Not All Reviews are Equal:
Towards Addressing Reviewer Biases for Opinion Summarization

Wenyi Tay
RMIT University, Australia
CSIRO Data61, Australia
`wenyi.tay@student.rmit.edu.au`

Abstract

Consumers read online reviews for insights which help them to make decisions. Given the large volumes of reviews, succinct review summaries are important for many applications. Existing research has focused on mining for opinions from only review texts and largely ignores the reviewers. However, reviewers have biases and may write lenient or harsh reviews; they may also have preferences towards some topics over others. Therefore, not all reviews are equal. Ignoring the biases in reviews can generate misleading summaries. We aim for summarization of reviews to include balanced opinions from reviewers of different biases and preferences. We propose to model reviewer biases from their review texts and rating distributions, and learn a *bias-aware opinion representation*. We further devise an approach for *balanced opinion summarization* of reviews using our bias-aware opinion representation.

1 Introduction

Consulting online reviews on products or services is popular among consumers. Opinions in reviews are scrutinised to make an informed decision on which product to buy, what service to use, or which point-of-interest to visit. An opinion is *a view or judgment formed about something, not necessarily based on fact or knowledge.*[1] In the context of online reviews, opinions contain information about the target ("something") and the sentiment ("view or judgment") that is associated with it. There can also be more than one opinion in a review.

Opinion mining research is dedicated to tasks that involves opinions (Pang and Lee, 2008). Current research in opinion mining mostly focuses only on review texts. Some key tasks include sentiment polarity classification (Hu and Liu, 2004b) at levels of words, sentences or documents, and opinion target (e.g., aspect) identification and classification.

Opinion summarization from reviews is an important task related to opinion mining. Early work on opinion summarization aims for structured representation of aspect-sentiment pairs (Hu and Liu, 2004a), where the positive and negative sentiment for each aspect are extracted from review texts and aggregated. Opinion summaries in natural language texts contain richer, detailed description of opinions and are easier for end users to understand. Existing studies mainly use the review texts for summarization.

However, reviewers are unique individuals with beliefs and preferences. Reviewers have preferences towards certain aspects, for example service or cleanliness in hotel reviews (Wang et al., 2010). Different reviewers can have different ways of expressing their opinions (Tang et al., 2015b). Also, some reviewers are lenient in their assessment of products or services, while others are harsher (Lauw et al., 2012). Overall, an opinion is a reflection of the reviewer as it encompasses their *biases*. Thus, not all reviews are equal.

Depending on the application, biases captured in the reviews can be amplified. Hu et al. (2006) suggest that reviewers write reviews when they are extremely satisfied or when they are extremely upset. Existing summarization techniques often treat all reviews equally by selecting salient opinions which may not necessarily be representative for different reviewers. We aim to compensate for biases in reviews, especially for review summarization. We focus on the following research questions:

1. How to model a reviewer's bias? What in-

---

[1]Oxford dictionary

formation from a reviewer should be used to model a reviewer's bias?

2. How to learn a representation for reviews that captures reviewer biases as well as the opinion?

3. How to generate a balanced opinion summary of reviews written by different reviewers?

Below, we outline the relevant past studies as well as our our research proposal to address these questions.

2 Related Work

Our research is related to two research areas summarized below.

2.1 Opinion and Reviewer Modeling

We identified two studies that jointly model opinions and reviewers. Wang et al. (2010) investigate the problem of decomposing the overall review rating into aspect ratings using a hotel domain dataset. The authors model opinions and reviewers using a generative approach. Reviewers are modeled to reflect their individual emphasis on various aspects. The authors demonstrate that despite giving the same overall review rating, two reviewers can value and rate aspects differently. Meanwhile, Li et al. (2014) present a topic model incorporating reviewer and item information for sentiment analysis. Through probabilistic matrix factorisation of reviewer-item matrix, the latent factors are included in a supervised topic model guided by sentiment labels. The proposed model outperforms baselines in predicting the sentiment label given the review text, reviewer and item on a movie review dataset and a microblog dataset.

