



EACL 2023

On the Role of Reviewer Expertise in Temporal Review Helpfulness Prediction

Mir Tafseer Nayeem and Davood Rafiei
University of Alberta



What is a Helpful Review?

- Provides **useful** and **informative feedback** to potential customers or users

These are the greatest headphones ever



These is the gratest headphones ever!!! Super high-quality sound. You can connect without worry to your Bluetooth (including Bose SimpleSync technology). The headphones deliver up to 20 hours of wireless play. you should definitely buy one

0 people found this helpful

Great value headphones with one caveat



These phones sound really nice, for being under \$20. They have much better sound than you might have expected at this price point. Frequency response is good for the price. Clear treble and adequate bass response to let you "feel" the drums and bass guitar. Soprano and contralto vocals sound nice and clear, without that "muddy" sound. The only caveat: the cord is only 3 feet long, not really long enough for home use with a desktop computer. Overall, though, great pair at a fair price.

89 people found this helpful [Verified Purchase](#)



Exaggerated headline



Short, poorly written review that mentions product details



Few or no helpful votes, not a verified purchase



Balanced review



Descriptive, well-written review with subjective details



Many helpful votes, verified purchase

What is a Helpful Review?

- Provides **useful** and **informative feedback** to potential customers or users
- Contains **specific details** about the product or service

These are the greatest headphones ever



These is the gratest headphones ever!!! Super high-quality sound. You can connect without worry to your Bluetooth (including Bose SimpleSync technology). The headphones deliver up to 20 hours of wireless play. you should definitely buy one

0 people found this helpful

Great value headphones with one caveat



These phones sound really nice, for being under \$20. They have much better sound than you might have expected at this price point. Frequency response is good for the price. Clear treble and adequate bass response to let you "feel" the drums and bass guitar. Soprano and contralto vocals sound nice and clear, without that "muddy" sound. The only caveat: the cord is only 3 feet long, not really long enough for home use with a desktop computer. Overall, though, great pair at a fair price.

89 people found this helpful [Verified Purchase](#)



Exaggerated headline



Short, poorly written review that mentions product details



Few or no helpful votes, not a verified purchase



Balanced review



Descriptive, well-written review with subjective details



Many helpful votes, verified purchase

What is a Helpful Review?

- Provides **useful** and **informative feedback** to potential customers or users
- Contains **specific details** about the product or service
- Usually includes both **positive and negative aspects** of the product or service

These are the greatest headphones ever



These is the gratest headphones ever!!! Super high-quality sound. You can connect without worry to your Bluetooth (including Bose SimpleSync technology). The headphones deliver up to 20 hours of wireless play. you should definitely buy one

0 people found this helpful

Great value headphones with one caveat



These phones sound really nice, for being under \$20. They have much better sound than you might have expected at this price point. Frequency response is good for the price. Clear treble and adequate bass response to let you "feel" the drums and bass guitar. Soprano and contralto vocals sound nice and clear, without that "muddy" sound. The only caveat: the cord is only 3 feet long, not really long enough for home use with a desktop computer. Overall, though, great pair at a fair price.

89 people found this helpful [Verified Purchase](#)



Exaggerated headline



Short, poorly written review that mentions product details



Few or no helpful votes, not a verified purchase



Balanced review



Descriptive, well-written review with subjective details



Many helpful votes, verified purchase

What is a Helpful Review?

- Provides **useful** and **informative feedback** to potential customers or users
- Contains **specific details** about the product or service
- Usually includes both **positive and negative aspects** of the product or service
- Helpful review may include **suggestions for improvement** or alternative products

These are the greatest headphones ever



These is the gratest headphones ever!!! Super high-quality sound. You can connect without worry to your Bluetooth (including Bose SimpleSync technology). The headphones deliver up to 20 hours of wireless play. you should definitely buy one

0 people found this helpful

Great value headphones with one caveat



These phones sound really nice, for being under \$20. They have much better sound than you might have expected at this price point. Frequency response is good for the price. Clear treble and adequate bass response to let you "feel" the drums and bass guitar. Soprano and contralto vocals sound nice and clear, without that "muddy" sound. The only caveat: the cord is only 3 feet long, not really long enough for home use with a desktop computer. Overall, though, great pair at a fair price.

