Helpfulness Prediction of Online Product Reviews

Md. Enamul Haque, Mehmet Engin Tozal, and Aminul Islam
University of Louisiana at Lafayette
School of Computing and Informatics
Lafayette, LA 70504 USA
{enamul,metozal,aminul}@louisiana.edu

ABSTRACT

The simple question "Was this review helpful to you?" increases an estimated $2.7B revenue to Amazon.com annually [1]. In this paper, we propose a solution to the problem of electronic product review accumulation using helpfulness prediction. The popularity of e-commerce and online retailers such as Amazon, eBay, Yelp, and TripAdvisor are largely relying on the presence of product reviews to attract more customers. The major issue for the user submitted reviews is to quantify and evaluate the actual effectiveness by combining all the reviews under a particular product. With the varying size of reviews for each product, it is quite cumbersome for the customers to get hold of the overall helpfulness.Therefore, we propose a feature extraction technique that can quantify and measure helpfulness for each product based on user submitted reviews.

CCS CONCEPTS

• **Computer systems organization → Document Content Analysis**; *Linguistic and semantic (content) analysis*; classification;

KEYWORDS

product review, market analysis, helpfulness, semantic analysis

1 INTRODUCTION

Online marketplaces have rapidly grown their popularity among customers over the last few years. For instance, Amazon earned sales revenue of $107 billions in 2015 [2]. The major driving force of this recent boost in purchase behavior is vastly influenced by the customer reviews. Users tend to verify the comments from other users who already purchased the product and shared their thoughts. A major drawback for this system is the delay in generating the review feedback. A user is required to wait for a certain time before a new review of a particular product accumulates enough feedback to make a decision. In this work, we are specifically focused on analyzing the helpfulness phenomena on Amazon product review. Figure 1 shows an example of a review with moderate helpfulness count. This particular method is known as *X of Y* approach where *X* refers to the number of users who clicked the review as helpful out of total *Y* users. To illustrate the objective of this paper, let us consider a simple example from Amazon where a typical user review exploits different parameters such as username, star-rating, review description, and timestamps. Customers are asked to submit a review and a star rating once they buy a product from Amazon. The review count becomes quite large for a particular product when large numbers of customers leave reviews and ratings. Therefore a new customer faces difficulty in reading all those reviews and finding whether they are helpful for decision making. Additionally, the star rating is mostly not aligned with the review text.

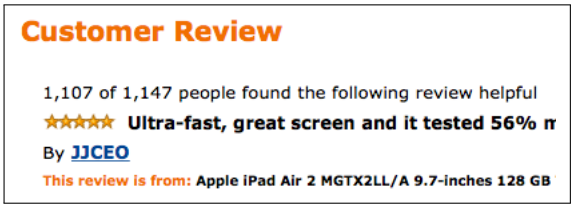

Figure 1: Use of X of Y helpfulness review system on Amazon. This particular review has 1147 customer review feedback from the "was this review helpful?" button.

Recent trend in product purchase from online marketplaces (Amazon, eBay, newegg) shows that users tend to buy recommended products more than the non-recommended ones. Thus, the product reviews from the users play a big role in the financial aspects for these big companies. We can consider the product reviews as a form of passive recommendation process or exposure of user sentiment for their purchase history. The major drawback for this review is to confidently match them with the star-rating and helpfulness. Additionally, there are very few products that have adequate review information for helpfulness prediction. These two uncontrollable factors play a key part to obtain the overall sense of helpfulness from every review text and associated score. To avoid this, we propose Lexical, Structural, and Semantic, Features (LS2F for short) to predict the helpfulness via automated summarization and semantic analysis of review texts.

To summarize, our work makes the following contributions. Firstly, we formalize the problem of review helpfulness without using *X of Y* score. Secondly, we introduce a hybrid feature generation technique (LS2F) that makes use of lexical, structural, and semantic analysis of review texts. Thirdly, we use word embedding to generate fixed size feature vectors from reviews of arbitrary lengths. Finally, we use *Flesch reading ease* [4] score to support the feature generation technique in review helpfulness prediction. The experimental results show that our proposed method can significantly improve the decision making process for online purchasing. In addition, our proposed method can be adopted to other domains such as restaurants, doctors, and movies to assist users during decision making processes.

