# Did the Writer Actually Visit the Location? Analysis of Location Reviews from Visit Experience

**Aitaro Yamamoto**♣,* **Hiroki Ouchi**♣
**Kota Tsubouchi**♠,†, **Tatsuo Yamashita**♠,‡
**Ryo Tsujimoto**♣,§ **Yuki Matsuda**♦,♣ **Hirohiko Suwa**♣
♣ NAIST ♠ LY Corporation ♦ Okayama University
yamamoto.aitaro.xv6@is.naist.jp, hiroki.ouchi@is.naist.jp,
ktsubouc@lycorp.co.jp, tayamash@lycorp.co.jp,
tsujimoto.ryo.tq0@is.naist.jp, yukimat@okayama-u.ac.jp, h-suwa@is.naist.jp

## Abstract

We investigate the characteristics of location review texts written on the basis of actual visit experiences or without any visit experiences. Specifically, we formalize this as a binary classification task and propose a data construction framework that labels reviews as `Visit` or `NotVisit` by linking them with users' GPS-based movement data. We train a logistic regression model on the dataset and evaluate it alongside human annotators and a large language model (LLM). The results show that the task is more challenging for humans and LLMs than for the simple trained model.

## 1 Introduction

Online platforms such as YELP[1] and TRIPADVISOR[2] allow users to post and share reviews of various locations, which play a crucial role in decision-making (Duan et al., 2008; Zhu and Zhang, 2010; Cheung and Thadani, 2012; Bing et al., 2016; Ocampo Diaz and Ng, 2018). These reviews have been widely studied in tasks such as helpfulness prediction (Kim et al., 2006; Chen et al., 2018; Liu et al., 2021; Chen et al., 2022), sentiment analysis, and utility scoring, using diverse textual features (e.g., TF-IDF, length, POS tags) (Liu et al., 2007; Tsur and Rappoport, 2009; Yang et al., 2015) and metadata (e.g., ratings (Einar Bjering and Moen, 2015), images (Nguyen et al., 2022), user demographics (Pezenka and Weismayer, 2020)).

Among various types of reviews, location-based reviews, such as those for restaurants or tourist spots, or hotels, play a particularly important role in guiding users' real-world decisions. A common assumption in previous work involving such reviews is that *the reviewer has actually visited the location they write about*. However, this assumption does not always hold: some reviews are fake (Liu et al., 2010; Luca and Zervas, 2016), or inaccurate due to memory decay. Although Bu (Bu et al., 2021) attempted to filter unreliable reviews via sentiment-rating mismatches, no prior work explicitly examines whether a review genuinely reflects an actual visit experience.

In this paper, we propose a new task, *Visit Experience Judgement*, which determines whether a review was based on a real-world visit. Figure 1 shows the task setting: given a review text, a model predicts whether the writer actually visited the reviewed location. To support this task, we propose a data construction framework that links review texts with GPS-based user movement data and label them as `Visit` or `NotVisit` (Section 4). To our knowledge, this is the first work to connect textual reviews with real-world user behavior.

Our research contributions are threefold: (1) formalizing a new task, (2) proposing a data construction framework, and (3) evaluating models on the task (Section 5). Our results show that fine-tuned models outperform humans in this task, and lexical analysis highlights key predictive features.

## 2 Related Work

Reviews on online platforms have been the subject of many studies, as they can provide information on a wide range of user preferences and general characteristics of reviews. Review texts have various aspects, such as length (Liu et al., 2007; Yue Lu and Polanyi, 2010), word-based features (such as TFIDF (Kim et al., 2006; Tsur and Rappoport, 2009)), and word-category features (such as part-of-speech (Yang et al., 2015; Zhang and Varadarajan, 2006)). The target of research is not limited to the texts themselves; any information related to a review may be useful. For example, im-

*This work was done while he belonged to NAIST.
†Corresponding author.
‡This work was done while he belonged to LY Corporation.
§This work was done while he belonged to NAIST.
[1]https://www.yelp.com/
[2]https://www.tripadvisor.jp/



*Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1331–1337
December 20-24, 2025 

