Student Project Exposé: Evaluating Review Helpfulness Prediction across Domains

Hermann Stolte
<stoltehe@informatik.hu-berlin.de>

Humboldt-University Berlin

March 1, 2019

Academic advisors: Prof. Ulf Leser <leser@informatik.hu-berlin.de>
M.Sc. Mario Sänger <saengema@informatik.hu-berlin.de>

1 Introduction

When deciding whether to buy a product or use a service, a valuable source of information are other peoples experiences and opinions. In online marketplaces such as Amazon or iTunes, every user can read and write reviews about the offered products and services. Due to the huge amount of reviews and opinions, it is very time-consuming to build an informed opinion. It is impossible to read every single review before making a decision. One approach to circumvent the problem is to let customers vote on the *helpfulness* of reviews. The customer can answer the question "Was this review helpful? (yes/no)" and the respective vote count is displayed below the review (e.g. "7 of 8 people found this helpful.") All existing reviews for a product can then sorted by this score, resulting in the most "helpful" reviews being listed at the top. The described voting process, however, is not optimal. As noted by [1], the amount of votes per review ends up being unbalanced, with a small proportion of all reviews having the majority of all votes. As a consequence, only a small subset of reviews will have a high helpfulness score and thus be shown to the customers. Many, potentially helpful reviews will almost never be considered in the customers decision-making process.

A *review helpfulness prediction* algorithm, applied in a system that automatically sorts all reviews by their predicted helpfulness score, can help to overcome the described problem. Furthermore, it could be used to predict the helpfulness of a review-draft in the review writing process and to gain an understanding

of what makes a review helpful. Ultimately, this could improve of the overall quality and helpfulness of online reviews.

In this student project, state of the art approaches to review helpfulness prediction will be implemented and validated on reviews from a broad selection of product categories. As the product type has an influence on review helpfulness [1], a special focus will lie on the difference between *search* and *experience* goods: For comparing search goods (e.g. products from the category "Electronics"), objective attributes are more relevant, whereas for experience goods (e.g. products from the category "Movies and Tv"), subjective attributes are more relevant [2, 3]. A subsequent master thesis will focus on the interpretability and explainability of helpfulness prediction by, on the one hand, developing an approach that allows certain insight into the sample-specific reasoning behind a prediction and, on the other hand, performing a qualitative analysis of important textual features across domains.

The exposé continues as follows. In section 2 an overview of the related work is given, including relevant domains, datasets and existing approaches to the review helpfulness prediction task. Section 4 introduces the planned experiments for the student project based on existing approaches.

2 Problem Setting

Given a set of reviews R, let r_{pos} be the number of positive and r_{neg} the number of negative helpfulness votes for a review $r \in R$. Then, for every review r that has received at least a single helpfulness vote ($r_ps + r_{neg} \geq 1$), the helpfulness score $h(r)$ is defined as

$$h(r) = \frac{r_{pos}}{r_{pos} + r_{neg}},\ r_{pos}, r_{neg} \in [0, 1] \tag{1}$$

The helpfulness score can directly be used as label for a *regression* problem. It can also be transformed into discrete class labels by binning the value range of h (e.g. $[0, 0.5]$ corresponding to "not helpful" and $]0.5, 1.0]$ corresponding to "helpful"), resulting in lables for a *classification* problem. Based on the helpfulness score, it is also possible to retrieve a *ranking* of a set of reviews, by sorting the reviews by the helpfulness score.

Common evaluation metrics for classification problems are the F1-Score, precision and recall. For regression, common metrics are root mean squared error (RMSE), mean squared error (MSE), mean absolute error (MAE) and the Pearsson correlation between the ground truth scores and predicted scores.

In 2018, Diaz and Ng [1] published a comprehensive survey on review helpfulness prediction. Before giving a summary of several approaches, the following subsection provides an overview of available datasets for helpfulness prediction.

