Analyzing the Support-Level for Tips Extracted from Product Reviews

Miriam Farber*
mfarber@fb.com
Facebook, Israel

David Carmel
dacarmel@amazon.com
Amazon, Israel

Lital Kuchy
litalku@amazon.com
Amazon, Israel

Avihai Mejer
amejer@amazon.com
Amazon, Israel

Abstract

Useful tips extracted from product reviews assist customers to take a more informed purchase decision, as well as making a better, easier, and safer usage of the product. In this work we argue that extracted tips should be examined based on the amount of support and opposition they receive from all product reviews. A classifier, developed for this purpose, determines the degree to which a tip is supported or contradicted by a single review sentence. These support-levels are then aggregated over all review sentences, providing a global support score, and a global contradiction score, reflecting the support-level of all reviews to the given tip, thus improving the customer confidence in the tip validity. By analyzing a large set of tips extracted from product reviews, we propose a novel taxonomy for categorizing tips as highly-supported, highly-contradicted, controversial (supported and contradicted), and anecdotal (neither supported nor contradicted).

CCS Concepts

• **Information systems → Information extraction**.

Keywords

Tip extraction, Support-level analysis

ACM Reference Format:
Miriam Farber, David Carmel, Lital Kuchy, and Avihai Mejer. 2022. Analyzing the Support-Level for Tips Extracted from Product Reviews. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531805

1 Introduction

Product reviews on e-commerce websites such as *amazon.com*, or *ebay.com*, have been shown to play a key role in customers' decision process [5]. However, many popular products have hundreds, if not thousands of reviews, making the information seeking task tedious and time-consuming. Hence, several approaches have been investigated for mining useful information from the product reviews, for example helpful sentence detection [6], or *tip* extraction [10].

There have been a long line of research on tips extraction, in domains such TripAdvisor [8, 27], restaurants [7, 27], hotels [12], and even from a community question answering site [25]. The definition of customer tips varies across different research directions; some works concentrate on identifying common and re-occurring advices (e.g. [6]), while others on practical advices (e.g [8, 10]). Aligned with prior work [10], we define a product tip as a concrete, self-contained, often actionable, and non-obvious piece of advice.

Tip extraction methods vary from classical machine learning (ML) approaches ([7], [8]), up to more modern transformers-based approaches [10], predicting whether a candidate review sentence can serve as a useful tip, hence mostly capturing the linguistic aspects of the sentence that makes it 'looks like' a tip. However, it is not clear whether a tip is anecdotal, i.e. representing one customer point of view, whether there are many customers that actually support, or oppose it, or whether it is controversial, i.e. it is supported and opposed simultaneously. We believe that customers could greatly benefit when a tip is accompanied by support or contradiction evidence, based on other customers' experience, as reflected in all product reviews.

In order to measure how much a tip is supported, or opposed, we fine-tune an off-the-shelf classifier, originally trained to identify the support-level between two text fragments, to predict the amount of support or opposition that a given tip attains from a review sentence. We show that the model trained on product review data outperforms competing models trained on general benchmark datasets. We then aggregate the support-level of all review sentences to a single support score, and a single contradiction score, representing the amount of support and contradiction the tip attracts from all product reviews. These scores can be used to rank the extracted tips, and can be exposed to customers, strengthening their confidence in recommended tips. In Figure 1 we present a potential product tips widget design which incorporates the support-level per tip, and links to concrete related reviews, in order to help customers estimate the tip's validity. Furthermore, we propose a new taxonomy for tip classification based on the tip's support-level, classifying tips to highly supported (high support, low opposition), highly contradicted (high opposition, low support), anecdotal (low support, low opposition), and controversial (high support, high opposition). To summarize, the main contributions of this work are:

*This work was done while Miriam was at Amazon.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531805

Figure 1: Product tips widget

- A learning framework that assesses the support-level that a given tip attains from a given review sentence, and support and contradiction scores based on the amount of support-level it receives across all product reviews.
- A novel tip taxonomy based on the tip's support and contradiction scores.
- **TipSu**[1] – a new manually annotated dataset with ~10K (tip, review sentence) pairs, labeled by support / contradiction / neutrality.

