Can You Distinguish Truthful from Fake Reviews? User Analysis and Assistance Tool for Fake Review Detection

Jeonghwan Kim* **Junmo Kang*** **Suwon Shin*** **Sung-Hyon Myaeng**
School of Computing, KAIST
Daejeon, Republic of Korea
{jeonghwankim123, junmo.kang, ssw0093, myaeng}@kaist.ac.kr

Abstract

Customer reviews are useful in providing an indirect, secondhand experience of a product. People often use reviews written by other customers as a guideline prior to purchasing a product/service or as a basis for acquiring information directly or through question answering. Such behavior signifies the authenticity of reviews in e-commerce platforms. However, fake reviews are increasingly becoming a hassle for both consumers and product owners. To address this issue, we propose *You Only Need Gold (YONG)*, an assistance tool for detecting fake reviews and augmenting user discretion. Our experimental results show the poor human performance on fake review detection, substantially improved user capability given our tool, and the ultimate need for user reliance on the tool.

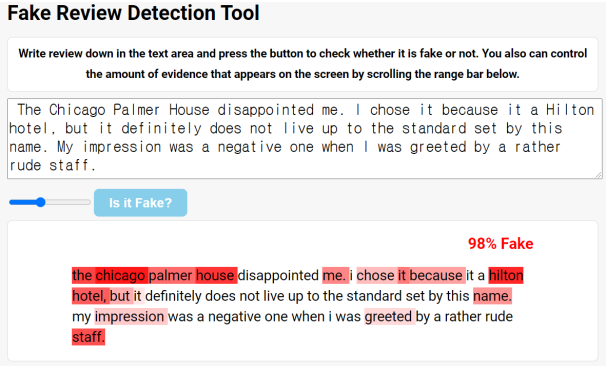

Figure 1: YONG - a prototype interface. Given the review input by user, YONG shows 1) whether it is gold or fake with 2) the probability (%), and 3) evidence that shows how much each word contributes to model's final decision. The highlighted evidences show the top p% proportion of the contributors, where the proportion can be adjusted using the horizontal slider bar.

1 Introduction

The increasing prominence of e-commerce platforms gave rise to numerous customer-written reviews. The reviews, given their authenticity, provide important secondhand experience to other potential customers or to information-seeking functions such as search and question answering. Meanwhile, fake reviews are increasingly becoming a social problem in e-commerce platforms (Chakraborty et al., 2016; Rout et al., 2018; Ellson, 2018). Such deceptive reviews are either incentivized by the beneficiaries (i.e., sellers, marketers) or motivated by those with malicious intention to damage the reputation of the target product.

To date, there have been many studies (Kim et al., 2015; Wang et al., 2017; Aghakhani et al., 2018; You et al., 2018; Kennedy et al., 2019) that address fake review detection in the field of natural language processing (NLP). The use of high-performance deep neural networks such as BERT (Devlin et al., 2018), fast and scalable anomaly detection algorithms like DenseAlert (Shin et al., 2017) have made effective and promising contributions in detection of fraudulent reviews. Despite such contributions, these approaches only focus on better modeling to improve the accuracy in fake review detection, instead of its practical applications such as assisting users to distinguish fake reviews (i.e., an assistance tool) or filtering out deceptive texts for review-based question answering (QA) (Gupta et al., 2019).

In the line of Human-Computer interaction (HCI), there have been a variety of studies on customer reviews (Wu et al., 2010; Alper et al., 2011; Yatani et al., 2011; Zhang et al., 2020). While there are *gold* reviews, which are authentic, real-user written reviews, there are *fake*, deceptive reviews as well. All of the previous works on review visualization and interaction implicitly assume the authenticity of collected reviews. Furthermore, the previous works mentioned above confine the scope

*Equal contribution.

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 53–59
April 20, 2021.

of research on reviews to interaction and visualization.

We claim that user capabilities of distinguishing fake reviews are seriously unreliable as suggested in (Lee et al., 2016), and this motivates our work as practical research for helping humans discern fake reviews. While the two lines of research in NLP and HCI have focused on improving the fake review detection models and developing effective visualization of reviews, respectively, the actual victims (the users) of fake reviews are being neglected. The challenges of carefully curated deceptive reviews on the Web necessitate the need for an assistance tool that helps users avoid fraudulent information.