2.1 Datasets and Domains

A number of user review datasets including helpfulness votes have already been published:

Product category	# reviews	# products	# min3 votes	# min3 ratio
Books	22.507.155	2.370.585	5.875.289	26,10%
Electronics	7.824.482	498.196	1.401.347	17,91%
Movies and TV	4.607.047	208.321	1.442.223	31,30%
CDs and Vinyl	3.749.004	492.799	1.393.862	37,18%
Clothing, Shoes and Jewelry	5.748.920	1.503.384	590.227	10,27%
Home and Kitchen	4.253.926	436.988	780.478	18,35%
Kindle Store	3.205.467	434.702	398.381	12,43%
Sports and Outdoors	3.268.695	532.197	520.921	15,94%
Cell Phones and Accessories	3.447.249	346.793	281.342	8,16%
Health and Personal Care	2.982.326	263.032	588.159	19,72%
Toys and Games	2.252.771	336.072	356.802	15,84%
Video Games	1.324.753	50.953	377.471	28,49%
Tools and Home Improvement	1.926.047	26.912	330.374	17,15%
Beauty	2.023.070	259.204	327.828	16,20%
Apps for Android	2.638.173	61.551	402.135	15,24%
Office Products	1.243.186	134.838	209.052	16,82%
Pet Supplies	1.235.316	110.707	159.615	12,92%
Automotive	1.373.768	33.109	149.807	10,90%
Grocery and Gourmet Food	1.297.156	17.176	208.047	16,04%
Patio, Lawn and Garden	993.490	109.094	198.266	19,96%
Baby	915.446	71.317	139.350	15,22%
Digital Music	836.006	279.899	135.873	16,25%
Musical Instruments	500.176	84.901	106.343	21,26%
Amazon Instant Video	583.933	30.648	65.172	11,16%
Total	80.737.562	8.693.378	16.438.364	20,36%

Table 1: The number of reviews , products, reviews with at least 3 helpfulness votes (in both the absolute and percentage measure) for every category of the amazon review dataset [4]

McAuley [4] shares a dataset of 142.8 million (deduplicated: 80.7 million) product reviews of 24 different product categories from Amazon.com spanning May 1996 - July 2014 (ARD)[1]. See Table 1 for a list of categories and more statistics. Yang et al. [5] share a subset of 400 reviews from ARD with alternative, manually annotated helpfulness labels. The music technology group of the University Pompeu Fabra in Barcelona [6] provides a subset of McAuleys ARD dataset, the Multimodal Album Reviews Dataset (MARD)[2]. It contains 263,525 reviews of music albums, filtered from selected amazon categories and enriched with various musical metadata (including artist and album identifiers) and audio descriptors.

Yelp offers a dataset of 6,685,900 reviews on 192,609 businesses across 10

[1] Available at http://jmcauley.ucsd.edu/data/amazon, last retrieved on January 23rd, 2019
[2] Available at https://www.upf.edu/web/mtg/mard, last retrieved on January 23rd, 2019

metropolitan areas [7]. Besides a vote count concerning the helpfulness of a review, the Yelp dataset also contains votes on whether a review is "funny" or "cool". Sobkowicz et al. share a dataset of 6.4 million reviews on games from the Steam platform including user votes on review helpfulness [8].

Before training a helpfulness prediction model on a dataset, the reviews are typically filtered by a minimum number of votes [1]. Increasing this threshold makes the ground truth more stable and reliable, since the opinion of more voters is represented in the helpfulness score. However, with a higher minimum number of votes, the size of the available training data shrinks. We decided to only take reviews with a minimum of 3 helpfulness votes into account. About 16.4 million of the deduplicated reviews from ARD have at least 3 helpfulness votes in total. Table 1 contains statistics about the amount of reviews with at least 3 helpfulness votes per category. The MARD dataset has 87,657 reviews with at least 3 helpfulness votes in total.

3 Related Work

Several approaches on review helpfulness prediction have already been investigated [1, 9]. For both classification and regression, Support Vector Machines have been a popular method [5, 10–14]. For classification, thresholded linear regression [15], Naive Bayes [14,16,17] and Random Forests [15–17] and for regression, linear regression [18], probabilistic matrix factorization [19] and extended tensor factorization models [20] have been used as well. Recent approaches often utilize neural networks to tackle the problem , including multilayer perceptrons networks [14, 21] and convolutional neural networks [22, 23].

Following [1], features used for helpfulness prediction can be grouped in two categories: *Content features* concern the actual (textual) review content and *context features* concern everything else, such as data about the reviewer, the product to be reviewed or the time and date of publication. In this student project we focus on content features. However, if the review text (indirectly) contains contextual information, it is relevant for this work. The following is a summary of frequently used content features:

Review Length The length of review has shown to correlate positively with it's helpfulness [24]. It can be represented as a feature utilizing as the number of characters or words in a review [5, 11, 18, 24, 25].

