Master Thesis Exposé: Drawing Attention to Interpretable Review Helpfulness Prediction

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

Humboldt-University Berlin

July 2, 2019

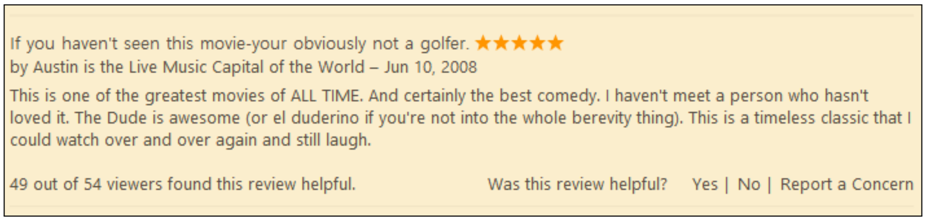
Academic advisors:	Prof. Ulf Leser	<leser@informatik.hu-berlin.de>
	Dr. Roman Klinger	<roman.klinger@ims.uni-stuttgart.de>
	M.Sc. Mario Sänger	<saengema@informatik.hu-berlin.de>

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. 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. "49 out of 54 viewers found this review helpful.", see Figure 1). All existing reviews for a product can then be sorted by this score, resulting in the most "helpful" reviews being listed at the top.

This 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 read by 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 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-



If you haven't seen this movie-your obviously not a golfer. ★★★★★
by Austin is the Live Music Capital of the World – Jun 10, 2008

This is one of the greatest movies of ALL TIME. And certainly the best comedy. I haven't meet a person who hasn't loved it. The Dude is awesome (or el duderino if you're not into the whole berevity thing). This is a timeless classic that I could watch over and over again and still laugh.

49 out of 54 viewers found this review helpful.

Figure 1: A screenshot of a review of the movie *The Big Lebowski* with 49 positive out of 54 total helpfulness votes, taken from Apple iTunes on Windows Desktop on April 12th, 2019.

draft in the review writing process. The predictor could also be analyzed to gain a general understanding of what makes a review helpful. Ultimately, this could improve the overall quality and helpfulness of online reviews.

This master thesis will focus on the interpretability and explainability of state-of-the-art helpfulness prediction models. In comparison with approaches based on hand-crafted features, such as SVMs [2–7], neural-network based predictors deliver improved performance in helpfulness prediction [8, 9] and other NLP tasks [10–16]. However, individual hand-crafted features and their importance for a prediction task can be compared better (e.g. in ablation studies) than features learned by a neural network. With this master thesis, we want to shed light on the explainability and interpretability of state-of-the-art neural-network models utilizing the *attention* mechanism [17, 18]. We will extract learned parameters to outline the (overall and sample-specific) reasoning and compare that to feature importance metrics obtained from traditional, hand-crafted feature models. As attention-based models are expected to deliver better performance than previous approaches, we hope to find further answers to the question what makes a review helpful and whether there are differences across product categories. Simultaneously, we will investigate if attention-based models learn a different concept of review helpfulness than hand-crafted feature models.

The exposé continues as follows. After a formal definition of the review helpfulness prediction task in Section 2.1, an overview of the relevant review domains and datasets is given in Section 2.2. Section 2.3 contains a brief overview of existing review helpfulness prediction approaches and their most relevant features. Section 3 introduces the attention mechanism. Our approach is described in Section 4 followed by our evaluation methods in Section 5.

2 Review Helpfulness Prediction

2.1 Problem Definition

Depending on the available data, the helpfulness prediction problem can be defined in multiple ways. For a dataset with positive and negative helpfulness votes, the helpfulness score can be defined as follows:

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

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

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 obtain a *ranking* of a set of reviews, by sorting the reviews by the helpfulness score.

For a dataset that includes only positive helpfulness votes, due to the lack of negative votes, the helpfulness score from Equation 1 cannot be used. Instead, either a dataset-specific alternative metric (e.g. the ratio of "useful" votes to the sum of "useful", "cool" and "funny" votes in the Yelp dataset, as applied in [19]) or a measure based solely on positive helpfulness votes can be used. An example for the latter would be defining a helpfulness-vote threshold t based on the data and considering all reviews with at least t helpfulness votes as helpful, reviews with less than t votes as non-helpful, resulting in labels for binary classification. For regression, the vote count could be used either directly or scaled (e.g. logarithmic). These approaches are not optimal, since the voting-process of the review-platform is shown to have an influence on the amount of total votes a review receives [1]. A high amount of positive votes for a review can not only be the result of it being more helpful than others, but also be the result of it being listed at the top and therefore getting read and voted for more often. However, the approaches represent review helpfulness to some extend and for such datasets there exists no alternative.