Opinion modeling Opinion can be represented as a aspect-sentiment tuple (Hu and Liu, 2004b). In order to obtain the components of the opinion, aspect identification and sentiment classification are key. Both tasks can be treated separately or combined. For aspect identification, aspects can be identified with the help of experts (Hu and Liu, 2004b; Zhang et al., 2012). The drawback is that it requires input from experts and is specific to a domain. This triggered studies that seek to discover aspects in an unsupervised manner using topic models (Brody and Elhadad, 2010; Moghaddam and Ester, 2010). However, such methods may not always produce interpretable aspects. Subsequent

models are developed to discover interpretable aspects (McAuley and Leskovec, 2013; Titov and McDonald, 2008a,b). To determine opinion polarity, lexicon-based (Hu and Liu, 2004b) and classification (Dave et al., 2003) approaches are often used. However, modeling opinions based on aspects and sentiment separately is not sufficient as the sentiment words can depend on the aspect. More recent models focus on incorporating context to model opinions. Such approaches include joint aspect-sentiment models (Lin and He, 2009), word embeddings (Maas et al., 2011), and neural network models (He et al., 2017).

Alternatively, opinions can potentially be represented as a high-dimensional vector. Opinion representation in this form is a relatively unexplored space. However, in the closely related area of sentiment classification, sentences and documents are represented as vectors to be used as inputs for classification (Conneau et al., 2017; Tang et al., 2015a). The idea is to model a sequence of words as a high-dimensional vector that captures the relationships of words. Similarly, opinions are sequences of sentences, thus it is appropriate to build on the work in sentence and document representation. One of the earliest work is an extension of word2vec (Mikolov et al., 2013) to learn a distributed representation of text (Le and Mikolov, 2014). More recently, pre-trained sentence encoders trained on a large general corpus aim to capture task-invariant properties that can be fine-tuned for downstream tasks (Cer et al., 2018; Conneau et al., 2017; Kiros et al., 2015). On another front, progress in context-aware embeddings (Peters et al., 2018) and pre-trained language models (Devlin et al., 2018; Howard and Ruder, 2018) provide other options to capture context that can be used to obtain sequence representation. All these studies focus on encoding topical semantics of text sequences, where opinions are not explicitly modeled.

Reviewer modeling Various reviewer characteristics that are modeled include expertise (Liu et al., 2008), reputation (Chen et al., 2011; Shaalan and Zhang, 2016), characteristics of language use (Tang et al., 2015b) and preferences (Zheng et al., 2017). Some of these modelings are achieved using reviewer aggregated statistics and review meta-data. Reviewer expertise is modeled by number of reviews, where larger number of reviews suggests higher expertise (Liu et al., 2008).

Reviewer reputation can be modeled by the number of helpfulness votes and total votes received by the reviewer. A higher ratio of helpfulness votes to total votes suggests a better reputation (Shaalan and Zhang, 2016). In another reviewer reputation model, reviewers are modeled to have domain expertise which corresponds to the product categories that the reviewer reviewed on (Chen et al., 2011).

Review text is also used in reviewer modeling. When predicting ratings from review text, the same sentiment bearing word, for example "good", can mean different sentiment intensity to different reviewers. Tang et al. (2015b) model reviewers' word use by using review text and its corresponding review rating. The resulting reviewer-modified word representations capture variations in reviewers' word use that translates to better rating prediction. Recently, review text is used in addition to review ratings to model users and items together for recommendation (Zheng et al., 2017). Using all the reviews written by the reviewer, the model learns a latent representation of the reviewer. All the above approaches focus on modeling the reviewer. However, our focus is to model opinions, where reviewer information is to be used as a factor during the process of modeling.

For our proposed work, we explore using review text, review ratings and meta data to model reviewers except for helpfulness votes. The helpfulness mechanism is shown to be biased (Liu et al., 2007) and it is still not well understood what we can infer from such votes (Ocampo Diaz and Ng, 2018).

2.2 Opinion Summarization

Opinion summarization aims to capture salient opinions within a collection of document, in our case online reviews. Key challenges in summarizing opinions from a collection of documents are highlighted by Pang and Lee (2008): (1) How to identify documents and parts of the document that are of the same opinion; and (2) How to decide two sentences or texts have the same semantic meaning.

