \\ \mathbf{u} & \text{otherwise.} \end{cases}$$

An interpretation v is the *grounded* interpretation of D iff v is the least fixed point of Γ_D.

Example 1. An example of an ADF $D = (S, L, C)$ is shown in Fig. 1. To each argument a propositional formula is associated, the acceptance condition of the argument. For instance, the acceptance condition of c, namely $\varphi_c : \neg b \wedge d$, states that c can be accepted in an interpretation where b is denied and d is accepted. The interpretation $v = \{a, b, \neg c, \neg d\}$ is a unique grounded interpretation in D.

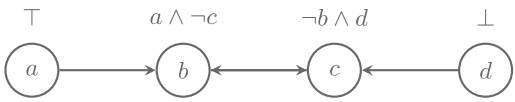

Fig. 1. ADF of Examples 1

4 Labeling of Reviews

The objective of this section is to categorize reviews based on their helpfulness. In Sect. 4.1, we introduce the characteristics of a helpful review that we aim to analyze. Then, in Sect. 4.2, to label a given review, we first construct an ADF D that corresponds to the review (r, n). Using the grounded interpretation of ADFs, we evaluate the truth values of the arguments, including the output arguments. Finally, we apply the label-function, as defined in Definition 7, to assign a label to the review.

4.1 Characters of a Helpful Review

The characteristics of a helpful review may differ, but the most typical feature is clarity, that is, The review should be easy to comprehend, utilizing simple and concise language [22].

To evaluate the clarity of reviews, we use the Dale-Chall readability formula [2], as defined in Definition 3.

Definition 3. *Let r be a review. The read() function takes a review r as input and returns its readability score, which is calculated using the Dale-Chall readability formula [2]. The readability score is a real number between 0 and 10, where a higher score indicates that the review is easier to read.*

In our work, we propose the concept of a coherence review, which aims to ensure that the content of a review is in line with its assigned score. Our definition, presented in Definition 5, is based on the intuitive idea that a review should include specific details and reasoning that support its assigned score. To establish the concept of coherence review in Definition 5, we adhere to the following guidelines.

– A good 5-star review emphasizes the positive aspects of a product or service, with any negative aspects mentioned being of minor importance.

- A 4-star review is generally positive, but with some reservations or criticisms. A good 4-star review should highlight the positive aspects while also mentioning areas for improvement.
- A 3-star review is considered neutral or mixed. A good 3-star review should provide an overall evaluation that takes both positive and negative aspects into account.
- A 2-star review typically has a negative assessment of the product, while allowing for minor positive aspects to be mentioned.
- The aim of a 1-star review is to warn others about a product. Therefore, the review should provide sufficient details and evidence to support the overall negative evaluation. However, it may also present some positive aspects to maintain balance.

Before introducing the concept of coherence review, the first step is to determine whether a review expresses a positive, negative, or mixed point of view about a product. This involves evaluating the sentiment of the content, and in our work, we use the terms 'polarity' and 'sentiment' interchangeably. Specifically, polarity refers to the orientation of the sentiment, such as 'positive', 'negative', or 'mixed point of view'.

In our study, we utilized Flair for sentiment analysis of the Amazon product review dataset. Flair is an open-source natural language processing (NLP) library that provides a simple and efficient way to develop state-of-the-art NLP models, including models for sentiment analysis [1]. Flair models are based on a combination of deep neural networks and traditional machine learning techniques, and are trained on large amounts of annotated data to achieve high accuracy and performance. The ease of use and versatility of Flair make it a valuable tool for researchers and practitioners in the field of NLP.

Definition 4. *Let r be a review. The $pol()$ function takes review as input and returns its sentiment, which is calculated using Flair*

$$pol(r) = \begin{cases} 1 & \text{if } r \text{ has a positive sentiment} \\ 0 & \text{if } r \text{ has a negative sentiment} \\ \frac{1}{2} & \text{if } r \text{ is mixed} \end{cases}$$

Since our goal is to define the coherency of a review based on its polarity, we have divided the level of coherency into two categories: positive coherency and negative coherency, as defined in Definition 5.

