Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews

Yinfei Yang
Amazon Inc.
Seattle, WA 98121
yangyin7@ gmail.com

Yaowei Yan
Dept. of Electrical & Computer Engineering
University of Akron
Akron, OH 44325-3904
yy28@ uakron.edu

Minghui Qiu
Alibaba Group
Hangzhou, China 311121
minghuiqiu@ gmail.com

Forrest Sheng Bao
Dept. of Electrical & Computer Engineering
University of Akron
Akron, OH 44325-3904
forrest.bao@ gmail.com

Abstract

Predicting the helpfulness of product reviews is a key component of many e-commerce tasks such as review ranking and recommendation. However, previous work mixed review helpfulness prediction with those outer layer tasks. Using non-text features, it leads to less transferable models. This paper solves the problem from a new angle by hypothesizing that helpfulness is an internal property of text. Purely using review text, we isolate review helpfulness prediction from its outer layer tasks, employ two interpretable semantic features, and use human scoring of helpfulness as ground truth. Experimental results show that the two semantic features can accurately predict helpfulness scores and greatly improve the performance compared with using features previously used. Cross-category test further shows the models trained with semantic features are easier to be generalized to reviews of different product categories. The models we built are also highly interpretable and align well with human annotations.

1 Introduction

Product reviews have influential impact to online shopping as consumers tend to read product reviews when finalizing purchase decisions (Duan et al., 2008). However, a popular product usually has too many reviews for a consumer to read. Therefore, reviews need to be ranked and recommended to consumers. In particular, review helpfulness plays a critical role in review ranking and recommendation (Ghose and Ipeirotis, 2011; Mudambi and Schuff, 2010; Danescu-Niculescu-Mizil et al., 2009). The simple question "Was this review helpful to you?" increases an estimated $2.7B revenue to Amazon.com annually[1].

However, existing literature solves helpfulness prediction together with its outer layer task, the review ranking (Kim et al., 2006; O'Mahony and Smyth, 2010; Liu et al., 2008; Martin and Pu, 2014). Those studies use features not contributing to helpfulness, such as date (Liu et al., 2008), or features making the model less transferable, such as product type (Mudambi and Schuff, 2010). Models built in these ways are also difficult to interpret from linguistic perspective.

Therefore, it is necessary to isolate review helpfulness prediction from its outer layer tasks and formulate it as a new problem. In this way, models can be more robust and generalizable. Beyond predicting *whether* a review is helpful, we can also understand *why* it is helpful. In our approach, the results can also facilitate many other tasks, such as review summarization (Xiong and Litman, 2014) and sentiment extraction (Hu and Liu, 2004).

Recent NLP studies reveal the connection between text style and its properties, include readability (Agichtein et al., 2008), informativeness (Yang and Nenkova, 2014) and trustworthiness (Pasternack and Roth, 2011) of text. Hence, we hypothesize that helpfulness is also an underlying property of text.

To understand the essence of review text, we leverage existing linguistic and psychological dictionaries and represent reviews in semantic dimensions. Two semantic features that are new to solving this problem, LIWC (Pennebaker et al., 2007) and INQUIRER (Stone et al., 1962), are employed in this work. The intuition behind is that people usually embed semantic meanings, such as emotion and reasoning, into text. For example, the re-

[1] http://www.uie.com/articles/magicbehindamazon/

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 38–44,
Beijing, China, July 26-31, 2015.

view *"With the incredible brightness of the main LED, this light is visible from a distance on a sunny day at noon.* is more helpful than the review *"I ordered an iPad, I received an iPad. I got exactly what I ordered which makes me satisfied. Thanks!"* because the former mentions user experience and functionality of the product while the latter has emotional statements only.

Previous work approximates the ground truth of helpfulness from users' votes using "X of Y approach": if X of Y users think a review is helpful, then the helpfulness score of the review is the ratio X/Y. However, not many reviews have statistically abundant votes, i.e., a very small Y. Fewer than 20% of the reviews in Amazon Review Dataset (McAuley and Leskovec, 2013) have at least 5 votes (Table 1) while only 0.44% have 100+ votes. In addition, the review voting itself may be biased (Danescu-Niculescu-Mizil et al., 2009; Cao et al., 2011). Therefore, we proactively recruited human annotators and let them score the helpfulness of reviews in our dataset.

We model the problem of predicting review helpfulness score as a regression problem. Experimental results show that it is feasible to use text-only features to accurately predict helpfulness scores. The two semantic features significantly outperform baseline features used in previous work. In cross-category test, the two semantic features show good transferability. To interpret the models, we analyze the semantic features and find that Psychological Process plays an important role in review text helpfulness. Words reflecting thinking and understanding are more related to helpful reviews while emotional words are not. Lastly, we validate the models trained on "X of Y approach" data on human annotated data and achieve highly correlated prediction.

