Identifying Helpful Online Reviews with Word Embedding Features

Jie Chen[1], Chunxia Zhang[2], and Zhendong Niu[1(✉)]

[1] School of Computer Science and Technology,
Beijing Institute of Technology, Beijing, China
{bit_chenjie,zniu}@bit.edu.cn
[2] School of Software, Beijing Institute of Technology, Beijing, China
cxzhang@bit.edu.cn

Abstract. The advent of Web 2.0 has enabled users to share their opinions via various social media websites. People's decision-making process is strongly influenced by online reviews. Predicting the helpfulness of reviews can help to save time and find helpful suggestions. However, most of previous works focused on exploring new features with external data source, such as user's profile, semantic dictionaries, etc. In this paper, we maintain that the helpfulness of an online review can be predicted by knowing only word embedding information. Word embedding information is a kind of word semantic representation computed with word context. We hypothesize that word embedding information would allow us to accurately predict the helpfulness of an online review. The experiments were conducted to prove this hypothesis and the results showed a substantial improvement compared with baselines of features previously used.

Keywords: Automatic helpfulness voting · User preference · Helpfulness classification

1 Introduction

The internet contains a wealth of reviews and opinions on any topics. User-generated contents come in various forms and sizes, objective opinions and subjective opinions. Postings in internet forums and user comments in websites are the important sources of information. The decision-making process of people is affected by the opinions of others in the information age [4]. When a person wants to change a job, he or she will start by searching for reviews and opinions written by the employees and former employees regarding the companies in his or her wish list. However, the number of reviews is often very large, which causes lots of reviews and opinions to be unnoticed, even though some of them are very helpful. As a result, predicting the helpfulness of a review is very important.

Many websites rank reviews by their published time, product rating, user voting, etc. Compared to sort by published time and product rating, the user voting method seems to be better and more helpful, since its results are cumulative from

F. Lehner and N. Fteimi (Eds.): KSEM 2016, LNAI 9983, pp. 123–133, 2016.
DOI: 10.1007/978-3-319-47650-6_10

lots of visitors. For example, in Amazon.com, they employ a voting system to collect the feedback by asking "was this review helpful to you? Yes/no". It would be useful to rank reviews based on the quality as soon as these reviews are shown. This would save lots of time on surfing the web-pages and finding helpful reviews. However, user voting mechanisms are controversial, including the imbalance vote bias, the winner cycle bias and the early bird bias [14]. These kinds of bias show that voting system is not the best choice for ranking user-generated contents.

Previous works approximate the ground truth of helpfulness from users' voting results. If there are X of Y users who consider a review to be helpful, then the helpfulness score is X/Y. However, it is hard to collect the right value of Y. For example, when a user opens a product details page with many reviews, he just read the basic information about the product and leaves. It's hard to decide whether we should add 1 to all the reviews in this page. In addition, the review voting itself can be influenced by many factors, such as page structure adjustment, review recommendation, etc.

In this paper, we model the problem of predicting review helpfulness score as a regression problem and analysis performance of different features used in previous researches. Many researches [3,5,10,18,25,26] focus on exploring new features to model review sentences, then gain better results on the task. However, novel features are limited by the data resources, language of sentence, third-party tools, etc. In order to overcome these limitations, word embedding features are introduced to model sentences. Experimental results show that word embedding features outperform other features used in previous research. From the point view of dimensionality reduction, we also compared the Unigram features with Latent Semantic Analysis (LSA) to other features. Result showed that LSA technology with unigram features gain better performance.

The following section discusses related works about review helpfulness prediction. The definition of helpfulness prediction and the format of data used in our experiments are given in Sect. 3. Details about features used in our approach are introduced in Sect. 4. Experiments and evaluation metrics are described in Sect. 5. In Sect. 6 we discuss and analysis the results. We conclude and present directions for future research in the last section.

2 Related Works

Presenting the helpful content to visitors is an important component for any content-centric websites. Engineers of such kind of website have been committed to improve the click rate of reviews, either using normal ranking mechanism or carefully improved mechanism. Consequently, there has been plenty of researches on various aspects of ratings and the quality of review contents.

Some of them focus on finding the most helpful features for predicting the quality of review content [9,10,14,20,25]. Meanwhile, there are also some researches focus on exploring new algorithms [5,13,22,26,27].