Definition 5. *Let r be a review, let n be the assigned score of r, and let $pol(r)$ be the sentiment of r.*

- *A review r is **positive coherent** if $n \in 4, 5$ and r has a positive polarity, i.e., $pol(r) = 1$.*
- *A review r is **negative coherent** if $n \in 1, 2$ and r has a negative polarity, i.e., $pol(r) = 0$. Alternatively, if $n = 3$ and r has a mixed polarity, i.e., $pol(r) = \frac{1}{2}$.*

a review is coherent if it is positive or negative coherent.

Definition 5 is based on the intuitive notion that a review should be coherent with its assigned score, meaning that the language used in the review should align with the sentiment conveyed by the score.

4.2 Associated ADF

The objective of this section is to categorize reviews into three classes based on two features: readability and coherence between the review text and the assigned score. The goal of this labeling problem is to determine the helpfulness of a review based on these two factors. We assign each review to one of the following categories: very helpful (denoted by H2), helpful (denoted by H1), or unhelpful (denoted by H0). In this task, we utilize Abstract Dialectical Frameworks as a means for categorizing reviews. By using ADFs, we can graphically depict the relationship between the features of a review and the label nodes. Since our goal is to categorize the reviews into one of three categories,i.e., $H0, H1, H2$, we have created three output variables, i.e., A_{H0}, A_{H1}, A_{H2}, respectively.

In labeling review, we consider two categorical features: readability, and coherence. We transform each feature into an argument, with readability transformed to A_{read}, and coherence to A_{coh}. We have introduced additional variables to show coherence between the review and the assigned score. These variables include A_{pol} to indicate the polarity of the review. For the assigned score, we have introduced variables A_s.

The aim of the classifier presented in Definition 6 is that if a review meets all criteria, i.e., readability and coherency, it belongs into category $H2$. If it meets any of the criteria, it belongs into category $H1$. If it fails to meet any of the criteria, it is in category $H0$ and is considered to be of unhelpful. Let $A_{out} = \{A_{H0}, A_{H1}, A_{H2}\}$ be the set of output arguments associated to the labels.

Definition 6. *Given review (r, n). The ADF (review labeling) associated with (r, n) is $D = (A, L, C)$, where:*

- $A = A_{out} \cup \{A_{pol}, A_s, A_{coh}, A_{read}\}$
- *The acceptance condition of any of the input arguments can be either $\top$ or $\bot$ w.r.t. the given review (r, n).*
  - $\varphi_{A_s} \equiv \top$ *if s is either 4 or 5. Otherwise, it is $\bot$.*
  - $\varphi_{A_{pol}} \equiv \top$ *if $pol(r) = 1$, and $\varphi_{A_{pol}} \equiv \bot$ otherwise.*
  - $\varphi_{A_{read}} \equiv \top$ *if $\mu - \sigma < read(r) < \mu + \sigma$, where μ is the the mean and σ is the standard deviation of the readability reviews;*
  - $\varphi_{A_{coh}} \equiv (A_{pol} \wedge A_s) \vee (\neg A_{pol} \wedge \neg A_s)$.
  - $\varphi_{A_{H0}} \equiv \neg A_{coh} \wedge \neg A_{read}$, $\varphi_{A_{H1}} \equiv (A_{coh} \wedge \neg A_{read}) \vee (\neg A_{coh} \wedge A_{read})$, *and* $\varphi_{A_{H2}} \equiv A_{coh} \wedge A_{read}$.

A review labeling ADF is depicted in Fig. 2

Note that in the ADF associated with a given review, the acceptance conditions indicate the set of links, so there is no need to explicitly present the set of links and their specific type. For instance, the acceptance condition of A_{coh},

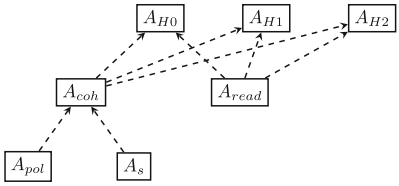

Fig. 2. A review labeling ADF

which is $A_{pol} \wedge A_s \vee \neg A_{pol} \wedge \neg A_s$, implies that A_{coh} is acceptable if the given review has either a positive sentiment and is rated 4 or 5-star, or a negative sentiment and is rated 1/2 or 3-star. Thus, to construct ADF D corresponding to review (r, n) there is no need of learning the type of links between arguments. All that is required is to specify the acceptance condition for the input arguments, which includes checking the sentiment and readability of a given review, as well as using the review score.

