# Using Argument-based Features to Predict and Analyse Review Helpfulness

Haijing Liu[1,2], Yang Gao[1], Pin Lv[1], Mengxue Li[1,2], Shiqiang Geng[3], Minglan Li[1,2] and Hao Wang[4]

[1]Institute of Software, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]School of Automation, Beijing Information Science and Technology University
[4]Qihoo 360 Search Lab
{*haijing,mengxue,minglan*}*2015@iscas.ac.cn*
{*gaoyang,lvpin*}*@iscas.ac.cn,* {*gsq_study,cashenry*}*@126.com*

## Abstract

We study the helpful product reviews identification problem in this paper. We observe that the evidence-conclusion discourse relations, also known as *arguments*, often appear in product reviews, and we hypothesise that some *argument*-based features, e.g. the percentage of argumentative sentences, the evidences-conclusions ratios, are good indicators of helpful reviews. To validate this hypothesis, we manually annotate arguments in 110 hotel reviews, and investigate the effectiveness of several combinations of argument-based features. Experiments suggest that, when being used together with the argument-based features, the state-of-the-art baseline features can enjoy a performance boost (in terms of F1) of 11.01% in average.

## 1 Introduction

Product reviews have significant influences on potential customers' opinions and their purchase decisions (Chatterjee, 2001; Chen et al., 2004; Dellarocas et al., 2004). Instead of reading a long list of reviews, customers usually are only willing to view a handful of *helpful reviews* to make their purchase decisions. In other words, helpful reviews have even greater influences on the potential customers' decision-making processes and thus on the sales; as a result, the automatic identification of helpful reviews has received considerable research attentions in recent years (Kim et al., 2006; Liu et al., 2008; Mudambi, 2010; Xiong and Litman, 2014; Martin and Pu, 2014; Yang et al., 2015, 2016).

Existing works on helpful reviews identification mostly focus on designing efficient features.

Widely used features include *external features*, (e.g. date (Liu et al., 2008), product rating (Kim et al., 2006) and product type (Mudambi, 2010)) and *intrinsic features* (e.g. semantic dictionaries (Yang et al., 2015) and emotional dictionaries (Martin and Pu, 2014)). Compared to external features, intrinsic features can provide some insights and explanations for the prediction results, and support better cross-domain generalisation. In this work, we investigate a new form of intrinsic features: the *argument* features.

An argument is a basic unit people use to persuade their audiences to accept a particular state of affairs (Eckle-Kohler et al., 2015). An argument usually consists of a *claim* (also known as *conclusion*) and some *premises* (also known as *evidences*) offered in support of the claim. For example, consider the following review excerpt: "*The staff were amazing, they went out of their way to help us*"; the texts before the comma constitute a claim, and the texts after the comma give a premise supporting the claim. Argumentation mining (Moens, 2013; Lippi and Torroni, 2016) receives growing research interests in various domains (Palau and Moens, 2009; Contractor et al., 2012; Park and Cardie, 2014; Madnani et al., 2012; Kirschner et al., 2015; Wachsmuth et al., 2014, 2015). Recent advances in automatic arguments identification (Stab and Gurevych, 2014), has stimulated the usage of argument features in multiple domains, e.g. essay scoring (Wachsmuth et al., 2016) and online forum comments ranking (Wei12 et al., 2016).

The motivation of this work is a hypothesis that, the helpfulness of a review is closely related to some argument-related features, e.g. the percentage of argumentative sentences, the average number of premises in each argument, etc. To validate our hypothesis, we manually annotate arguments in 110 hotel reviews so as to use

these "ground truth" arguments to testify the effectiveness of argument-based features for detecting helpful hotel reviews. Empirical results suggest that, for four baseline feature sets we test, their performances can be improved, in average, by 11.01% in terms of F1-score and 10.40% in terms of AUC when they are used together with some argument-based features. Furthermore, we use the effective argument-based features to give some insights into which product reviews are more helpful.

## 2   Corpus

We use the Tripadvisor hotel reviews corpus built by (O'Mahony and Smyth, 2010) to test the performance of our helpful reviews classifier. Each entry in this corpus includes the review texts, the number of people that have viewed this review (denoted by Y) and the number of people that think this review is helpful (denoted by X).

