RESEARCH ARTICLE

Feature selection for helpfulness prediction of online product reviews: An empirical study

Jiahua Du[1], **Jia Rong**[1,2]*, **Sandra Michalska**[1], **Hua Wang**[1], **Yanchun Zhang**[1]

1 Institute of Sustainable Industries & Liveable Cities, Victoria University, Melbourne, VIC, Australia, **2** Faculty of Information Technology, Monash University, Clayton, VIC, Australia

* jiarong@acm.org

Abstract

Online product reviews underpin nearly all e-shopping activities. The high volume of data, as well as various online review quality, puts growing pressure on automated approaches for informative content prioritization. Despite a substantial body of literature on review helpfulness prediction, the rationale behind specific feature selection is largely under-studied. Also, the current works tend to concentrate on domain- and/or platform-*dependent* feature curation, lacking wider generalization. Moreover, the issue of result comparability and reproducibility occurs due to frequent data and source code unavailability. This study addresses the gaps through the most comprehensive feature identification, evaluation, and selection. To this end, the 30 most frequently used content-based features are first identified from 149 relevant research papers and grouped into five coherent categories. The features are then selected to perform helpfulness prediction on six domains of the largest publicly available Amazon 5-core dataset. Three scenarios for feature selection are considered: (i) individual features, (ii) features within each category, and (iii) all features. Empirical results demonstrate that semantics plays a dominant role in predicting informative reviews, followed by sentiment, and other features. Finally, feature combination patterns and selection guidelines across domains are summarized to enhance customer experience in today's prevalent e-commerce environment. The computational framework for helpfulness prediction used in the study have been released to facilitate result comparability and reproducibility.

Citation: Du J, Rong J, Michalska S, Wang H, Zhang Y (2019) Feature selection for helpfulness prediction of online product reviews: An empirical study. PLoS ONE 14(12): e0226902. https://doi.org/10.1371/journal.pone.0226902

Editor: Farhan Hassan Khan, College of EME, NUST, PAKISTAN

Received: May 11, 2019

Accepted: December 6, 2019

Published: December 23, 2019

Data Availability Statement: This paper used the Amazon 5-core dataset which is publicly available on http://jmcauley.ucsd.edu/data/amazon/. The configuration of training-validation-testing splits and source code for our experiments containing pre-processing, feature extraction, and forward feature selection, are provided with open access on https://github.com/tokawah/Helpfulness-Feature-Selection.

Funding: The author(s) received no specific funding for this work.

Introduction

Customer product reviews play a significant role in today's e-commerce world, greatly assisting in online shopping activities. According to a survey conducted in 2016 [1], 91% of online shoppers read product reviews while searching for goods and services, and 84% of them believe that the reviews are equally trustworthy as recommendations from their friends. Online reviews do not only enhance the customer purchasing experience through valuable feedback provision, but also facilitate future product development activities by better understanding the customer needs.

Competing interests: The authors have declared that no competing interests exist.

Online product reviews are also highly susceptible to quality control [2], which can potentially harm online shopping experience. A recent study [3] shows that users tend to limit their attention to only first few reviews, regardless of their helpfulness. It is generally viewed that helpful reviews have more impact on customers' final decisions. However, the large and overwhelming nature of online product reviews makes it difficult for customers to efficiently locate useful information. Although the majority of online platforms enable review helpfulness assessment through user voting, the large proportion of records does not contain any votes. The scarcity of user votes is even more noticeable for less popular products.

Automatic helpfulness prediction helps consumers identify high-quality reviews, which has attracted substantial attention. The mainstream approach follows a procedure of careful feature curation from multiple data sources [4]. Still, the features are frequently domain- and/or platform-dependent, substantially inhibiting wider application. Also, the features are selected arbitrarily without solid justification. Furthermore, prior research mainly focuses on the predictive power of the entire feature set, while little is known on the contribution and necessity of using individual or subsets of features. Since identical feature set is rarely used among existing studies, the reported results prove challenging for fair comparison. Finally, the existing studies are often conducted on publicly unavailable ad-hoc datasets, hampering result reproducibility.

To address the aforementioned gaps, this study comprehensively identifies, evaluates, and selects representative features for helpfulness prediction. Specifically, frequently used domain- and platform-*independent* features (i.e., content-based features) are first identified from considerable recent literature. The predictive power of the identified features is then evaluated on six domains of large-scale online product reviews. Instead of evaluating the entire feature set, the study allows for performance-oriented feature selection under multiple scenarios. Such flexibility can effectively justify (not) selecting certain features. As a result, feature combination patterns and selection guidelines across domains are summarized, offering valuable insights into general feature selection for helpfulness prediction. The publicly available source code and datasets ensure result comparability and reproducibility of the study.

This study contributes to existing literature in four aspects:

- First, the study conducts one of the most comprehensive literature reviews on helpfulness analysis to identify frequently used content-based features.
- Second, the study conducts the first and most extensive empirical validation on large-scale publicly available online product reviews to report feature behaviors in multiple scenarios (individual and combinations) and domains.
- Third, a holistic computational framework is developed for helpfulness prediction from scratch, including data pre-processing, extracting the identified features, and evaluating the predictive power of individual features and feature combinations.
- Fourth, the source code, dataset splits, pre-processed reviews, and extracted features have been released for result reproducibility, benchmark studies, and further improvement.

The remaining of the study is organized as follows. The Related work section surveys recent literature regarding the use of features and feature selection for review helpfulness prediction. The Methodology section introduces steps for approaching feature-based helpfulness prediction, including feature identification, feature extraction, and feature selection strategies used in the study. Substantial analysis is conducted in the Empirical analysis section. Empirical results are reported and discussed to evaluate and locate optimal feature combinations, followed by frequent pattern discovery. Subsequently, the study summarizes the implications and discusses the limitations in the Implications and Limitations section, respectively. Finally, the Conclusions and future works section encapsulates the findings and outlines future directions of the study.

Related work

The automatic prediction of reviews helpfulness is majorly approached via feature engineering. Previous studies have curated a large body of features derived from (i) review content [5–10] and (ii) review contexts such as reviewers [11, 12], social networks among reviewers [13, 14], review metadata [15, 16], and product metadata [17, 18]. Some other less frequent contextual features include review photos [19, 20], manager responses [21], travel distances [22], to name a few. This study focuses on content-based features due to the ubiquitous use in literature and the ability of review texts to generalize across online platforms.

Recent studies regarding helpfulness prediction and feature selection have been identified and summarized. Kim et al. [5] investigated the effect of ten features spanning four main categories (i.e., lexical, structural, semantic, syntactic and metadata), and their combinations on helpfulness prediction. The authors found out that the most useful features were review length, unigrams, and product ratings. Zeng et al. [23] reported the results of individual features and all-minus-one feature combinations. They introduced "the degree of detail" feature as a function of review length and *n*-grams, alongside seven other features. The introduced feature proved to be the most important in helpfulness prediction, leading to a significant drop in accuracy after its exclusion. Yang et al. [8] evaluated the impact of review structure, unigrams, and three sentiment features: Geneva Affect Label Coder, Linguistic Inquiry and Word Count, and General Inquirer. The latter two features not only improved the prediction performance, but also provided a useful interpretation to what makes a review helpful.

Akbarabadi et al. [24] focused on 12 features from the review characteristics category, including review length, review age, part-of-speech, richness, sentiment and readability. The title characteristics category was also introduced, which did not improve the performance of helpfulness prediction. Vo et al. [25] investigated the four feature categories, namely anatomical, metadata, lexical and added feature group, which included (i) the number of helpfulness votes, and (ii) the number of positive and negative words. The impact of (i) on prediction accuracy proved to depend on both datasets and the choice of classifiers. The results for (ii) demonstrated a similar pattern.

Haque et al. [26] analyzed the performance of lexical, structural, semantic and readability feature groups. The last group was added in order to unfold the complexity of the review content, and showed significant impact on helpfulness prediction. Chen et al. [27] adopted the features related to text surface (i.e., the number of words, sentences, exclamation marks, question marks, and uppercase, lowercase), unigrams, part-of-speech, and word embeddings. The word embedding features trained using the Skip-gram model outperformed unigrams on an opinion spam detection dataset collected from amazon.

In terms of neural network-based models, Fan et al. [28] conducted helpful review identification based on recurrent neural networks, using the metadata of their target products. Saumya et al. [29] developed a two-layer convolutional model upon both the Skip-gram and Global Vectors model. Still, such approaches lack interpretability, making it difficult to identify what particular aspects of the reviews are good indicators of helpfulness.

As presented above, the numerous analysis tasks have been conducted to extract the most useful features for helpfulness prediction. However, the research within domain is often fragmented and heterogeneous, which challenges the objective comparison and findings synthetization. For example, the categorization of features differs among the studies, impacting finding generalizability. Also, the features selected in prior research frequently lacks justification behind particular feature selection, leading to the potential bias in results interpretation. Moreover, most of the existing studies suffer from result reproducibility due to the unavailability of ad-hoc datasets and implementation details.

Given the limitations identified, this study (1) provides the most comprehensive and generalizable content-based feature set evaluation on large-scale publicly available datasets, (2) conducts the empirical validation of the most effective feature selection in an objective manner, and (3) releases the datasets and source code describing the implementation details used in this study.

To the best of our knowledge, this study is the first to address the reproducibility and transferability issue of review helpfulness prediction, as well as the first work that provides the justification-driven feature selection process regardless of the platform and domain of applications. The complete and systematic literature review proves practically infeasible given largely fragmented state of the research in helpfulness prediction domain. Still, the study has made best efforts to report the latest state-of-art and identify the gaps to fill with the current work.

Methodology

Feature-based helpfulness prediction entails three steps. To start with, the procedure and criteria are described to collect recent relevant literature, from which frequently cited content-based feature candidates are identified. Each of the identified feature candidates is then introduced and the feature construction process is specified. Finally, the evaluation protocols and feature selection strategies are provided to locate optimal feature combinations for review helpfulness prediction.

Feature identification

The study identifies frequently cited feature candidates from recent literature to provide wide generalization and fair comparison with the majority of studies on the topic. To this end, a collection of most recent relevant studies are first collected and filtered, from which feature candidates are identified.

Paper acquisition The collection of relevant papers is based on (i) the references of the three most recent survey papers from the review helpfulness field [4, 30, 31] and (ii) the top 50 relevant studies retrieved from the Google Scholar database and published before 2019, using the following search query:

(“*online reviews*” OR “*product reviews*” OR “*user review*” OR “*customer review*” OR “*consumer reviews*”) AND (“*useful*” OR “*helpful*” OR “*usefulness*” OR “*helpfulness*”).

Given the scope of the study, the 149 collected papers are filtered based on the following criteria: (i) *automated* prediction of online product review helpfulness; (ii) inclusion of *factors* influencing review helpfulness; and (iii) *English-written* review analysis only. As a result, 74 papers (See the “Literature” column in Table 1) are identified.

Feature acquisition Features mentioned in the 74 identified papers are collected, along with the frequency of feature mentions. The following rules are adopted for feature list compilation: (i) features mentioned at least three times over the entire paper collection to exclude rare features, (ii) removal of human-annotated features due to expensive manual annotation process, and (iii) inclusion of only *content-based* features to support platform-independent generalizability and transferability. As a results, 27 feature candidates are identified.

Table 1. Features used in the analysis.