[1] http://www.uie.com/articles/magicbehindamazon
[2] http://www.marketwatch.com/investing/stock/amzn/financials

DocEng, August 2018, Halifax, Nova Scotia, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

2 RELATED WORK

Online product reviews have strong and influential impact on the consumers as they tend to use the reviews for purchasing decisions [2, 3, 10]. The drawback is that a popular product usually has too many reviews for the consumers to read. Additionally, new products might have too few reviews. Thus, the review text, ranking, and rating must be recommended to the consumers. Specifically, review helpfulness plays a vital role in product rankings and recommendations [5, 9, 11, 14, 16]. Another major aspect of modeling these reviews is to capture effective feature information from the dataset [1, 7]. We can relate our problem with finding the truthfulness of any comment used in online platforms as well. In [15, 17, 19], truth assessment and semantic analysis for product review text is performed on Amazon.com data.

Aspect based product review helpfulness is proposed in [18]. As related products share common aspects such as shipping and warranty, the authors propose an aspect extraction model making use of product category information to balance the aspects of a general category and those of subcategories under them. On top of this, a two layer regressor is trained for helpfulness prediction. Experimental results demonstrate 7% additional prediction accuracy compared to baseline methods on five product category data collected from Amazon.

In [8], the authors used both qualitative and quantitative factors as helpfulness prediction of reviews. Their findings suggest that word count has a threshold in its effects on review helpfulness. Additionally, reviewer experience and their impact were not statistically significant predictors for helpfulness. Past helpfulness records tends to predict future helpfulness ratings. Finally they conclude that characteristics of reviewers and review messages have a varying degree of impact on review helpfulness.

Most of the previous works either use vector space model or traditional semantic analysis with document features to improve the prediction. Considering that in mind, we use several linguistic (lexical, semantic), anatomical, and metadata information with word embedding for feature extraction.

3 PROPOSED APPROACH

Our proposed approach (LS2F) is depicted in Figure 2 that consists of five major steps. Firstly, review with star-rating score is fed into the feature generation engine that includes four different functions such as lexical, structural, semantic and combined. Secondly, generated features are used as input to standard machine learning classification algorithm. In the third step, four different trained models are built using the classification algorithm. In the fourth and fifth steps, test reviews are evaluated based on the pre-trained classification models for review helpfulness detection. We consider the review helpfulness prediction as a binary classification problem once reviews are labeled as helpful or not helpful with the aid of *Flesch reading ease*.

3.1 Dataset Description

We collected a subset of Amazon product review data used in [6]. We focus on three product categories such as *automotive, musical instruments, and sports & outdoors*. The dataset statistics is provided in Table 1. We use the first three data categories (* marked) to test our approach. Note that the additional datasets are used to generate word embedding models of different vector lengths that ensures a fixed length feature vector from variable length review text. Data and code are available online [3].

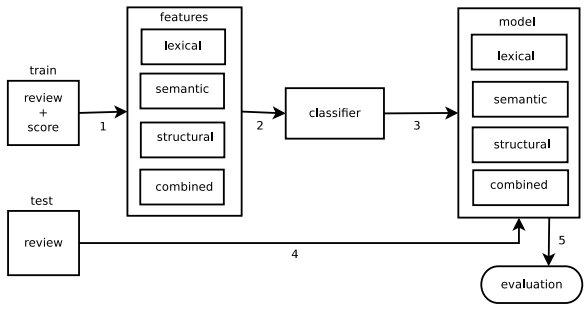

Figure 2: High level overview of LS2F process blocks consisting of feature generation and classification steps.

Every review contains user id, product category, review texts, helpfulness counts (X of Y), and star-rating. We only use review text and star rating to create training classification models depicted in Figure 2 at step 3.

Table 1: Review dataset statistics

Dataset	Review count
Automotive *	20473
Musical instruments *	10261
Sports and Outdoors *	296337
Instant video	37126
Patio, Lawn and Garden	13272
Office products	53258

3.2 Feature Generation

We use three individual and a combined categories of features for the helpfulness prediction task.