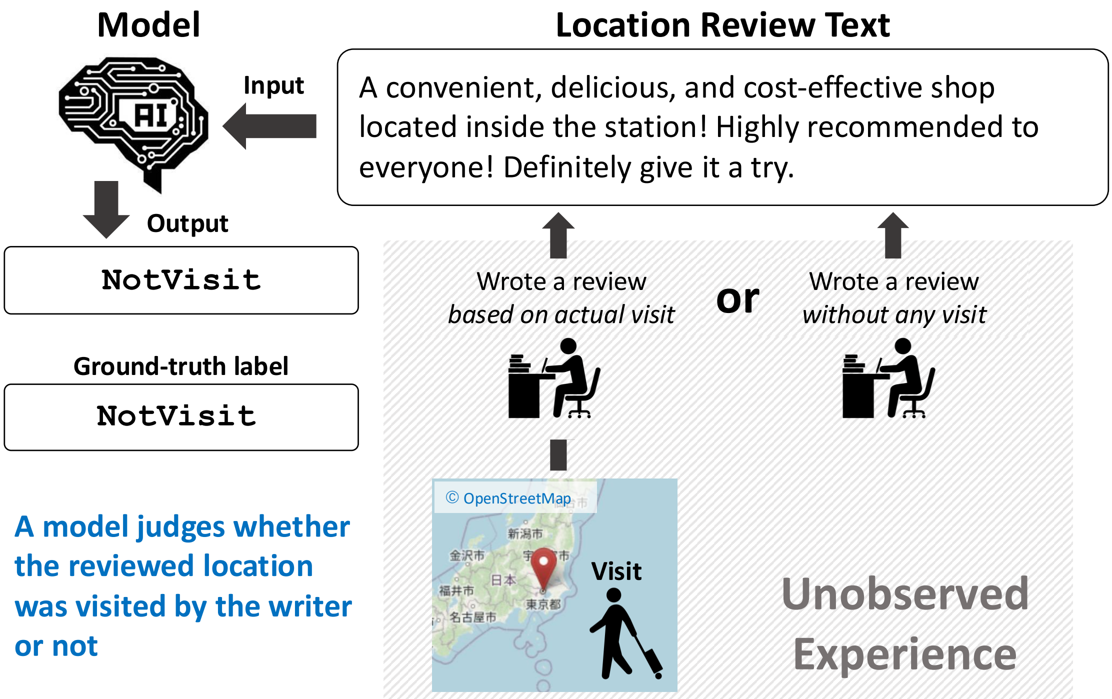


Figure 1: Overview of our proposed task, *Visit Experience Judgement*. Here, a model seeks to judge whether the reviewed location was visited by the writer or not. Each input text reviews a certain location. Some of the texts were written by the writers based on the actual visit experiences, and others were written without any visits. The model has to distinguish them from only the textual information.

ages (Nguyen et al., 2022), ratings (Einar Bjering and Moen, 2015), user's hometown (Pezenka and Weismayer, 2020), have been studied.

There is an implicit assumption that reviews are posted by users who have actually visited the location. However, it should be noted that, in reality, it is not uncommon to find false or inaccurate reviews. For these problems, Bu (Bu et al., 2021) focused on the discrepancy between sentiment and rating by conducting aspect sentiment analysis in order to exclude unreliable reviews. However, to the best of our knowledge, there are no studies that focus on whether review posters actually visited the location they are reviewing.

## 3 Task

The proposed task, *Visit Experience Judgement*, requires models (or humans) to predict whether a location review text was written on the basis of an actual visit experience or not.

The task is formalized as a binary classification problem: given a location review of $n$ tokens, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, the goal is to predict whether the writer actually visited the location or not. The probability of the actual visit is defined as follows:

$$P(y = 1|\boldsymbol{x}) = \sigma(f_\theta(\boldsymbol{x})) \quad (1)$$

where $y \in \{0, 1\}$ represents visit information; i.e., $y = 1$ is an actual visit and $y = 0$ is not. $f_\theta$ is a model (scoring function) with its parameters $\theta$ that returns a real value, and $\sigma$ is a sigmoid function.