89 people found this helpful [Verified Purchase](#)



Exaggerated headline



Short, poorly written review that mentions product details



Few or no helpful votes, not a verified purchase



Balanced review



Descriptive, well-written review with subjective details



Many helpful votes, verified purchase

Problems in User Reviews

- User reviews may **contain spam, excessive appraisal, or unexpected biases.**
- Multiple factors that affect the quality of a review. These factors are not usually explicit in the review text.
 - **Reviewers' life experience**
 - **Educational background**
 - **Motive for writing the review**

Problems in User Reviews

- User reviews may **contain spam, excessive appraisal, or unexpected biases.**
- Multiple factors that affect the quality of a review. These factors are not usually explicit in the review text.
 - **Reviewers' life experience**
 - **Educational background**
 - **Motive for writing the review**
- **Customers usually have limited patience for reading reviews** – most customers read **less than 10 reviews** before making a purchase decision (Murphy, 2016).
- **Large volume of reviews and their unpredictable quality** and the limited customer patience demand better review utilization strategies.

Helpfulness Votes

- **One standard method to identify more informative reviews** is to ask for feedback from customers.
 - *“Was this review helpful to you?”* or *“Did you find this review helpful?”*
- **User reviews that gain the most helpful votes** are shown first to the potential buyers to make the decision easier.

Helpfulness Votes

- **One standard method to identify more informative reviews** is to ask for feedback from customers.
 - *“Was this review helpful to you?”* or *“Did you find this review helpful?”*
- **User reviews that gain the most helpful votes** are shown first to the potential buyers to make the decision easier.
- **Problems**
 - **The voting data suffers from scarcity** (Siersdorfer et al., 2010) since only a tiny proportion of customers are willing to cast helpfulness votes.
 - The scarcity is even more severe in **reviews of less popular products and more recently submitted reviews** (*a.k.a., cold-start reviews*)

Can we automatically identify helpful reviews? - Related Works

- **Text Only**

- Extracted hand-crafted features from the review text.
 - **Structural** (Susan and David, 2010; Xiong and Litman, 2014),
 - **Lexical** (Kim et al., 2006; Xiong and Litman, 2011),
 - **Syntactic** (Kim et al., 2006),
 - **Emotional** (Martin and Pu, 2014),
 - **Semantic** (Yang et al., 2015),
 - **Arguments** (Liu et al., 2017)
- Chen et al. (2018) uses a text-based CNN model to automatically capture the **character-level**, **word-level**, and **topic-level** features.

These methods heavily rely on manual feature engineering, which is labor-intensive and time-consuming.

Can we automatically identify helpful reviews? - Related Works



- **Text and Star Rating**

- Fan et al. (2018) uses an end-to-end multi-task neural architecture with the help of an auxiliary task, such as rating regression.


- **Text and Image**



- Recently, Liu et al. (2021) and Han et al. (2022) use both text and images to guide the review helpfulness prediction.
- **Image field is usually optional in reviews**, a large volume of reviews contain only text, for which these multimodal models would **produce inconsistent results**.

Our Focus [Text + Metadata]

 Reviewed December 30, 2019  via mobile

Best view in town
"What can I say .. this is my best place in town. Average food, but the view pays the price. Breathtaking London View, lovely staff, Love to"


 41 Helpful votes

(a)

 Reviewed December 21, 2021

COVID restricted
"The room was clean and comfortable. We were looking forward to the breakfast buffet, but due to COVID, it wasn't available. We didn't dine in for other meals"

 0 Helpful vote

(b)

 Reviewed November 12, 2016  via mobile

HORRIBLE Service
"Terrible food! Overpriced, Cold, and flavorless. Shocking Service!! Undoubtedly the WORST place I have ever been! Call +1 437 *** **** OR visit this restaurant *****..."