After constructing the ADF D associated with a given review (r, n), we evaluate its grounded semantics, denoted by $grd(D)$. The use of grounded semantics is justified by the following reasons: 1. In general, every ADF has a unique grounded interpretation. 2. Moreover, the grounded interpretation accepts arguments that cannot be rejected, rejects those that cannot be accepted, and does not take a stance on other arguments. 3. Specifically, Proposition 2 states that any output argument is either accepted or rejected in $grd(D)$, where $grd(D)$ is the grounded interpretation of the ADF D associated with a given review. Additionally, Proposition 3 asserts that only one output argument is accepted in $grd(D)$.

Proposition 1. *An argumentation framework (ADF) is considered acyclic if its corresponding graph does not contain any cycles. For a review (r, n), let D be its corresponding ADF. It holds that D is acyclic.*

Proof. Let (r, n) represent a review and let D be its corresponding ADF. The corresponding graph of D is denoted by G, which takes input arguments to indicate the readability, sentiment, and score of the review. Each of the input arguments is an initial argument, that is, it has no parents. The purpose links in D is to connect input arguments to output arguments, so edges in G can only be directed to deeper layers. Therefore, the associated graph G is acyclic.

Proposition 2. *Let (r, n) be a review, and let D be its corresponding ADF with the grounded interpretation $grd(D)$. In $grd(D)$, every output argument, namely A_{H0}, A_{H1}, and A_{H2}, is assigned either* **t** *or* **f**. *That is, any output argument is either accepted or rejected $grd(D)$.*

Proof. Proposition 1 proves that the ADF D corresponding to a review (r, n) is acyclic. According to [13], any acyclic ADF has a grounded interpretation in which each argument is assigned either **t** or **f**. Therefore, in the grounded

interpretation of D, every argument must be assigned either **t** or **f**, including every output argument.

Proposition 3. *Let $D = (A, L, C)$ be an ADF corresponding to a review (r, n). In the grounded interpretation of D, exactly one of the output arguments is assigned to* **t**.

Proof. Proposition 1 shows that D is an acyclic ADF. Therefore, according to [13], each argument in $grd(D)$ is either accepted or rejected. Specifically, A_{coh} and A_{read} are assigned either **t** or **f**.

Based on the truth table, there are four combinations of any truth values for A_{coh} and A_{read} where exactly one row has the value **t** and all other rows have **f**. On the other hand, the acceptance condition of each argument A_{H_0}, A_{H_1}, and A_{H_2} is a logical combination of its parents. Thus, the truth value of each A_{H_0}, A_{H_1}, and A_{H_2}, corresponds to one line of the truth table. Hence, exactly one output argument is assigned **t** and all other output arguments are assigned **f**.

After assessing the grounded interpretation of ADF D corresponding to review (r, n), we introduce a label function in Definition 7 to label a given review based on its helpfulness.

Definition 7. *(label-function) Let (r, n) be a review, and let $D = (A, L, C)$ be the associated ADF of (r, n). The label of (r, n) is the output of label-function $\mathcal{LF}(r, n)$, where the function $\mathcal{LF}(-, -)$ is as follows:*

$$\mathcal{LF}(r,n) = \begin{cases} H_0(\textit{i.e., unhelpful}) & \textit{if } grd(D)(A_{H_0}) \mapsto \mathbf{t} \\ H_1(\textit{i.e., helpful}) & \textit{if } grd(D)(A_{H_1}) \mapsto \mathbf{t} \\ H_2(\textit{i.e., very helpful}) & \textit{if } grd(D)(A_{H_2}) \mapsto \mathbf{t} \end{cases}$$

We need to show that the label-function is a well-defined function. Theorem 1 says that function $\mathcal{LF}$ is well-defined function.

Theorem 1. *Let (r, n) be a review, and let $D = (A, L, C)$ be the associated ADF of (r, n). Let $\mathcal{LF}(-, -)$ be a label-function, as Definition 7. It holds that $\mathcal{LF}(-, -)$ is a well-defined function.*

Proof. To demonstrate that $\mathcal{LF}(-, -)$ is a well-defined function, we need to establish that for any input (r, n), the function $\mathcal{LF}$ assigns it to exactly one of the outputs, namely H_0, H_1, or H_2. We leverage Proposition 3, which asserts that only one of these output arguments is acceptable in the grounded interpretation. In other words, either A_{H0}, A_{H1}, or A_{H2} is assigned to **t** in the grounded interpretation of D. Therefore, we can confirm that $\mathcal{LF}(r, n)$ has at least one output. Moreover, if $\mathcal{LF}(r, n)$ were to have more than one output, it would contradict Proposition 3. Thus, we conclude that $\mathcal{LF}$ is indeed a well-defined function.