2 Related Work

Product tips extracted from customer reviews were introduced by Hirsch et al. [10]. Tips are extracted from Amazon reviews, considering each review sentence separately and labelling it as a tip or not. Abstractive tip generation [14, 15] is a complementary approach for tip extraction, where tips are automatically generated from accumulated user reviews.

Extracting supportive evidence for a claim, from external resources, attracted a lot of attention in the argument mining and fact checking literature. Rinott et al. [21] detected arguments in unstructured text (e.g. Wikipedia pages) for supporting claims raised during a debate contest. Habernal et al. [9] predicted the convincingness of such arguments related to a certain topic. Braunstain et al. [2] extracted passages from Wikipedia to support human answers for advice-seeking questions in CQA sites. Popat et al [20] measured the credibility of arbitrary claims by automatically finding supportive sources in news and social media. Karadzhov et al.[11] described a fully-automatic fact checking model using external sources to confirm or reject a claim. This model was applied for rumor detection and for fact checking of answers in CQA forums. Mihaylova et al. [19] checked the veracity of answers to particular CQA questions, proposing a multi-faceted model which captures information from the answer, the author, and from external authoritative sources of information. The SemEval-2019 task [18] focused on predicting whether an answer to a factual question is true, false or not a proper one. An answer is considered to be true if it can be proven with an external resource. The detection of contradiction in opinionated texts (such as reviews) was studied in some works [1, 13, 24], where contradiction is usually detected by the existence of discrepancy in opinions around the same aspect or topic.

[1]The dataset is freely available at https://registry.opendata.aws under the name TipSu

3 Useful Tips – Support-Level Detection

In this work we borrow the notion of product tips as introduced in [10]. We follow their approach in order to extract tips from reviews, create a database of labeled tips, and train a RoBERTa-based classifier that labels each candidate sentence as a tip or not. For more details we refer the reader to [10].

Given a product associated with a list of customer reviews, and a tip, our goal is to measure the amount of support and contradiction the tip receives from all reviews. Table 1 provides examples of supporting and contradicting (tip, review sentence) pairs.

A popular approach for support-level detection [3, 23, 26] is training a specialised ML model for this task based on annotated data. The model classifies a pair of sentences into one of three classes: support / contradiction / neutrality. As baselines, we trained transformer-based classification models on two benchmark datasets[2]. The first model is trained on the SNLI+MultiNLI datasets[3] that contain together about one million manually labelled sentence pairs. The second one is derived from the first one via 5-epoch fine-tuning over the SICK dataset [17] which contains about 10K sentence pairs. This fine-tuning increased the model accuracy to 90.8% from 88.8%. The trained models are denoted by m_{NLI} and m_{SICK} respectively. For an in depth comparison between these two datasets, we refer the reader to [22].

The models' ability to detect a tip's support-level was evaluated by randomly selecting 724 popular products and extracting 9195 tips from these products' reviews[4]. Then, for each tip and for each of the support-level detection models (m_{NLI} and m_{SICK}), the following process was performed:

- We applied the Universal Sentence Encoder (USE) model[5] to create an embedding vector for each tip and for all customer review sentences that belong to the tip's product. Using cosine similarity between these embedding vectors, 500 review sentences that obtained the highest similarity score were extracted for each tip. This step was done in order to ensure that there is some topical relationship between the tip and the review sentences, and in order to expedite the next steps, which rely on heavier models.
- The support-level detection model was then applied for each (tip, review sentence) pair. The review sentence with the highest contradiction score, and the sentence with the highest support score, were sent to manual annotation[6].
- Annotators were asked to label the support-level of each pair by one of three labels: (a) support, (b) contradiction, or (c) neutrality. As the annotation task is non-trivial and subjective, we collected five annotations per pair. Only pairs labeled the same by at least 3 annotators were assigned the agreed label (equivalent to majority voting). Any other pair was assigned as neutral. In order to assess the quality of this majority voting (Fleiss Kappa = 0.38), we compared it to annotations performed by in-house experts. We found agreement in 62% of the cases, in only 5% there was

[2]The classification models were implemented using https://huggingface.co/cross-encoder/nli-roberta-base, and are based on the pre-trained RoBERTa model [16]. We set the batch size to 64 and we use the Adam optimizer with a learning rate of $2e-5$.
[3]https://nlp.stanford.edu/projects/snli/, https://cims.nyu.edu/~sbowman/multinli/
[4]To extract the tips we used Bert-based model, following the best performing model described in section 4.2 in [8].
[5]https://tfhub.dev/google/universal-sentence-encoder/4
[6]We used the Appen crowdsourcing service https://appen.com/ for manual annotation.