Common evaluation metrics for classification problems are the F1 score, precision and recall. For regression, common metrics are root mean squared error, mean squared error, mean absolute error 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.

2.2 Datasets and Review Domains

A number of user review datasets including helpfulness votes have already been published. McAuley [20] 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]. It includes both positive and negative helpfulness vote counts. The music technology group of the University Pompeu

[1] See `https://jmcauley.ucsd.edu/data/amazon`, accessed on June 15th, 2019

Fabra in Barcelona [21] 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. Blitzer et al. [22] share the Multi-Domain Sentiment Dataset, which is another dataset of reviews from Amazon.com. Compared to ARD, it contains significantly less reviews (1,422,530). Yelp offers a dataset of 6,685,900 reviews on 192,609 businesses across 10 metropolitan areas [23]. Besides the count of votes marking a review useful, the Yelp dataset also contains vote counts on a review being "funny" or "cool". Sobkowicz et al. [24] published a dataset of 6.4 million reviews on games from the Steam platform including positive helpfulness votes. Gräßner et al. [25] published a dataset of 215,063 reviews[3] from drugs.com, including drug name, condition and the number of positive helpfulness votes. Tang et al. [26] published a dataset of 302,232 reviews[4] from Ciao.com, a no longer existing product review website. In contrast to ARD and MARD, the helpfulness votes in the Ciao dataset are not binary, but in the range of 0 to 5, and not anonymous, but coupled to user ids. According to Diaz and Ng [1], the data "allowed researchers to make observations [. . .] that cannot be made on Amazon.com" by offering "information on a social trust network, where users choose to connect to reviewers if they find their reviews consistently helpful".

2.3 Existing Approaches

Several approaches on review helpfulness prediction have already been investigated [1, 27]. For both classification and regression, Support Vector Machines have been a popular method [2–7]. For classification, thresholded linear regression [28], Naive Bayes [7,29,30] and Random Forests [28–30] and for regression, linear regression [31], probabilistic matrix factorization [26] and extended tensor factorization models [32] have been used as well. Recent approaches often utilize neural networks to tackle the problem , including multilayer perceptrons networks [7, 33] and convolutional neural networks [8, 34].

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 all other metadata, such as data about the reviewer, the product to be reviewed or the time and date of publication. Frequently used content features represent the review length [3, 6, 31, 35, 36], readability [7,28,37,38], sentiment [6,7], emotions [7,39,39], a reviews language style [2,6], linguistic attributes [7,38] or the mention of certain words [3,4,9,35] and aspects [9]. For neural network based models, co-occurrence based character, word and topic embeddings have been used as an alternative encoding to traditional, hand-crafted features [8,34]. Most previous work is based on reviews in English language from Amazon.com or Ciao.com with some exceptions. Zhang et al. [40]

[2] See `https://www.upf.edu/web/mtg/mard`, accessed on June 15th, 2019

[3] See `https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+(Drugs.com)`, accessed on June 6th, 2019

[4] See `https://www.cse.msu.edu/~tangjili/trust.html`, accessed on June 19th, 2019

evaluated a model using on reviews from the Yelp dataset in German and English language, Liu et al. [35] used reviews from IMDB.com and O'Mahony et al. [41] used reviews from Tripadvisor.com.

3 Text Classification with Attention

Attention is a mechanism in neural networks that is widely used nowadays and inspired by a capability in human perception [42]: Imagine trying to predict the missing word in the following sentence, with `[MASKED]` being a placeholder for the missing word:

She is eating a green `[MASKED]`*.*

Some words in that sentence (especially *eating* and *green*) provide more contextual information about the missing word than others. To make a prediction, these words will be accounted for the most, while other, less important words (e.g. *She*, *is*) will be mostly ignored. More generally, attention weights can be used to model the relationship of words in a sentence. A high attention weight corresponds to a strong relationship. An example of high and low attention weights is depicted in Figure 2.