In the labeling of the reviews we make two assumptions here, one is that the features are independent. That is, if the review is readable, it does not necessarily mean that the there exist a coherency between the text and the score. Another

assumption made here is that both of the features have equal effect on the outcome. That is, readability of the review does not have more importance in deciding the helpfulness of the review. We use bottom-up approach for assessing the label of the review in the associated ADF and we use top-bottom approach to explain the reason behind the chosen label.

Example 2. Let r be a 4-star review that says 'Largely my fault for not reading carefully, but the high synthetic fiber content made the shirt uncomfortable to the touch. Not the manufacturers fault, of course.' To construct the associated ADF for $(r, 4)$ as defined in Definition 6, we only need to specify the acceptance conditions of input arguments, as the acceptance conditions of other arguments are fixed. Since r expresses a negative sentiment and is a 4-star review, we have $\varphi_{A_{pol}} \equiv \bot, \varphi_{A_s} \equiv \top$. Additionally, considering the mean and standard deviation of the readability scores in the reviews, we have decided to set a threshold of 4.5. This means that if the readability of a review, denoted as $read(r)$, is less than 4.5, we define $\varphi_{A_{read}} \equiv \bot$.

The associated ADF $D = (A, L, C)$ for r is depicted in Fig. 3 (on the left). After indicating the acceptance conditions of all input arguments, we can assess the grounded interpretation of D. In the grounded interpretation of D, A_s, and A_{H_0} are assigned to **t**, and all other arguments to **f**. In Fig. 3, the arguments that are assigned to **t** in the grounded interpretation are colored blue, and the arguments assigned to **f** are depicted in gray.

Note that in the grounded interpretation of D, A_{coh} is also assigned to **f** in $grd(D)$ because $\varphi_{A_{coh}}$ is equivalent to $(A_{pol} \wedge A_s) \vee (\neg A_{pol} \wedge \neg A_s)$. Furthermore, since $\varphi_{A_{H_0}}$ is equivalent to $\neg A_{coh} \wedge \neg A_{read}$, A_{H_0} is assigned to **t** in $grd(D)$.

After evaluating the grounded interpretation of D, we apply the label-function, as defined in Definition 7, to determine the label of $(r, 4)$. Since A_{H_0} is assigned to **t** in the grounded interpretation, the result of $\mathcal{LF}(r, n)$ is H_0, indicating that $(r, 4)$ is labeled as unhelpful.

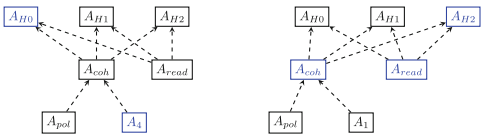

Fig. 3. The ADFs corresponding to the label evaluation of $(r, 4)$ (on the left) and $(r', 1)$ (on the right) in Example 2. The arguments that are assigned to **t** in the grounded interpretation are colored blue, while the arguments assigned to **f** are depicted in gray. (Color figure online)

Here is an example of our labeling method assigning a "very helpful" label to a 1-star review. The review in question is r' : 'After using this shoes seven times for regular activities like walking or dancing the sole came out. I am very disappointed. I don't recommend this product at all. The worse part is that I can not return this item because the return window closed 4 days ago.'