2 Dataset

Two subsets of reviews are constructed from Amazon Review Dataset (McAuley and Leskovec, 2013), which includes nearly 35 million reviews from Amazon.com between 1995 and 2013. A subset of 696,696 reviews from 4 categories: Books, Home (home and kitchen), Outdoors and Electronics, are chosen in this research. For each category, we select the top 100 products with the most reviews and then include all reviews related to the selected products for analysis. Each review comes with users' helpfulness votes and hence helpfulness score can be approximated using "X of Y approach." Finally, 115,880 reviews, each of which has at least 5 votes, form the **automatic labeled** dataset (Table 1).

Table 1: Number of Reviews for Each Category

Category	Total number of reviews	Number of reviews with at least 5 votes, selected for experiments
Books	391,666	81,014 (20.7%)
Home	116,194	13,331 (11.5%)
Outdoors	52,838	6,158 (11.7%)
Electronics	135,998	15,377 (11.3%)
Overall	696,696	115,880 (16.6%)

In addition, we also create the **human labeled** dataset. As mentioned earlier, the X of Y approach may not be a good approximation to helpfulness. A better option is human scoring. We randomly select 400 reviews outside of the automatic labeled dataset, 100 from each category. Eight students annotated these reviews in a fashion similar to that in (Bard et al., 1996) by assigning real-value scores ($\in [0, 100]$) to each review. Review text was the only information given to them. The average helpfulness score of all valid annotations is used as the ground truth for each review. We have released the human annotation data at `https://sites.google.com/site/forrestbao/acl_data.tar.bz2`.

3 Features

Driven by the hypothesis that helpfulness is an underlying feature of text itself, we consider text-based features only. Features used in previous related work, namely Structure (STR) (Kim et al., 2006; Xiong and Litman, 2011), Unigram (Kim et al., 2006; Xiong and Litman, 2011; Agarwal et al., 2011) and GALC emotion (Martin and Pu, 2014), are considered as baselines.

We then introduce two semantic features LIWC and General Inquirer (INQUIRER) for easy mapping from text to human sense, including emotions, writing styles, etc. Our rationale for the two semantic features is that a helpful review includes opinions, analyses, emotions and personal experiences, etc. These two features have been proven effective in other semantic analysis tasks and hence we are here giving them a try for studying review helpfulness. We leave the study of using more sophisticated features like syntactic and discourse representations to future work. All features except UGR are independent of training data.

STR Following the (Xiong and Litman, 2011), we use the following structural features: total number of tokens, total number of sentences, average length of sentences, number of exclamation marks, and the percentage of question sentences.

UGR Unigram feature has been demonstrated as a very reliable feature for review helpfulness prediction in previous work. We build a vocabulary with all stopwords and non-frequent words ($df < 3$) removed. Each review is represented by the vocabulary with $tf - idf$ weighting for each appeared term.

GALC (Geneva Affect Label Coder) (Scherer, 2005) proposes to recognize 36 effective states commonly distinguished by words. Similar to (Martin and Pu, 2014), we construct a feature vector with the number of occurrences of each emotion plus one additional dimension for non-emotional words.

LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2007) is a dictionary which helps users to determine the degree that any text uses positive or negative emotions, self-references and other language dimensions. Each word in LIWC is assigned 1 or 0 for each language dimension. For each review, we sum up the values of all words for each dimension. Eventually each review is represented by a histogram of language dimensions. We employ the LIWC2007 English dictionary which contains 4,553 words with 64 dimensions in our experiments.

INQUIRER General Inquirer (Stone et al., 1962) is a dictionary in which words are grouped in categories. It is basically a mapping tool which maps each word to some semantic tags, e.g., *absurd* is mapped to tags NEG and VICE. The dictionary contains 182 categories and a total of 7,444 words. Like for LIWC representation, we compute the histogram of categories for each review.

4 Experiments

Up to this point, we are very interested in first whether a prediction model learned for one category can be generalized to a new category, and second what elements make a review helpful. In other words, we want to know the robustness of our approach and the underlying reasons.

In this section we will evaluate the effectiveness of each of the features as well as the combination of them. For convenience, we use Fusion$_{Semantic}$ to denote the combination of GALC, LIWC and INQUIRER, and Fusion$_{All}$ to denote the combination of all features. Because STR and UGR are widely used in previous work, we use them as two baselines. GALC has been introduced for this task as an emotion feature before, so we use it as the third baseline. STR, URG and GALC are used as 3 baselines. For predicting helpfulness scores, we use SVM regressor with RBF kernel provided by LibSVM (Chang and Lin, 2011).