Table 1: Examples of supporting and contradicting (tip, review sentence) pairs.

product	tip	review sentence	relationships
	Don't fully tighten most bolts until it's in place and even.	They were right about 'not tightening it all up' until the end basically.	support
	Just make sure you have several inches on either side of the doorway to accommodate the handles.	You will need about six inches of clearance on each side of the door for it to fit.	support
	This is a great shirt, but you should order a size up.	The shirt is true to size.	contradiction
	The seat fabric is textured and rough, so sitting on the chairs for extended periods of time while wearing shorts can cause mild irritation.	I can sit all day with no issues.	contradiction

flipped label between support and contradiction, and in the rest of the cases disagreement between a natural label to support or contradiction.

The precision of m_{SICK} was $Prec@1 = 51.5\%$ for support detection and 27% for contradiction detection. m_{NLI}'s performance was worse, with $Prec@1 = 48.7\%$ for support detection and 23.8% for contradiction detection. Note that since not all tips have supporting or contradicting sentence in the customer review corpus, we do not have an upper bound on $Prec@1$. Still, for the purpose of comparison between models' performance, this metric is viable. In addition, based on the $Prec@1$ scores, we can see that at least half of the tips have at least one supporting review sentence, and at least quarter have at least one contradicting review sentence.

Following the annotation task we now have manually labeled data of (tip, sentence) pairs classified each as supporting/ contradicting/ neutral. From this data, we selected pairs with high annotator agreement (at least 4 out of 5 annotators agreed on a given class). We refer to this dataset as **TipSu** (**Tip Su**pport); it contains 9609 sentence pairs - 4481 supporting, 3468 neutral, and 1657 contradicting. Using TipSu we fine-tuned m_{SICK} (the better among the models) for 5 additional epochs, creating a model which we denote by m_{PROD}. The training process of m_{PROD} is the same as for m_{SICK}, in terms of batch size, learning rate, etc. To assess its performance and compare it with m_{SICK}, we repeated steps (1) and (2) from above on a set of 400 tips, belonging to 238 products, that do not belong to the training set of m_{PROD}, this time annotating the top-5 review sentences that got highest supporting/contradicting scores. Table 2 presents $Prec@K$ of the two models for $K = 1, .., 5$; for the support detection task (left) and contradiction detection task (right).

As can be seen from the table, m_{PROD} greatly outperforms m_{SICK} for the two tasks at all K levels. We can also see that the precision of contradiction detection is lower than the precision of support detection. This could stem from either contradiction examples are being more rare, or that contradiction detection might be just more difficult task in general (see [4] for more insight on the reasons for this difficulty). We believe this probably stems from the combination of the two.

Table 2: $Prec@k$ of the models m_{SICK} and m_{PROD}. Left: the support detection task. Right: the contradiction detection task.

	support		contradiction	
k	m_{SICK}	m_{PROD}	m_{SICK}	m_{PROD}
1	49.7	56.6	30.2	36.9
2	48.3	56.6	28.4	34.6
3	46.2	55.3	27.4	33.9
4	45.1	54.6	26.8	32.6
5	44.6	54.2	25.7	32.2

4 Tip support and contradiction scores

We now take sentence-based support-level analysis one step further, assessing the support level that a tip attains from all product reviews. Let T_p be a given tip for a product p, and $Sent(p)$ be the set of all review sentences of p. We introduce $Support(T_p)$ and $Contradict(T_p)$ scores for a tip T_p by averaging the support/ contradiction scores over all review sentences:

$$Support(T_p) = \frac{1}{|Sent(p)|} \sum_{s \in Sent(p)} m_{PROD}(T_p, s)_{.sup}$$

$$Contradict(T_p) = \frac{1}{|Sent(p)|} \sum_{s \in Sent(p)} m_{PROD}(T_p, s)_{.contr}$$

$|Sent(p)|$ is the number of review sentences for product p, while $m_{PROD}(T_p, s)_{.sup}$ and $m_{PROD}(T_p, s)_{.contr}$ are the support and contradiction scores obtained via applying the model m_{PROD} on the pair (T_p, s). In practice, since $Sent(p)$ might be very large for popular products, it may not be feasible to perform the calculations involved on a large scale set of reviews. Thus, we restrict $Sent(p)$ to include only the top 500 review sentences that achieve the highest USE-similarity score with T_p. Such sentences are also more likely to be related to the tip, and hence to support or oppose it. Table 3 presents some tips with very high support score (top 1 percentile across support scores), as well as tip with very high contradiction score (top 1 percentile across contradiction scores).