Category	Feature	Dimension	Description	Literature
Semantics	UGR	V	Unigram TF-IDF	[5, 7, 8, 13, 23, 26, 32–43, 88]
	BGR	V^2	Bigram TF-IDF	[5, 23, 26, 40, 42, 44]
	LDA	100	LDA topic distribution	[37, 39, 45–48, 88]
	SGNS	300	Skip-gram Negative Sampling	—
	GV	300	Global Vectors	—
Sentiment	LIWC	93	Linguistic Inquiry and Word Count dictionary	[8, 35, 49–55, 88]
	GI	182	General Inquirer	[5, 8, 18, 33, 39, 42, 56, 88]
	GALC	21	Geneva Affect Label Coder	[7, 8, 33]
	OL	3	Opinion lexicon	[13, 38, 40, 57–59]
	SWN	3	SentiWordNet	[37, 59–62]
	SS	3	SentiStrength	[10, 63, 64]
	VADER	3	VADER lexicon	—
Readability	FKRE	1	Flesch–Kincaid Reading Ease score	[7, 10, 18, 29, 35, 38–40, 48, 51, 54, 59, 62, 65–67]
	FKGL	1	Flesch–Kincaid Grade level	[9, 10, 35, 38, 39, 65, 68]
	GFI	1	Gunning Fog Index	[9, 10, 16, 38, 40, 50, 52, 53, 65, 66, 68]
	SMOG	1	Simple Measure of Gobbledygook	[9, 10, 38–40, 65]
	ARI	1	Automated Readability Index	[9, 10, 38, 52, 56, 65, 66, 68, 69]
	CLI	1	Coleman–Liau Index	[9, 10, 15, 38, 49, 52, 65, 66, 68, 70]
Structure	CHAR	1	Number of characters	[26, 35, 45, 52, 65, 71–75]
	WORD	1	Number of words	[5, 7, 8, 11, 13, 15, 16, 18–21, 23, 26, 29, 32, 33, 35, 37–43, 45, 48–50, 52, 53, 56–58, 60, 62, 63, 65–71, 76–82, 88]
	SENT	1	Number of sentences	[5, 7, 8, 13, 21, 26, 32, 33, 35, 37, 39–43, 45, 52, 57, 58, 65, 71, 79, 88]
	AVG	1	Average number of words per sentence	[5, 8, 13, 33, 35, 39, 40, 43, 45, 52, 57, 58, 71, 79, 83, 88]
	EXCLAM	1	Number of exclamatory sentences	[5, 8, 33, 41, 42, 88]
	INTERRO	1	Number of interrogative sentences	[5, 8, 26, 33, 41, 42, 88]
	MIS	1	Number of misspelling words	[38, 39, 54, 59, 65, 71]
Syntax	NOUN	1	Number of nouns	[5, 7, 10, 13, 29, 32, 39, 42, 43, 52, 59]
	VERB	1	Number of verbs	[5, 7, 10, 13, 29, 32, 39, 41–43, 59, 84]
	ADJ	1	Number of adjectives	[5, 7, 10, 13, 29, 32, 39, 41–43, 58, 59]
	ADV	1	Number of adverbs	[5, 7, 10, 13, 32, 39, 41–43, 58]
	COMP	1	Number of comparative sentences	[13, 23, 32, 43, 55, 85]

V indicates the vocabulary size of the training set of a corpus.

https://doi.org/10.1371/journal.pone.0226902.t001

As a novelty, the study additionally incorporates two semantic features and one sentiment feature that are gaining more recent attention. Such features have been proved robust in numerous text mining and natural language processing applications but are so far under-studied in review helpfulness prediction.

Table 1 presents the 30 content-based features identified from recent literature. The features are further grouped into five coherent categories (i.e., semantics, sentiment, readability, structure, and syntax) following the convention in the research field.

Note that context-based features such as reviewer characteristics are currently excluded from the feature pool since they are domain- and/or platform-dependent, and thus not always available.

Feature extraction

The description and construction process of the identified features in groups is presented as follows. It is worth noting that some features overlap functionally, for instance, all sentiment features compute the emotional composition of reviews via different lexicons. Some features are constituents of others, such as readability scores resulting from different linear transformations of certain structural features. Following the convention in the research field, features in both cases are treated as individual ones.

Semantics. Semantic features refer to the meaning of words and topical concepts from the review content by modelling terms statistics into vectors. The five semantic features for the helpfulness prediction task are as follows:

UGR and BGR The unigram bag-of-words representation of a review uses the term frequency-inverse document frequency (TF-IDF) weighting scheme [86], where each element of a vector corresponds to a word in the vocabulary. Similarly, the bigram bag-of-words representation encodes all possible word pairs formed from neighboring words in a corpus. Both UGR and BGR ignore terms that have a document frequency value below 10 when building the vocabulary. The vector representations are then transformed into unit vectors via the L2 normalization.

LDA Latent Dirichlet Allocation representation learns the topic distribution of a review. Topic modeling considers corpus as a mixture of topics, and each topic consists of a set of words. In the case of online product reviews, the topics can be different product properties, emotional expressions, etc. The original LDA algorithm [87] is adopted to learn the probability distribution of latent topics for each review. Following [88], the number of topics is set to 100 during training.

SGNS and GV As a novelty, the study also uses the two most recent types of *word embeddings* as features. The Skip-Gram with Negative Sampling [89] and Global Vectors [90] aim at learning the distributed representations of words. Under this setting, each word is mapped into a dense vector space, where similar terms display closer spatial distance. Thus, each review can be simply converted into a vector by averaging the embeddings of its constituent words, where out-of-vocabulary words are skipped.

Sentiment. Sentiment features analyze the subjectivity, valence, and emotion status of content written by customers. Previous works [22, 91] have shown relevance between helpfulness of a review and the sentiments expressed through its words. The study constructs sentiment features using the seven most frequently-used lexicons. The first three lexicons are category-based, each estimating the probability of a review belonging to its predefined lexicon categories. The remaining lexicons are valence-based, each looking up the valence (i.e., positive, neutral, and negative) of words in a review where possible. Note that both the categories and word valence are defined differently among lexicons. As a result, the seven sentiment features will lead to different vector representations due to various measurement criteria.

LIWC The Linguistic Inquiry and Word Count dictionary [92] classifies contemporary English words into 93 categories, including social and psychological states. The dictionary covers almost 6, 400 words, word stems, and selected emoticons.

GI General Inquirer [93] attaches syntactic, semantic, and pragmatic information to part-of-speech tagged words. It contains 11, 788 words collected from the Harvard IV-4 dictionary and Lasswell value dictionary, which are assigned to 182 specified categories.

GALC Geneva Affect Label Coder [94] recognizes 36 emotion categories of affective states commonly distinguished by 267 word stems. The Geneva Emotion Wheel model [7, 8] is followed, and the 20 of the GALC categories plus an additional dimension for non-emotional words are adopted.

OL The Opinion Lexicon [95] is widely used by researchers for opinion mining. It consists of 2, 006 positive and 4, 783 negative words, along with the misspellings, morphological variants, slang, and social media markups.

SWN SentiWordNet [96] is a lexical resource for sentiment and opinion mining. It assigns to each synset of WordNet [97] three sentiment scores: positivity, negativity, and objectivity, in terms of probability.

SS SentiStrength [98] is a tool for automatic sentiment analysis on short social web texts written in informal language, incorporating intensity dictionaries, words with non-standard spellings, emoticons, slang and idioms.

VADER As a novelty, the study also adopts the Valence Aware Dictionary and sEntiment Reasoner [99]. VADER is a lexicon specifically attuned for social media texts. It has 3, 345 positive and 4, 172 negative terms, and is enhanced with general heuristics for capturing sentiment intensity.

Sentiment features are built as follows. For each categorical lexicon, a sentiment feature is represented by the histogram of all its predefined categories. Take LIWC as an instance, the generated feature vector of 93 dimensions contains numeric statistics of a review corresponding to each predefined category. Similarly, a feature vector derived from GI and GALC contains 182 and 21 elements encoding information of a review towards individual predefined categories, respectively.

As for valence-based lexicons, a review is described using a three-dimensional vector: the percentage of positive, neutral, and negative sentences in a review. Given a sentence, all its words are looked up in a lexicon, and the corresponding valence values are subsequently summed up. A sentence is considered positive if the total valence is greater than zero, negative if less than zero, and neutral otherwise. During the valence lookup, VADER heuristics are applied to OL and SWN to improve the detection accuracy [100]. The heuristics does not apply to SS since the toolkit offers a similar built-in mechanism for sentiment intensity evaluation.

The aforementioned sentiment features differ one another. In category-based lexicons, the sentiment of a review is described using predefined categories, similar to an opinion is understood from different perspectives. Meanwhile, valence-based lexicons detect the polarity of review words differently. For example, the term “clean” can be positive in some lexicons but neutral in others. As a result, the same review will obtain different vector representations due to various sentiment measurement criteria. Further details of the lexicon composition, such as the predefined categories and vocabulary can be found in the corresponding literature of individual lexicon and the survey papers [100, 101].

Readability. Readability measures the ease of reading texts. As pointed out by [102], even a minor increase in readability largely improves review readership, leading to more opportunities for reviews to receive helpful votes. Thus, readability has been frequently addressed in the past papers on helpfulness prediction. The following six formulas are used to construct the readability features, taking advantage of the number of characters, syllables, words, complex words, and sentences.

$$\text{FKRE [103]} = 206.835 - 1.015\left(\frac{\#\text{words}}{\#\text{sentences}}\right) - 84.6\left(\frac{\#\text{syllables}}{\#\text{words}}\right) - 15.59, \tag{1}$$

$$\text{FKGL [104]} = 0.39\left(\frac{\#\text{words}}{\#\text{sentences}}\right) + 11.8\left(\frac{\#\text{syllables}}{\#\text{words}}\right), \tag{2}$$

$$\text{GFI [105]} = 0.4\left[\left(\frac{\#\text{words}}{\#\text{sentences}}\right) + 100\left(\frac{\#\text{complex words}}{\#\text{words}}\right)\right], \tag{3}$$

$$\text{SMOG [106]} = 1.0430\sqrt{\#\text{complex words} \times \frac{30}{\#\text{sentences}}} + 3.1291, \tag{4}$$

$$\text{ARI[107]} = 4.71\left(\frac{\#\text{characters}}{\#\text{words}}\right) + 0.5\left(\frac{\#\text{words}}{\#\text{sentences}}\right) - 21.43, \tag{5}$$

$$\text{CLI [108]} = 0.0588L - 0.296S - 15.8, \tag{6}$$

where complex words have at least three syllables, $L = \frac{\#\text{characters}}{\#\text{words}} \times 100$, $S = \frac{\#\text{sentences}}{\#\text{words}} \times 100$. The *z*-score is calculated for each feature for normalization.

Similar to the sentiment category, the six readability features used in the study will obtain different vector representations. While referring to the same underlying concept (ease of readiness), the use of different formulas, namely linear transformations of the counting statistics, reflects different focuses on understanding the readability of a review. Interested readers can access detailed information regarding readability tests in [109].

Structure. Structural features count the length and occurrence of specific language unit types. The following six features are selected to represent the structure of a review. The first three features are self-explanatory, including the number of characters (CHAR), tokens (WORD), and sentences (SENT). Similarly to Xiong et al. [110], the percentage of exclamatory (EXCLAM) and interrogative (INTERRO) sentences is taken into account. Finally, the number of misspelling words (MIS) in a review is considered.