Two kinds of labels are used: automatic labels obtained in "X of Y approach" from votes, and human labels made by human annotators. Performance is evaluated by Root Mean Square Error (RMSE) and Pearson's correlation coefficients. Ten-fold cross-validation is performed for all experiments.

4.1 Results using Automatic Labels

Before studying the transferability of models, we first need to make sure that models work well on reviews of products of the same category.

4.1.1 RMSE

RMSE and correlation coefficient using automatic labels are given in Table 2 and Table 3 respectively. Each row corresponds to the model trained by a feature or a combination of features, while each column corresponds to one product category. The lowest RMSE achieved using every single feature in each category is marked in bold.

The two newly employed semantic features, LIWC and INQUIRER, have 8% lower RMSE on average than UGR, the best baseline feature. Fusion$_{All}$ has the best overall RMSE, ranging from 0.200 to 0.265. Fusion$_{Semantic}$ has the second best performance on average. It achieves the lowest RMSE in Books category.

Table 2: RMSE (the lower the better) using automatic labels

	Books	Home	Outdoors	Electro.	Average
STR	0.239	0.289	0.314	0.307	0.287
UGR	0.242	0.260	0.284	0.286	0.268
GALC	0.266	0.290	0.310	0.308	0.365
LIWC	**0.188**	0.256	0.279	0.278	0.250
INQUIRER	0.193	**0.248**	**0.274**	**0.273**	**0.247**
Fusion$_{Semantic}$	0.187	0.248	0.272	0.268	0.244
Fusion$_{All}$	0.200	0.247	0.261	0.265	0.243

Table 3: Correlation coefficients (the higher the better) using automatic labels. All correlations are highly significant, with $p < 0.001$.

	Books	Home	Outdoors	Electronics
STR	0.500	0.280	0.333	0.351
UGR	0.507	0.467	**0.458**	0.471
GALC	0.239	0.216	0.255	0.274
LIWC	**0.742**	0.439	0.424	0.475
INQUIRER	0.720	**0.487**	0.455	**0.498**
Fusion$_{Semantic}$	0.744	0.490	0.467	0.527
Fusion$_{All}$	0.682	0.525	0.535	0.539

4.1.2 Correlation Coefficient

In line with RMSE measurements, the semantic feature based models outperform the baseline

features in terms of correlation coefficient (Table 3). In each category, the highest correlation coefficient is achieved by using LIWC or INQUIRER, with only one exception (Outdoors). The two fusion models further improve the results. Fusion$_{Semantic}$ has the highest coefficients in Books category while Fusion$_{All}$ has the highest coefficients in other 3 categories.

4.2 Cross Category Test

One motivation of introducing semantic features is that, unlike UGR which is category-dependent, they can be more transferable. To validate the transferability of semantic features, we perform cross category test by using the model trained from one category to predict the helpfulness scores of reviews in other categories. GALC is excluded in this analysis due to its poor performance earlier.

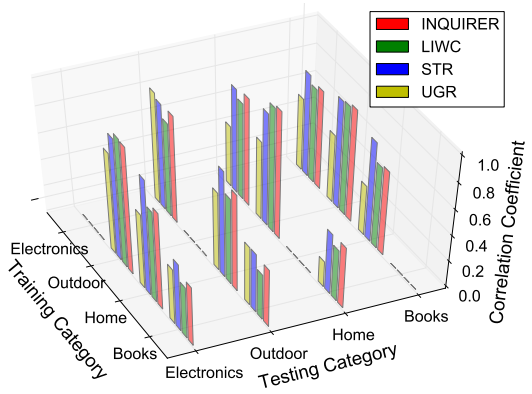

Figure 1: Normalized cross-category correlation coefficients

Model transferability from Category A to Category B cannot be measured simply by the performance when using A as the training set and B as the test set. Instead, it should be compared relatively with the performance when using A as both the training and test sets. There are 4 categories in our dataset, and the performances on the 4 categories vary (Tables 2 and 3). In order to provide a fair comparison, we normalize cross-category correlation coefficients by the corresponding same-category ones, i.e., cross-category correlation coefficient / correlation coefficient on training category. For example, the 3 cross-category correlation coefficients of using Books category as training set are all normalized by the correlation coefficient when using Books as both training and test sets earlier. A normalized correlation coefficient of 0 means the prediction on the test category is random, and thus the model has no transferability, while 1 means as accurate as predicting on the training category, and thus the model is fully transferable.