To identify documents and parts of documents of the same opinion, one strategy is to use review ratings as a means to identify similar opinion. However, review ratings have drawbacks such as rating scales differ for different review sources, different assessment criteria among reviewers and reviewers may not share the same opinion despite giving the same overall rating. Review ratings can be adjusted to correct for different assessment criteria by comparing the reviewers' rating behaviour relative to the community rating behaviour (Lauw et al., 2012; Wadbude et al., 2018). The review rating only captures the overall sentiment polarity of the review but not the individual opinions that make up the review. As such, the authors propose to decompose the review rating into aspect ratings according to the review text (Wang et al., 2010). Alternatively, the same opinions can be found by mining aspects and sentiment polarity of each review. Opinion summarization can be seen as a task that builds on top of the opinion mining task.

In deciding if two sentences or texts have the same semantic meaning, the crux lies in the representation of sentences and text. Sentences with the same meaning have good overlap in words (Ganesan et al., 2010). More recent approaches adopt representing sentences or texts as high-dimensional vectors such that similar representations have similar meaning (Le and Mikolov, 2014; Tang et al., 2015a).

The presentation of the opinion summary depends on two considerations, (1) the needs of the reader; and (2) the approach to construct opinion summaries. An opinion summary can be presented in different ways, catering to the different needs of readers. The summary can be on one product (Angelidis and Lapata, 2018; Hu and Liu, 2004a), comparing two products (Sipos and Joachims, 2013) or generate a summary in response to a query (Bonzanini et al., 2013).

There are two main ways of constructing opinion summaries. The extractive opinion summaries are summaries put together by selecting sentences or word segments (Angelidis and Lapata, 2018; Xiong and Litman, 2014). For abstractive summaries, the summary is generated from scratch (Ganesan et al., 2010; Wang et al., 2010).

An early work in opinion summarization proposed an aspect-based summary by organising all opinions according to aspects and their sentiment polarity (Hu and Liu, 2004a). Although there is no textual summarization involved, it inspired future work to focus on including aspects into the generated summary regardless whether it is extractive or abstractive.

For extractive summarization, the objective is to identify salient sentences, at the same time reducing redundancy in the selected sentences. An-

gelidis and Lapata (2018) score opinion segments according to the aspect and the sentiment polarity. In another work, sentences in the review are scored according to a combination of textual features and latent topics discovered by helpfulness votes (Xiong and Litman, 2014). To reduce redundancy in selected sentences, a greedy algorithm can be applied to add one sentence at a time to form the summary. The greedy algorithm imposes the criterion that the selected sentence must be different from the sentences that are already in the summary (Angelidis and Lapata, 2018). As most extractive summarization techniques are closely coupled with identifying opinions from review texts, the outcome is a set of sentences that are salient in terms of topic coverage, but they may not necessarily be the most representative opinions from reviewers.

On the other hand, abstractive methods first learn to identify the salient opinions before generating a shorter text to reflect the opinion. A graph-based method is proposed by Ganesan et al. (2010) which models a word with its Part-of-Speech (POS) tag as nodes and directed edges to represent the order of words. The edge weights increases when the sequence of words is repeated. The summary is generated by capturing the paths with high edge weights. In a recent study, an encoder-decoder network is employed to generate an abstractive summary of movie reviews (Wang and Ling, 2016).

3 Proposed Methodology

The intuition for our research is that summarization techniques that rely on similarity between opinions to identify salient opinions benefit from clustering similar opinions together and separating different opinions into different clusters. By modeling reviewers with opinions, we aim to capture biases reviewers bring into their opinions. We next elaborate our approaches to modeling user biases, learning *bias-aware opinion representations* and *balanced opinion summarization*.

3.1 Bias-Aware Opinion Representation

To achieve a bias-aware opinion representation, we model opinions and reviewer biases for each sentence in a review. We assume that one sentence contains one opinion (Hu and Liu, 2004b). We envision two possible approaches to learn a bias-aware opinion representation: (1) Two-step process by modeling opinions then adjust the opinions according to reviewer biases; and, (2) Generative approach using text, rating and reviewer information.

Using a two-step process, our main objective is to first learn a representation of the sentences to capture the opinion and this is not a trivial task. Ideally, we expect our opinion representation to exhibit two key characteristics: (1) Similar opinions need to be close in its representation. Using opinions for restaurant reviews as an example, "The soup is rich and creamy" and "Delicious food" are similar opinions but expressed differently; and, (2) Opinion models should be able to tease apart different opinions.