The acceptance conditions for the input arguments in the associated are as follows: $\varphi_{A_{neg}} \equiv \bot$ because the review has a negative sentiment, $\varphi_{A_1} \equiv \bot$ because r is a 1-star review. Furthermore, $\varphi_{A_{read}} \equiv \top$ because $read(r) = 4.566721$. Based on the acceptance conditions specified in the associated ADF, the grounded interpretation assigns the following arguments to the **t**, namely, A_{read}, A_{coh}, and A_{H_2}. All other arguments are assigned to the **f**. Therefore, the output label of $\mathcal{LF}(r', 1)$ is H_2, indicating that the review r' is labeled as 'very helpful' according to our labeling method.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F1Macro	TT (Sec)
lr	Logistic Regression	0.9074	0.6426	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0420
ridge	Ridge Classifier	0.9074	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0300
dummy	Dummy Classifier	0.9074	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0280
svm	SVM - Linear Kernel	0.9035	0.0000	0.0000	0.0000	0.0000	-0.0065	-0.0083	0.4746	0.0340
lda	Linear Discriminant Analysis	0.8958	0.6485	0.0000	0.0000	0.0000	-0.0197	-0.0265	0.4725	0.0620
knn	K Neighbors Classifier	0.8957	0.5315	0.0400	0.2000	0.0667	0.0380	0.0558	0.5057	0.0520
rf	Random Forest Classifier	0.8919	0.5927	0.0000	0.0000	0.0000	-0.0243	-0.0301	0.4714	0.4060
lightgbm	Light Gradient Boosting Machine	0.8918	0.6600	0.0400	0.1000	0.0571	0.0278	0.0311	0.4998	0.0660
xgboost	Extreme Gradient Boosting	0.8803	0.6019	0.0400	0.0500	0.0444	0.0033	0.0022	0.4902	0.0940
gbc	Gradient Boosting Classifier	0.8727	0.6387	0.0000	0.0000	0.0000	-0.0519	-0.0596	0.4660	0.2040
et	Extra Trees Classifier	0.8726	0.5998	0.0500	0.0500	0.0500	-0.0032	-0.0102	0.4908	0.3380
ada	Ada Boost Classifier	0.8650	0.5782	0.0000	0.0000	0.0000	-0.0590	-0.0654	0.4637	0.2060
qda	Quadratic Discriminant Analysis	0.8569	0.5187	0.0400	0.1000	0.0571	-0.0044	-0.0008	0.4896	0.0540
dt	Decision Tree Classifier	0.8263	0.5139	0.1300	0.1019	0.1133	0.0187	0.0193	0.5085	0.0300
nb	Naive Bayes	0.8183	0.6762	0.1200	0.1200	0.1200	0.0221	0.0202	0.5093	0.0320

Fig. 4. Performance Metrics and Comparison of Classification Models without Incorporating the Outcomes of Sect. 4 as features.

5 Statistical Analysis and Experimental Results

We evaluated our proposed method using the Amazon Fashion 5 reviews dataset by McAuley et al. [18]. This dataset comprises a collection of fashion product reviews. After eliminating duplicates, we were left with a total of 371 reviews. Among these, only 35 received votes from readers. We consider these voted reviews to be valuable indicators of helpfulness due to their resonance with readers.

Our primary objective was to assess how our labeling methods affect classifier performance. To achieve this, we initially evaluated the impact of various statistical measures, such as readability, sentiment analysis, and polarity scores, on review classification. The results are illustrated in Fig. 4, revealing that all classifiers performed poorly when classifying voted reviews.

Subsequently, we introduced two new features, "helpful reviews" and "coherent reviews", as they are presented in this work, to the classifiers. Our investigation, depicted in Fig. 5, indicated that by incorporating these features, the classifiers exhibited improved performance in labeling helpful reviews.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	F1Macro	TT (Sec)
ridge	Ridge Classifier	0.9074	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0520
qda	Quadratic Discriminant Analysis	0.9074	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0300
dummy	Dummy Classifier	0.9074	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.4757	0.0280
knn	K Neighbors Classifier	0.9073	0.5996	0.0400	0.2000	0.0667	0.0558	0.0776	0.5089	0.0840
lr	Logistic Regression	0.9035	0.6940	0.0000	0.0000	0.0000	-0.0066	-0.0091	0.4746	0.0840
svm	SVM - Linear Kernel	0.8959	0.0000	0.0000	0.0000	0.0000	-0.0183	-0.0222	0.4725	0.0500
lda	Linear Discriminant Analysis	0.8881	0.7496	0.0400	0.2000	0.0667	0.0345	0.0563	0.5034	0.0320
ada	Ada Boost Classifier	0.8845	0.7444	0.1800	0.3067	0.2233	0.1677	0.1752	0.5804	0.1180
gbc	Gradient Boosting Classifier	0.8804	0.6614	0.1800	0.3467	0.2056	0.1486	0.1727	0.5704	0.1120
xgboost	Extreme Gradient Boosting	0.8804	0.6903	0.0900	0.1667	0.1143	0.0619	0.0639	0.5251	0.0580
rf	Random Forest Classifier	0.8802	0.6848	0.0000	0.0000	0.0000	-0.0377	-0.0407	0.4681	0.3660
lightgbm	Light Gradient Boosting Machine	0.8689	0.7151	0.1300	0.1567	0.1416	0.0765	0.0744	0.5352	0.0720
et	Extra Trees Classifier	0.8610	0.6522	0.0400	0.1000	0.0571	-0.0096	-0.0066	0.4910	0.1980
dt	Decision Tree Classifier	0.8032	0.4962	0.1200	0.1050	0.1108	0.0065	0.0035	0.4996	0.0580
nb	Naive Bayes	0.3820	0.7738	0.9600	0.1277	0.2250	0.0737	0.1783	0.3547	0.0560

