

Identifying Consumer Preferences from User-Generated Content (UGC) on Amazon.com

by Leveraging Machine Learning

Jikhan Jeong

(jikhan.jeong@wsu.edu)

Washington State University (WSU)

Jan 3-5, 2021 AEA

Motivation

- Online product reviews can be useful

- They allow inexperienced consumers to reduce search costs and uncertainty about the quality of products.

- Some reviewers on Amazon.com mention the usefulness of previous reviews


“We bought this model because of the exceptional Consumer Products review/ratings”

“After reading some of the negative reviews, I was hesitant to purchase these units.”


Motivation

- Limitations of one-side review systems (e.g., Amazon.com)

Create Review

 Nest Learning Thermostat, 2nd Generation, Works with Amazon Alexa

Overall rating

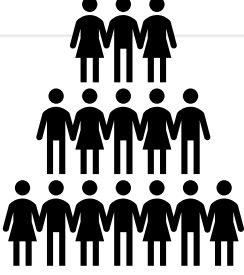


Add a headline

What's most important to know?

Write your review

What did you like or dislike? What did you use this product for?



One-side review system: only buyers can write reviews and information about reviewers is limited

→ **Asymmetric information** problems arise

- **Unobserved** consumer characteristics
- **Unobserved** product content dimensions
- **Unobserved** consumer sentiments

Research Questions

1. Can latent consumers' preferences be identified?
2. Can potential individual consumers' preferences be predicted?
3. Can consumers' sentiments be classified?

- Raw data: 141 million Amazon reviews (R. He, J. McAuley 2016)
- Target Products: Home Energy Control Devices



Nest Learning Thermostat, 2nd Generation, Works with Amazon Alexa

Brand: NEST ABOVE

★★★★☆ 4,492 ratings

Price: \$294.00 + \$5.37 shipping

Get \$150 off instantly: Pay \$144.00 upon approval for the Amazon Prime Rewards Visa Card.

Not eligible for Amazon Prime. Available with free Prime shipping from other sellers on Amazon.

WORKS WITH ALEXA
Add voice control by combining with an Alexa device



This item Required

Total Price: \$333.99

[Add both to Cart](#)

[^ See Less](#)

- Works with Alexa for voice control (Alexa device sold separately).
- Nest saves energy by automatically turning itself down when you're away
- 2nd generation design - nest is now 20-percent thinner and works in 95-percent of homes with low Voltage systems
- Auto-Away: Nest saves energy by automatically turning itself

- Experience goods
 - Programmable Thermostats (PTs)
- A new firm entering the market
 - The Nest

- This study focuses on consumers who write reviews on the review system

Total Consumer Space, $S_t = S_a \cup S_{na}$

Not Using Amazon, $S_{na} = S_t \cap S_a^c$

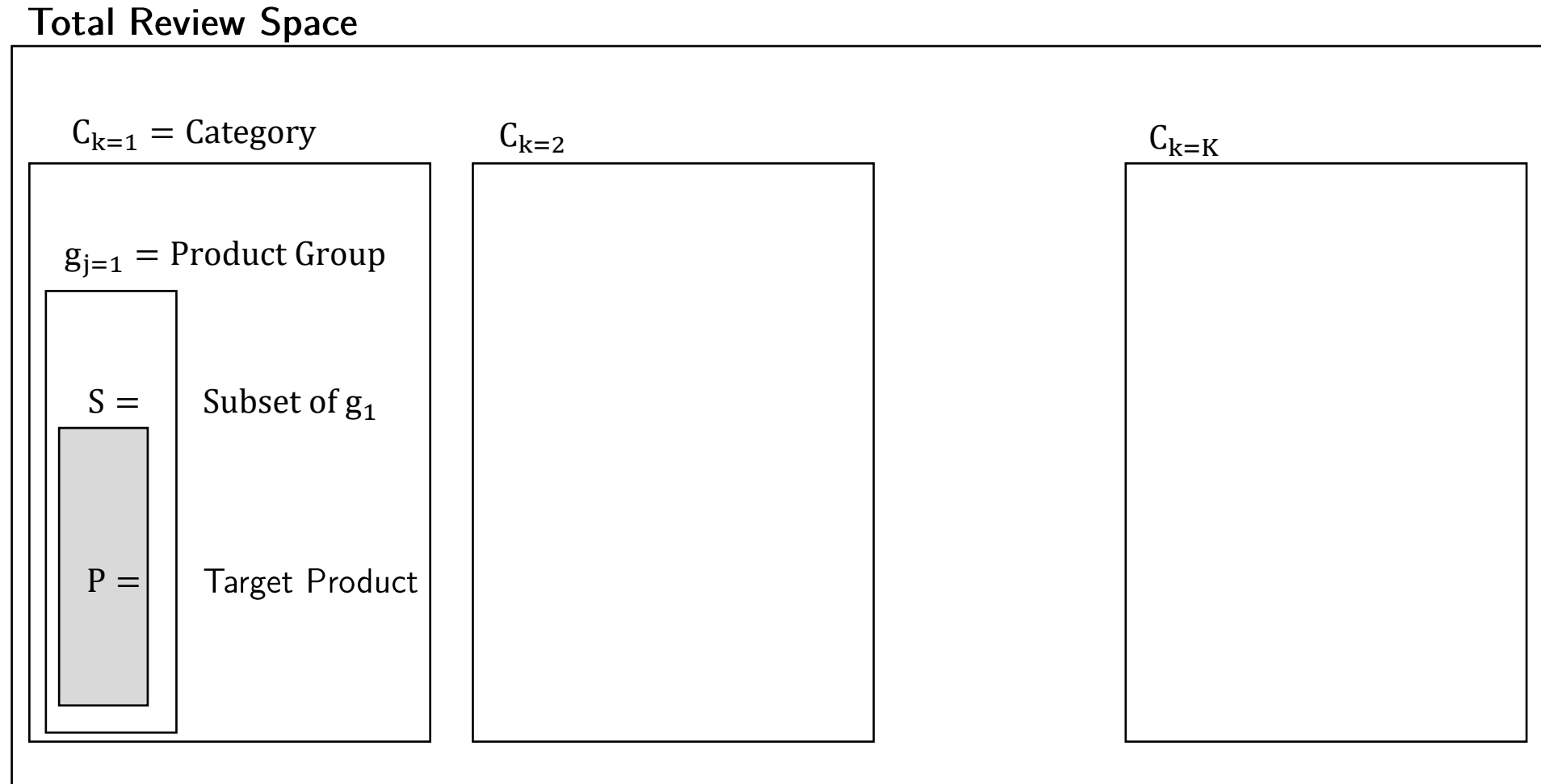
Using Amazon, $S_a = S_{aw} \cup S_{anw}$