In this work, we propose *You Only Need Gold (YONG)* (Figure 1), a simple assistance tool that augments user discretion in fake review detection. YONG is built upon the body of previous studies by fine-tuning BERT and providing a self-explanatory *gold indicator* to assist users in exploring customer reviews. Through a series of user evaluations, we reveal the over-confident nature of people despite their poor performance in distinguishing fake reviews from real ones and the need to implement an explainable, human-understandable features to guide user decisions in fake review detection. We also demonstrate that the application of YONG effectively augments user performance.

Our contributions is two-fold, the tool and an extensive user understanding, as follows:

- An easy-to-use tool for fake review detection with the intuitive *gold indicator*, which consists of the following three features: (i) Model Decision, (ii) Percentage Indicator (%), and (iii) Evidence.
- User analysis on fake review detection. Our work sheds a light on how susceptible human judgment is to deceptive text. Understanding human decisions and behaviors in discerning fake reviews provides an insight into the design considerations of a system or tool for fake reviews.

2 Related Work

2.1 Fake Review Detection Using Deep Neural Networks

Fake review detection is an extensively researched topic among deep learning researchers. (Kennedy et al., 2019) provide a list of analytic comparison of non-neural and neural network models. In their work, a fine-tuned BERT performs the best on crowd sourced *Deceptive Opinion Spam* (Ott et al., 2013, 2011) data set and automatically obtained Yelp (Rayana and Akoglu, 2015) data set. Another work proposes a generative framework for fake review generation (Adelani et al., 2020). It proposes a pipeline of GPT-2 (Radford et al., 2019) and BERT to generate fake reviews. Due to the remarkable fluency of these reviews, its participants failed to identify fake reviews, and surprisingly, gave higher fluency score to the generated reviews than to the gold reviews. Similar result is evidenced in (Donahue et al., 2020), where humans have difficulty identifying machine-generated sentences. Other related works (Lee et al., 2016) use probabilistic methods such as Latent Dirichlet Analysis (LDA) to discover word choice patterns and linguistic characteristics of fake reviews.

2.2 Review Interaction and Visualization

The growth of customer reviews sparked research on its interaction and visualization. Visualization tools like OpinionBlocks (Alper et al., 2011) provide an interactive visualization for better organization of the reviews in bi-polar sentiments. Similar works like OpinionSeer (Wu et al., 2010) and Review Spotlight (Yatani et al., 2011) are also focused on providing an accessible and interactive visualization of reviews. The major drawback of these works is that they naively assume the *authenticity* of the reviews. From the previous lines of work, we argue the threat of fake reviews and their imminence with the rise of generative models (e.g., GPT-2) necessitate the use of our tool.

3 Method

We build a tool that provides straightforward, intuitive features to guide user decision in fake review detection. The tool is built upon a state-of-the-art NLP model fine-tuned on the OpSpam data set.

3.1 You Only Need Gold (YONG)

The tool we propose is a prototype for receiving a review text as an input and returns whether it is "Gold" or "Fake". YONG is an easy-to-use tool with three features collectively referred to as the *gold indicator* - namely, (i) Model Decision, (ii) Probability (%) and (iii) Evidence. Model Decision is the model output on the top right corner of Figure 1 as either Gold or Fake. The Probability (%) is the softmax output, which is also the model

Models	*Accuracy*
SVM	0.864
FFN	0.888
CNN	0.669
LSTM	0.876
BERT (Kennedy et al., 2019)	0.891
BERT (Ours)	**0.896**

Table 1: Accuracy performance comparison of conventional classification models against BERT on OpSpam.

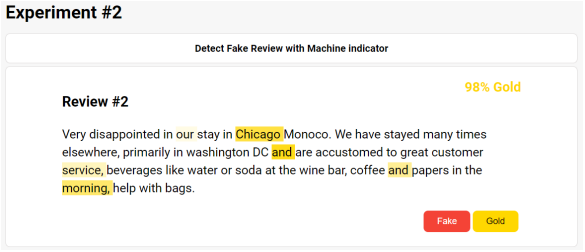

Figure 2: Experiment Test Bed for YONG - Experiment #2

confidence for its decision. The word highlights, which we also define as the Evidence, is a visualization of the attention weights from the last layer of BERT to provide an interpretable medium to model's decision for its users.