In terms of representing opinions that are similar, a promising technology for us is to make use of pre-trained sentence encoders and language models (Cer et al., 2018; Devlin et al., 2018; Peters et al., 2018; Conneau et al., 2017). These pre-trained models have the advantage of transferring the learned information from large corpora. However, we hypothesize that even with the use of pre-trained models, we are unable to capture sentiment polarity of opinions accurately. It will be similar to the problem that word embeddings are not able to capture sentiment polarity (Maas et al., 2011). One potential direction is to adopt supervised learning using labeled aspect and sentiment polarity labels to improve our opinions representation. But labeled data is expensive to acquire and the granularity of aspect can vary with different aspect annotation guidelines. We propose to use review ratings as supervision signal to improve our opinion representation as ratings can provide a guide to sentiment polarity of opinions.

Towards learning bias-aware opinion representations, we further refine the learnt opinion vectors via modeling reviewer biases from their reviews and ratings. Reviewer biases can influence their star rating and textual expressions. The key to model reviewer biases is learning a distribution of latent factors and sentiment polarity from the reviews and their rating distributions for the reviewer. The refinement will be a user matrix that learn weights corresponding to the opinion representation. This can also be seen as the matrix that represents the biases of reviewers. We plan to explore different ways to learn this matrix. One option to model reviewers' biases is to learn representations from their past reviews such as using

techniques in recommender systems literature to model reviewers using review text (Zheng et al., 2017). Alternatively, other associated review information such as review ratings and even metadata of reviews can possibly guide the modeling of biases. We can also explore textual features of review such as the position of opinions may also provide clues to model reviewers.

For our second possible approach, we adopt a generative approach to model opinions as topics using reviewer information as latent factors (Li et al., 2014; Wang et al., 2010). However, the topic model approach is restricted to using words as tokens. The neural topic model (Cao et al., 2015) is a potential technique to utilise word embeddings to improve the learning of topics in the collection of reviews.

3.2 Balanced Opinion Summarization

Summaries generated by the existing summarization techniques are accurate to the collection of reviews it summarizes. They are not a reflection of the true opinion towards the product. In view that opinions capture reviewer biases, we propose a novel way of summarizing opinions.

Instead of the usual summary that is presented as a paragraph of selected sentences, we are inspired by the work of Paul et al. (2010) and Wang et al. (2010), where opposing opinions are contrasted. We propose a *balanced opinion summary*, where we summarize and contrast the opinions of reviewers having different biases. For example, we contrast opinions of a reviewers who are lenient against reviewers who are critical. This allows us to present a balanced summary to the reader. The biases can be latent factors that will be discovered during the modeling process.

We propose to achieve a balanced summary that selects salient opinions from reviewers with different biases. We hypothesize that the bias-aware opinion representation will form clusters of similar opinions from reviewers with similar biases. Building on a graph-based approach to summarization like LexRank (Erkan and Radev, 2004) and Yin and Pei (2015), opinions can be represented as nodes and edges as the similarity between bias-aware opinion representation. The density of the graph can be adjusted by the similarity threshold imposed on the graph. The saliency of the opinions can then be obtained by applying PageRank on the graph. In doing so, we also model the similar opinions that signals agreement or consensus among reviewers. After ranking opinions based on its salience, we can utilise a diversity objective through a greedy approach or Maximal Marginal Relevance (MMR) to select salient opinions that are different.

4 Evaluation

Datasets Suitable datasets in the restaurant domain for our research questions are: (1) NY city search (Ganu et al., 2013); (2) SemEval 2016 ABSA Restaurant Reviews in English (Pontiki et al., 2016); and, (3) Yelp dataset challenge[2]. All datasets contain user ID, product ID, review text and review rating, which will allow us to model opinions. In addition, datasets (1) and (2) are labeled with aspect and sentiment polarity. Although we choose to work in the restaurant domain for our proposed work, our models are not domain-specific. Other potential review datasets are on product and hotel reviews (McAuley et al., 2015; Wang et al., 2010).

We approach evaluation in a two part process. First, we evaluate our proposed model on how well it learns a representation of opinion sentence. Next, we compare summaries generated with our bias-aware opinion representation with selected baseline models.