3.2.1 Structural features. Structural features of a review mostly exploit different kinds of content frequency measures. Our intuition behind using structural features is to capture the impact of individual tokens, sentences, and paragraphs. The structural features used in our experiments include length of review, number of sentences, character count, number of all capitalized words, and number of question marks. Length of a review intuitively tells us whether the review is important or not. Usually, shorter reviews are less informative than the longer ones. All capitalized words are kept as we think it can represent the high importance on some specifics in the review. Additionally, frequent question marks present in the text is generally not useful too as it may ask for suggestions instead of an honest feedback.

3.2.2 Lexical features. Both unigram and bigram features are computed for each review using *tf-idf* method as it provides more weight to the less frequent words compared to high frequent words.

The *tf-idf* statistics for word w in review r is computed using $tf - idf = tf \times idf$, where tf refers to the normalized frequency of the token in the whole review data,

$$tf(w, r) = 0.5 + 0.5 \frac{f(w, r)}{max(f(t, r))} \quad (1)$$

[3] https://github.com/enamul-haque/DocEng2018

Table 2: Precision, recall, and macro F1 metrics for all features and three datasets.

Datasets	Lexical			Structural			Semantic			LS2F		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Automotive	0.58±0.030	0.73±0.046	0.64±0.014	0.67±0.018	0.66±0.016	0.66±0.007	0.73±0.006	0.73±0.006	0.73±0.006	0.76±0.006	0.76±0.006	0.76±0.008
Musical	0.57±0.016	0.74±0.047	0.64 ±0.019	0.65±0.012	0.64±0.018	0.65±0.014	0.73±0.004	0.73±0.004	0.73±0.005	0.75±0.005	0.75±0.005	0.75±0.006
Sports	0.58±0.007	0.80±0.017	0.67±0.004	0.69±0.006	0.67±0.005	0.68±0.002	0.74±0.008	0.74±0.008	0.74±0.007	0.77±0.007	0.77±0.007	0.77±0.006

and the inverse document frequency is computed using Equation 2.

$$idf(w, R) = \log \frac{N}{1 + |r \in R : w \in r|} \tag{2}$$

where $N = |R|$ and t refers to all the words present in review r.

3.2.3 Semantic features. First, we use skip-gram model [13] on all the review datasets present in Table 1 to build word embedding. The basic skip-gram formulation defines $p(w_{k+j}|w_k; \theta)$ using the softmax function (Equation 3):

$$p(w_{k+j}|w_k; \theta) = \frac{exp(e^T_{w_{k+j}} e_{w_k})}{\sum_{w=1}^{N} e^T_w e_{w_k})} \tag{3}$$

where, θ is the model parameter that needs to be learned, $e_{w_k}, e_{w_{k+J}}$ are the word embedding learned from word2vec [12]. Then the log-likelihood $H(\theta)$ of the whole training data is maximized as follows using Equation 4

$$H(\theta) = \sum_{w_k, w_{k+j}} \log p(w_{k+j}|w_k; \theta) \tag{4}$$

Next, we create features from the above model representing semantic embedding vectors for reviews.

3.2.4 Combined features. We combine all the lexical, structural, and semantic features and use them as a separate feature setting during classification process as well. In natural language processing and other linguistic research, one major requirement is to find out how readable a text is before doing any preprocessing and normalization. There are standard methods to find out how much a sentence or whole article is readable. The readability index indicates the difficulty level of comprehension for a particular sentence or article. We add reading ease score using *Flesch reading ease* (F_{fres}) to label each review as helpful or not-helpful using Equation 5.

$$F_{fres} = 206.835 - 1.015 \times \frac{C_{words}}{C_{sentences}} - 84.6 \times \frac{C_{syllables}}{C_{words}} \tag{5}$$

where C_{words} denotes total word count, $C_{sentences}$ refers to the total number of sentences, and $C_{syllables}$ refers to number of syllables in each review. We consider a training review helpful if F_{fres} and the original review scores are at least 80 and 4, respectively. One can tune these parameters for different domains based on empirical evaluations on training reviews.

3.3 Classification

We use standard decision tree classifier on training review $D = \{X^n, Y^n\}$ where $X^n = \{x_1, x_2, \ldots, x_n\}$ refers to the features of n samples and $Y^n \in \{0, 1\}$ refers to the label vector. We use *gini impurity* as criterion function that measures the quality of a split during the training process. The number of features to consider for the best split is set as all available features. Additionally, we set the minimum number of samples to be at the leaf nodes as 1. The minimum number of samples required to split an internal node is set as 2.