Syntax. Syntactic features consider specific types and patterns of parts-of-speech within the review content. The percentage of the most prevalent open-class word categories, namely nouns (NOUN), adjectives (ADJ), verbs (VERB), and adverbs (ADV) is estimated. Additionally, the percentage of comparative sentences (COMP) is calculated. The procedure for comparative sentence detection follows the work by Jindal et al. [111], which employs a list of keywords and patterns to match the review sentences. Given that comparisons can take place implicitly, only explicit expressions are captured.

Feature selection for helpfulness prediction

Feature-based helpfulness prediction is formulated as a binary classification (either helpful or unhelpful) problem. Most existing studies approach the task either by classification or regression. This study adopts the former due to its intuitive and simple output to customers.

The task of feature-based helpfulness prediction is formally defined as follows. Let $\mathcal{D} = \{(D_1, u_1), \ldots, (D_n, u_n)\}$ be a collection of *n* product reviews, where *D* is the content of a review and *u* the accompanying helpfulness information ($u = 1$ helpful and $u = 0$ unhelpful). Each review content $D \in \mathcal{D}$ is associated with a set of features, denoted by

$\mathcal{F}(D) = \{f_1(D), \ldots, f_m(D)\}$, via m different feature extractors $\{f\}$. The goal of the task is to train a binary classifier C that searches for the optimal feature combination $\hat{\mathcal{F}}$ from the feature pool $\mathcal{F}$ to approximate the helpfulness u such that:

$$\hat{\mathcal{F}} = \underset{\mathcal{F}' \subseteq \mathcal{F}(D)}{\arg\max} \sum_{D \in \mathcal{D}} \mathbb{1}(u = C(\mathcal{F}')), \tag{7}$$

where $\mathbb{1}(\cdot)$ is an indicator function. Ideally, the search of $\hat{\mathcal{F}}$ would exhaust all possible feature combinations. Though, such a scenario is not suitable for $m = 30$ features due to the exponential complexity of calculation.

Instead, the search is fulfilled by a wrapper method, specifically, the step forward feature selection. Given the feature pool, the search starts with the evaluation of each feature and selects the one with the highest performance. Subsequently, all possible combinations between the selected feature and each of the remaining features are evaluated, and the second feature is selected. The iteration continues until adding features cannot improve the prediction performance. As a result, the selected features together form the optimal feature combination.

As for the classifier C, the linear Support Vector Machine (SVM) algorithm is chosen given its wide adoption and high performance in previous studies on the task [5, 9, 23, 84]. Using the most common linear SVM classifier also facilitates fair comparison between the studies within the same field.

Empirical analysis

This section conducts substantial helpfulness prediction analysis using the 30 identified content-based features. The large-scale publicly available datasets and implementation details is introduced. The empirical results are discussed and further summarized to obtain insights into feature selection for helpfulness prediction.

Datasets

The analysis is conducted on the largest publicly available Amazon 5-core dataset [112]. Amazon is the largest Internet retailer, which has accumulated large-scale user-generated reviews. The helpfulness of such reviews is rated by online customers, which makes it an ideal candidate for review helpfulness prediction task. In fact, Amazon product reviews are predominantly used and analyzed in previous studies. Thus, adopting Amazon reviews allows for fair comparisons with previous studies.

The original dataset consists of 24 domains, covering 142.8 million reviews collected between May 1996 and July 2014. The six domains with the highest number of reviews are selected for the study. Table 2 presents the *helpful* versus *unhelpful* review examples, where each review contains (i) a summary headline, (ii) a review text commenting in detail on the product, and (iii) the helpfulness information, namely the number of helpful and unhelpful votes given by online customers. During analysis, content-based features are extracted from the combination of the summary headline and review text.

The following pre-processing steps are performed: (1) All blank and non-English reviews are filtered out; and (2) regarding identical and nearly identical reviews [113] common on Amazon, only the ones with the highest number of votes are retained (two reviews are nearly identical if more than 80 percent of their bigram occurrence is shared [5]); and (3) the reviews with less than 10 votes are skipped to alleviate the effect of words of few mouths [114]; and (4) the remaining reviews are lowercased, tokenized, and the articles are removed.

The helpfulness label of the pre-processed reviews is determined via human assessment. For each domain, a review is labeled as unhelpful if its ratio of helpful votes is fewer than a pre-

Table 2. Example helpful and unhelpful reviews.

Helpful	Unhelpful
actually free!!!	are you serious?!!?
i thought for sure after downloading this app that i would have to creat an account and give away unnessecary information but surprisingly the only thing that was required was an email address and a password and i was instantly watching movies also from other reviews i was worried this app wouldnt have much to choose from but was again delightfully surprised they had more than a few good movie titles	I love hangman personally. when I saw this app I was like hmmm. . . and then NINETYNINE DOLLARS ARE YOU KIDDING ME? I don't care how ultimate this is. no app is worth 99$ especially hangman! I mean the graphics are cool! I could see like 2.99 at the most but I would never buy any app for 100 dollars. really? I thought you guys were really stupid but this is ridiculous! God I cant believe it! no no no no.
825 of 906 people think this review is helpful.	3 of 131 people think this review is helpful.

Typos in the reviews are intentionally preserved.

https://doi.org/10.1371/journal.pone.0226902.t002

defined threshold and otherwise helpful. The threshold is set to 0.6, which is the most commonly used threshold in prior research [9, 10, 65]. To avoid the class imbalance problem, which is outside the scope of this study, the same number of helpful and unhelpful reviews are sampled.

Finally, reviews in each domain are partitioned using random stratified split: 80%, 10%, and 10% of the reviews are randomly (with a fixed seed) selected respectively for training, validation, and testing, while preserving the percentage of samples for each class. During analysis, all feature combinations are trained on the training set, compared and selected on the validation set, and evaluated on the test set serving as unseen data in reality.

Table 3 demonstrates the descriptive statistics and out-of-vocabulary (OOV) rate of the six domains sorted by data size in ascending order. The vote distributions are further presented in Fig 1, displaying a similar pattern for each domain that high frequency of reviews have a relatively low number of votes.

Implementation

All analysis tasks are implemented with Python 3.6 and run on Ubuntu 16.04. Text pre-processing, part-of-speech tagging, and feature extraction are done using NLTK [115]. Specifically, both SGNS trained on 100 billion words from Google News and GV trained on 840 billion words from Common Crawl are publicly available online. Regarding the sentiment category, LIWC 2015, the commercial version (February 2017) of SentiStrength, and VADER 3.2.1 are employed. The remaining lexicons are acquired as per the papers. All the readability

Table 3. Descriptive statistics of the balanced datasets after pre-processing.

	Domain	#Reviews	#Tokens	#Sentences	$\frac{\#Tokens}{\#Reviews}$	$\frac{\#Tokens}{\#Sentences}$	$\frac{\#Sentences}{\#Reviews}$	OOV Rate
D1	Apps for Android	20,416	1,184,650	107,702	58.03	11.00	5.28	1.68% / 1.74%
D2	Video Games	23,100	7,522,835	469,856	325.66	16.01	20.34	0.86% / 0.86%
D3	Electronics	33,962	8,255,411	537,996	243.08	15.34	15.84	0.84% / 0.91%
D4	CDs and Vinyl	105,934	23,096,933	1,468,718	218.03	15.73	13.86	0.77% / 0.76%
D5	Movies and TV	164,052	40,549,434	2,510,899	247.17	16.15	15.31	0.62% / 0.64%
D6	Books	306,430	71,632,822	4,405,047	233.77	16.26	14.38	0.51% / 0.51%

OOV rate indicates the ratio of terms in the vocabulary of the training set is missing from the validation/test set.

https://doi.org/10.1371/journal.pone.0226902.t003

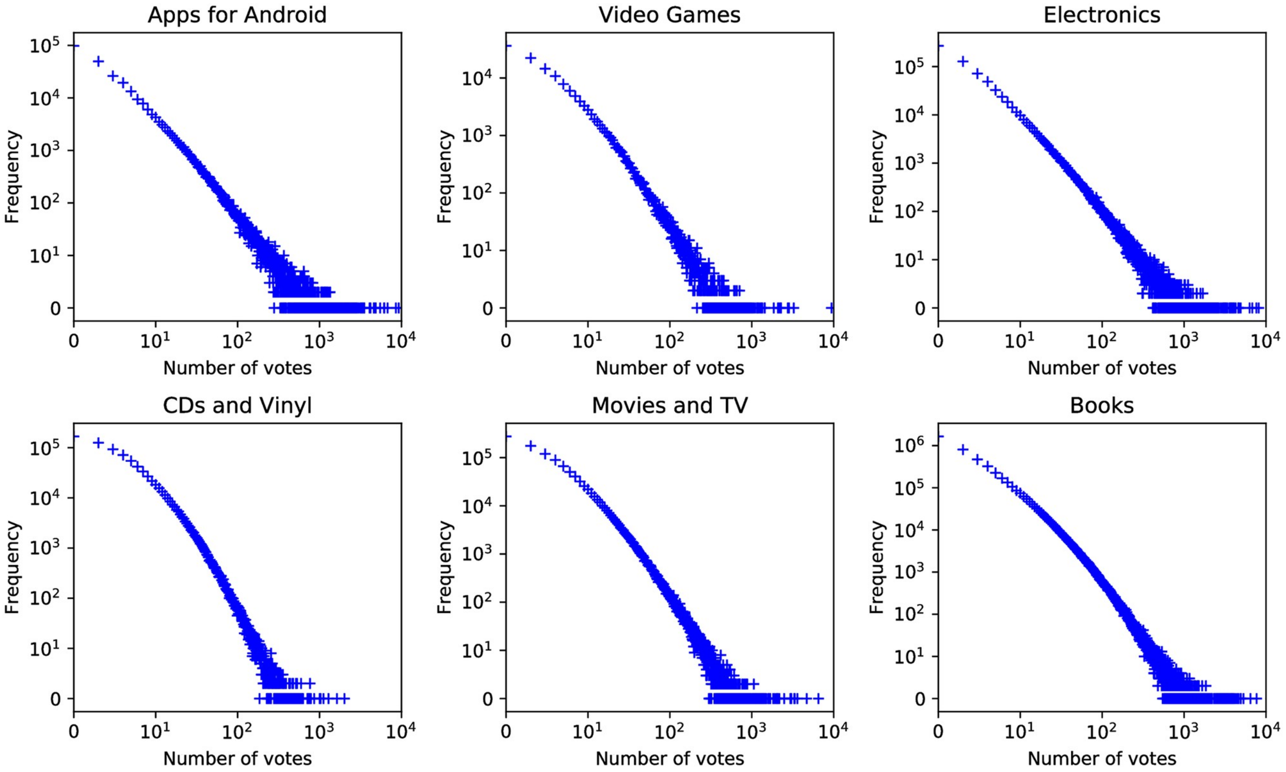

Fig 1. Vote distributions.

https://doi.org/10.1371/journal.pone.0226902.g001

scores are computed via the `textstat` library. The Hunspell spell checker is used to detect misspelling words. To enable the detection for product brands and contemporary language expressions, Hunspell is extended with Wikipedia titles (Retrieved February 13, 2019, from Wikimedia dump service). The linear SVM classifier [116] is developed using Scikit-learn [117]. For reproducibility, all randomization processes involved in the study are initialized with the same random seed.