Fig. 5. Performance Metrics and Comparison of Classification Models with Incorporating the Outcomes of Sect. 4 as features.

6 Explanation in the Form of Dialogic Argumentation

Explanation involves the process of providing clarification or elucidation for a concept or argument. While an expert in formal argumentation can swiftly grasp the logic behind the labeling through the structure of the associated ADF, the labeling method may prove intricate for agents unfamiliar with it.

Our approach to labeling reviews involves utilizing a bottom-up approach to construct the associated ADF of a given review while taking a top-down approach to explain the decision. ADFs serve as the underlying structure for generating dialogical explanations. In each ADF the acceptance condition of each argument is presented using a propositional formula, which provides a powerful approach for explanation. However, it is important to present the explanation process in a clear and understandable manner for users who may not be familiar with ADFs. One of the main reasons for explaining the reasoning behind a labeling decision is to provide feedback to the review writer.

In order to address the explanation gap in connecting review labels to their features, we have defined explanation function. A user can interact with the

system by requesting an explanation about the label of a review. To help illustrate our explanation method before presenting formal definitions, we provide an example in Example 3.

Example 3. Consider the review $(r, 4)$ presented in Example 2, where r is 'Largely my fault for not reading carefully, but the high synthetic fiber content made the shirt uncomfortable to the touch. Not the manufacturers fault, of course.' The labeling function assigned an unhelpful label to $(r, 4)$. A dialogic argumentation to explain the reason behind the labeling is as follows.

Agent: Why was $(r, 4)$ labeled as unhelpful?

System: The labeling function assigned the label H_0 to $(r, 4)$ because A_{H_0}, referring to unhelpful reviews, was accepted in the grounded interpretation of the associated ADF of $(r, 4)$.

Agent: Why was the argument A_{H_0}, referring to unhelpful reviews, accepted for $(r, 4)$?

System: The acceptance condition for A_{H_0} is $\neg A_{coh} \wedge \neg A_{read}$, which means that A_{H_0} is acceptable if and only if the review is neither readable nor coherent. In this case, $(r, 4)$ was neither readable nor coherent, hence argument A_{H_0} was accepted for $(r, 4)$.

Agent: Why was r considered not readable?

System: r was considered not readable because its readability score was below the chosen threshold, which in this case was lower than the mean mineus standard deviation of the readability score of the reviews.

Agent: Why was $(r, 4)$ considered not coherent?

System: $(r, 4)$ was considered not coherent because it expressed a negative sentiment while being a 4-star review, which implies that the review should focus on the positive aspects of the product. Therefore, the review did not meet the coherence criterion set by the labeling function.

Note that an unfamiliar agent may not be interested in the intricate details of the formal framework that leads to the system decision. Instead, they may be more interested in understanding the factors that influence the system's final decision. To address this, we introduce two different functions: the Parents Explanation Function and the Factors Explanation Function.

The Parents Explanation Function provides detailed information about each step of the decision-making process, as demonstrated in Example 3. On the other hand, the Factors Explanation Function focuses on presenting the specific factors that play a role in the final labeling decision.

The purpose of the *parents explanation* function is to elucidate the connection between the label of review r and the truth value of its parents. Additionally, the following function can be employed to explain the truth value of the parents of A_{coh}. This means that an agent can utilize the parents function to inquire about the rationale behind the labeling of a given review, as well as to inquire about the coherence of the review.

Note that, in the following $Par(a)$ is the set of parents of a as it is presented in Sect. 3, and $grd(a)$ indicate the truth value of a in the grounded interpretation.