4.1 Linguistic style differences between supported and contradicted tips

Looking at the few examples in Table 3, we can see that tips with high contradiction score are subjective in nature, while highly supported tips may cover common topics like heating related warnings

Table 3: Tips that belong to the top percentile of support scores (left) or contradiction scores (right)

Tips with very high support score		Tips with very high contradiction score	
category	tip	category	tip
home	I definitely recommend putting it in the dryer for 10 minutes after opening the sealed bag to fluff up the pillow.	apparel	Good Weight, should purchase one size smaller
kitchen	It heats up quickly so don't leave it alone for long.	beauty	Love, but DO NOT use these brushes while blow drying your hair.
wireless	It charges very quick, but make sure you use the correct quick charge adapter 3.0.	kitchen	Doesn't keep things cold though so need to add ice packs
pc	Recommend using provided driver installation as Windows used a generic driver but that seemed to work fine as well.	sports	I'd suggest buying a size smaller bc they run a little big.

for kitchen products or charging for wireless products. To make a more quantitative analysis, we investigated the textual discrepancy between supported and contradicted tips by training two language models; one for tips that belong to the top 5 percentile via support score and another for tips that belong to the top 5 percentile via the contradiction score (each one of the two sets contained 2400 tips). Table 4 displays the most prominent terms (n-grams) in supported and contradicted tips. These terms are those who contribute the most to the KL divergence between the two language models. We can see that the most prominent terms in the supported tips reflect common imperative provisions ('make sure', 'be careful'). However, most prominent terms in contradicted tips deal with product size, probably due to the high sensitivity and controversy surrounding apparel-related tips (See Section 4.2 on the diversity of tip support-level patterns across different product categories).

Table 4: Examples of frequent terms (n-grams) in supportive/ contradicting sentences.

Support:		
'adjust'	'measure'	'check'
'be careful'	'clean the'	'be sure'
'easy to'	'works great'	'very easy'
'plug it'	'remove the'	'be careful when'
'make'	'make sure'	'just be sure'
Contradict:		
'size'	'order'	'larger
'size down'	'size up'	'size smaller'
'your normal'	'I recommend ordering'	'I recommend going'
'one size'	'same size'	'suggest'
'smaller'	'size larger'	'I would recommend'

4.2 Support-level Taxonomy

We next inspect the distribution of Support and Contradict scores over tips extracted from different product categories, as depicted in Figure 2. We can see that there is a variety in tip support-level among categories; the apparel category especially stands out as it has both high support and high contradiction scores on average, while PC has both low support and low contradiction scores.

In order to shed some light on possible explanation to this phenomenon, we looked into the average tip length. Short tips may be more simple and easier to support/contradict, while long tips may be more specialized, hence discussing more obscure topics which have less support and contradiction. Overall, the average tip length is 19.9 words. Looking at the average amount of words in tips per category, we observed that apparel are the shortest tips on average (17 words per tip), while PC are the longest (20.7 words per tip). PC tips also tend to be rather specific, e.g *"Run the Adrenalin software package to unpack the drivers, then cancel the installation."*. These insights could explain, at least partially, the discrepancy in the level of support between these two categories, and the overall high support and contradiction scores for apparel tips.

To get more insight regarding the interaction between support and contradiction scores, we split each score range into low (bottom 20% percents), high (top 20% percents) and medium (middle 60% percents). We create a 3-by-3 heat-map, splitting the tips according to the cells they belong to (support-score percentile represented by the y-coordinate, while contradiction score percentiles represented by the x-coordinate). Figure 3 depicts the heat-map of the entire data (left), as well as the apparel category (right). Calculating the Pearson correlation between support and contradiction scores across 50K tips shows a (surprising) positive correlation of 0.34. At first, this could seem unlikely, however, it is easy to show that the sum of support, contradiction and neutrality scores per tip is 1[7]. Hence, controversial tips (high support and high contradiction scores) can be characterized by a low neutrality score. Moreover, support and contradiction scores depend heavily on the tip's topic. Obscure topics will tend to have both low support and low contradiction scores as they do not attract the attention of many customers. On the other hand, hotly debated topics (e.g. cloth size) are expected to attract both high support and high contradiction, as they are essential to a large number of customers.