In the research of Kim et al. [9], lexical, structural, syntactic, semantic and meta-data related features were used for automatic helpfulness prediction. Text

surface features and unigrams are proved to be the most helpful features and widely used in later researches.

Zhang and Varadarajan [27] built a regression model by incorporating a diverse set of features, and achieved highly competitive performance of utility scoring on three real-world data sets. Their experiments also proved that the shallow syntactic features turned out to be the most influential predictors.

Liu [14] worked on how to detect low quality reviews. They introduced features to model the informativeness, subjectiveness and readability of a review and classified them into high or low qualities.

Yang et al. [25] hypothesized that helpfulness is an internal property of text and introduced LIWC and INQUIRER semantic features to model the review text. Their experiments showed that two semantic features could accurately predict helpfulness scores and greadly improve the performance compared with features previously used.

RevRank is an unsupervised algorithm to ranking helpfulness of online book reviews [22]. They first constructed a lexicon of dominant terms across reviews, then a virtual core review based on this lexicon was created. They used the distance between the virtual review and each real review to determine overall helpfulness ranking.

Hong et al. [5] developed a binary helpfulness classification system. The system used a set of novel features based on needs fulfillment, information reliability and sentiment divergence measure. Their system outperformed some earlier researches with the same dataset.

Lee and Choeh [13] proposed a helpfulness prediction neural network model and made use of products, review characteristics, and reviewer information as features. This is the first study to predict helpfulness using neural networks. The authors proved that their model outperform the conventional linear regression model analysis in predicting helpfulness.

Rong Zhang et al. [26] proposed a comment-based collaborative filtering approach which captures correlations between hidden aspects in review comments and numeric ratings. They also estimated the aspects of comments based on profiles of users and items, the model outperformed baseline system in Chinese review dataset.

Srikumar [10] proposed a predictive model extracts novel linguistic category features by analysing the textual content of review. He made use of review metadata, subjectivity and readability related features for helpfulness prediction. He proved that the proposed linguistic category features were better predictors of review helpfulness for experience goods.

3 Task Definition

In this section, we defined the task of review helpfulness prediction (RHP), and we introduce the data format of Amazon.com reviews. This data have been successfully used in related review helpfulness prediction tests. All the data analysis, illustrations and experiments are based on the dataset.

Table 1. An example of reviews in Amazon dataset

Tag	Value
Member id	A1004AX2J2HXGL
Product id	B00064LJVE
Date	January 13, 2005
Number of helpful feedbacks	5
Number of feedbacks	15
Rating	1.0
Title	Into the woods
Body	M. is a hack, a second-place magician in a high school talent show. He's drawn comparisons to Hitchcock and Spielberg - in the same sentence no less? Resting on the laurels of exactly ONE good movie, he manages to eek out a career for himself. Since THE SIXTH SENSE, his movies have gotten progressively worse. UNBREAKABLE was fair at best. An interesting idea with a dull, rumbling ride to the conclusion. SIGNS was a very rough movie to watch. The characters were cookie-cutter samples of human emotion and conflict - toss in a guy in an alien suit and you have what exactly?...

3.1 Task of RHP

The task of RHP aims to automatically predict the helpfulness score of a specific product's reviews. In this task, in order to eliminate the interference of external information, only text information is considered rather than any other human interaction information, such as user background, user level etc. The RHP should assign a high score to a review which gains a high manual voting score and assigns a low value to a review which gains a low manual voting score.

Therefore, given a set of reviews, the RHP should output a score list of each review's helpfulness score. We treat this as a regression task of reviews regarding their helpfulness.

3.2 Amazon.com Data Format

We use the Amazon review data which was prepared for Opinion Spam Detection [6]. This dataset provides 5.8 million reviews about products sold in Amazon. Each review contains product number, date, number of helpful feedback, number of feedbacks, rating, title and body. An example of reviews in this dataset was shown in Table 1.