Definition 8. *(parents explanation) For any review (r, n) with corresponding ADF $D = (A, L, C)$ and label $\mathcal{LF}(r, n)$, an argumentation dialogue can be initiated to explain the label of the review or the coherence of the review. This dialogue involves a user requesting an explanation using $\mathcal{Q}(a)$, to which the system responds with an explanation $\mathcal{X}_a$, where $a \in \{A_{H_0}, A_{H_1}, A_{H_2}, A_{coh}\}$. The request takes the form of, $\mathcal{Q}(a) = \{$ Why is grd$(a)?\}$*

The label explanation $\mathcal{X}_{\mathcal{LF}}(a)$ clarifies that $\{a$ is grd(a) because grd$(Par(a))\}$.

As an agent, it may seem vague to hear about the truth value of the argument in the grounded interpretation of D. An agent may ask about the factors that has effect on the labeling by using factor explanation function, presented in Definition 9.

Definition 9. *Let $\mathcal{LF}(r, n)$ be the label of review (r, n). The inquiry can be formulated as $\mathcal{QF}(\mathcal{LF}(r, n)) = \{$ Why is (r, n) labeled as $\mathcal{LF}(r, n)?\}$*
The factor explanation $\mathcal{F}_{\mathcal{LF}}(\mathcal{LF}(r, n))$ clarifies that $\{(r, n)$ is labeled as $\mathcal{LF}(r, n)$ because grd(A_{pol}), grd(A_s), and grd$(A_{read})\}$.

7 Conclusion and Future Work

The vast number of online reviews can significantly impact a user's decision-making process. However, utilizing online information presents challenges, particularly in determining which reviews are genuinely helpful. In this study, we address this challenge by employing formal argumentation, specifically Abstract Dialectical Frameworks (ADFs), to identify helpful reviews. The labeling of a review as helpful is based on two key features: the coherence between the review and its score, and the readability of the review. We show that our labeling method has a demonstrably positive effect on the performance of the classifiers.

One of the advantages of ADFs is their ability to incorporate additional arguments and associated relations easily. For our future work, we aim to incorporate additional features of a review, such as text complexity, to enable more precise labeling of reviews. Moreover, in the current study, we classify reviews into three categories: very useful, helpful, and unhelpful. By considering further features of the reviews, we can potentially introduce additional categories of helpfulness or even present the degree of helpfulness in a continuous manner. However, it is important to acknowledge that a more precise labeling entails a more complex combination of the input factors.

Furthermore, in our current work, we employ the Flair library to determine the sentiments of the reviews. We are specifically interested in splitting the reviews into smaller chunks and evaluating the sentiment of each chunk within a review. This approach enables us to utilize the sentiment of each chunk as an input argument. By doing so, we can define the notion of coherence between the review and its assigned score more precisely. Additionally, this allows us to provide a more detailed and vivid explanation of the labeling process to agents, enhancing their understanding of how reviews are labeled.

*We have chosen ADFs as the formalism to model the relationship between review features and their corresponding labels. ADFs offer the advantage of representing various types of relations between arguments. Unlike other formalisms, ADFs do not require explicit indication of argument relations. Instead, these relations can be implicitly represented by attaching acceptance conditions to arguments in the form of propositional formulas.

In Definition 6, we present the extended form of ADFs specifically tailored for reviews. For a given review, we only need to specify the acceptance conditions of input arguments based on the sentiment, readability, and score of the review, while the rest of the model remains fixed. By using the grounded semantics of ADFs, we can evaluate the label arguments.

We used propositional formulas attached to the arguments in ADFs to provide an explanation of the label of a given review. This explanation can serve as feedback to a writer, especially when their review is labeled as unhelpful in our setting. In Example 3, we provide an explanation for a 4-star review that is labeled as unhelpful in out setting.

While we employed a bottom-up reasoning approach to assessing the label of a review based on the grounded interpretation of the associated ADF, we used a top-down approach to explain the reasons behind the labeling. Labeling reviews can be the first step in assessing the trustworthiness of reviews. After excluding unhelpful reviews, we aim to use formalisms of argumentation to identify the set of consistent reviews. Additionally, we are interested in evaluating product strengths and weaknesses using argumentation frameworks.

As our current research has focused on labeling online product reviews, we are interested in investigating whether the same methods can be applied to other types of online information such as tweets.