Motivated by Figure 3, we come-up with the following tip taxonomy:

- **Highly supported** – Tips with many supporting and almost no contradicting sentences. 0.6% of the tips belong to this category.

[7]Note that $\forall p, \forall s \in Sent(p), m_{PROD}(T_p, s)_{.sup} + m_{PROD}(T_p, s)_{.contr} + m_{PROD}(T_p, s)_{.neut} = 1$.

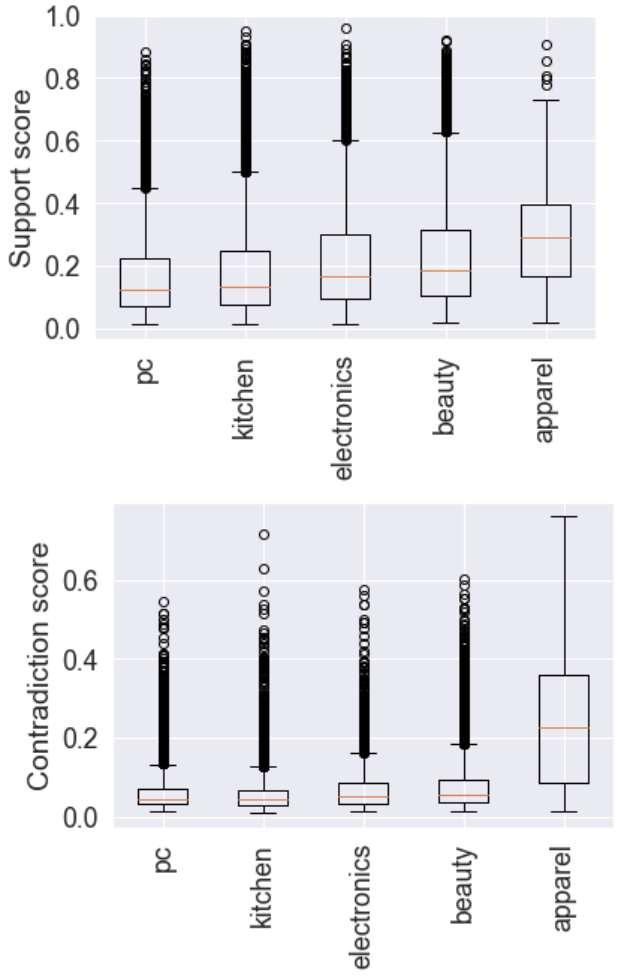

Figure 2: Distribution of support (top) and contradiction (bottom) scores across product categories

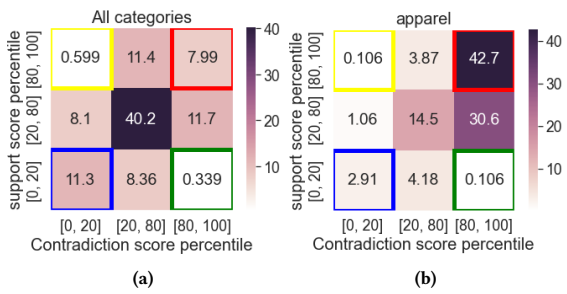

Figure 3: Heat-maps of tips support-level. (a) whole data; (b) Apparel product category. Color map: Yellow- highly supported; Blue- anecdotal; Red- controversial; Green- highly contradicted.

- **Highly contradicted** – Tips with many contradicting and almost no supporting sentences. 0.34% of the tips belong to this category.
- **Controversial** – Tips with many supporting and many contradicting sentences. 8% of the tips belong to this category.
- **Anecdotal** – Tips with almost no support and no contradiction sentences. 11% of the tips belong to this category.