Readability How easy a review text is to read has an influence on it's helpfulness, as shown in [15, 26, 27]. There exist many readability metrics that have been utilized as features for helpfulness prediction (see [28] for details on each of them). The Automated Readability Index (ARI), Simple Measure of Gobbledygook (SMOG), Flesch-Kincaid Grade Level (FKGL), Gunning Fog Index (GFI), Coleman-Liau Index (CLI) and Flesch Reading Ease (FRE) have been applied by [14,15,27]. Furthermore, the Dale Chall Formula (DCF) and its component Percentage of Difficult Words (PDW) were used by [29].

Sentiment Yang et al. [5] found a positive correlation between a review's sentiment positivity and helpfulness. It can be derived by counting the occurrences of positive and negative sentiment words from a dictionary (e.g. the General Inquirer (INQUIRER) [30]) . Besides that, the overall sentiment polarity (being the percentage of both postive or negative words) can be used as feature as well [14].

Emotions Different types of emotions (e.g. anxiety, anger, sadness, joy, surprise and anticipation) in reviews also influence it's helpfulness [31]. Multiple word-emotion dictionaries have been used in review helpfulness prediction: The Geneva Affect Label Coder (GALC) [14] and Linguistic Inquiry and Word Count (LIWC)[3] [31].

Language Style

Syntactic Tokens Singh et al. [29] have shown that the amount of nouns, verbs and adjectives have an influence on review helpfulness. Malik et al. [14] applied the percentage of nouns, verbs, adjectives and adverbs as a feature for review helpfulness prediction.

Subjectivity The amount of subjective words correlates weakly with helpfulness as shown by [10]. Ghose et al. [15] trained a classifier to mark the sentences of a review as subjective or objective and applied these features for review helpfulness prediction. Yang et al. [5] used a list of subjective words from LIWC to calculate the percentage of subjective words of a review and use that as a feature.

Token-Based Features The presence of certain tokens (i.e. words or n-grams), modeled with a TF-IDF vectorization approach, is a frequently used feature in helpfulness prediction [11, 12, 25, 33].

For neural network based models, co-occurrence based character, word- and topic- embeddings (e.g. GloVe [34]) are used as an alternative encoding to traditional, hand-crafted features [22, 23].

4 Planned Experiments

In the student project, review helpfulness prediction experiments will be carried out using both a regression binary classification problem formulations. The data is taken from multiple categories of the ARD dataset. As mentioned in the end of section 1, a special focus will lie the comparison of search and experience goods, with the product categories "Electronics" and "Movies and Tv" being respective examples. Two main types of models, a SVM-based model with hand-crafted features (including review length, emotions, sentiment, readability, language style and token-based features) and a convolutional neural network model based on word embeddings will be implemented. Both models will be evaluated and compared with the metrics and the corresponding related work from Section 2.

[3] See [32] for details on LIWC.

References

[1] G. O. Diaz and V. Ng, "Modeling and Prediction of Online Product Review Helpfulness: A Survey," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 698–708, 2018.

[2] P. Nelson, "Information and Consumer Behavior," *Journal of Political Economy*, vol. 78, no. 2, pp. 311–329, 1970.

[3] P. Nelson, "Advertising as Information," *JOURNAL OF POLITICAL ECONOMY*, p. 26.

[4] R. He and J. McAuley, "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), pp. 507–517, International World Wide Web Conferences Steering Committee, 2016.

[5] Y. Yang, Y. Yan, M. Qiu, and F. S. Bao, "Semantic analysis and helpfulness prediction of text for online product reviews," in *The 53rd Annual Meeting of the Association for Computational Linguistics (ACL-2015)*, 2015.

[6] S. Oramas, L. Espinosa-Anke, A. Lawlor, X. Serra, and H. Saggion, "Exploring customer reviews for music genre classification and evolutionary studies," in *17th International Society for Music Information Retrieval Conference (ISMIR 2016)*, (New York), pp. 150–156, 07/08/2016 2016.

[7] "Yelp open dataset." https://www.yelp.com/dataset. Accessed: 2019-02-26.

[8] A. Sobkowicz and W. Stokowiec, "Steam Review Dataset - new, large scale sentiment dataset," p. 5.

[9] M. Arif, U. Qamar, F. H. Khan, and S. Bashir, "A Survey of Customer Review Helpfulness Prediction Techniques," in *Intelligent Systems and Applications* (K. Arai, S. Kapoor, and R. Bhatia, eds.), vol. 868, pp. 215–226, Cham: Springer International Publishing, 2019.