The model parameters $\boldsymbol{\theta}$ are trained by minimizing the binary cross-entropy loss:

$$\ell(\theta) = -\log P(y = 1|\boldsymbol{x}) + \log(1 - P(y = 1|\boldsymbol{x}))$$

We explain the model $f_\theta$ that we used in more detail in Section 5.4.

## 4 Data Construction Framework

In this section, we introduce our framework for constructing a review dataset with visit experience.

### 4.1 Flow of Data Construction

Figure 2 illustrates the flow of our data construction. The goal is to determine a ground-truth label, `Visit` or `NotVisit`, for each review. Specifically, the flow is as follows:

1. Extract each review from the review database.
2. Use the reviewed location ID of the target review as a query for searching the map database and obtain its coordinates (i.e., latitude and longitude).

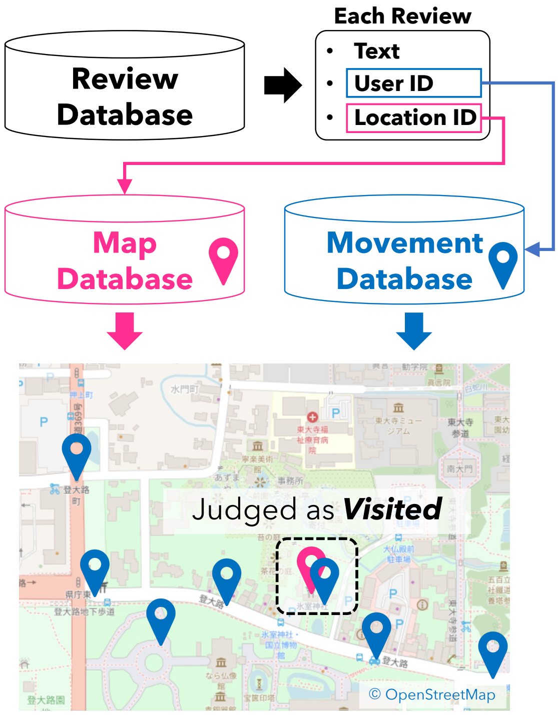


Figure 2: Flow of our dataset construction.

3. Use the user ID of the review as a query for searching the movement database and obtain a set of the coordinates where the user stayed.
4. Determine the label, `Visit` or `NotVisit`, for the review[3]:
    - If the reviewed location point is close to any one of the movement points, we judge it as `Visit`.
    - Otherwise, `NotVisit`.

The constructed labeled dataset is used for training models (Equation 1). In real-world situations, trained models can be used for arbitrary unseen and unlabeled location review texts.

### 4.2 Databases

As we saw, our framework assumes to use (i) a review database $\mathcal{D}^{\text{review}}$, (ii) a map database $\mathcal{D}^{\text{map}}$, (iii) a movement database $\mathcal{D}^{\text{map}}$. In this subsection, we explain them in more detail.

- Review database $\mathcal{D}^{\text{review}}$ consists of $N$ reviews: $\mathcal{D}^{\text{review}} = \{r_i\}_{i=1}^{N}$. Each review $r$ has the following two types of information:
    - User ID $r^{\text{user\_id}}$,
    - Location ID $r^{\text{loc\_id}}$.
- Map database $\mathcal{D}^{\text{map}}$ stores map coordinates (i.e., latitude and longitude) of points of interest (e.g., cities, shops, and temples/shrines). By searching the database with a location ID, we can obtain its latitude $g$ and longitude $h$: $\langle g, h\rangle = \mathcal{D}^{\text{map}}(r^{\text{loc\_id}})$.
- Movement database $\mathcal{D}^{\text{move}}$ stores spatial-temporal information on where and when a user stayed. By searching the database with a user ID, we can obtain a set of points: $\boldsymbol{p} = \mathcal{D}^{\text{move}}(r^{\text{user}})$, where $\boldsymbol{p} = (p_1, p_2, \dots)$. Each point is a triple, $p_k = \langle g, h, t\rangle$, where $g$ is latitude, $h$ is longitude, $t$ is time (date).

The use of the map and movement databases allows us to associate review texts with visit experiences.