Results and discussion

The study considers three scenarios for feature selection: (i) individual features, (ii) features within each category, and (iii) all features. The research questions investigated can be formulated as follows:

RQ1: *What is the effect of individual features on review helpfulness prediction across domains?*

RQ2: *What are the optimal combinations of features within a category for review helpfulness prediction across domains?*

RQ3: *What are the optimal combinations of all features for review helpfulness prediction across domains?*

RQ4: *Are there any patterns of features/feature combinations for review helpfulness prediction that perform well in general?*

RQ1, RQ2, and RQ3 are answered one in a subsection. As for RQ4, the combination patterns and selection guidelines (if any) are discussed at the end of each subsection.

Table 4. The classification accuracy and ranking (In-category/Overall) of individual features.

Category	Feature	D1		D2		D3		D4		D5		D6	
		Acc.	Rank.	Acc.	Rank.	Acc.	Rank.	Acc.	Rank.	Acc.	Rank.	Acc.	Rank.
Semantics	UGR	66.06	3 / 4	74.98	2 / 2	74.1	1 / 1	80.83	3 / 3	78.06	1 / 1	75.02	1 / 1
	BGR	60.72	4 / 10	70.74	4 / 5	69.07	4 / 5	76.89	4 / 5	74.37	4 / 4	71.03	4 / 4
	LDA	55.39	5 / 12	67.23	5 / 7	63.1	5 / 10	65.49	5 / 11	63.39	5 / 10	61.67	5 / 9
	SGNS	67.58	2 / 2	75.15	1 / 1	73.28	3 / 3	81.02	2 / 2	77.26	3 / 3	73.91	3 / 3
	GV	68.66	1 / 1	74.94	3 / 3	73.34	2 / 2	81.06	1 / 1	77.41	2 / 2	74.04	2 / 2
Sentiment	LIWC	66.16	1 / 3	73.94	1 / 4	70.78	1 / 4	76.99	1 / 4	72.25	1 / 5	68.58	1 / 5
	GI	63.76	2 / 5	69.18	2 / 6	67.07	2 / 6	72.07	2 / 6	70.75	2 / 6	67.04	2 / 6
	GALC	55.88	7 / 11	53.81	7 / 28	58.21	7 / 21	56.14	7 / 27	54.86	7 / 27	55.17	7 / 23
	OL	62.78	4 / 7	62.99	4 / 17	66.83	3 / 7	70.13	3 / 7	66.57	3 / 7	63.64	3 / 7
	SWN	61.41	5 / 8	63.12	3 / 16	62.54	5 / 12	65.29	6 / 13	63.87	5 / 9	60.92	5 / 10
	SS	60.77	6 / 9	58.05	6 / 22	60.86	6 / 19	65.6	5 / 10	61.23	6 / 17	60.6	6 / 11
	VADER	63.42	3 / 6	62.55	5 / 18	64.6	4 / 8	67.68	4 / 8	65.32	4 / 8	62.51	4 / 8
Readability	FKRE	52.99	4 / 23	58.44	3 / 21	55.36	5 / 26	62.38	4 / 20	59.15	5 / 24	53.64	6 / 27
	FKGL	51.22	6 / 28	56.84	5 / 25	55.71	4 / 25	62.12	5 / 22	59.25	4 / 23	54.92	4 / 25
	GFI	53.48	2 / 21	56.15	6 / 26	53.18	6 / 29	61.66	6 / 23	58.48	6 / 26	54.16	5 / 26
	SMOG	53.23	3 / 22	61.34	2 / 20	59.12	1 / 20	64.8	1 / 15	61.32	1 / 16	56.45	1 / 21
	ARI	51.37	5 / 27	57.1	4 / 24	55.83	3 / 24	63.15	3 / 19	59.46	3 / 22	55.57	2 / 22
	CLI	53.82	1 / 19	62.08	1 / 19	56.45	2 / 22	64.58	2 / 16	60.72	2 / 19	54.98	3 / 24
Structure	CHAR	54.36	3 / 16	65.24	1 / 9	62.07	1 / 13	64.85	1 / 14	62.56	1 / 13	58.79	2 / 14
	WORD	54.55	2 / 15	64.76	2 / 11	61.86	2 / 14	63.95	2 / 17	62.17	2 / 14	58.87	1 / 13
	SENT	52.69	5 / 26	63.94	3 / 15	61.8	3 / 15	61.64	3 / 24	60.51	3 / 20	57.94	3 / 18
	AVG	52.94	4 / 24	57.71	4 / 23	56.18	4 / 23	59.83	4 / 26	58.87	4 / 25	56.59	4 / 20
	EXCLAM	51.13	6 / 29	52.81	7 / 30	53.86	6 / 28	53.99	6 / 29	52.79	6 / 29	53.05	5 / 28
	INTERRO	55.29	1 / 13	53.16	6 / 29	51.32	7 / 30	53.43	7 / 30	53.24	5 / 28	51.8	6 / 29
	MIS	50.78	7 / 30	54.29	5 / 27	55.27	5 / 27	55.72	5 / 28	52.35	7 / 30	50.51	7 / 30
Syntax	NOUN	54.06	2 / 17	64.94	2 / 10	61.77	3 / 16	65.38	2 / 12	63	2 / 12	58.92	1 / 12
	VERB	53.72	4 / 20	64.2	4 / 13	61.09	5 / 18	62.27	4 / 21	60.81	4 / 18	58.41	4 / 17
	ADJ	55.14	1 / 14	65.41	1 / 8	63.27	1 / 9	65.63	1 / 9	63.11	1 / 11	58.65	2 / 15
	ADV	52.89	5 / 25	64.16	5 / 14	61.74	4 / 17	61.04	5 / 25	60.24	5 / 21	57.76	5 / 19
	COMP	53.97	3 / 18	64.33	3 / 12	62.89	2 / 11	63.6	3 / 18	61.56	3 / 15	58.56	3 / 16

https://doi.org/10.1371/journal.pone.0226902.t004

Throughout the analysis, the performance of review helpfulness prediction is measured by classification accuracy and its ranking. The latter is provided as another prioritization measure to capture the general trend of feature performance since the accuracy of a feature (set) can largely vary in domain.

RQ1: The predictive power of individual features. Table 4 demonstrates the classification accuracy, in-category ranking, and overall ranking of individual features, respectively. As shown, the semantics and sentiment category in general perform better than the other three categories.

Semantics The semantics category consists of most of the globally top-five features. The best overall performance lies in semantic features directly modeling review content, leading to more dimensions for encoding information. In particular, UGR sets a strong baseline in all domains, which indicates that specific term occurrences differ between helpful and unhelpful reviews. Both GV and SGNS show comparable or higher performance than UGR, with about 1% in accuracy lower than UGR in the worst case. The promising performance demonstrates the efficacy of traditional machine learning algorithms trained on general-purpose distributed word representations for helpfulness prediction. GV outperforms SGNS in all domains except D2, being a preferable option. In contrast, BGR scores 4%–5% lower compared with UGR, suggesting increased data sparsity while using bigram features. LDA consistently ranks the lowest within the category and is even lower than several features in the sentiment and syntax category. The inferior performance can be attributed to short product reviews hindering the training of topic distributions, which explains the lowest (highest) overall LDA ranking on D1 (D2).

Sentiment The sentiment category shows mixed performance. As for the categorical lexicons, LIWC, GI, and GALC rank respectively first, second, and last in all domains. LIWC outperforms UGR in D1 but is beaten by other domains. The accuracy gap, ranging from 1% to 6%, is proportional to data size. As such, LIWC can substitute for semantics when applied to small datasets. While the drastic low performance of GALC results from its few predefined categories and low vocabulary coverage compared with LIWC, GI shows that having almost double the size of predefined lexicon categories and words does not necessarily bring higher performance. On the other hand, the valence-based lexicons perform variously depending on data size. In most cases, OL and VADER produce higher accuracy than SWN and SS. Starting from D3, a more precise pattern that OL>VADER>SWN>SS is observed. OL generally performs better than other valence-based lexicons because it is originally generated from Amazon reviews, and thus more related to the tested domains. The results from the category show that the predictive power of lexicon-based features highly depends on the definition of lexicon categories, vocabulary coverage, as well as data size.

Readability, Structure and Syntax Features from the remaining three categories generally have less individual predictive power. The majority of the features have lower rankings, with the accuracy about 10%-27% inferior to UGR. The low performance indicates the indistinguishable nature among classes. In the readability category, for instance, similar scores are observed regardless of helpfulness of a review. Likewise, both helpful and unhelpful reviews are characterized by similar ratio of exclamatory and interrogative sentences, as well as misspellings. As a result, such features are less preferable in the helpfulness prediction task when used individually. Still, the slightly improved accuracy in the syntax category indicates that helpfulness is more related to the proportion of open-class words. In particular, ADJ generally performs better than other syntactic features due to the descriptive nature of products or general purchase satisfaction/dissatisfaction.

To better understand the behaviour of individual features across domains, the mean and standard deviation of the overall ranking of each feature are produced in Fig 2. The former describes the average performance of a feature, whereas the latter describes the stability of feature performance. As demonstrated, GV, SGNS, UGR, LIWC, and GI are the most ideal features with both excellent performance and stability. Those features show the feasibility of helpfulness prediction by modeling semantics and sentiment of product reviews. The remaining features, however, have either less satisfactory or stable performance.

Summary The findings and guidelines for review helpfulness prediction using individual features are summarized below:

1. Consider UGR, GV, and SGNS in the semantics category with higher priority since they are the most distinctive for informative reviews. In particular, GV performs better than SGNS in most cases.

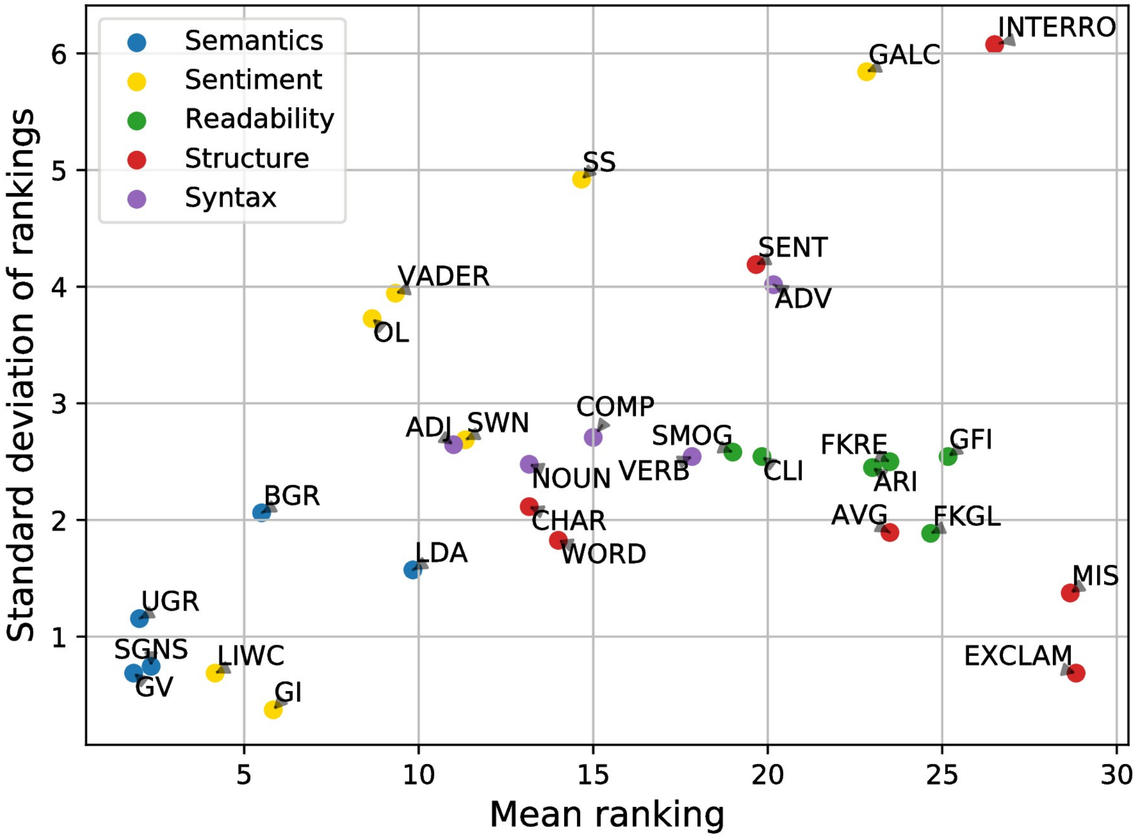

Fig 2. Feature performance versus stability.

https://doi.org/10.1371/journal.pone.0226902.g002

2. Features in the sentiment category are less effective in review helpfulness prediction compared with the three semantic counterparts. However, it is worth trying to replace the semantics with LIWC in small datasets.
3. Most features in the structure, readability, and syntax category are of minor predictive power and not suggested to use individually.