3.2 Fake Review Detection Model

To validate the claim made by a previous work (Kennedy et al., 2019) that BERT outperforms other baselines in fake review detection and to employ the most effective existing approach to fake review detection in our tool, we compare the performance of BERT against other classification models on the OpSpam (Ott et al., 2011, 2013) data set. In Table 1, we report the results of non-neural and neural network models under the 5-fold cross validation setting from (Kennedy et al., 2019). Our implementation of BERT outperforms all the other baseline models, reaching a 90% accuracy. Based on the model performances on OpSpam data set in Table 1 and the renowned language understanding capability of BERT (Devlin et al., 2018), we decide to employ BERT in our tool. We also fine-tune our BERT with the HuggingFace (Wolf et al., 2020) version of `bert-base-uncased`, where the `[CLS]` embedding is used for binary sequence classification.

3.3 Training & Dataset

To fine-tune BERT and to choose carefully curated data on fake review detection, we use the Deceptive Opinion Spam corpus, also known as the OpSpam data set (Ott et al., 2011, 2013). This data set consists of a total of 1,600 review instances, which is divided into truthful (i.e., gold) and deceptive reviews. The gold reviews are extracted from the 20 most popular hotels in Chicago (800 gold reviews) and the set of fake reviews are built using Amazon Mechanical Turk (AMT) (800 fake reviews). The gold reviews are not artificially generated, but are carefully chosen based on the criteria of determining deception as an effort to ensure truthfulness in (Ott et al., 2011). The data set is divided into a training and test set ratio of 8:2 in this work.

4 Experiment and Result

4.1 Research Question

To validate the usefulness of our tool and further the understanding of users in the application of our tool, we define the following research questions (RQ):

RQ1. "How do humans fare against machines on fake review detection?"

RQ2. "Does YONG augment human performance on the task?"

RQ3. "Can we increase the level of human trust on YONG by injecting prior knowledge about human and model performance?"

RQ4. "How much influence does each feature of the gold indicator have on human trust?"

Here, we define the term "trust" as the level of human reliance on the tool's decision. To be specific, as we evaluate in Section 4.5, the level of trust is calculated on the participant-machine agreement (regardless of the ground-truth label). Through the experiments that correspond to the RQs, we build a concrete understanding of user behaviors in the presence of customer-generated reviews and the gold indicator through human performance evaluation on fake review detection.

4.2 Experimental Design

We provided 10 reviews from the data set to a total of 24 participants. The 10 reviews were randomly sampled, resulting in correct-to-wrong ratio of the model prediction to 7:3 (i.e., 7 correct and 3 wrong model predictions; score = 0.70), and the Gold-to-Fake ratio to 5:5. From Experiment 1 to 4 we assess the following criteria: human discretion, tool

Experiment	*Condition*	*Score*	*Trust (low / high group)*
Exp. 1	w/o tool	0.41	-
Exp. 2	w/ tool	0.54	0.81 / 0.79
Exp. 3	w/ tool & score	0.56	0.83 / 0.71
-	Model	0.70	-

Table 2: Average scores (i.e., the number of correct decision w.r.t. the ground truth) of participants and the model, and the level of trust (i.e., participant-model agreement). Here, *low* and *high* groups are those with below and above average scores in Exp. 1, respectively.

helpfulness, trust level on the tool, and feature-level influence on decisions made:

Experiment 1. Users are required to classify fake reviews given 10 reviews without our tool.

Experiment 2. Users conduct the same task provided our tool (i.e. gold indicator).

Experiment 3. Users are shown their score and the model score for Experiment 1, and conduct the same task as in Experiment 2.

Experiment 4. Users are asked to score how each of the three features in our tool influences their decision.

We designed and built a separate Experiment test bed (Figure 2) using React [1] and FastAPI [2] to conduct an extensive quantitative and qualitative experiments on the participants of our study. To alleviate the learning effect, the participants are made unaware of the ground-truth answers and stay oblivious to both their and model scores (until Experiment 3).

4.3 Human Discretion Assessment

In Experiment 1, we assess the human performance on fake review detection. The participants are not given any hint - only the raw text. At the end of Experiment 1, we ask the participants enter their expected score. Here, the "score" corresponds to the number of correct decisions with respect to the ground-truth. In Table 2, we see that the average score of the participants are at 0.41. This result contrasts with the high accuracy score of our model (Table 1). An interesting observation made from Experiment 1 is the level of confidence participants had. Their expected score was 0.65 while their actual score was, on average, 0.41. They assumed that they got on average of 2.4 more problems (out of 10) correct than their actual score.