In the context of natural language processing, attention was originally introduced to improve the performance of sequence to sequence (or encoder-decoder) models [43] for tasks like machine translation or abstractive summarization. These models were often based on recurrent neural networks, with an encoder modeling the entire input sequence in a single context vector and the decoder generating an output sequence from the context vector. Its shortcomings were, next to the difficult training procedure, it's lack of modeling long-range dependencies within an input sequence. The modeling of such long range dependencies can be improved by the attention mechanism. Every output token can access information about the input sequence through a separate context vector that is built as a weighted sum of input token vectors. The attention weights are trained with the model and capture how much the current, to be predicted output-token is influenced by (or "attends to") tokens from the input sequence. Since then, different variants of the attention mechanism (e.g. self-attention [44–46], pointer-networks [47] or the Simple Neural Attention Meta-Learner [48]) have been proposed and successfully applied in other NLP tasks like text classification, named entity recognition and relationship extraction [17,18,46,49,50].

In the context of text classification, there are two specific approaches we want to highlight: To reflect the structure of a text document Hierarchical Attention Networks (HAN) [18] capture sentence-level attention next to token-level attention. The recently published model Bidirectional Encoder Representations from Transformers (BERT) [17] has improved the state-of-the-art performance for text classification and other NLP tasks considerably. The key innovation in BERT, compared to previous approaches (e.g. GPT [51] and ELMO [14]), is the joint training of left and right (attention-based) word context with masked language modeling. This allows the model to capture the bidirectional con-

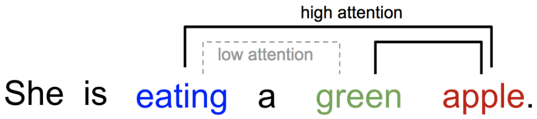


Figure 2: An illustration of how a word pays different amounts of attention to other words. (Source: [42], accessed on May 9th, 2019)

text of words in a sentence or document. Different pre-trained model variants are available for download and can be fine-tuned for several tasks, including text classification. According to our knowledge, attention-based neural network models have not yet been evaluated on the review helpfulness prediction task.

4 Approach

In this master thesis the central research question will be: *What makes a review text helpful and what are the differences across product review domains?* To find answers to the question, a series of review helpfulness prediction experiments will be carried out using both regression and binary classification problem formulations. The data for the experiments will primarily consist of reviews taken from multiple Amazon product categories of the ARD dataset [20]. Furthermore, reviews from the Drugs.com dataset [25] will be evaluated. The definition of helpfulness will equal the helpfulness score for ARD and an only-positive-vote based metric for Drugs.com (see Section 2.1) To point out differences between helpful and non helpful reviews and differences across product categories, a dataset analysis utilizing statistical measures and visualizations will be carried out upfront.

Two main types of prediction models, attention-based neural models (e.g. BERT [17] and HAN [18]) and a traditional model based on hand-crafted features will be compared. For the latter we will use a Support Vector Machine [52]. The hand-crafted features will include review length, emotions, sentiment, readability and other linguistic and token-based features. To highlight review domain differences, their performance will be compared in ablation studies across product categories.

A subsequent research question concerning attention-based models will be: *What is represented in attention weights of attention-based neural network models and can they be mapped to hand-crafted review helpfulness features?* We will therefore evaluate at least two attention-based models (e.g. BERT [17] and HAN [18]). To find answers to the question, we will extract attention weights from the models that represent token importance for review helpfulness. To gain insight into the sample-specific relevance of certain tokens for review helpfulness, these weight-token pairs will be visualized for selected reviews similar to visualizations in previous work [18, 45].

To get more insight into what attention-based models learn to be relevant for review helpfulness, we will further analyze the weight-token pairs throughout all reviews of a product category or dataset. More specifically, we will summarize the weight-token pairs for a set of reviews (e.g. by calculating the sum of attention weights for each token). Next, we will aggregate these values for word groups corresponding to linguistic features (e.g. emotion, sentiment) based on dictionaries (e.g. INQUIRER, NRC Word-Emotion Association Lexicon, MPQA Subjectivity Lexicon). From that we will obtain a ranking of feature categories, that will be evaluated across product categories. This ranking will be compared to the ranking of hand-crafted features in the aforementioned ablation studies. With this analysis, we intend to investigate whether attention-based models learn a different concept of review helpfulness than hand-crafted feature models.