RQ2: Optimal combinations of features within each category. Table 5 presents the accuracy and ranking of the optimal feature combination in each category. As shown, BGR is the only feature not being selected in any scenarios due to the associated sparsity. Also, all domains demonstrate an identical ranking of feature categories, with the semantics, sentiment, and structure category playing the dominant role in helpfulness prediction.

To evaluate the benefit of combining multiple features within the same category, the optimal feature combination is compared with the most promising individual feature. As Fig 3 illustrates, in all but one cases, using multiple features achieves better performance on a category level. The rationale is that combining features provides new descriptive information of reviews and allows the information to complement one another. The improvement, depending on domains, tends to be more noticeable in the sentiment, readability, and structure category. On D1, GV alone reports higher accuracy than the optimal combination GV+SGNS in the semantics category since the domain has a large proportion of OOV words. As shown in Table 3, D1 has the shortest average length but highest OOV rate, which is about twice as

Table 5. Optimal combinations of features within each category.

	Category	Accuracy	Ranking	Combination
D1	Semantics	67.53	1	GV+SGNS
	Sentiment	67.29	2	LIWC+OL+GI
	Readability	54.06	5	SMOG+CLI+FKGL
	Structure	56.51	3	SENT+INTERRO+EXCLAM+AVG+WORD+CHAR
	Syntax	55.14	4	ADJ+ADV+COMP
D2	Semantics	76.67	1	SGNS+GV+LDA
	Sentiment	74.81	2	LIWC+VADER
	Readability	64.46	5	CLI+SMOG+GFI+ARI+FKGL
	Structure	66.67	3	CHAR+WORD+MIS+INTERRO
	Syntax	65.54	4	NOUN+ADJ+ADV+VERB
D3	Semantics	74.10	1	GV+SGNS
	Sentiment	73.98	2	LIWC+OL+GI+GALC+VADER
	Readability	59.12	5	SMOG
	Structure	64.10	3	CHAR+INTERRO+WORD
	Syntax	63.60	4	ADJ+ADV
D4	Semantics	81.32	1	UGR+GV
	Sentiment	78.81	2	LIWC+OL+GI+SS+SWN+VADER
	Readability	65.52	5	SMOG+ARI+FKGL
	Structure	68.68	3	CHAR+INTERRO+WORD+MIS+SENT
	Syntax	67.36	4	ADJ+ADV+VERB+NOUN
D5	Semantics	78.18	1	UGR+GV
	Sentiment	74.95	2	LIWC+GI+OL+VADER+SS+SWN
	Readability	62.05	5	SMOG+CLI+GFI+ARI+FKRE
	Structure	65.92	3	CHAR+INTERRO+WORD+EXCLAM+MIS
	Syntax	63.89	4	NOUN+VERB+ADV+COMP
D6	Semantics	75.49	1	UGR+SGNS+LDA
	Sentiment	71.45	2	LIWC+GI+OL+SWN+SS+GALC
	Readability	59.05	5	SMOG+ARI+FKRE+CLI+FKGL
	Structure	61.59	3	WORD+INTERRO+MIS+EXCLAM+SENT
	Syntax	59.11	4	NOUN+COMP+VERB+ADV

https://doi.org/10.1371/journal.pone.0226902.t005

much as other domains. Further manual inspection reveals that many OOV words are domain-specific terms such as names of mobile applications and mobile games. Moreover, only 53% of the OOV words overlap between the validation and test set. When converting reviews into embeddings, the OOV issue in the pre-trained SGNS model further affects the performance, which explains why GV+SGNS is worse than GV and less robust on D1.

The average number of features within each category used for helpfulness prediction is provided in Table 6. Frequent feature combination patterns that occur at least four times across domains are extracted via the PrefixSpan algorithm. The constant use of LIWC, SMOG, ADV, and INTERRO+WORD is observed, and thus it is recommended to include them for optimal feature combinations within the corresponding categories. As for the sentiment category, adding GI+OL (VADER alone) on top of LIWC can achieve higher performance in five (four) of six domains. Similarly, using INTERRO+WORD in conjunction with CHAR (MIS) can improve the structure category in five (four) domains. Furthermore, including one of ARI, CLI, and FKGL in addition to SMOG in the readability category helps to increase the accuracy in four domains. The same applies to ADJ and NOUN+VERB for ADV in syntax category.

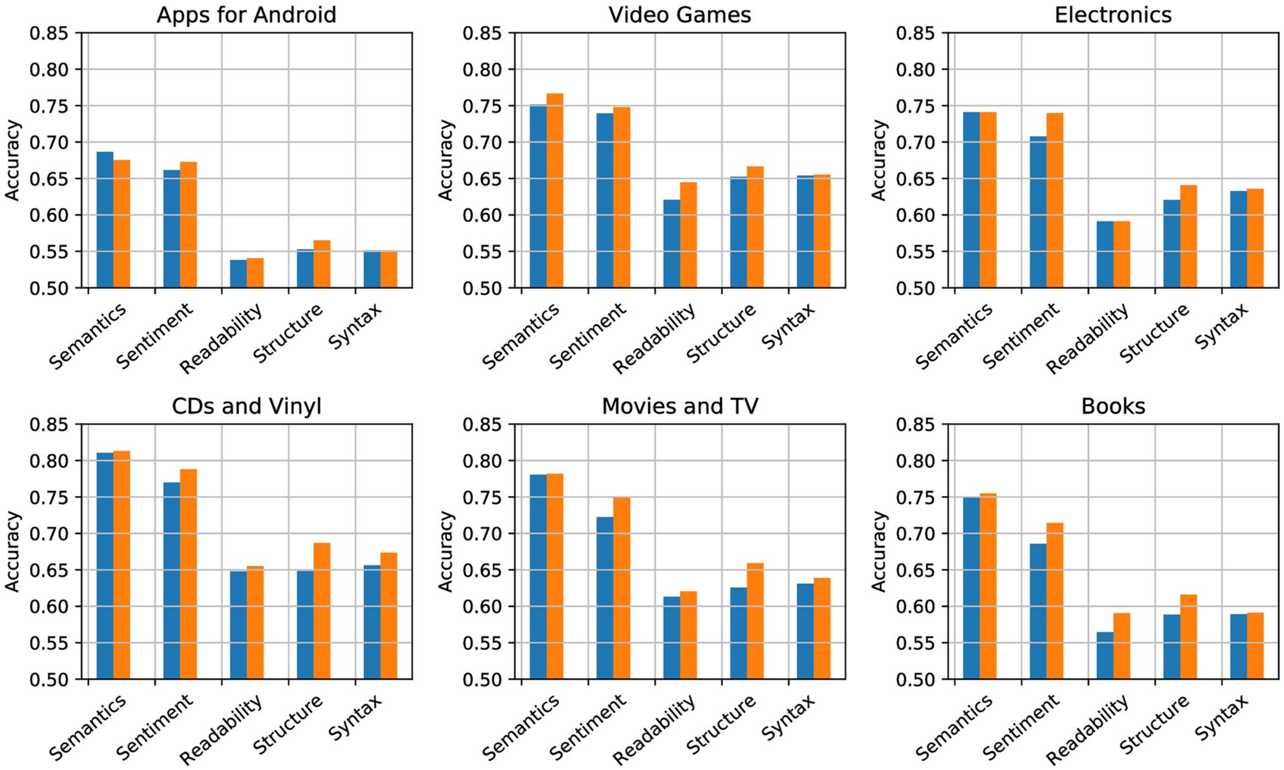

Fig 3. Comparison between max-performance individual feature (Blue) and max-performance combination of features (Orange) within each category.

https://doi.org/10.1371/journal.pone.0226902.g003

Finally, the semantics category tends to have various feature combinations, with GV and SGNS prevalent in most cases.

Summary The findings and guidelines for review helpfulness prediction using multiple features within each category are summarized below:

1. The optimal combination of semantic features consistently outperforms those in other categories in helpfulness prediction. Specifically, it is suggested that the combination includes GV as the first feature.

2. Regarding the sentiment (structure) category, it is recommended the optimal combination base on LIWC (INTERRO+WORD), and subsequently follow an addition of OL+GI (CHAR alone) to the corresponding category since performance gains are reported in most cases.

Table 6. Frequent feature combination patterns within each category.

Category	#Features	Pattern (Frequency)
Semantics	2.33 ± 0.47	GV (5);
Sentiment	4.67 ± 1.60	LIWC (6); GI+OL+LIWC (5); LIWC+VADER (4)
Readability	3.67 ± 1.49	SMOG (6); ARI+SMOG (4); CLI+SMOG (4); FKGL+SMOG (4)
Structure	4.67 ± 0.94	INTERRO+WORD (6); CHAR+INTERRO+WORD (5); INTERRO+MIS+WORD (4)
Syntax	3.50 ± 0.76	ADV (6); ADJ+ADV (4); ADV+NOUN+VERB (4)

https://doi.org/10.1371/journal.pone.0226902.t006

Table 7. Optimal combinations of all features.

	Accuracy	Sem.	Sen.	Read.	Str.	Syn.	Combination
D1	69.78	✓	✓			✓	GV+VADER+ADV+SWN
D2	76.80	✓			✓	✓	SGNS+GV+LDA+MIS+INTERRO+ADJ
D3	75.96	✓	✓	✓			GV+SGNS+LIWC+OL+GALC+ARI
D4	83.09	✓	✓	✓	✓	✓	UGR+LIWC+ARI+INTERRO+CLI+ADJ+VERB
D5	79.72	✓	✓		✓	✓	UGR+CHAR+LIWC+ADJ
D6	76.20	✓	✓	✓		✓	UGR+SMOG+ADV+VADER

https://doi.org/10.1371/journal.pone.0226902.t007

3. In regard to the readability (syntax) category, it is suggested the optimal feature combination base on SMOG (ADV), and subsequently follow an addition of one of ARI, CLI, and FKGL (ADJ, NOUN+VERB) to the corresponding category as this generally leads to visible performance gains.