4 EXPERIMENTAL RESULTS

In this section, we explain the experimental results of our proposed LS2F method. First, we use all the datasets from Table 1 to create word embeddings. We use varying vector lengths to train the word embeddings starting from 5 to 100 with intervals of 5. Then we use only autmotive, musical instruments, and sports category data for evaluation. We use *Decision tree* as the classifier for evaluating our feature extraction process and classifying the review into either helpful or not helpful classes. We use 10-fold cross validation for all experimental settings. Figure 3 shows the accuracy of different feature selection approaches for three datasets. For all datasets, the combined feature performs better than other individual features. Note that, the accuracy of combined and semantic feature selection methods are averaged over all vector lengths.

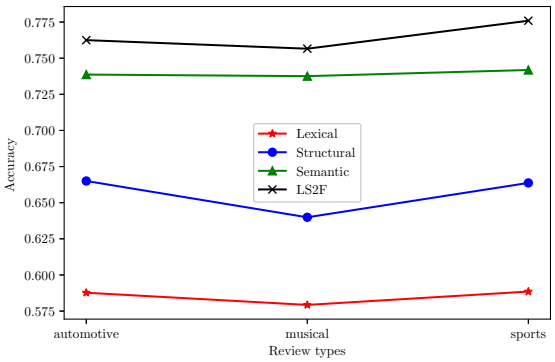

Figure 3: Accuracy of different feature selection methods.

Table 2 demonstrates average precision, recall, and macro F1 with standard deviation for all feature selection methods on three datasets. In all datasets, the combined method performs better for helpfulness classification. However, both semantic and combined feature selection require to generate features from already built word embedding models.

Next, we present the classification performance between semantic and combined (LS2F) methods using accuracy, precision, recall, and macro F1 metrics with respect to different vector lengths. We show the metrics on multiple vector lengths to demonstrate the sensitivity of these methods. Figure 4, 5, and 6 present all those metrics for Amazon automotive, musical instruments, and sports data. Note that, semantic method tends to perform better with the increased vector length. On the other hand, combined (LS2F) method performs better with smaller vector length. Therefore, we suggest using LS2F with maximum vector length of 30 to avoid additional computational complexity.