As can be seen in Figure 3, controversial tips are very common in apparel, where 43% of the tips are of this type. They are often sizing-related, e.g *"Order a size bigger than what you would normally wear."*. When considering anecdotal tips, many of them are quite specific, e.g: *"If you're serious about lifting weights you should have a set of 35s to gently add weight in 10 lb increments rather than jumping from more common 25s to 45s on the typical lifts."*. We suspect that in many cases, highly supported tips are quite obvious, while anecdotal tips, even without support, might be of interest to customers since they are less obvious and contain non-trivial information.

5 Conclusions

With the motivation to increase the customers' confidence in product tips, we proposed a framework for determining the support and opposition a tip receives from other product reviews. We proposed a taxonomy that identifies four types of tips: Highly supported, highly contradicted, controversial, and anecdotal. We also publish a new manually annotated dataset with 9609 (tip, review sentence) pairs, each associated with a manually judged support-level label.

In the close future we intend to experiment with various ways of exposing customers to extracted tips together with their support-level scores. In addition, we plan to explore how the support-level framework we presented may be applied for other types of product related claims. As examples, a claim about the product's functionality, or a usage-related advice, expressed by the seller (in the product description) or by another customer (in the product reviews or Q&As).

References

[1] Ismail Badache, Sébastien Fournier, and Adrian-Gabriel Chifu. 2018. Predicting Contradiction Intensity: Low, Strong or Very Strong?. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1125–1128.

[2] Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in CQA sites. In *European Conference on Information Retrieval*. Springer, 129–141.

[3] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *NeurIPS*.

[4] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In *Proceedings of ACL-08: HLT*. 1039–1047.

[5] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. Do online reviews matter?—An empirical investigation of panel data. *Decision support systems* 45, 4 (2008), 1007–1016.

[6] Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying Helpful Sentences in Product Reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 678–691.

[7] Dmitri Goldenberg, Gal Kampel, and Yehoshua Shuki Cohen. 2016. Needle in a Haystack: Tips Extraction from YELP Reviews. (2016).

[8] Ido Guy, Avihai Mejer, Alexander Nus, and Fiana Raiber. 2017. Extracting and Ranking Travel Tips from User-Generated Reviews. In *Proceedings of the 26th International Conference on World Wide Web (WWW)* (Perth, Australia) *(WWW '17)*. 987–996.

[9] Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1589–1599.

[10] Sharon Hirsch, Slava Novgorodov, Ido Guy, and Alexander Nus. 2021. Generating Tips from Product Reviews. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*. 310–318.

[11] Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully Automated Fact Checking Using External Sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. 344–353.

[12] Shivendra Kumar and C Ravindranath Chowdary. 2020. Semantic model to extract tips from hotel reviews. *Electronic Commerce Research* (2020), 1–19.

[13] Chuqin Li, Xi Niu, Ahmad Al-Doulat, and Noseong Park. 2018. A computational approach to finding contradictions in user opinionated text. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 351–356.

[14] Piji Li, Zihao Wang, Lidong Bing, and Wai Lam. 2019. Persona-Aware Tips Generation?. In *Proceedings of the 28th International Conference on World Wide Web (WWW)*. 1006–1016.

[15] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. 345–354.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[17] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 216–223.

[18] Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 Task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 860–869.

[19] Tsvetomila Mihaylova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[20] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2016. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM international on conference on information and knowledge management (CIKM)*. 2173–2178.

[21] Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP)*. 440–450.

[22] Aarne Talman and Stergios Chatzikyriakidis. 2018. Testing the generalization power of neural network models across NLI benchmarks. *arXiv preprint arXiv:1810.09774* (2018). https://arxiv.org/abs/1810.09774

[23] Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. 2018. Multiway attention networks for modeling sentence pairs. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. 4411–4417.

[24] Mikalai Tsytsarau, Themis Palpanas, and Kerstin Denecke. 2011. Scalable detection of sentiment-based contradictions. *DiversiWeb, WWW* 1 (2011), 9–16.

[25] Ingmar Weber, Antti Ukkonen, and Aris Gionis. 2012. Answers, not links: extracting tips from yahoo! answers to address how-to web queries. In *Proceedings of the fifth ACM international conference on Web search and data mining (WSDM)*. 613–622.

[26] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (NAACL-HLT)*. Association for Computational Linguistics, 1112–1122.

[27] Di Zhu, Theodoros Lappas, and Juheng Zhang. 2018. Unsupervised tip-mining from customer reviews. *Decision Support Systems* 107 (2018), 116–124.