RQ3: Optimal combinations of all features. The final result of review helpfulness prediction using the optimal feature combination from all categories are presented in Table 7. The optimal combinations contain four to seven features selected from only 18 out of the 30 features. Some of the 12 excluded features have excellent individual performance or are popular in category-level combinations, such as GI and WORD. The exclusion is due to features selected earlier (partly) contain information provided by those later. Despite no clear-cut patterns across domains are observed from the combinations, the semantics, sentiment, and syntax category play more important role in forming the optimal feature combinations. Especially, GV, UGR, LIWC and ADJ are used on half of the occasions.

The accuracy among the max-performance individual feature, optimal feature combination within each category, and optimal combination of all features is further compared in Fig 4. As

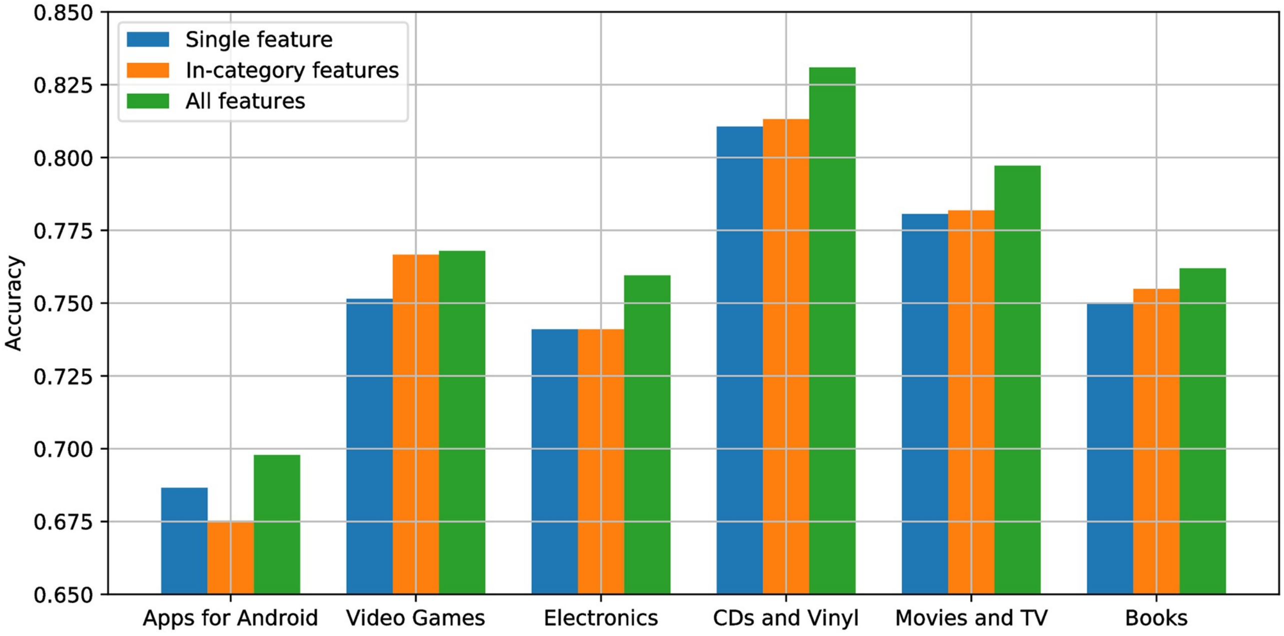

Fig 4. Comparison among the max-performance individual feature (Blue), max-performance in-category feature combination (Orange), and max-performance combination of all features (Green).

https://doi.org/10.1371/journal.pone.0226902.g004 shown, using features from multiple categories consistently achieves the highest performance. Similar to using multiple features within a category, the improvement lies in features from different categories together describe a review from multiple perspectives, making the vector representations more comprehensive.

Summary The findings and guidelines for review helpfulness prediction using features from multiple categories are summarized below:

1. Initialize the optimal combination with no more than three (usually one or two) semantic features, starting with GV or UGR, followed by SGNS.
2. Extend the combination with the remaining features in a forward selection manner. It is suggested that features mentioned in Table 4 have higher priority than those that are not.
3. Finalize the search by integrating the unused features into the combination using forward selection.

Implications

The last section has presented a series of optimal feature combinations across domains, along with their predictive power. The general guidelines for feature selection under the three scenarios are also summarized for future researchers. Extensive analysis shows that appropriately increasing the number of features can increase the performance of helpfulness prediction in almost all cases, regardless of feature categories and feature selection scenarios. As discussed, the performance gains lie in multiple features helping model a review's helpfulness information in a more comprehensive manner.

Nevertheless, it is worth noting that the semantics category contributes largely to the final performance. Throughout the study, using UGR alone accounts for 97.96% ± 0.35% of the accuracy compared with the optimal combination of all features across domains. The exclusive use of SGNS and GV can also yield comparable prediction performance. The empirical results demonstrate that combining many of the selected features, while leading to various performance gains, does not significantly improve helpfulness prediction. This contradicts prior studies largely combining multiple features without solid and sufficient justification. The extensive feature evaluation conducted in this study fills the gap of currently arbitrary feature selection process to review helpfulness evaluation.

The success of the semantics category can be explained from two perspectives: the encoding dimensionality and encoding methods. UGR, SGNS, and GV encode review content using more dimensions than other features. For many features that only have single dimension, encoding all text information into limited vector space can be challenging. On the other hand, both SGNS and GV achieve comparable performance to UGR with far fewer dimensions, showing that the information density of a feature varies from encoding methods. Even when used jointly, features beyond the semantics category are still less representative.

The dominance of review semantics also proves the feasibility of a new helpfulness prediction direction: Instead of laborious feature engineering, potential performance gains can be hopefully achieved by modeling sole semantic features from reviews via more advanced techniques, for example, state-of-the-art deep learning algorithms.

The authors hope that the exploration of potential factors behind the helpfulness evaluation process will deepen the insights obtained and contribute toward improved prediction system development.

Limitations

In terms of limitations, only the content-based features are considered due to their wide availability across various platforms. Also, the simplified forward selection search process for optimal feature combinations is adopted, thus not all possible scenarios are exhausted. Finally, the potential customer bias for the review helpfulness judgement (assertion of an initial belief), the common fraudulence issue (positive/negative review manipulation), as well as the sequential bias (early reviews receive disproportionately higher number of votes due to positive feedback loop [118]) are not taken into consideration due to the complex nature of such assessment.

Conclusions and future works

With the rapid development of Web 2.0, online product reviews have become an essential source of knowledge for most customers when making e-purchase decisions. In the deluge of data, to identify and recommend the informative reviews, rather than those of random quality is an important task. Feature-based methods have long been the paradigm of helpfulness prediction due to relatively simple implementation and effective interpretability. In the study presented, the 30 most frequent content-based features from five categories have been identified, and their extensive evaluation is conducted on six top domains of the largest publicly available Amazon 5-core dataset. The individual features, feature combinations within each category, and all feature combinations that lead to optimal performance have been studied. As stated by Charrada [31], the usefulness of a review is likely to depend on numerous factors that are difficult to isolate and study. The empirical results set comparable and reproducible baselines for review helpfulness prediction, and more importantly, highlight the feature combination patterns that lead to general good prediction performance, regardless of application domain and/or source platform.

Several significant findings and guidelines in feature selection are worth highlighting. Among many features, unigram TF-IDF and the two more recent pre-trained word embeddings yield strong predictive power across all domains, demonstrating the effectiveness of encoding semantics for helpfulness prediction. The LIWC dictionary achieves the closest performance to the three semantic features with far fewer feature dimensions, showing the feasibility of helpfulness prediction with fine-grained categorical sentiments. Another important finding is that appropriately combining features from multiple categories effectively improves the performance over individual features, or features from one single category. A good rule of thumb for feature selection is to initialize the search with semantic features, followed by features mentioned in Table 4, and finally the remaining content-based features. The findings and guidelines of this work can facilitate feature selection in review helpfulness prediction.

As final contribution, the authors have open sourced the computational framework that implements a holistic solution for feature-based helpfulness prediction. The dataset split configurations, pre-processed reviews, and extracted features used in the study have also been publicly released within research community for result reproducibility, fair comparison, and future improvement. The framework can be extended to support additional methods for feature extraction, feature selection, classification, and parameter tuning, allowing for more flexible investigation on the feature behaviors. Meanwhile, the off-the-shelf extracted features can help future researchers efficiently explore many possible feature combinations for the task.

The following directions will be addressed in the future. (1) Selected context-based features and less popular content-based features that are currently excluded will be taken into account to validate their predictive power. Especially, the social connection among reviewers and reviewer characteristics (e.g., reviewer age, the number of history reviews posted by a reviewer) will be emphasized. (2) The potential extension to other domains in the 5-core Amazon dataset and other platforms such as Yelp and TripAdvisor will be included following the holistic view on helpfulness prediction task. (3) The moderating factors will be explored, such as the product type and sequential bias. As stated by Ocampo et al. [4], it is perfectly sensible to expect the helpful reviews of different product types to be different. Given the context of a review, Sipos et al. [119] found that the helpfulness votes are often the consequence of its nearest neighbours. (4) More robust and sophisticated machine learning models will be employed to select representative features for helpfulness prediction. For example, recent explainable deep learning techniques can be employed to model semantic features from review content to free helpfulness prediction studies from heavy feature engineering.

Author Contributions

Conceptualization: Jiahua Du.

Data curation: Jiahua Du.

Investigation: Jiahua Du.

Methodology: Jiahua Du, Jia Rong.

Software: Jiahua Du.

Supervision: Jia Rong, Hua Wang, Yanchun Zhang.

Validation: Jiahua Du, Jia Rong, Sandra Michalska.

Visualization: Jiahua Du, Sandra Michalska.

Writing – original draft: Jiahua Du.

Writing – review & editing: Jia Rong, Sandra Michalska.

References

1. BrightLocal. Local Consumer Review Survey; 2016. Available from: https://www.brightlocal.com/learn/local-consumer-review-survey-2016/.
2. Momeni E, Cardie C, Diakopoulos N. How to Assess and Rank User-Generated Content on Web. In: Companion Proceedings of the The Web Conference 2018. WWW'18. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2018. p. 489–493. Available from: https://doi.org/10.1145/3184558.3186239.
3. Askalidis G, Malthouse EC. The Value of Online Customer Reviews. In: Proceedings of the 10th ACM Conference on Recommender Systems. RecSys'16. New York, NY, USA: ACM; 2016. p. 155–158. Available from: http://doi.acm.org/10.1145/2959100.2959181.
4. Ocampo Diaz G, Ng V. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 698–708. Available from: http://www.aclweb.org/anthology/P18-1065.
5. Kim SM, Pantel P, Chklovski T, Pennacchiotti M. Automatically Assessing Review Helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. EMNLP'06. Stroudsburg, PA, USA: Association for Computational Linguistics; 2006. p. 423–430. Available from: http://dl.acm.org/citation.cfm?id=1610075.1610135.
6. Li M, Huang L, Tan CH, Wei KK. Helpfulness of Online Product Reviews as Seen by Consumers: Source and Content Features. International Journal of Electronic Commerce. 2013; 17(4):101–136. https://doi.org/10.2753/JEC1086-4415170404
7. Martin L, Pu P. Prediction of Helpful Reviews Using Emotions Extraction. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI'14. AAAI Press; 2014. p. 1551–1557. Available from: http://dl.acm.org/citation.cfm?id=2892753.2892768.
8. Yang Y, Yan Y, Qiu M, Bao F. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short

Papers). Beijing, China: Association for Computational Linguistics; 2015. p. 38–44. Available from: http://www.aclweb.org/anthology/P15-2007.