[10] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," in *Proceedings of the 15th ACM international conference on Information and knowledge management - CIKM '06*, (Arlington, Virginia, USA), p. 51, ACM Press, 2006.

[11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, (Sydney, Australia), p. 423, Association for Computational Linguistics, 2006.

[12] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou, "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*, (Portland, Oregon, USA), p. 495, ACM Press, 2012.

[13] Y.-C. Zeng, T. Ku, S.-H. Wu, L.-P. Chen, and G.-D. Chen, "Modeling the Helpful Opinion Mining of Online Consumer Reviews as a Classification Problem," in *International Journal of Computational Linguistics & Chinese Language Processing*, p. 16, 2014.

[14] M. Malik and A. Hussain, "Helpfulness of product reviews as a function of discrete positive and negative emotions," *Computers in Human Behavior*, vol. 73, pp. 290–302, Aug. 2017.

[15] A. Ghose and P. G. Ipeirotis, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, pp. 1498–1512, Oct. 2011.

[16] M. P. O'Mahony, P. Cunningham, and B. Smyth, "An Assessment of Machine Learning Techniques for Review Recommendation," in *Artificial Intelligence and Cognitive Science* (L. Coyle and J. Freyne, eds.), vol. 6206, pp. 241–250, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[17] S. Krishnamoorthy, "Linguistic features for review helpfulness prediction," *Expert Systems with Applications*, vol. 42, pp. 3751–3759, May 2015.

[18] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in *Proceedings of the 19th international conference on World wide web - WWW '10*, (Raleigh, North Carolina, USA), p. 691, ACM Press, 2010.

[19] J. Tang, H. Gao, X. Hu, and H. Liu, "Context-aware review helpfulness rating prediction," in *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*, (Hong Kong, China), pp. 1–8, ACM Press, 2013.

[20] S. Moghaddam, M. Jamali, and M. Ester, "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*, (Seattle, Washington, USA), p. 163, ACM Press, 2012.

[21] S. Lee and J. Y. Choeh, "Predicting the helpfulness of online reviews using multilayer perceptron neural networks," *Expert Systems with Applications*, vol. 41, pp. 3041–3046, May 2014.

[22] C. Chen, M. Qiu, Y. Yang, J. Zhou, J. Huang, X. Li, and F. Bao, "Review Helpfulness Prediction with Embedding-Gated CNN," *arXiv:1808.09896 [cs]*, Aug. 2018. arXiv: 1808.09896.

[23] C. Chen, Y. Yang, J. Zhou, X. Li, and F. S. Bao, "Cross-Domain Review Helpfulness Prediction Based on Convolutional Neural Networks with Auxiliary Domain Discriminators," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, (New Orleans, Louisiana), pp. 602–607, Association for Computational Linguistics, 2018.

[24] Mudambi and Schuff, "Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, vol. 34, no. 1, p. 185, 2010.

[25] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and predicting the helpfulness of online reviews," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 443–452, 12 2008.

[26] Y.-H. Hu and K. Chen, "Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings," *International Journal of Information Management*, vol. 36, pp. 929–944, Dec. 2016.

[27] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," *Electronic Commerce Research and Applications*, vol. 11, pp. 205–217, May 2012.

[28] W. DuBay, "Principles of Readability," p. 77.

[29] J. P. Singh, S. Irani, N. P. Rana, Y. K. Dwivedi, S. Saumya, and P. Kumar Roy, "Predicting the "helpfulness" of online consumer reviews," *Journal of Business Research*, vol. 70, pp. 346–355, Jan. 2017.

[30] P. J. Stone, R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie, "The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information," *Behavioral Science*, vol. 7, pp. 484–498, Jan. 2007.

[31] University of Missouri, D. Yin, S. D. Bond, Georgia Institute of Technology, H. Zhang, and Georgia Institute of Technology, "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *MIS Quarterly*, vol. 38, pp. 539–560, Feb. 2014.

[32] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The Development and Psychometric Properties of LIWC2007," p. 22, 2007.

[33] Y. Yang, C. Chen, and F. S. Bao, "Aspect-Based Helpfulness Prediction for Online Product Reviews," in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, (San Jose, CA, USA), pp. 836–843, IEEE, Nov. 2016.

[34] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.