In this paper, we only consider the body part of each review as the available local resources for RHP. The 'body' part gives the content of a review. Other items, such as 'title', 'ratings', are not totally available in this dataset for each

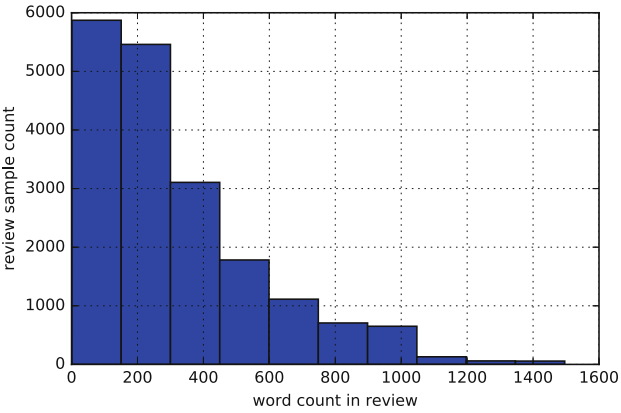

Fig. 1. Word count distribution in the corpus

product. In order to avoid dealing with missing information, we do not use the title and other fields which are optional in this experiment. The length of 'body' part of reviews in this dataset is various. Word count distribution about review sample in the corpus is given in Fig. 1.

4 Features

To make the experiment reproducible, only text-based features are used and discussed in this work. Text surface features [9,15,17,24], Unigram features [1,9, 24], Part-of-speech (POS) features [9,10,15] are widely used in previous research work, then we considered them as baselines.

4.1 Surface Features

Following previous researches [24,25], text surface features used are shown in Table 2. These features have proven effectiveness and are easy to implement for a new corpus.

4.2 Unigram Features

It is proved that the unigram feature is a reliable feature for review helpfulness prediction in previous work [25]. After removing all the stop words and word frequency lower than 10, we build a word dict. Each review is represented as a word vector, in which the value is *TF-IDF* weight.

In addition, for getting the semantic features and saving the training time, we also employ the LSA [11] technology to perform dimensionality reduction of vector space. Each review represented with unigram features is re-represented in a lower dimension vector space.

Table 2. The description of surface features

Number	Feature description
1	The number of sentences in the review
2	The number of words in the review
3	The average length of sentences
4	The number of exclamation marks
5	The percentage of question sentences
6	The ratio of uppercase to lowercase characters in the review text

4.3 POS Features

The efficiency of part of speech (POS) features has been proved in previous research and there is not much difference among ways of implementing of POS features, which made it to be a reasonable feature in RHP. We use the following POS features: number of Noun words, number of Adjective words, number of Verb words, and number of Adverb words.

4.4 Word Embedding Features

We use the Genism tool[1] to learn the word embeddings from the provided 5.8 M Amazon product reviews, with the following settings:

1. we removed non-english reviews, which reduces the corpus to 5.5 M reviews.
2. we used the skip-gram model with window size 5 and filtered words with a frequency less than 10.

We use word embeddings of size 100, which means the dimension of output vector is 100. This setting is same with default settings of other tools, such as *word2vec*. The details of computing word embedding features are introduced in previous researches [16,19].

5 Experiments and Results

We empirically evaluate our approach, described in Sect. 4, by comparing the performance of different features combination. Below, we describe our experimental setup, choose evaluation metric, present our results and analyze different features' performance.

5.1 Evaluation Setup and Evaluation Metrics

In order to predict the helpfulness score of reviews, we focus on reviews with helpful feedback voting in the Amazon dataset. For removing duplicate reviews in the dataset, we use Hong's [5] deduplication method to filter the redundant reviews. There are too many reviews without voting information or feedback information. For this, we filter out the reviews with feedbacks count lower than 100.

[1] http://radimrehurek.com/gensim.

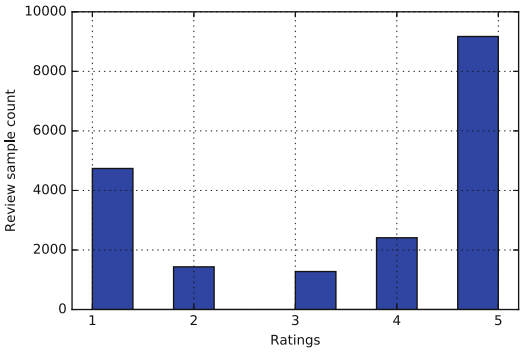

Fig. 2. Distribution of review ratings

The final dataset involves 19,030 reviews on 9805 products. The distribution of review ratings is shown in Fig. 2. To obtain the helpfulness voting score, we follow the annotation of review quality defined by Liu [14]. On the basis, we tested each group of feature combination on the whole dataset.