 1 Helpful vote

(c)

- Incorporating the **reviewer's expertise** and **temporal information** in reviews to predict the helpfulness, especially for **unreliable** and **cold-start reviews**.
- People who **post more reviews** and **earn more helpful votes** are more likely to be better reviewers.
- Recently submitted reviews may contain more relevant and **time-sensitive information** (e.g., "New COVID Restrictions" or "Dirty Pool Area") but no helpfulness vote.

Dataset Construction

- **No human-annotated dataset available** with the reviewers' attributes and review date.
 - We build our dataset by scraping reviews from TripAdvisor.
 - **Reviews**
 - Review Text
 - Total Review Helpful Votes
 - Review Posting Time
 - **Reviewers**
 - Total Number of Reviews Contributed
 - Cumulative Helpful Votes

Dataset Construction

- We leverage a **logarithmic scale** to categorize the reviews based on the **number of votes received**.
- We map the number of votes into five intervals (*i.e.*, **[1,2)**, **[2, 4)**, **[4, 8)**, **[8, 16)**, **[16, ∞)**), each corresponding to a helpfulness score $Y \in \{1, 2, 3, 4, 5\}$, where the higher the score, the more helpful the review.

	Train	Valid	Test
Total #Samples	145,381	8,080	8,080
Avg. #Sentences	7.82	7.80	7.81
Avg. #Words	152.37	152.25	148.90

Table 1: Our dataset statistics.

Review Helpfulness Prediction (RHP)

- Review Helpfulness Prediction (RHP) can be modeled as a supervised machine learning task where the input contains information about the reviews (R) and the reviewers (U).

$$\mathcal{R}_i = ([s_1, \dots, s_N], t_i)$$

$$\mathcal{U}_i = (n_i, m_i)$$

- S = Review Sentence
- t = Review time
- n = num of review posted
- m = num of helpful votes received

- We formulate the task where we seek to find a model f that minimizes the loss function.

$$\min_{\theta} \mathcal{L}(f(\theta, \mathcal{R}, \mathcal{U}), Y)$$

Review Helpfulness Prediction (RHP)

- We encode the review sentences using BERT.

$$[h^{[\text{CLS}]}, h^{(1)}, h^{(2)}, \dots] = \text{BERT}([\text{CLS}] s_1, \dots, s_N [\text{SEP}]),$$

$$x_h = \Theta (\text{MLP} (h_l^{[\text{CLS}]})).$$



- In this work, we also integrate **reviewer expertise** and **temporal information** of the reviews.

Integrating Reviewer Expertise and Time

- **Expertise**

- Reviewers who **post more reviews** and **earn more helpful votes** are likely to be better reviewers.

Integrating Reviewer Expertise and Time

- **Expertise**

- Reviewers who **post more reviews** and **earn more helpful votes** are likely to be better reviewers.

- **Temporal Information**

- Older reviews are **more likely to accumulate more helpfulness votes** than newer reviews but are not necessarily the most relevant describing the current conditions (*e.g., new COVID restrictions*).

Integrating Reviewer Expertise and Time

- We define the **term reviewer expertise** as the mean number of helpful votes received per review.

$$h_s = \text{MLP}(e_s)$$

- Let td be the **relative age of a review in days**, for example, as of the day the reviews are scraped.

$$h_t = \text{MLP}(td)$$

- Both the review age and the reviewer expertise are **normalized to a fixed range $[a, b]$**

$$z_i = (b - a) \frac{x_i - \min(\mathcal{X})}{\max(\mathcal{X}) - \min(\mathcal{X})} + a$$

RHP Model

- We concatenate the textual representation, expertise representation, and temporal representation to get the final embedding.