9. Krishnamoorthy S. Linguistic features for review helpfulness prediction. Expert Systems with Applications. 2015; 42(7):3751–3759. https://doi.org/10.1016/j.eswa.2014.12.044.

10. Malik MSI, Hussain A. Helpfulness of product reviews as a function of discrete positive and negative emotions. Computers in Human Behavior. 2017; 73:290–302. https://doi.org/10.1016/j.chb.2017.03.053.

11. Cheng YH, Ho HY. Social influence's impact on reader perceptions of online reviews. Journal of Business Research. 2015; 68(4):883–887. https://doi.org/10.1016/j.jbusres.2014.11.046.

12. Hu YH, Chen K. Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings. International Journal of Information Management. 2016; 36(6, Part A):929–944. https://doi.org/10.1016/j.ijinfomgt.2016.06.003.

13. Lu Y, Tsaparas P, Ntoulas A, Polanyi L. Exploiting Social Context for Review Quality Prediction. In: Proceedings of the 19th International Conference on World Wide Web. WWW'10. New York, NY, USA: ACM; 2010. p. 691–700. Available from: http://doi.acm.org/10.1145/1772690.1772761.

14. Tang J, Gao H, Hu X, Liu H. Context-aware Review Helpfulness Rating Prediction. In: Proceedings of the 7th ACM Conference on Recommender Systems. RecSys'13. New York, NY, USA: ACM; 2013. p. 1–8. Available from: http://doi.acm.org/10.1145/2507157.2507183.

15. Zhu L, Yin G, He W. Is this opinion leader's review useful? Peripheral cues for online review helpfulness. Journal of Electronic Commerce Research. 2014; 15(4):267.

16. Fang B, Ye Q, Kucukusta D, Law R. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. Tourism Management. 2016; 52:498–506. https://doi.org/10.1016/j.tourman.2015.07.018.

17. Willemsen LM, Neijens PC, Bronner F, de Ridder JA. "Highly Recommended!" The Content Characteristics and Perceived Usefulness of Online Consumer Reviews. Journal of Computer-Mediated Communication. 2011; 17(1):19–38. https://doi.org/10.1111/j.1083-6101.2011.01551.x

18. Kuan K, Hui KL, Prasarnphanich P, Lai HY. What Makes a Review Voted? An Empirical Investigation of Review Voting in Online Review Systems. Journal of the Association for Information Systems. 2015; 16:48–71. https://doi.org/10.17705/1jais.00386

19. Yang SB, Shin SH, Joun Y, Koo C. Exploring the comparative importance of online hotel reviews' heuristic attributes in review helpfulness: a conjoint analysis approach. Journal of Travel & Tourism Marketing. 2017; 34(7):963–985. https://doi.org/10.1080/10548408.2016.1251872

20. Karimi S, Wang F. Online review helpfulness: Impact of reviewer profile image. Decision Support Systems. 2017; 96:39–48. https://doi.org/10.1016/j.dss.2017.02.001.

21. Kwok L, Xie KL. Factors contributing to the helpfulness of online hotel reviews: Does manager response play a role? International Journal of Contemporary Hospitality Management. 2016; 28(10):2156–2177. https://doi.org/10.1108/IJCHM-03-2015-0107

22. Lee PJ, Hu YH, Lu KT. Assessing the helpfulness of online hotel reviews: A classification-based approach. Telematics and Informatics. 2018; 35(2):436–445. https://doi.org/10.1016/j.tele.2018.01.001.

23. Zeng YC, Ku T, Wu SH, Chen LP, Chen GD. Modeling the helpful opinion mining of online consumer reviews as a classification problem. International Journal of Computational Linguistics & Chinese Language Processing, Volume 19, Number 2, June 2014. 2014; 19(2).

24. Akbarabadi M, Hosseini M. Predicting the helpfulness of online customer reviews: The role of title features. International Journal of Market Research. 2018; p. 1470785318819979. https://doi.org/10.1177/1470785318819979

25. Vo C, Duong D, Nguyen D, Cao T. From Helpfulness Prediction to Helpful Review Retrieval for Online Product Reviews. In: Proceedings of the Ninth International Symposium on Information and Communication Technology. ACM; 2018. p. 38–45.

26. Haque ME, Tozal ME, Islam A. Helpfulness Prediction of Online Product Reviews. In: Proceedings of the ACM Symposium on Document Engineering 2018. DocEng'18. New York, NY, USA: ACM; 2018. p. 35:1–35:4. Available from: http://doi.acm.org/10.1145/3209280.3229105.

27. Chen J, Zhang C, Niu Z. Identifying Helpful Online Reviews with Word Embedding Features. In: Lehner F, Fteimi N, editors. Knowledge Science, Engineering and Management. Cham: Springer International Publishing; 2016. p. 123–133.

28. Fan M, Feng C, Guo L, Sun M, Li P. Product-Aware Helpfulness Prediction of Online Reviews. In: The World Wide Web Conference. WWW'19. ACM. New York, NY, USA: ACM; 2019. p. 2715–2721. Available from: http://doi.acm.org/10.1145/3308558.3313523.

29. Saumya S, Singh JP, Baabdullah AM, Rana NP, Dwivedi YK. Ranking online consumer reviews. Electronic Commerce Research and Applications. 2018; 29:78–89. https://doi.org/10.1016/j.elerap.2018.03.008.

30. Hoffait AS, Ittoo A, Schyns M. Assessing and predicting review helpfulness: Critical review, open challenges and research agenda. In: 29ème conférence européenne sur la recherche opérationnelle (EURO2018); 2018.

31. Charrada EB. Which One to Read? Factors Influencing the Usefulness of Online Reviews for RE. In: 2016 IEEE 24th International Requirements Engineering Conference Workshops (REW); 2016. p. 46–52.

32. Zhang Z, Varadarajan B. Utility Scoring of Product Reviews. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management. CIKM'06. New York, NY, USA: ACM; 2006. p. 51–57. Available from: http://doi.acm.org/10.1145/1183614.1183626.

33. Liu H, Gao Y, Lv P, Li M, Geng S, Li M, et al. Using Argument-based Features to Predict and Analyse Review Helpfulness. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics; 2017. p. 1358–1363. Available from: https://www.aclweb.org/anthology/D17-1142.

34. Ngo-Ye TL, Sinha AP. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. Decision Support Systems. 2014; 61:47–58. https://doi.org/10.1016/j.dss.2014.01.011.

35. Ngo-Ye TL, Sinha AP, Sen A. Predicting the helpfulness of online reviews using a scripts-enriched text regression model. Expert Systems with Applications. 2017; 71:98–110. https://doi.org/10.1016/j.eswa.2016.11.029.

36. Ma Y, Xiang Z, Du Q, Fan W. Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning. International Journal of Hospitality Management. 2018; 71:120–131. https://doi.org/10.1016/j.ijhm.2017.12.008.

37. Hwang SY, Lai CY, Jiang JJ, Chang S. The identification of Noteworthy Hotel Reviews for Hotel Management. In: Pacific Asia Journal of the Association for Information Systems (PACIS); 2014. p. 1–17.

38. Zhang Z, Ma Y, Chen G, Wei Q. Extending associative classifier to detect helpful online reviews with uncertain classes. In: Proceedings of the 2015 Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology; 2015. p. 1134–1139.

39. Zheng X, Zhu S, Lin Z. Capturing the essence of word-of-mouth for social commerce: Assessing the quality of online e-commerce reviews by a semi-supervised approach. Decision Support Systems. 2013; 56:211–222. https://doi.org/10.1016/j.dss.2013.06.002.

40. Dong R, Schaal M, Smyth B. Topic Extraction from Online Reviews for Classification and Recommendation. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI 13); 2013. p. 1310–1316.

41. Liu Y, Jiang C, Ding Y, Wang Z, Lv X, Wang J. Identifying helpful quality-related reviews from social media based on attractive quality theory. Total Quality Management & Business Excellence. 2017; 0(0):1–20.

42. Xiong W, Litman D. Automatically Predicting Peer-review Helpfulness. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers—Volume 2. HLT'11. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011. p. 502–507. Available from: http://dl.acm.org/citation.cfm?id=2002736.2002836.

43. Maroun LB, Moro MM, Almeida JM, Silva APC. Assessing Review Recommendation Techniques Under a Ranking Perspective. In: Proceedings of the 27th ACM Conference on Hypertext and Social Media. HT'16. New York, NY, USA: ACM; 2016. p. 113–123. Available from: http://doi.acm.org/10.1145/2914586.2914598.

44. Mertz M, Korfiatis N, Zicari RV. Using Dependency Bigrams and Discourse Connectives for Predicting the Helpfulness of Online Reviews. In: Hepp M, Hoffner Y, editors. E-Commerce and Web Technologies. Cham: Springer International Publishing; 2014. p. 146–152.

45. Cao Q, Duan W, Gan Q. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems. 2011; 50(2):511–521. https://doi.org/10.1016/j.dss.2010.11.009.

46. Moghaddam S, Jamali M, Ester M. ETF: Extended Tensor Factorization Model for Personalizing Prediction of Review Helpfulness. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. WSDM'12. New York, NY, USA: ACM; 2012. p. 163–172. Available from: http://doi.acm.org/10.1145/2124295.2124316.

47. Mukherjee S, Popat K, Weikum G. Exploring Latent Semantic Factors to Find Useful Product Reviews. In: Proceedings of the 2017 SIAM International Conference on Data Mining; 2017. p. 480–488. Available from: https://epubs.siam.org/doi/abs/10.1137/1.9781611974973.54.

48. Xiang Z, Du Q, Ma Y, Fan W. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tourism Management. 2017; 58:51–65. https://doi.org/10.1016/j.tourman.2016.10.001.

49. Yin D, Bond SD, Zhang H. Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews. MIS Q. 2014; 38(2):539–560. https://doi.org/10.25300/MISQ/2014/38.2.10

50. Yin D, Mitra S, Zhang H. Research Note—When Do Consumers Value Positive vs. Negative Reviews? An Empirical Investigation of Confirmation Bias in Online Word of Mouth. Information Systems Research. 2016; 27(1):131–144. https://doi.org/10.1287/isre.2015.0617

51. Agnihotri A, Bhattacharya S. Online Review Helpfulness: Role of Qualitative Factors. Psychology & Marketing. 2016; 33(11):1006–1017. https://doi.org/10.1002/mar.20934

52. Chua AYK, Banerjee S. Helpfulness of user-generated reviews as a function of review sentiment, product type and information quality. Computers in Human Behavior. 2016; 54:547–554. https://doi.org/10.1016/j.chb.2015.08.057.

53. Jing L, Xin X, Ngai E. An examination of the joint impacts of review content and reviewer characteristics on review usefulness-the case of Yelp. com. In: 22nd Americas Conference on Information Systems (AMCIS); 2016. p. 1–5.

54. Li ST, Pham TT, Chuang HC. Do reviewers' words affect predicting their helpfulness ratings? Locating helpful reviewers by linguistics styles. Information & Management. 2019; 56(1):28–38. https://doi.org/10.1016/j.im.2018.06.002.