## 5 Experiments

### 5.1 Research Questions

We address the following questions:

**RQ1** How difficult is the task for humans?

**RQ2** How accurately can machine learning models judge visit experiences?

For RQ1, we asked humans to judge whether each review is written on the basis of the writer's visit experience or not.[4] Through the comparison between the accuracy of humans and machine learning models, we demonstrated the difficulty level of the task.

For RQ2, we investigated two types of models: (i) a Logistic Regression model and (ii) a large language model (LLM) (Section 5.4 in more detail). The logistic regression model is based on word-frequency-based features, so it is much easier to reveal how important each word is for the prediction. Note that our aim is NOT to achieve higher prediction accuracy with more sophisticated models, which is left for the future research.

[3] Our labels are operationally defined by GPS proximity and should be understood as proxy ground-truth rather than perfect truth. Also, to further reduce noise, we only included users with a sufficient volume of mobility logs and applied trajectory reconstruction (interpolating GPS points into continuous paths). Specifically, to reduce GPS noise, we reconstructed user trajectories following standard smoothing methods: consecutive GPS points were connected and implausible points were removed. For each three consecutive points ($a \rightarrow b \rightarrow c$), we measured the angle between $ab$ and $bc$; if the angle exceeded 60°, the point was removed as unrealistic. This rule filters out sudden zigzag jumps caused by GPS drift or signal loss, not normal turns such as right-angle movements at intersections. These steps make the dataset more reliable while keeping realistic movement patterns intact.

[4] The annotation was conducted by the authors themselves.

| Label | Model | P | R | F1 |
|---|---|---|---|---|
| Visit | Human | 0.57 | 0.83 | 0.66 |
| | LogReg | **0.71** | 0.74 | **0.73** |
| | Llama3 | 0.50 | **0.95** | 0.65 |
| NotVisit | Human | **0.79** | 0.33 | 0.39 |
| | LogReg | 0.73 | **0.70** | **0.72** |
| | Llama3 | 0.54 | 0.05 | 0.09 |

Table 1: Performance for each class. Best values in bold. "LogReg" stands for Logistic Regression.

| | Visit |
|---|---|
| Nouns | staff, customer service, lunch, park |
| Verbs | enter, buy, sell, put, give |
| Adjectives | delicious, near, bright, cold, hard to do |
| Adverbs | a little, not much, soon, always |
| | NotVisit |
| Nouns | hot spring, sightseeing, trip, scenery |
| Verbs | enjoy, go, visit, become, stop |
| Adjectives | excellent, good, wide, easy, difficult |
| Adverbs | very, by all means, variously, always |

Table 2: Top contributing words for each label.

### 5.2 Data Construction and Filtering

As described in Section 4, labels were assigned based on proximity between the reviewed location and user movement points. Due to data sparsity, we filtered for users with sufficient GPS records by applying daily and monthly thresholds, and used trajectories to better capture actual visits.

### 5.3 Dataset

For the review database, we collected over 500,000 Japanese review texts posted by over 60,000 users on Yahoo!Loco in January 2023. For the map database, we collected over 10 million locations and facilities registered in YAHOO!LOCO. For the movement database, we collected over 5 million location points in January 2023 from various services provided by YAHOO!. After labeling, we obtained 45,943 `Visit` and 3,498 `NotVisit` reviews. To balance the dataset, we sampled 3,498 from each class, resulting in 6,996 reviews. We used 10-fold cross-validation with an 8:1:1 train/valid/test split.

### 5.4 Model Details

As a model $f_\theta$ in Eq. 1, we used logistic regression model, which takes as input a feature vector of each text. As the vector, we created a feature vector using TF-IDF, which reflects the relative importance of words across the review texts. The details of the preprocessing of each text for creating the TF-IDF vectors are written in Appendix A. As our LLM, we used Llama-3-ELYZA-JP-8B (Hirakawa et al., 2024), approximately 8 billion parameters, in a 0-shot setting. Given a review, the model outputs a label without additional fine-tuning. The prompt is shown in Figure 3 in Appendix.

## 6 Results and Analysis

### 6.1 Results

Human participants showed a strong bias toward over-predicting `Visit`, resulting in low recall for `NotVisit` (0.33). This suggests that humans tend to recognize most of the review texts as `Visit`. As the example of Figure 1, although many texts with the label `NotVisit` do not mention visit experiences, they are likely to be misunderstood as `Visit` without careful reading.