We randomly sample 110 hotel reviews from this corpus to annotate the "ground truth" argument structures [1]. In line with (Wachsmuth et al., 2015), we view each sub-sentence in the review as a *clause* and ask three annotators independently to annotate each clause as one of the following seven argument components:

*Major Claim:* a summary of the main opinion of a review. For instance, *"I have enjoyed the stay in the hotel", "I am sad to say that i am very disappointed with this hotel"*;

*Claim:* a subjective opinion on a certain aspect of a hotel. For example, *"The staff was amazing", "The room is spacious"*;

*Premise:* an objective reason/evidence supporting a claim. For instance, *"The staff went out of their way to help us"*, it supports the first example claim above; *"We had a sitting room as well as a balcony"*, it supports the second example claim above;

*Premise Supporting an Implicit Claim (PSIC):* an objective reason/evidence that supporting an *implicit claim*, which does appear in review. For instance, *"just five minutes' walk to the down town"* supports some implicit claims like *"the location of the hotel is good"*, although this implicit claims has never appeared in the review;

*Background:* an objective description that does not give direct opinions but provides some back-

---

[1] The annotated corpus can be obtained by contacting the first author

| Component Type | Number | Kappa |
|---|---|---|
| Major claim | 143 | 0.86 |
| Claim | 581 | 0.77 |
| Premise | 206 | 0.65 |
| PSIC | 121 | 0.94 |
| Background | 80 | 0.89 |
| Recommendation | 70 | 1.00 |
| Non-argumentative | 145 | 0.78 |

Table 1: The number and Fleiss' kappa for each argument component type we annotate.

ground information. For example, *"We checked into this hotel at midnight", "I stayed five nights at this hotel because i was attending a conference at the hotel"*;

*Recommendation:* a positive or negative recommendation for the hotel. For instance, *"I would definitely come to this hotel again the next time I visit London", "Do not come to this hotel if you look for some clean places to live"*;

*Non-argumentative:* for all the other clauses.

We use the Fleiss' kappa metric (Fleiss, 1971) to evaluate the quality of the obtained annotations, and the results are presented in Table 1. We can see that the lowest Kappa scores (for Premise) is still above 0.6, suggesting that the quality of the annotations are *substantial* (Landis and Koch, 1977); in other words, there exist little noises in the ground truth argument structures. We aggregate the annotations using majority voting.

## 3   Features

In line with (Yang et al., 2015), we consider the helpfulness as an intrinsic characteristic of product reviews, and thus only consider the following four intrinsic features as our baseline features.

Structural features (**STR**) (Kim et al., 2006; Xiong and Litman, 2014): we use the following structural features: total number of tokens, total number of sentences, average length of sentences, number of exclamation marks, and the percentage of question sentences.

Unigram features (**UGR**) (Kim et al., 2006; Xiong and Litman, 2014): we remove all stopwords and non-frequent words ($tf < 3$) to build the unigram vocabulary. Each review is represented by the vocabulary with *tf-idf* weighting for each appeared term.

Emotional features (**GALC**) (Martin and Pu, 2014): the Geneva Affect Label Coder (GALC) dictionary proposed by (Scherer, 2005) defines 36 emotion states distinguished by words. We build a

real feature vector with the number of occurrences of each emotional word plus one additional dimension for the number of non-emotional words.

Semantic features (**INQUIRER**) (Yang et al., 2015): the General Inquirer (INQUIRER) dictionary proposed by (Stone et al., 1962) maps each word to some semantic tags, e.g. word *absurd* is mapped to tags NEG and VICE; similar to the GALC features, the semantic features include the number of occurrences of each semantic tag.

### 3.1 Argument-based Features

The argument-based features can have different granularity: for example, the number of argument components can be used as features, and the number of tokens (words) in the argument components can also be used as features. We consider four granularity of argument features, detailed as follows.

**Component-level argument features**. A natural feature that we believe to be useful is the ratio of different argument component numbers. For example, we may be interested in the ratio between the number of premises and that of claims; a high ratio suggests that there are more premises supporting each claim, indicating that the review gives many evidences. To generalise this component ratio feature, we propose *component-combination ratio* features: we compute the ratios between any two argument components combinations. For example, we may be interested in the ratio between the number of MajorClaim+Claim+Premise and that of Background+Non-argumentative. As there are 7 types of labels, the number of possible combinations is $2^7 - 1 = 127$, and thus the possible number of combination ratio pairs is $127 \times 126 = 16002$. In other words, the component-level feature is a 16002-dimensional real vector.

**Token-level argument features**. In a finer-granularity, we consider the number of tokens in argument components to build features: for example, suppose a review has only two claims, one has 10 words and the other has 5 words; we may want to know the average number of words contained in each claim, the total number of words in claims, etc. In total, for each argument component type, we consider 5 types of token-level statistics: the total number of words in the given component type, the length (in terms of word) of the shortest/longest component of the given type, and the mean/variance of the number of words in each component of the given type. Thus, there are in total $7 \times 5 = 35$ features to represent the token-level statistics.