[1] https://reactjs.org

[2] https://fastapi.tiangolo.com

4.4 Helpfulness Assessment

To evaluate the helpfulness of our tool in augmenting user performance on fake review detection, we provide the gold indicator as in Figure 2. For a fair comparison, the participants were given the same reviews as in Experiment 1. In Table 2, the accuracy increases substantially from 0.41 to 0.54 by simply providing the gold indicator with the review text. Before proceeding to Experiment 3 - *trust level assessment*, we provide the participants their own scores on Experiment 1 and the model score to see if the injection of prior bias (i.e., being aware of the performance gap) could influence the participants to more align their answers with those of the model.

4.5 Trust Level Assessment

As previously stated, the participant is shown both scores prior to entering the experiment. The result shows an increase in the average score in Experiment 3 compared to that in Experiment 1. With the user awareness of their and model scores shown to improve the average participant score on the task, we decide calculate the change in the level of user trust on our tool. The trust level is calculated based on the proportion of overlapping answers (Table 2). The analysis on the trust level reveals two disparate groups: (i) Those who earned lower score than average in Experiment 1 and (ii) those who earned higher score than average in the same experiment. While the latter shows a drop in trust level from Experiment 2 to 3 ($0.79 \rightarrow 0.71$), the former shows an increase in the trust level ($0.81 \rightarrow 0.83$).

From Experiment 1, 2 to 3, we see the increase in average accuracy from 0.41 to 0.54, and to 0.56, respectively. To evaluate the statistical significance of the changes, we apply one-way repeated ANOVA with a post-hoc test to see that there are significant differences between Experiments 1 and 2 *($p < 0.005$)* and Experiments 1 and 3 *($p < 0.005$)*, and no statistically significant difference between Experiments 2 and 3. The result suggests that there is a large gap between human reasoning and data-driven model decision, and thus providing YONG contributes to an augmented human performance on fake review detection.

4.6 Feature-Level Influence Assessment

Aside from evaluating performance, we also measure the *feature-wise influence* on decision. We ask each participant to rate the three features in a

Feature	*Decision*	*Probability*	*Evidence*
Influence	3.69	3.91	1.87

Table 3: Influence score for each feature of the gold indicator.

1-to-5 scale for how much influence did each feature have on their decision. Here, the scale at 1 represents "No Influence", while scale at 5 means "Major Influence". In Table 3, we show the average rating per gold indicator feature. This result shows that the Probability (%) plays the primary role in convincing users that model prediction is correct. In other words, the higher the probability score, the more "trustworthy" the model decision becomes.

5 Discussion

There are three major findings in this work. In a fake review detection setting, (i) human capability is unreliable and needs machine assistance, (ii) the interpretable hidden weights are hardly explicable, and (ii) for DNN-powered assistive tools like YONG, it is essential to provide faith-gaining features for its users.

A notable finding is the difference between *interpretability* and *explainability*. In this work, we distinguish the two terms to ease their equivocal use. Interpretability refers to the property of being able to describe the cause and effect of the model through observation, whereas explainability refers to the human-understandable qualities that can be accepted on the terms of human logic and intuition. Although numerous existing works (Lee et al., 2017; Vig, 2019; Kumar et al., 2020) have repeatedly provided an interpretable look into the model's decision making process through layer-wise hidden vector or attention weight visualization, most stop at showing the colored vectors and matrices. In Figure 2, we can see how the model places much attention on proper and common nouns like "Chicago," and "morning," that fail to disambiguate and explain the reasoning process of our model for a human to understand. We can also observe in Table 3, where the evidence (i.e., highlighted words based on the attention weights) shows the lowest influence score on the users of our tool. This result implies that users found the feature either obsolete or inexplicable, leading to the low reference to the respective feature. These findings are also supported by a number of previous works on attention weights' explainability (Jain and Wallace, 2019; Kobayashi et al., 2020). (Jain and Wallace, 2019) show that the much touted attention weights' transparency for model decision is unaccounted for and do not provide meaningful explanations. (Kobayashi et al., 2020), furthermore, shows that focusing only on the parallels between the attention weights and linguistic phenomena within the model is insufficient and thus requires a norm-based analysis. Based on such observation, a possible addition to our tool can be generating textual explanations (Liu et al., 2018) or providing not only the token-level highlights as in Figure 2, but also more high-level (e.g., sentence-, paragraph-level information) highlights that show a comprehensible process to model decision.