Llama3 showed the same tendency as the humans. The model tends to generate `Visit` for most of the texts, resulting in very low recall for `NotVisit`, less than 0.1 recall. This means that Llama3 cannot grasp charastiristics of `NotVisit` texts with just a few examples.[5]

By contrast, the logistic regression model achieved the best results: 0.73 F1 for `Visit` and 0.72 for `NotVisit`. Nevertheless of the simplicity, the performance was much better than humans and Llama3. This suggests that if models are trained on enough numbers of training examples, they acquire ability to distinguish the texts with `Visit` and `NotVisit`.

### 6.2 Lexical Analysis

We analyzed the logistic regression model to identify important lexical cues (Table 2). Visited reviews included concrete nouns (e.g., "staff," "customer service"), experiential verbs (e.g., "enter," "buy"), and impression-related adjectives/adverbs (e.g., "delicious," "a little"). On the other hand, non-visited reviews were characterized by abstract expressions (e.g., "sightseeing," "can enjoy"), and emphatic adverbs (e.g., "very," "by all means"), suggesting second-hand descriptions. There findings suggest that `Visit` reviews reflect detailed,

[5] Even though the model was given 3-shot examples, the performance was not improved.

concrete personal experiences while `NotVisit` reviews are more general or descriptive, sometimes copied or paraphrased from external sources.

## 7 Conclusion

We introduced the task of *Visit Experience Judgement*, which aims to determine whether a location review was written based on an actual visit. To support this task, we proposed a data construction framework that links review texts with user movement data. Our experiments showed that the task is challenging for humans and LLMs alike, both tending to over-predict `Visit`. In contrast, a simple logistic regression model achieved strong performance (F1 > 0.7), demonstrating that concrete, experience-based vocabulary plays a key role in distinguishing visited reviews. In future work, we plan to refine our framework by incorporating verified visitation records, and explore fine-tuning LLMs for improved performance on this task.

## Limitations

**Language** In this paper, we used review texts written in Japanese. Therefore, our experiments are limited to the Japanese language. However, our proposed task and data construction framework are designed to be language-agnostic.

**Potential Misclassification of Not-Visited Reviews** In our data construction, the data classified as not-visited reviews might include some visited ones. In our experiments, we only targeted users with sufficient movement information and thus collected reviews that were likely to be in the not-visited one (Section 5.2). However, the collected data might not be perfect. Although definitive confirmation is not possible, the likelihood that the writers actually visited the location is considered low, as our data selection was limited to users with sufficient movement information.

**Lack of Fine-Tuning and Use of Advanced Models** In this paper, we did not fine-tune LLMs or explore state-of-the-art deep learning models. However, the main contribution of this work lies not in building sophisticated models, but in proposing a novel task and a data construction framework. We consider model optimization, such as fine-tuning and leveraging more advanced architectures, to be an important direction for future work.

**Optimization of Model Performance** We primarily used the default hyperparameter settings provided by each framework and conducted only a limited hyperparameter search due to time and computational constraints. Therefore, more systematic optimization may lead to further improvements in model performance.

## Ethical Considerations

**License of Used Resources** MECAB, a Japanese part-of-speech and morphological analyzer, is available under GPL (the GNU General Public License), LGPL(Lesser GNU General Public License), or BSD License. SCIKIT-LEARN is available under BSD license. Llama3-ELYZA is available under Meta Llama 3 Community License.[6]

**Privacy Policy of Movement Database** Our movement database complies with the privacy policy and has been properly anonymized and securely stored. Our research has also been approved by an ethics review board.

## Acknowledgments

We would like to thank the anonymous reviewers for their constructive comments.

## References

Lidong Bing, Tak-Lam Wong, and Wai Lam. 2016. Unsupervised extraction of popular product attributes from e-commerce web sites by considering customer reviews. *ACM Transactions on Internet Technology*, 16(2):1–17.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. Asap: A chinese review dataset towards aspect category sentiment analysis and rating prediction. In *North American Chapter of the Association for Computational Linguistics*.

Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Sheng Bao. 2018. 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)*, pages 602–607, New Orleans, Louisiana. Association for Computational Linguistics.

Zaiqian Chen, Daniel Verdi do Amarante, Jenna Donaldson, Yohan Jo, and Joonsuk Park. 2022. Argument mining for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8922,

[6]https://llama.meta.com/llama3/license/

Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Christy MK Cheung and Dimple R Thadani. 2012. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. *Decision support systems*, 54(1):461–470.

Wenjing Duan, Bin Gu, and Andrew B. Whinston. 2008. The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of Retailing*, 84(2):233–242.

Lars Jaakko Havro Einar Bjering and Oystein Moen. 2015. An empirical investigation of self-selection bias and factors influencing review helpfulness. *International Journal of Business and Management*.

Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Daisuke Oba, Sam Passaglia, and Akira Sasaki. 2024. elyza/llama-3-elyza-jp-8b.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 423–430, Sydney, Australia. Association for Computational Linguistics.

Bing Liu and 1 others. 2010. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666.

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 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 334–342, Prague, Czech Republic. Association for Computational Linguistics.

Junhao Liu, Zhen Hai, Min Yang, and Lidong Bing. 2021. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Thong Nguyen, Xiaobao Wu, Anh Tuan Luu, Zhen Hai, and Lidong Bing. 2022. Adaptive contrastive learning on multimodal transformer for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: A survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Ilona Pezenka and Christian Weismayer. 2020. Which factors influence locals' and visitors' overall restaurant evaluations? *International Journal of Contemporary Hospitality Management*, 32(9):2793–2812.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *In International AAAI Conference on Web and Social Media*.

Yinfei Yang, Yaowei Yan, Minghui Qiu, and Forrest Bao. 2015. Semantic analysis and helpfulness prediction of text for online product reviews. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 38–44, Beijing, China. Association for Computational Linguistics.

Alexandros Ntoulas Yue Lu, Panayiotis Tsaparas and Livia Polanyi. 2010. Exploiting social context for review quality prediction. *In Proceedings of the 19th International Conference on World Wide Web*.

Zhu Zhang and Balaji Varadarajan. 2006. Utility scoring of product reviews. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, page 51–57, New York, NY, USA. Association for Computing Machinery.

Feng Zhu and Xiaoquan Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*.

## A Preprocessing for TD-IDF feature vectors

In creating the word frequency and TF-IDF features, the texts are morphologically analyzed by MECAB[7]. We use their surface form, leaving only those words whose parts of speech are nouns, verbs, adjectives and adverbs. As the implementation, we use COUNTVECTORIZER[8] and TFIDFVECTORIZER[9] from SCIKIT-LEARN.

## B Prompt for Llama3

**Prompt Template for Llama3 with 3-shot learning**

You are an expert in analyzing reviews to determine if the reviewer has actually visited the place they are reviewing visited the place they are reviewing. Given the following reviews, determine whether the reviewer has actually visited the place (output "Visit") or not (output "NotVisit")

Review 1: { `Review1_Text` }
Output: { `Review1_Label` }

Review 2: { `Review2_Text` }
Output: { `Review2_Label` }

Review 3: { `Review3_Text` }
Output: { `Review3_Label` }

Review 4: { `Review4_Text` }
Output: { `Review4_Label` }

Review 5: { `Review5_Text` }
Output: { `Review5_Label` }

Review 6: { `Review6_Text` }
Output: { `Review6_Label` }

Here is the review: { `Input_Review_Text` }

Please respond with only "Visit" or "NotVisit".

Figure 3: Prompt template used for Llama3 with 3-shot examples. Note that, in the case of 3-shot learning, we give a model six examples, i.e., three positive examples (`Visit`) and three negative examples (`NotVisit`).

Figure 3 illustrates the prompt template used for Llama3 with 3-shot learning. In the case of 3-shot learning, we randomly sample three positive examples (`Visit`) and three negative examples (`NotVisit`) from the training set.

## C AI Assistant Use

We used an AI assistant for tasks such as correcting grammatical errors and improving phrasing during the writing process.

[7] https://taku910.github.io/mecab/
[8] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html