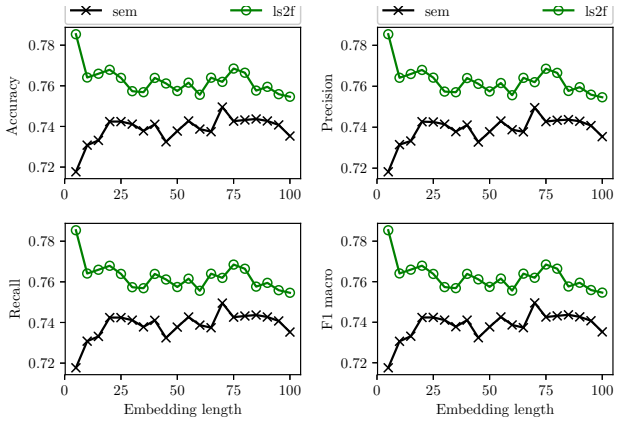

Figure 4: Automotive

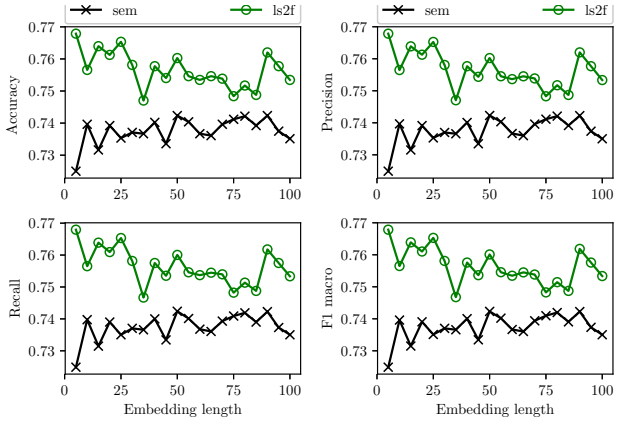

Figure 5: Musical instruments

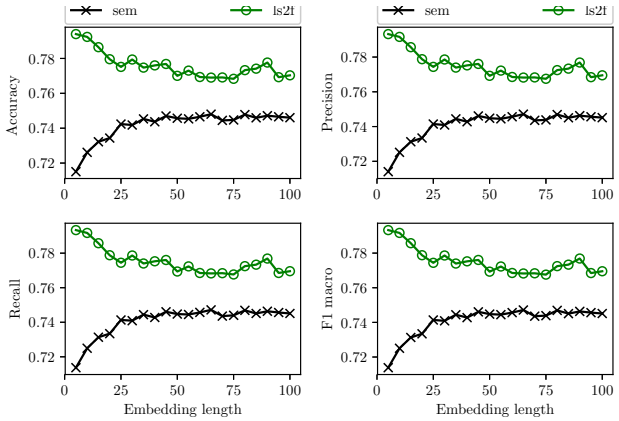

Figure 6: Sports and outdoors

5 CONCLUSION

In this paper, we present several feature extraction methods along with *Flesch reading ease* score to better understand a product review with respect to helpfulness. We evaluated the methods on large datasets consisting of amazon product reviews of different domains. The results clearly show that both *Flesch reading ease* and LS2F feature selection methods have significant impact on the overall helpfulness decisions. In addition to more accurately predicting product reviews as helpful, LS2F exploits efficiency with smaller vector embeddings. Traditional models (based on ratings alone) are difficult to apply to new products, that have too few ratings to model the helpfulness measure. In contrast, our proposed method allows us to uncover such factors from even a single review. We assume that a review with complex sentence structure would baffle the readers even if it contains important information. *Flesch reading ease* helps to unfold the complexity in that aspect. Therefore, our proposed combined feature extraction can greatly improve the helpfulness prediction for online product reviews.

REFERENCES

[1] Laura Connors, Susan M Mudambi, and David Schuff. 2011. Is it the review or the reviewer? A multi-method approach to determine the antecedents of online review helpfulness. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 1–10.

[2] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. The dynamics of online word-of-mouth and product salesâĂŤAn empirical investigation of the movie industry. *Journal of retailing* 84, 2 (2008), 233–242.

[3] Bin Fang, Qiang Ye, Deniz Kucukusta, and Rob Law. 2016. Analysis of the perceived value of online tourism reviews: Influence of readability and reviewer characteristics. *Tourism Management* 52 (2016), 498–506.

[4] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[5] Anindya Ghose and Panagiotis G Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on* 23, 10 (2011), 1498–1512.

[6] Ruining He and Julian McAuley. 2016. 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*. International World Wide Web Conferences Steering Committee, 507–517.

[7] Yu Hong, Jun Lu, Jianmin Yao, Qiaoming Zhu, and Guodong Zhou. 2012. 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*. ACM, 495–504.

[8] Albert H Huang, Kuanchin Chen, David C Yen, and Trang P Tran. 2015. A study of factors that contribute to online review helpfulness. *Computers in Human Behavior* 48 (2015), 17–27.

[9] Haijing Liu, Yang Gao, Pin Lv, Mengxue Li, Shiqiang Geng, Minglan Li, and Hao Wang. 2017. Using Argument-based Features to Predict and Analyse Review Helpfulness. *arXiv preprint arXiv:1707.07279* (2017).

[10] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2007. ARSA: a sentiment-aware model for predicting sales performance using blogs. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 607–614.

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

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[14] Susan M Mudambi and David Schuff. 2010. What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly* 34, 1 (2010), 185–200.

[15] Ndapandula Nakashole and Tom M Mitchell. 2014. Language-Aware Truth Assessment of Fact Candidates.. In *ACL (1)*. 1009–1019.

[16] Aika Qazi, Karim Bux Shah Syed, Ram Gopal Raj, Erik Cambria, Muhammad Tahir, and Daniyal Alghazzawi. 2016. A concept-level approach to the analysis of online review helpfulness. *Computers in Human Behavior* 58 (2016), 75–81.

[17] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. 2014. 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*. ACM, 337–348.

[18] Yinfei Yang, Cen Chen, and Forrest Sheng Bao. 2016. Aspect-Based Helpfulness Prediction for Online Product Reviews. In *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*. IEEE, 836–843.

[19] Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Sheng Bao. [n. d.]. Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews. *Volume 2: Short Papers* ([n. d.]), 38.