6 Conclusion

Our work proposed You Only Need Gold (YONG), an assistant tool for fake review detection. From a series of experiments, we deepened our understanding of human capability in fake review detection. We observed that people were generally overconfident with their ability to discern fake reviews from real ones, and we discovered that the model far outperforms its human counterparts, suggesting the need for effective design to convince users to trust the model decision. Furthermore, our work reveals the need to develop more "explainable" tools and promotes collaboration of users and the machine for fake review detection. For future work, expanding the scope of our tool to other fields such as products and restaurants would likely contribute to its generalizability.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services).

References

David Ifeoluwa Adelani, Haotian Mai, Fuming Fang, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2020. Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications*, pages 1341–1354. Springer.

Hojjat Aghakhani, Aravind Machiry, Shirin Nilizadeh, Christopher Kruegel, and Giovanni Vigna. 2018. Detecting deceptive reviews using generative adversarial networks. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 89–95. IEEE.

Basak Alper, Huahai Yang, Eben Haber, and Eser Kandogan. 2011. Opinionblocks: Visualizing consumer reviews. In *IEEE VisWeek 2011 Workshop on Interactive Visual Text Analytics for Decision Making*.

Manajit Chakraborty, Sukomal Pal, Rahul Pramanik, and C. Ravindranath Chowdary. 2016. Recent developments in social spam detection and combating techniques: A survey. *Information Processing Management*, 52(6):1053–1073.

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*.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Andrew Ellson. 2018. 'a third of tripadvisor reviews are fake' as cheats buy five stars.

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019. Amazonqa: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Stefan Kennedy, Niall Walsh, Kirils Sloka, Andrew McCarren, and Jennifer Foster. 2019. Fact or factitious? contextualized opinion spam detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 344–350, Florence, Italy. Association for Computational Linguistics.

Seongsoon Kim, Hyeokyoon Chang, Seongwoon Lee, Minhwan Yu, and Jaewoo Kang. 2015. Deep semantic frame-based deceptive opinion spam analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1131–1140.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075.

Avinash Kumar, Vishnu Teja Narapareddy, Veerubhotla Aditya Srikanth, Aruna Malapati, and Lalita Bhanu Murthy Neti. 2020. Sarcasm detection using multi-head attention based bidirectional lstm. *IEEE Access*, 8:6388–6397.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126.

Kyungyup Daniel Lee, Kyungah Han, and Sung-Hyon Myaeng. 2016. Capturing word choice patterns with LDA for fake review detection in sentiment analysis. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, pages 1–7.

Hui Liu, Qingyu Yin, and William Yang Wang. 2018. Towards explainable nlp: A generative explanation framework for text classification. *arXiv preprint arXiv:1811.00196*.

Myle Ott, Claire Cardie, and Jeffrey T Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 497–501.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319, Portland, Oregon, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, pages 985–994.

Jitendra Kumar Rout, Amiya Kumar Dash, and Niranjan Kumar Ray. 2018. A framework for fake review detection: issues and challenges. In *2018 International Conference on Information Technology (ICIT)*, pages 7–10. IEEE.

Kijung Shin, Bryan Hooi, Jisu Kim, and Christos Faloutsos. 2017. Densealert: Incremental dense-subtensor detection in tensor streams. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1057–1066.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Xuepeng Wang, Kang Liu, and Jun Zhao. 2017. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 366–376.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yingcai Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, and Huamin Qu. 2010. Opinionseer: interactive visualization of hotel customer feedback. *IEEE transactions on visualization and computer graphics*, 16(6):1109–1118.

Koji Yatani, Michael Novati, Andrew Trusty, and Khai N Truong. 2011. Review spotlight: a user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1541–1550.

Zhenni You, Tieyun Qian, and Bing Liu. 2018. An attribute enhanced domain adaptive model for cold-start spam review detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1884–1895.

Xiong Zhang, Jonathan Engel, Sara Evensen, Yuliang Li, Çağatay Demiralp, and Wang-Chiew Tan. 2020. Teddy: A system for interactive review analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.