Results on transferrability are visualized in Figure 1 with same-category correlation coefficients ignored as they are always 1. Correlation coefficients of 4 features are clustered for each pair of training and testing categories and are color-coded.

It is shown that INQUIRER and STR are two best features in cross category test, leading in most of the category pairs. LIWC follows, achieving at least 70% of the same-category correlation coefficients in most cases. The UGR feature, however, performs poorly in this test. In most cases, the correlation coefficients have been halved, compared with same-category results.

According to the results, we can conclude that semantic features are accurate and transferable, UGR is accurate but is not transferable, and STR is transferable but not accurate enough (Figure 2).

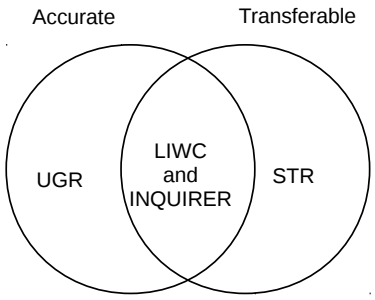

Figure 2: Classification of features based on experimental results

4.3 What Makes a Review Helpful: A Semantic Interpretation

LIWC and INQUIRER not only have better performances than previously used features but also provide us a good semantic interpretation to what makes a review helpful. We analyze the correlation coefficients between helpfulness and each language dimension in the two dictionaries. The top 5 language dimensions that are mostly correlated to helpfulness from LIWC and INQUIRER are given in Figure 3.

The top 5 dimensions from LIWC are: Relativ (Relativity), Time, Incl (Inclusive), Posemo (Positive Emotion), and Cogmech (Cognitive Processes). All of them belong to *Psychological Processes* categories in LIWC, indicating that people are more thoughtful when writing a helpful review.

The top 5 dimensions from INQUIRER are: Vary, Begin, Exert, Vice and Undrst. Words with

Vary, Begin or Exert tags belong to *process or change words*, such as *start, happen* and *break*. Vice tag contains words indicating an assessment of moral disapproval or misfortune.Undrst (Understated) tag contains words indicating deemphasis and caution in these realms, which often reflects the lack of emotional expressiveness. Accordingly, we can infer that consumers perfer critical reviews with personal experience and a lack of emotion.

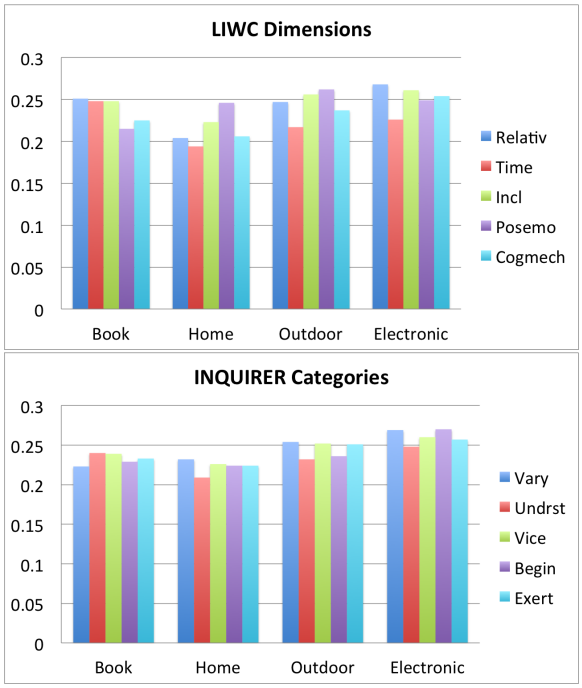

Figure 3: Language dimensions with highest correlation coefficients. Top: LIWC's; Bottom: INQUIRER's.

The discovery that helpful reviews are less emotional is consistent with the weak performance of GALC (Tables 2, 3 and 4), which is emotion focused. However, we notice that one of the top 5 dimensions in LIWC, PosEmo, is an emotional feature. This is partially because some words appear in both emotional and rational expressions, such as LIWC PosEmo words: *love*, *nice*, *sweet*. For example, the sentence *"I used to love linksys, but my experience with several of their products makes me seriously think that their quality is suspect"* is a rational statement. But the word "love" appears in it.

4.4 Prediction Results on Human Labels

A better ground truth for helpfulness is human rating. We further evaluate the prediction models on human annotated data to evaluate whether the predictions indeed align with human perceptions of review helpfulness by reading text only.

The model we built indeed aligns with human perceptions of review helpfulness when text is the only data. Table 4 shows the correlation coefficients between the predicted scores and human annotated scores. INQUIRER is the best feature, leading in 3 of 4 categories. It is followed by UGR and LIWC, which show comparable results.

