Attribute Level Rating System Using User Reviews and Ratings

KunduruVenkata Chaitanya Lakshmi,D.Kumar

DEPT OF CSE

Dr.Samuel George Institute of Engineering & Technology, Markapur, India

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.

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.

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.

1. **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.

2. **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.

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:

1. **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.
2. **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.

3. **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.

1. **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.
2. **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.
3. **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.
4. **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.
5. **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.
6. **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.
7. **VADER** As a novelty, the study also adopts the Valence Aware Dictionary and sEntimentReasoner [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].

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 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:

1. ***RQ1***: *What is the effect of individual features on review helpfulness prediction across domains?*
2. ***RQ2***: *What are the optimal combinations of features within a category for review helpfulness prediction across domains?*
3. ***RQ3***: *What are the optimal combinations of all features for review helpfulness prediction across domains?*

4. ***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.

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.

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.

References

1. BrightLocal. Local Consumer Review Survey;

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.

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.

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.

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

6. **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.

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

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:

Krishnamoorthy S. Linguistic features for review helpfulness prediction. Expert Systems with Applications. 2015;42(7):3751–3759.

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.

Cheng YH, Ho HY. Social influence's impact on reader perceptions of online reviews. Journal of Business Research. 2015;68(4):883–887.

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.

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:

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.

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.

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.

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.

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.

Student details:

Name:KunduruVenkata Chaitanya Lakshmi

Mail:chaitukun99@gmail.com

Dr.Samuel George Institute of Engineering & Technology, Markapur, India

Guide details:

D.Kumarreceived B.Tech (CSIT) Degree from JNT University in 2006 and M.Tech (CSE) Degree from JNTUK Kakinada in 2011. He has 11 years of teaching experience. He joined as Assistant Professor in Dr.Samuel George Institute of Engineering & Technology, Markapur, India in 2006. Presently he is working as Associate Professor in CSE Dept. His Interested research areas are Image Processing and Computer Networks. He attended Various National Workshops and Conferences.