In the training process, we use three regression methods including Linear Regression (LR), Linear Support Vector Regression (LSVR) and Support Vector Regression (SVR)[21]. In the evaluation process, we run 10-fold cross validation. The original Amazon ratings are not used as ground truth, because the ratings are stared by their author for the product not for the review text.

In our experiment, we use the Root Mean Square Error (RMSE) metric to evaluate the performance.

5.2 Results

In this experiment, we test the performance with single feature groups described in Sect. 4 and results are shown in Table 3. Different combinations of features are also tested and results are shown in Table 4.

Table 3. RMSE of single feature

Features	LR	LSVR	SVR
Surface features (SF)	0.314	0.591	0.341
POS features (PF)	0.323	0.349	0.337
Unigram features (UF)	0.376	0.283	0.343
LSA + Unigram features (LUF)	**0.245**	0.252	0.285
Embedding features (EF)	**0.248**	0.254	0.250

Feature Performance. The first group of results is baselines of this experiment. As described in previous researches, SF features focus on statistics information and they are used as the baseline.

The second group of results is about Unigram features with LSA technology and word embedding features. From the results, LUF gains better performance than word embedding features. However, the difference between them is not large. Compared to Unigram features without LSA, the LUF improves the performance a lot. The word embedding features also perform better than Unigram features.

Table 4. RMSE of feature combinations

Features	LR	LSVR	SVR
SF + PF	**0.305**	0.324	0.318
UF + SF	0.363	**0.280**	0.331
UF + PF	0.365	**0.282**	0.343
UF + SF + PF	0.362	**0.280**	0.331
LUF + SF	**0.244**	0249	0.278
LUF + PF	**0.245**	0.251	0.285
LUF + SF + PF	**0.243**	0.249	0.278
EF + SF	**0.243**	0.249	0.246
EF + PF	**0.247**	0.253	0.250
EF + SF + PF	**0.242**	0.248	0.246
UF + SF + PF + EF	**0.357**	0.274	0.294
LUF + SF + PF + EF	**0.238**	0.241	0.249

The first group in Table 4 is about combinations of UF, SF and PF, we use them as feature combination baselines.

The second group shows the performance about LUF with other features. Compared to combinations about Unigram features, this group makes notable improvements (about 13 %).

The third group shows the performance about word embedding features with other features. From the results, combinations with EF show better performance than UF combinations with UF and LUF. This can verify the efficiency of EF features.

The last group in Table 4 shows combinations with all the features. From the results, combinations with LUF, SF, PF, and EF perform better than UF, SF, PF and EF. The results show than LUF can improve the performance again. In addition, this combination shows the best performance among all the combinations.

Table 5. Model performance

	LR	LSVR	SVR
Single feature	4	1	0
Feature combination	8	4	0

Regression Model Performance. Furthermore, we try to find the relationship between features and the underlying model of helpfulness prediction. For the result of each feature in Table 3 and each feature combination in Table 4, we count the best performance of three regression models. The statistical results are shown in Table 5. Linear regression gets the best in both single feature and feature combination results. Linear SVR also performs better than SVR. It shows that linear relation exists between these features and helpfulness of reviews.

6 Conclusions and Future Work

Until now, the helpfulness of reviews has been well studied with kinds of features, including Unigram features, text structural features, part-of-speech features, semantic features etc. However, features used in previous research so far produce results that are too unreliable to become a basis of a discourse-level prediction. We assert that the helpfulness of an online review should be predicted with its hidden structural information and lexical information. In this paper, we first give the definition of review helpfulness prediction, and then introduce word embedding features to predict the helpfulness score. Our experiments show that the word embedding features can lead to a substantial improvement over previous features. In addition, we test the LSA technology on Unigram features and the results show that LSA can lead to a substantial improvement over Unigram features. As a result of different features combinations, we try to analyze the hidden relationship between features and helpfulness of a review.

In the future, we will test the prediction performance on different corpus and try to do prediction with deep learning [12]. Convolutional neural network (CNN) has been proved to be efficient in modeling sentences [8], text categorization [7,23] and machine reasoning [2]. Further, we will investigate how to bring CNN into this research and predict the helpfulness of reviews.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61370137, 61272361) and the 111 Project of Beijing Institute of Technology.