Acknowledgments. This research has been supported by the Netherlands eScience Center project "The Eye of the Beholder" (project number 027.020.G15), and the first researcher is also supported by the Netherlands Organisation for Scientific Research (NWO) through the Hybrid Intelligence Gravitation Programme with project number 024.004.022.

References

1. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: Flair: an easy-to-use framework for state-of-the-art nlp. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pp. 54–59 (2019)
2. Brauwers, G., Frasincar, F.: A survey on aspect-based sentiment classification. ACM Comput. Surv. **55**(4), 1–37 (2022)
3. Brewka, G., Ellmauthaler, S., Strass, H., Wallner, J.P., Woltran, S.: Abstract dialectical frameworks: an overview. In: Handbook of Formal Argumentation, pp. 237–285. College Publications (2018)
4. Brewka, G., Ellmauthaler, S., Strass, H., Wallner, J.P., Woltran, S.: Abstract dialectical frameworks: an overview. IFCoLog J. Logics Appl. (FLAP) **4**(8) (2017)

5. Brewka, G., Woltran, S.: Abstract dialectical frameworks. In: Proceedings of the Twelfth International Conference on the Principles of Knowledge Representation and Reasoning (KR 2010), pp. 102–111 (2010)
6. Ceolin, D., Primiero, G., Soprano, M., Wielemaker, J.: Transparent assessment of information quality of online reviews using formal argumentation theory. Inf. Syst. **110**, 102107 (2022)
7. Ceolin, D., Primiero, G., Wielemaker, J., Soprano, M.: Assessing the quality of online reviews using formal argumentation theory. In: Brambilla, M., Chbeir, R., Frasincar, F., Manolescu, I. (eds.) ICWE 2021. LNCS, vol. 12706, pp. 71–87. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-74296-6_6
8. Chi, H., Liao, B.: A quantitative argumentation-based automated explainable decision system for fake news detection on social media. Knowl. Based Syst. **242**, 108378 (2022)
9. Cocarascu, O., Rago, A., Toni, F.: Extracting dialogical explanations for review aggregations with argumentative dialogical agents. In: AAMAS, pp. 1261–1269. International Foundation for Autonomous Agents and Multiagent Systems (2019)
10. Cocarascu, O., Toni, F.: Detecting deceptive reviews using argumentation. In: PrAISe@ECAI, pp. 9:1–9:8. ACM (2016)
11. Čyras, K., et al.: Explanations by arbitrated argumentative dispute. Expert Syst. Appl. **127**, 141–156 (2019)
12. Dale, E., Chall, J.S.: A formula for predicting readability: instructions. Educ. Res. Bull., 37–54 (1948)
13. Diller, M., Keshavarzi Zafarghandi, A., Linsbichler, T., Woltran, S.: Investigating subclasses of abstract dialectical frameworks. Argum. Comput. **11**(1–2), 191–219 (2020)
14. García, A.J., Chesñevar, C.I., Rotstein, N.D., Simari, G.R.: Formalizing dialectical explanation support for argument-based reasoning in knowledge-based systems. Expert Syst. Appl. **40**(8), 3233–3247 (2013)
15. Zafarghandi, A.K., Ceolin, D.: Fostering explainable online review assessment through computational argumentation. In: ArgXAI@COMMA, vol. 3209 of CEUR Workshop Proceedings. CEUR-WS.org (2022)
16. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch (1975)
17. Lacave, C., Díez, F.J.: A review of explanation methods for heuristic expert systems. Knowl. Eng. Rev. **19**(2), 133–146 (2004)
18. 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)
19. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. arXiv preprint arXiv:1107.4557 (2011)
20. Prakken, H.: On dialogue systems with speech acts, arguments, and counterarguments. In: Ojeda-Aciego, M., de Guzmán, I.P., Brewka, G., Moniz Pereira, L. (eds.) JELIA 2000. LNCS (LNAI), vol. 1919, pp. 224–238. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-40006-0_16
21. Rago, A., Cocarascu, O., Toni, F.: Argumentation-based recommendations: Fantastic explanations and how to find them. In: IJCAI, pp. 1949–1955. ijcai.org (2018)
22. Wang, Y., Wang, J., Yao, T.: What makes a helpful online review? a meta-analysis of review characteristics. Electron. Commer. Res. **19**, 257–284 (2019)