4.1 Bias-Aware Opinion Representation

Our objective is to learn a bias-aware opinion representation such that similar opinions from reviewers with similar bias should cluster together and different opinions form different clusters. We apply the evaluation method used to evaluate vector representation of text sequences by Le and Mikolov (2014). We believe this evaluation method is applicable for our representation. We begin with a dataset of labeled opinions. From the labeled dataset, a triplet of opinions is created with the first and second opinions of the triplet to be of the same opinion, and first and third opinions to be of different opinions. We compute the similarity of opinion between a pairs of the triplet of representation. We expect the first and second opinion to produce a higher similarity as compared to the similarity of the first and third opinion. Of all the triplets we create, we will report the error rate. Error rate here refers to the number of triplets that

---

[2]https://www.yelp.com/dataset/challenge

first and third opinion is more similar than first and second opinion over the total number of triplets.

Our second evaluation will be a cluster analysis of opinion representations. We expect homogeneous clusters of similar opinions from reviewers with similar bias and different clusters for different opinions with reviewer biases. A potential approach will be to perform a k-means clustering where the number of clusters k can be determined by an elbow plot. The quality of clusters can be evaluated using the Silhouette Score.

In order to evaluate the bias-aware opinion representation, we look to answer a related question. Suppose each opinion captures the opinion target, the polarity and reviewer bias. Each opinion within the review contributes to the overall rating. The task is to predict the overall rating based on review text. The model will be trained on a training set of review text, reviewer information and rating. If the model accurately captures the opinion and reviewer bias in the representation, the representative should improve the ability to predict the overall rating of the review given the review text and reviewer information.

4.2 Summarization

Evaluating summaries is a challenging problem. There are two options to evaluate summaries. First, an automatic evaluation method using metrics such as ROUGE and BLEU. However, such method requires a gold standard summary. Obtaining a gold standard summary for our purpose is a challenging task. The second method of evaluation is a user-study type evaluation. Users are presented with generated summaries and are asked to judge the summary according to given criteria or to compare between different summaries. Some baseline models to compare against are Lexrank (Erkan and Radev, 2004) to represent word level models and DivSelect+CNNLM to represent vector representation models (Yin and Pei, 2015). We intend to evaluate our summaries using a user-study.

5 Summary

Not all reviews are equal as reviews capture biases of their reviewers. These biases can be amplified when we analyse a collection of reviews that is not representative of the consumers of the product. As such, analysis on the collection of reviews is not representative and can potentially impact readers who depend on the analysis for decision-making. To address this problem, we propose to model opinion with its reviewer using review text and review rating to obtain a bias-aware opinion representation. We plan to demonstrate the utility of the representation in opinion summarization. Specifically, the representation will be useful in the scoring the sentences for saliency and selection of sentences for generating a balanced summary. Although we focus on modeling opinions for opinion summarization, we believe the same modeling concepts can also be applied to recommendation. We leave evaluation of bias-aware opinion representation on recommendations to future work.

Acknowledgments

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing Opinions: Aspect Extraction Meets Sentiment Prediction and They Are Both Weakly Supervised. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium.

Marco Bonzanini, Miguel Martinez-Alvarez, and Thomas Roelleke. 2013. Extractive Summarisation via Sentence Removal: Condensing Relevant Sentences into a Short Summary. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 893–896, Dublin, Ireland.

Samuel Brody and Noemie Elhadad. 2010. An Unsupervised Aspect-sentiment Model for Online Reviews. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Los Angeles, CA.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A Novel Neural Topic Model and Its Supervised Extension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2210–2216, Austin, TX.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the*

Conference on Empirical Methods in Natural Language Processing, pages 169–174, Brussels, Belgium.

Bee-Chung Chen, Jian Guo, Belle Tseng, and Jie Yang. 2011. User Reputation in a Comment Rating Environment. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 159–167, San Diego, CA.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the International Conference on World Wide Web*, pages 519–528, Budapest, Hungary.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the International Conference on Computational Linguistics*, pages 340–348, Beijing, China.

Gayatree Ganu, Yogesh Kakodkar, and AméLie Marian. 2013. Improving the Quality of Predictions Using Textual Information in Online User Reviews. *Information Systems*, 38(1):1–15.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 388–397, Vancouver, Canada.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 328–339, Melbourne, Australia.

Minqing Hu and Bing Liu. 2004a. Mining and Summarizing Customer Reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, WA.

Minqing Hu and Bing Liu. 2004b. Mining Opinion Features in Customer Reviews. In *Proceedings of the National Conference on Artifical Intelligence*, pages 755–760, San Jose, CA.