Table 4: Correlation coefficients between predicted scores and human annotation, *: $p < 0.001$.

	Books	Home	Outdoors	Electronics
STR	0.539*	0.522*	0.471*	0.635*
UGR	0.607*	0.560*	0.579*	0.626*
GALC	0.214	0.405*	0.156	0.418*
LIWC	0.524*	0.553*	0.517*	**0.702***
INQUIRER	**0.620***	**0.662***	**0.620***	0.676*
Fusion$_{Semantic}$	0.556*	0.680*	0.569*	0.603*
Fusion$_{All}$	0.610*	0.801*	0.698*	0.768*

For Fusion$_{All}$ models, correlation coefficients are about or over 0.7 in 3 of 4 categories, indicating the successful prediction. The only exception is on Books category. We notice that reviews in Books are more subjective. Therefore, in Books reviews, consumers are more influenced by factors outside of the text, e.g., personal preference on the book. In this case, the approximate scores used in training may not reflect the real text helpfulness. This observation echoes with our speculation that the "X of Y approach" may not always be a good approximation for helpfulness due to the subjectivity. We will leave the analysis to this as a future work.

5 Conclusion

In this paper, we formulate a new problem which is an important component of many tasks about online product reviews: predicting the helpfulness of review text. We hypothesize that helpfulness is an underlying property of text and isolate helpfulness prediction from its outer layer problems, such as review ranking. Introducing two semantic features, which have been shown effective in other NLP tasks, we achieve more accurate and transferable prediction than using features used in existing related work. The ground truth is provided by votes on massive Amazon product reviews. We further explore a semantic interpretation to reviews' helpfulness that helpful reviews exhibit more reasoning and experience and less emotion. The results are further validated on human scoring to helpfulness.

References

[Agarwal et al.2011] Deepak Agarwal, Bee-Chung Chen, and Bo Pang. 2011. Personalized recommendation of user comments via factor models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 571–582, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Agichtein et al.2008] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194. ACM.

[Bard et al.1996] Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, 72(1):pp. 32–68.

[Cao et al.2011] Qing Cao, Wenjing Duan, and Qiwei Gan. 2011. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decis. Support Syst.*, 50(2):511–521, January.

[Chang and Lin2011] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Danescu-Niculescu-Mizil et al.2009] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: A case study on amazon.com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 141–150, New York, NY, USA. ACM.

[Duan et al.2008] Wenjing Duan, Bin Gu, and Andrew B. Whinston. 2008. The dynamics of online word-of-mouth and product sales-an empirical investigation of the movie industry. *Journal of Retailing*, 84:233242.

[Ghose and Ipeirotis2011] A. Ghose and P.G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. volume 23, pages 1498–1512, Oct.

[Hu and Liu2004] Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 755–760. AAAI Press.

[Kim et al.2006] Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 423–430, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Liu et al.2008] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, pages 443–452, Washington, DC, USA. IEEE Computer Society.

[Martin and Pu2014] Lionel Martin and Pearl Pu. 2014. Prediction of helpful reviews using emotions extraction. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI '14.

[McAuley and Leskovec2013] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: Understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, pages 165–172, New York, NY, USA. ACM.

[Mudambi and Schuff2010] Susan M. Mudambi and David Schuff. 2010. What makes a helpful online review? a study of customer reviews on amazon.com. *MIS Quarterly*, pages 185–200.

[O'Mahony and Smyth2010] Michael P. O'Mahony and Barry Smyth. 2010. Using readability tests to predict helpful product reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 164–167, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[Pasternack and Roth2011] Jeff Pasternack and Dan Roth. 2011. Making better informed trust decisions with generalized fact-finding. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three*, IJCAI'11, pages 2324–2329. AAAI Press.

[Pennebaker et al.2007] J. W. Pennebaker, Roger J. Booth, and M. E. Francis. 2007. Linguistic inquiry and word count: Liwc.

[Scherer2005] Klaus R. Scherer. 2005. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

[Stone et al.1962] P. J. Stone, R. F. Bales, J. Z. Namenwirth, and D. M. Ogilvie. 1962. The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. In *Behavioral Science*, pages 484–498.

[Xiong and Litman2011] Wenting Xiong and Diane Litman. 2011. Automatically predicting peer-review helpfulness. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 502–507, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Xiong and Litman2014] Wenting Xiong and Diane Litman. 2014. Empirical analysis of exploiting review helpfulness for extractive summarization of online reviews. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1985–1995.

[Yang and Nenkova2014] Yinfei Yang and Ani Nenkova. 2014. Detecting information-dense texts in multiple news domains. In *Proceedings of Twenty-Eighth AAAI Conference on Artificial Intelligence*.