55. Park YJ. Predicting the Helpfulness of Online Customer Reviews across Different Product Types. Sustainability. 2018; 10(6).

56. Siering M, Muntermann J, Rajagopalan B. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. Decision Support Systems. 2018; 108:1–12. https://doi.org/10.1016/j.dss.2018.01.004.

57. Liu J, Cao Y, Lin CY, Huang Y, Zhou M. Low-Quality Product Review Detection in Opinion Summarization. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 334–342. Available from: http://www.aclweb.org/anthology/D/D07/D07-1035.

58. Liu Y, Jin J, Ji P, Harding JA, Fung RYK. Identifying helpful online reviews: A product designer's perspective. Computer-Aided Design. 2013; 45(2):180–194. https://doi.org/10.1016/j.cad.2012.07.008.

59. Singh JP, Irani S, Rana NP, Dwivedi YK, Saumya S, Roy PK. Predicting the "helpfulness" of online consumer reviews. Journal of Business Research. 2017; 70:346–355. https://doi.org/10.1016/j.jbusres.2016.08.008.

60. Baek H, Ahn J, Choi Y. Helpfulness of Online Consumer Reviews: Readers' Objectives and Review Cues. International Journal of Electronic Commerce. 2012; 17(2):99–126. https://doi.org/10.2753/JEC1086-4415170204

61. Baek H, Lee S, Oh S, Ahn JH. Normative social influence and online review helpfulness: Polynomial modeling and response surface analysis. Journal of Electronic Commerce Research. 2015; 16:290–306.

62. Lee M, Jeong M, Lee J. Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. International Journal of Contemporary Hospitality Management. 2017; 29(2):762–783. https://doi.org/10.1108/IJCHM-10-2015-0626

63. Salehan M, Kim DJ. Predicting the performance of online consumer reviews: A sentiment mining approach to big data analytics. Decision Support Systems. 2016; 81:30–40. https://doi.org/10.1016/j.dss.2015.10.006.

64. Mousavizadeh M, Koohikamali M, Salehan M. The Effect of Central and Peripheral Cues on Online Review Helpfulness: A Comparison between Functional and Expressive Products. In: Thirty Sixth International Conference on Information Systems, Fort Worth (ICIS); 2015. p. 1–22.

65. Ghose A, Ipeirotis PG. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. IEEE Transactions on Knowledge and Data Engineering. 2011; 23(10):1498–1512. https://doi.org/10.1109/TKDE.2010.188

66. Liu Z, Park S. What makes a useful online review? Implication for travel product websites. Tourism Management. 2015; 47:140–151. https://doi.org/10.1016/j.tourman.2014.09.020.

67. Wu PF, Van Der Heijden H, Korfiatis N. The influences of negativity and review quality on the helpfulness of online reviews. In: International conference on information systems; 2011. p. 1–10.

68. 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. 2012; 11(3):205–217. https://doi.org/10.1016/j.elerap.2011.10.003.

69. Park S, Nicolau JL. Asymmetric effects of online consumer reviews. Annals of Tourism Research. 2015; 50:67–83. https://doi.org/10.1016/j.annals.2014.10.007.

70. Wang Y, Wang J, Yao T. What makes a helpful online review? A meta-analysis of review characteristics. Electronic Commerce Research. 2018.

71. Lee S, Choeh JY. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Systems with Applications. 2014; 41(6):3041–3046. https://doi.org/10.1016/j.eswa.2013.10.034.

72. Bjering E, Havro LJ, Moen Ø. An empirical investigation of self-selection bias and factors influencing review helpfulness. International Journal of Business and Management. 2015; 10(7):16.

73. Pan Y, Zhang JQ. Born Unequal: A Study of the Helpfulness of User-Generated Product Reviews. Journal of Retailing. 2011; 87(4):598–612. https://doi.org/10.1016/j.jretai.2011.05.002.

74. Einar B. Online Consumer Reviews: The Moderating Effect of Product Category. Norwegian University of Science and Technology; 2014.

75. Chevalier JA, Mayzlin D. The Effect of Word of Mouth on Sales: Online Book Reviews. Journal of Marketing Research. 2006; 43(3):345–354. https://doi.org/10.1509/jmkr.43.3.345

76. Mudambi SM, Schuff D. Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. MIS Quarterly. 2010; 34(1):185–200. https://doi.org/10.2307/20721420

77. Huang AH, Chen K, Yen DC, Tran TP. A study of factors that contribute to online review helpfulness. Computers in Human Behavior. 2015; 48:17–27. https://doi.org/10.1016/j.chb.2015.01.010.

78. O'Mahony MP, Smyth B. A Classification-based Review Recommender. In: Bramer M, Ellis R, Petridis M, editors. Research and Development in Intelligent Systems XXVI. London: Springer London; 2010. p. 49–62.

79. Otterbacher J.'Helpfulness' in Online Communities: A Measure of Message Quality. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI'09. New York, NY, USA: ACM; 2009. p. 955–964. Available from: http://doi.acm.org/10.1145/1518701.1518848.

80. Huang AH, Yen DC. Predicting the Helpfulness of Online Reviews—A Replication. International Journal of Human–Computer Interaction. 2013; 29(2):129–138. https://doi.org/10.1080/10447318.2012.694791

81. Wu J. Review popularity and review helpfulness: A model for user review effectiveness. Decision Support Systems. 2017; 97:92–103. https://doi.org/10.1016/j.dss.2017.03.008.

82. Lee S, Choeh JY. Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees. Management Decision. 2017; 55(4):681–700. https://doi.org/10.1108/MD-06-2016-0398

83. Zhao P, Wu J, Hua Z, Fang S. Finding eWOM customers from customer reviews. Industrial Management & Data Systems. 0;0(0):null.

84. Hong Y, Lu J, Yao J, Zhu Q, Zhou G. What Reviews Are Satisfactory: Novel Features for Automatic Helpfulness Voting. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR'12. New York, NY, USA: ACM; 2012. p. 495–504. Available from: http://doi.acm.org/10.1145/2348283.2348351.

85. Qazi A, Syed KBS, Raj RG, Cambria E, Tahir M, Alghazzawi D. A concept-level approach to the analysis of online review helpfulness. Computers in Human Behavior. 2016; 58:75–81. https://doi.org/10.1016/j.chb.2015.12.028.

86. Salton G, Buckley C. Term-weighting Approaches in Automatic Text Retrieval. Inf Process Manage. 1988; 24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0

87. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003; 3(Jan):993–1022.

88. Yang Y, Chen C, Bao FS. Aspect-Based Helpfulness Prediction for Online Product Reviews. In: 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI); 2016. p. 836–843.

89. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013. p. 3111–3119.

90. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–1543. Available from: http://www.aclweb.org/anthology/D14-1162.

91. Chatterjee S. Drivers of helpfulness of online hotel reviews: A sentiment and emotion mining approach. International Journal of Hospitality Management. 2019; p. 102356. https://doi.org/10.1016/j.ijhm.2019.102356.

92. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. The University of Texas at Austin; 2015.

93. J SP, F BR, Zvi NJ, M OD. The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science. 1966; 7(4):484–498.

94. Scherer KR. What are emotions? And how can they be measured? Social Science Information. 2005; 44(4):695–729. https://doi.org/10.1177/0539018405058216

95. Hu M, Liu B. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2004. p. 168–177.

96. Baccianella S, Esuli A, Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of LREC. vol. 10; 2010.

97. Fellbaum C. WordNet: An electronic lexical database (Language, Speech, and Communication). The MIT Press; 1998.

98. Thelwall M, Buckley K, Paltoglou G. Sentiment Strength Detection for the Social Web. J Am Soc Inf Sci Technol. 2012; 63(1):163–173. https://doi.org/10.1002/asi.21662

99. Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In: International AAAI Conference on Web and Social Media; 2014. p. 216–225. Available from: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

100. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench—a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Science. 2016; 5(1):23. https://doi.org/10.1140/epjds/s13688-016-0085-1

101. Yue L, Chen W, Li X, Zuo W, Yin M. A survey of sentiment analysis in social media. Knowledge and Information Systems. 2019; 60(2):617–663. https://doi.org/10.1007/s10115-018-1236-4

102. DuBay WH. Smart Language: Readers, Readability, and the Grading of Text. ERIC; 2007.

103. Farr JN, Jenkins JJ, Paterson DG. Simplification of Flesch Reading Ease Formula. Journal of applied psychology. 1951; 35(5):333. https://doi.org/10.1037/h0062427

104. Kincaid JP, Aagard JA, O'Hara JW, Cottrell LK. Computer readability editing system. IEEE Transactions on Professional Communication. 1981; PC-24(1):38–42. https://doi.org/10.1109/TPC.1981.6447821

105. Gunning R. The Fog Index After Twenty Years. Journal of Business Communication. 1969; 6(2):3–13. https://doi.org/10.1177/002194366900600202

106. Mc Laughlin GH. SMOG grading–a new readability formula. Journal of reading. 1969; 12(8):639–646.

107. Smith EA, Kincaid JP. Derivation and Validation of the Automated Readability Index for Use with Technical Materials. Human Factors. 1970; 12(5):457–564. https://doi.org/10.1177/001872087001200505

108. Coleman M, Liau TL. A computer readability formula designed for machine scoring. Journal of Applied Psychology. 1975; 60(2):283. https://doi.org/10.1037/h0076540

109. Benjamin RG. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. Educational Psychology Review. 2012; 24(1):63–88. https://doi.org/10.1007/s10648-011-9181-8

110. Xiong W, Litman D. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics; 2014. p. 1985–1995. Available from: http://www.aclweb.org/anthology/C14-1187.

111. Jindal N, Liu B. Mining Comparative Sentences and Relations. In: Proceedings of the 21st National Conference on Artificial Intelligence—Volume 2. AAAI'06. AAAI Press; 2006. p. 1331–1336. Available from: http://dl.acm.org/citation.cfm?id=1597348.1597400.

112. He R, McAuley J. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In: Proceedings of the 25th International Conference on World Wide Web. WWW'16. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee; 2016. p. 507–517. Available from: https://doi.org/10.1145/2872427.2883037.

113. Danescu-Niculescu-Mizil C, Kossinets G, Kleinberg J, Lee L. How Opinions Are Received by Online Communities: A Case Study on Amazon.Com Helpfulness Votes. In: Proceedings of the 18th International Conference on World Wide Web. WWW'09. New York, NY, USA: ACM; 2009. p. 141–150. Available from: http://doi.acm.org/10.1145/1526709.1526729.

114. Roy G, Datta B, Mukherjee S. Role of electronic word-of-mouth content and valence in influencing online purchase behavior. Journal of Marketing Communications. 2018; 0(0):1–24.

115. Bird S, Klein E, Loper E. Natural Language Processing with Python. 1st ed. O'Reilly Media, Inc.; 2009.

116. Vapnik VN. The Nature of Statistical Learning Theory. Berlin, Heidelberg: Springer-Verlag; 1995.

117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12:2825–2830.

118. Wan Y. The Matthew effect in social commerce. Electronic Markets. 2015; 25(4):313–324. https://doi.org/10.1007/s12525-015-0186-x

119. Sipos R, Ghosh A, Joachims T. Was This Review Helpful to You?: It Depends! Context and Voting Patterns in Online Content. In: Proceedings of the 23rd International Conference on World Wide Web. WWW'14. New York, NY, USA: ACM; 2014. p. 337–348. Available from: http://doi.acm.org/10.1145/2566486.2567998.