In addition, the ratio of some token-level statistics may also be of interests: for example, given a review, we may want to know the ratio between the number of words in Claims+MajorClaims and that in Premises. Thus, the combination ratio can also be applied here. We consider only the combination ratio for two statistics: the total number of words and the average number of words in each component-combination; hence, there are $16002 \times 2 = 32004$ dimensions for the combination ratio for the statistics. In total, there are $32004 + 35 = 32039$ dimensions for the token-level argument features.

**Letter-level argument features**. In the finest-granularity, we consider the letter-level features, which may give some information the token-level features do not contain: for example, if a review has a big number of letters and a small number of words, it may suggests that many long and complex words are used in this review, which, in turn, may suggests that the linguistic complexity of the review is relative high and the review may gives some very professional opinions. Similar to the token-level features above, we design 5 types of statistics and their combination ratios. Thus, the dimension for the letter-level features is the same to that of the token-level features.

**Position-level argument features**. Another dimension to consider argument features is the positions of argument components: for example, if the major claims of a review are all at the very beginning, we may think that readers can more easily grasp the main idea of the review and, thus, the review is more likely to be helpful. For each component, we use a real number to represent its position: for example, if a review has 10 sub-sentences (i.e. clauses) in total and the first sub-sentence the component overlaps is the second sub-sentence, then the position for this component is $2/10 = 0.2$. For each type of argument component, we may be interested in some statistics for its positions: for example, if a review has several premises, we may want to know the location of the earliest/latest appearance of premises, the average position of all premises and its variance, etc. Similar to the token- and letter-level features, we design the same number of features for position-

|  | Accuracy | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| AF | 0.617 | 0.625 | 0.617 | 0.620 | 0.611 |
| STR | 0.600 | 0.360 | 0.600 | 0.450 | 0.500 |
| STR+AF | 0.604 | 0.614 | 0.604 | 0.607 | 0.599 |
| UGR | 0.697 | **0.760** | 0.697 | 0.646 | 0.627 |
| UGR+AF | **0.718** | 0.718 | **0.719** | **0.717** | **0.706** |
| GALC | 0.621 | 0.605 | 0.621 | 0.579 | 0.560 |
| GALC+AF | 0.647 | 0.654 | 0.647 | 0.649 | 0.640 |
| INQUIRER | 0.533 | 0.512 | 0.533 | 0.517 | 0.493 |
| INQUIRER+AF | 0.657 | 0.664 | 0.657 | 0.659 | 0.651 |

Table 2: Helpful reviews identification performances using argument-based features and/or baseline features. AF stands for argument-based features.

level features.

## 4 Experiments

Following (O'Mahony and Smyth, 2010; Martin and Pu, 2014), we model the helpfulness prediction task as a classification problem; thus, we use accuracy, precision, recall, macro F1 and area under the curve (AUC) to as evaluation metrics. Similar to (O'Mahony and Smyth, 2010), we consider a review as helpful if and only if at least 75% opinions for the review are positive, i.e. $X/Y \geq 0.75$ (see X and Y in Sect. 2). For the features whose number of dimensions is more than 10k (i.e. the UGR features and argument-based features), to reduce their dimensions and to improve the performance, we only use the positive-information-gain features in these feature sets. In line with most existing works on helpfulness prediction (Martin and Pu, 2014; Yang et al., 2015), we use the LibSVM (Chang and Lin, 2011) as our classifier.

The performances of different features are presented in Table 2. Each number in the table is the average performance in 10-fold cross-validation tests. From the table we can see that, when being used together with the argument-based features, either of the four baseline features enjoys a performance boost in terms of all metrics we consider. To be more specific, in terms of accuracy, precision, recall, F1 and AUC, the average improvement for the baseline features are 4.33%, 10.30%, 4.32%, 11.01% and 10.40%, respectively. However, we observe that the precision of UGR+AF, although gives the second highest score among all feature combinations, is lower than that of UGR; we leave it for future work. Also, we notice that when using the argument-based features alone, its performance (in terms of Precision, F1 and AUC) is superior to those of STR, GALC and INQUIRER, and is only inferior to UGR. However, a major drawback of the UGR feature is its huge and document-dependent dimensionality, while the dimensionality of argument-based features is fixed, regardless of the size of the input documents. Moreover, the UGR features are sparse and problematic in online learning. To summarise, compared with the other state-of-the-art features, argument-based features are effective in identifying helpful reviews, and can represent some complementary information that cannot be represented in other features.