References

1. Agarwal, D., Chen, B.C., Pang, B.: Personalized recommendation of user comments via factor models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 571–582. Association for Computational Linguistics (2011)

2. Bottou, L.: From machine learning to machine reasoning. Mach. Learn. **94**(2), 133–149 (2014)
3. Chen, C.C., Tseng, Y.D.: Quality evaluation of product reviews using an information quality framework. Decis. Support Syst. **50**(4), 755–768 (2011)
4. Duan, W., Gu, B., Whinston, A.B.: The dynamics of online word-of-mouth andproduct salesan empirical investigation of the movie industry. J. Retail. **84**(2), 233–242 (2008)
5. 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 2012, pp. 495–504. ACM, New York (2012)
6. Jindal, N., Liu, B.: Opinion spam and analysis. In: International Conference on Web Search and Data Mining, pp. 219–230 (2008)
7. Johnson, R., Zhang, T.: Effective use of word order for text categorization with convolutional neural networks. arXiv preprint arXiv:1412.1058 (2014)
8. Kalchbrenner, N., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188 (2014)
9. Kim, S.M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006, pp. 423–430. Association for Computational Linguistics, Stroudsburg (2006)
10. Krishnamoorthy, S.: Linguistic features for review helpfulness prediction. Expert Syst. Appl. **42**(7), 3751–3759 (2015)
11. Landauer, T.K.: An introduction to latent semantic analysis. Discourse Process. **25**(2), 259–284 (1998)
12. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053 (2014)
13. Lee, S., Choeh, J.Y.: Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Syst. Appl. **41**(6), 3041–3046 (2014)
14. Liu, J., Cao, Y., Lin, C.Y., Huang, Y., Zhou, M.: Low-quality product review detection in opinion summarization. In: EMNLP-CoNLL, pp. 334–342 (2007)
15. Liu, Y., Jin, J., Ji, P., Harding, J.A., Fung, R.Y.K.: Identifying helpful online reviews: a product designer's perspective. Comput. Aided Des. **45**(2), 180–194 (2013)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Momeni, E., Tao, K., Haslhofer, B., Houben, G.J.: Identification of useful user comments in social media: a case study on flickr commons. In: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2013, pp. 1–10. ACM, New York (2013)
18. Otterbacher, J.: ̀helpfulnessín online communities: a measure of message quality. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 955–964. ACM, New York (2009)
19. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, vol. 14, pp. 1532–1543 (2014)
20. Siersdorfer, S., Chelaru, S., Nejdl, W., San Pedro, J.: How useful are your comments? Analyzing and predicting youtube comments and comment ratings. In: Proceedings of the 19th International Conference on World Wide Web, WWW 2010, pp. 891–900. ACM, New York (2010)
21. Smola, A.J., Scholkopf, B.: A tutorial on support vector regression. Stat. Comput. **14**(3), 199–222 (2004)

22. Tsur, O., Rappoport, A.: RevRank: a fully unsupervised algorithm for selecting the most helpful book reviews. In: AAAI Conference on Weblogs and Social Media - ICWSM 2009 (2009)
23. Wang, P., Xu, J., Xu, B., Liu, C.L., Zhang, H., Wang, F., Hao, H.: Semantic clustering and convolutional neural network for short text categorization. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, vol. 2, pp. 352–357 (2015)
24. 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, HLT 2011, vol. 2, pp. 502–507. Association for Computational Linguistics, Stroudsburg (2011)
25. 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: Short Papers, vol. 2, pp. 38–44. Association for Computational Linguistics, Beijing (2015)
26. Zhang, R., Gao, Y., Yu, W., Chao, P., Yang, X., Gao, M., Zhou, A.: Review Comment Analysis for Predicting Ratings. In: Dong, X.L., Yu, X., Li, J., Sun, Y. (eds.) WAIM 2015. LNCS, vol. 9098, pp. 247–259. Springer, Heidelberg (2015). doi:10.1007/978-3-319-21042-1_20
27. Zhang, Z., Varadarajan, B.: Utility scoring of product reviews. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, CIKM 2006, pp. 51–57. ACM, New York (2006)