For further comparison, we will extract weight-token pairs from a model solely based on tf-idf features. The same process as described above will be applied to obtain an alternative ranking of dictionary-categories. Both rankings will be compared with statistical methods.

The attention-weights will also be analyzed in terms of the relative position of high-attention tokens in a review. With that, we will investigate whether different parts of a review (e.g. the title, beginning or end) are equally important for review helpfulness. Existing, specialized visualization tools [53] will possibly be utilized supportively in the analyzes.

5 Evaluation

We will evaluate the performance of both the attention based and hand-crafted feature based models for the classification and regression task. The performance will be compared across product categories from ARD [20] and Drugs.com [25]. The importance of feature groups will be compared in ablation studies. A detailed error analysis will be carried out (e.g. by examining a potential correlation of low total helpfulness vote counts with high prediction errors). We will also investigate the models cross-domain performance by predicting on a different product category or dataset than it was initially trained on. The differences between dictionary-category rankings extracted from attention and tf-idf feature weights will be evaluated with statistical methods, as well as the differences in token weights directly.

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] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

[11] R. Johnson and T. Zhang, "Supervised and semi-supervised text categorization using lstm for region embeddings," in *International Conference on Machine Learning*, pp. 526–534, 2016.

[12] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, and R. Kurzweil, "Universal sentence encoder for english," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Brussels, Belgium), pp. 169–174, Association for Computational Linguistics, Nov. 2018.

[13] S. Gray, A. Radford, and D. P. Kingma, "Gpu kernels for block-sparse weights," *Technical report, OpenAI*, 2017.

[14] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[15] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in *Advances in Neural Information Processing Systems*, pp. 6294–6305, 2017.

[16] R. Johnson and T. Zhang, "Deep Pyramid Convolutional Neural Networks for Text Categorization," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Vancouver, Canada), pp. 562–570, Association for Computational Linguistics, 2017.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, Oct. 2018. arXiv: 1810.04805.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical Attention Networks for Document Classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1480–1489, Association for Computational Linguistics, 2016.

[19] M. Fan, C. Feng, L. Guo, M. Sun, and P. Li, "Product-aware helpfulness prediction of online reviews," in *The World Wide Web Conference*, WWW '19, (New York, NY, USA), pp. 2715–2721, ACM, 2019.

[20] 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.

[21] 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.

[22] J. Blitzer, M. Dredze, and F. Pereira, “Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification,” in *In ACL*, pp. 187–205, 2007.

[23] “Yelp open dataset.” `https://www.yelp.com/dataset`. Accessed: 2019-02-26.

[24] A. Sobkowicz and W. Stokowiec, “Steam Review Dataset - new, large scale sentiment dataset.” `https://www.researchgate.net/publication/311677831_Steam_Review_Dataset_-_new_large_scale_sentiment_dataset`, 2016. Accessed: 2019-05-22.

[25] F. Gräßer, S. Kallumadi, H. Malberg, and S. Zaunseder, “Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning,” in *Proceedings of the 2018 International Conference on Digital Health*, DH ’18, (New York, NY, USA), pp. 121–125, ACM, 2018.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Y. Zhang and Z. Lin, "Predicting the helpfulness of online product reviews: A multilingual approach," *Electronic Commerce Research and Applications*, vol. 27, pp. 1–10, Jan. 2018.

[41] M. O'Mahony and B. Smyth, "A classification-based review recommender," *Knowledge-Based Systems*, vol. 23, pp. 323–329, May 2010.

[42] L. Weng, "Attention? attention!." `https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html`. Accessed: 2019-05-13.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[44] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, 2015.

[45] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561, 2016.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv:1706.03762 [cs]*, June 2017. arXiv: 1706.03762.

[47] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *Advances in Neural Information Processing Systems*, pp. 2692–2700, 2015.

[48] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," *arXiv preprint arXiv:1707.03141*, 2017.

[49] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and u. Kaiser, "Universal Transformers," *arXiv:1807.03819 [cs, stat]*, July 2018. arXiv: 1807.03819.

[50] C. Du and L. Huang, "Text Classification Research with Attention-based Recurrent Neural Networks," *International Journal of Computers Communications & Control*, vol. 13, p. 50, Feb. 2018.

[51] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *Technical report, OpenAI*, 2018.

[52] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[53] J. Vig, "Visualizing attention in transformer-based language models," *arXiv preprint arXiv:1904.02679*, 2019.