## 5 What Makes a Review Helpful ?

Argument-based features can not only improve the performance of review helpfulness identification, but also can be used to interpret what makes a review helpful. We analyse the information gain ranking of the argument-based features and find that, among all the positive-information-gain argument features, 36% are from the token-level argument feature set, and 29% are from the letter-level argument feature set, suggesting that these two feature sets are most effective in identifying helpful reviews. Among all the token-level argument features with positive information gain, 69% are ratios of sum of token number between component-combinations, and the remaining are ratios of the mean token numbers between component-combinations. We interpret this observation as follows: given a review, the larger number of tokens it contains, and the more likely the review is helpful. In fact, helpful reviews are tend to occur in those long reviews, which generally provide with more experiences and comments about the product being reviewed. Among all the letter-level argument features, around three-quarters are ratios of the sum of the number of letters between component-combinations. This observation, again, suggests that the length of reviews plays an important role in the review helpfulness identification.

Moreover, among all the argument-based features with positive information gain values, a quarter of features are the position-level argument feature. This is because the position of each argument component influences the logic flow of reviews, which, in turn, influences the readability, convincingness and helpfulness of the reviews. This information can hardly be represented by all the baseline features we considered, and we believe this explains why the performances of the baseline features are improved when being used together with the argument-based features. However, among all the argument-based features with positive information gain values, only 10% are the component-level argument feature. This indicates that compared to three finer-granularity argument features above, the component-level argument feature provides less useful information in review helpfulness identification.

## 6 Conclusion and Future Work

In this work, we propose a novel set of intrinsic features of identifying helpful reviews, namely the *argument*-based features. We manually annotate 110 hotel reviews, and compare the performances of argument-based features with those of some state-of-the-art features. Empirical results suggest that, argument-based features include some complementary information that the other feature sets do not include; as a result, for each baseline feature, the performance (in terms of various metrics) of jointly using this feature and argument-based features is higher than using this baseline feature alone. In addition, by analysing the effectiveness of different argument-based features, we give some insights into which reviews are more likely to be helpful, from an argumentation perspective.

For future work, an immediate next step is to explore the usage of automatically extracted arguments in helpful reviews identification: in this work, all argument-based features are based on manually annotated arguments; deep-learning based argument mining (Li et al., 2017; Eger et al., 2017) has produced some promising results recently, and we plan to investigate whether the automatically extracted arguments can be used to identify helpful reviews, and how the errors made in the argument extraction stage will influence the performance of helpful reviews identification. We also plan to investigate the effectiveness of argument-based features in other domains.

## References

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Patrali Chatterjee. 2001. Online reviews: Do consumers use them? *ACR North American Advances*.

Pei-Yu Chen, Shin-yi Wu, and Jungsun Yoon. 2004. The impact of online recommendations and consumer feedback on sales. *ICIS 2004 Proceedings*, page 58.

Danish Contractor, Yufan Guo, and Anna Korhonen. 2012. Using argumentative zones for extractive summarization of scientific articles. In *coling*, volume 12, pages 663–678.

Chrysanthos Dellarocas, Neveen Awad, and Xiaoquan Zhang. 2004. Exploring the value of online reviews to organizations: Implications for revenue forecasting and planning. *ICIS 2004 Proceedings*, page 30.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *EMNLP*, pages 2236–2242.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430. Association for Computational Linguistics.

Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *ArgMining@ HLT-NAACL*, pages 1–11.

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

Minglan Li, Yang Gao, Hui Wen, Yang Du, Haijing Liu, and Hao Wang. 2017. Joint rnn model for argument component boundary detection. *arXiv preprint arXiv:1705.02131*.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Data mining, 2008. ICDM'08. Eighth IEEE international conference on*, pages 443–452. IEEE.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28. Association for Computational Linguistics.

Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI-14)*, EPFL-CONF-210749.

Marie-Francine Moens. 2013. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, page 2. ACM.

Susan M Mudambi. 2010. What makes a helpful online review? astudy of customer reviews on amazon. com. *MIS Quarterly*, 34(1):185–200.

Michael P O'Mahony and Barry Smyth. 2010. A classification-based review recommender. *Knowledge-Based Systems*, 23(4):323–329.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *ArgMining@ ACL*, pages 29–38.

Klaus R Scherer. 2005. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*, pages 46–56.

Philip J Stone, Robert F Bales, J Zvi Namenwirth, and Daniel M Ogilvie. 1962. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. *Behavioral Science*, 7(4):484–498.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*, pages 1680–1691.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow-a general model of web review argumentation. In *EMNLP*, pages 601–611.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, and Gregor Engels. 2014. Modeling review argumentation for robust sentiment analysis. In *COLING*, pages 553–564.

Zhongyu Wei12, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in the online forum. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 195.

Wenting Xiong and Diane J Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *COLING*, pages 1985–1995.

Yinfei Yang, Cen Chen, and Forrest Sheng Bao. 2016. Aspect-based helpfulness prediction for online product reviews. In *Tools with Artificial Intelligence (IC-TAI), 2016 IEEE 28th International Conference on*, pages 836–843. IEEE.

Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Sheng Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *ACL (2)*, pages 38–44.