$$o_{final} = h_s \oplus x_h \oplus h_t$$

$$\hat{Y} = \text{softmax}(W_r \cdot o_{final} + b_r),$$

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{Y}, Y)$$

Experimental Results

Baseline Models	Acc. (↑)	MAE (↓)	MSE (↓)
ARH	58.73	0.476	0.619
UGR + BGR	62.76	0.464	0.674
TextCNN	62.82	0.444	0.608
MTNL	62.77	0.458	0.653
BERTHelp	63.03	0.432	0.591
Our Ablations	Acc. (↑)	MAE (↓)	MSE (↓)
RHP (ours)	65.18[†]	0.393[†]	0.491[†]
- w/o Expertise	63.87	0.421 [†]	0.550 [†]
- w/o Temporal	63.40	0.437 [†]	0.592
- w/o Expertise + Temporal	62.92	0.446	0.617

Table 2: Performance compared to our baseline models and the result of our ablation study (↑ indicates higher values for a better performance and ↓ indicates lower values for a better performance). † reported results are statistically significant in paired t-test by taking BERTHelp (Xu et al., 2020) as a reference with the confidence of 95% (p -value < 0.05).

Analysis

Helpfulness Class	Unigram	Bigram
Class #1 Helpful Votes [1, 2)	'room'	'front desk'
	'staff'	'coffee maker'
	'location'	'breakfast buffet'
	'time'	'sofa bed'
	'service'	'swim pool'
Class #2 Helpful Votes [2, 4)	'room'	'front desk'
	'staff'	'shampoo conditioner'
	'service'	'customer service'
	'location'	'resort fee'
	'time'	'pool area'
Class #3 Helpful Votes [4, 8)	'room'	'front desk'
	'staff'	'resort fee'
	'time'	'customer service'
	'service'	'coffee maker'
	'view'	'city view'
Class #4 Helpful Votes [8, 16)	'room'	'front desk'
	'staff'	'resort fee'
	'service'	'customer service'
	'time'	'minute walk'
	'pool'	'life jacket'
Class #5 Helpful Votes [16, ∞)	'room'	'front desk'
	'time'	'resort fee'
	'service'	'bed bug'
	'staff'	'beach chair'
	'pool'	'cable car'

Table 3: Top 5 unigrams and bigrams extracted from five different classes of reviews divided according to helpfulness votes. For each column, **green** color indicates the overlap with all 5 classes, whereas **blue** for 4, **orange** for 3, and **red** for 2 overlaps.

- We randomly selected **m examples** for each class of reviews considering helpfulness votes.
- We extract **Top K (where K = 5)** n-grams from each class of reviews to identify the most relevant **keywords or topics** in reviews to assess what aspects are most talked about the items.

Case Study

[Free WiFi, Free parking, Location, Room, Staffs, Front Desk, Food, swimming pools, foods, Bar, Air conditioning, Non-smoking rooms, Fitness center, ATM on site, Shuttle service, Room service, Spa,]



Aspects / Facilities

[CLS] We could not have been happier with our choice for our family's 3 night stay in Las Vegas recently. The location was perfect. We stayed in a 2 bedroom villa, which was so spacious and had a great view of the Vegas lights and airportThe bathroom to the main bedroom had a fabulous big bath. The beds very comfortable. Dinner in the restaurant in the lobby one night, the food and service were both great. We particularly liked the restaurant and bar next to the pool on level 5, very relaxing for lunch [SEP]



Review Text

Figure 3: Top 10 ranked tokens of the RHP model shown in green colors with the color intensity indicating the importance of the tokens in the overall prediction.

- The top-ranked words are highly representative of the **aspects or facilities** listed on the restaurant page.
- We also notice that the use of **personal pronouns** (*e.g., I, we, they, etc.*), describing **personal experiences**, contributes to the helpfulness prediction.

Limitations & Future Work

- How to incorporate **personal preferences** for the helpfulness prediction task?
- We aim to extend this work to **support more languages**.

Thanks!