Nan Hu, Paul A. Pavlou, and Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-mouth Communication. In *Proceedings of the ACM Conference on Electronic Commerce*, pages 324–330, Ann Arbor, MI.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought Vectors. In *Proceedings of the International Conference on Neural Information Processing Systems - Volume 2*, pages 3294–3302, Montreal, Canada.

Hady W Lauw, Ee-Peng Lim, and Ke Wang. 2012. Quality and Leniency in Online Collaborative Rating Systems. *ACM Transactions on the Web (TWEB)*, 6(1):4.

Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the International Conference on International Conference on Machine Learning - Volume 32*, pages II–1188–II–1196, Beijing, China.

Fangtao Li, Sheng Wang, Shenghua Liu, and Ming Zhang. 2014. SUIT: A Supervised User-Item Based Topic Model for Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1636–1642, Quebec, Canada.

Chenghua Lin and Yulan He. 2009. Joint Sentiment/Topic Model for Sentiment Analysis. In *Proceedings of the ACM Conference on Information and Knowledge Management*, pages 375–384, Hong Kong, China.

Jingjing Liu, Yunbo Cao, Chin-Yew Lin, Yalou Huang, and Ming Zhou. 2007. Low-Quality Product Review Detection in Opinion Summarization. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 334–342, Prague, Czech Republic.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and Predicting the Helpfulness of Online Reviews. In *IEEE International Conference on Data Mining*, pages 443–452, Pisa, Italy.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, OR.

Julian McAuley and Jure Leskovec. 2013. Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text. In *Proceedings of the*

ACM Conference on Recommender Systems, pages 165–172, Hong Kong, China.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52, Santiago, Chile.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119, Lake Tahoe, NV.

Samaneh Moghaddam and Martin Ester. 2010. Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 1825–1828, Toronto, Canada.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and Prediction of Online Product Review Helpfulness: A Survey. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708, Melbourne, Australia.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(12):1–135.

Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing Contrastive Viewpoints in Opinionated Text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 66–76, Cambridge, MA.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, New Orleans, LA.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the International Workshop on Semantic Evaluation*, pages 19–30, San Diego, CA.

Yassien Shaalan and Xiuzhen Zhang. 2016. A time and opinion quality-weighted model for aggregating online reviews. In *Australasian Database Conference*, pages 269–282. Springer.

Ruben Sipos and Thorsten Joachims. 2013. Generating Comparative Summaries from Reviews. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, pages 1853–1856, San Francisco, CA.

Duyu Tang, Bing Qin, and Ting Liu. 2015a. Document Modeling with Gated Recurrent Neural Network for Sentiment Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, Lisbon, Portugal.

Duyu Tang, Bing Qin, Ting Liu, and Yuekui Yang. 2015b. User Modeling with Neural Network for Review Rating Prediction. In *Proceedings of the International Conference on Artificial Intelligence*, pages 1340–1346, Buenos Aires, Argentina.

Ivan Titov and Ryan McDonald. 2008a. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 308–316, Columbus, OH.

Ivan Titov and Ryan McDonald. 2008b. Modeling Online Reviews with Multi-grain Topic Models. In *Proceedings of the International Conference on World Wide Web*, pages 111–120, Beijing, China.

Rahul Wadbude, Vivek Gupta, Dheeraj Mekala, and Harish Karnick. 2018. User Bias Removal in Review Score Prediction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 175–179, Goa, India.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 783–792, Washington, DC.

Lu Wang and Wang Ling. 2016. Neural Network-Based Abstract Generation for Opinions and Arguments. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57, San Diego, CA.

Wenting Xiong and Diane Litman. 2014. Empirical Analysis of Exploiting Review Helpfulness for Extractive Summarization of Online Reviews. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*, pages 1985–1995, Dublin, Ireland.

Wenpeng Yin and Yulong Pei. 2015. Optimizing Sentence Modeling and Selection for Document Summarization. In *Proceedings of International Conference on Artificial Intelligence*, pages 1383–1389, Buenos Aires, Argentina.

Kunpeng Zhang, Yu Cheng, Wei-keng Liao, and Alok Choudhary. 2012. Mining Millions of Reviews: A Technique to Rank Products Based on Importance of Reviews. In *Proceedings of the International Conference on Electronic Commerce*, pages 12:1–12:8, Liverpool, United Kingdom.

Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, pages 425–434, Cambridge, United Kingdom.