Writing Reviews, S_{aw}

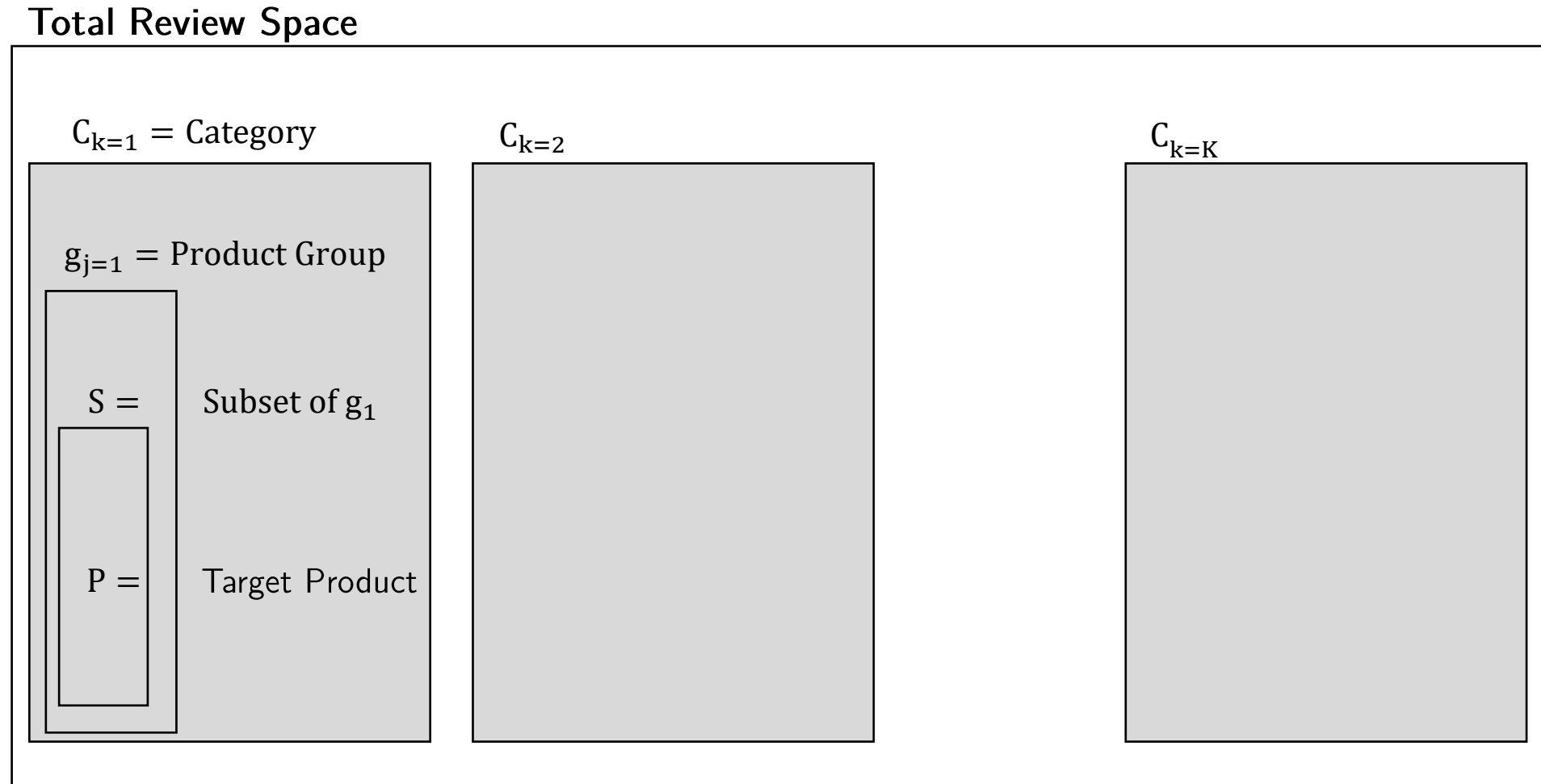
Biased Reviewer, S_{awb}

$S_{anw} = S_a \cap S_{aw}^c$

- This study focuses on the target consumers' reviews for **the target products**



- This study also considers the target consumers' prior reviews across all categories.

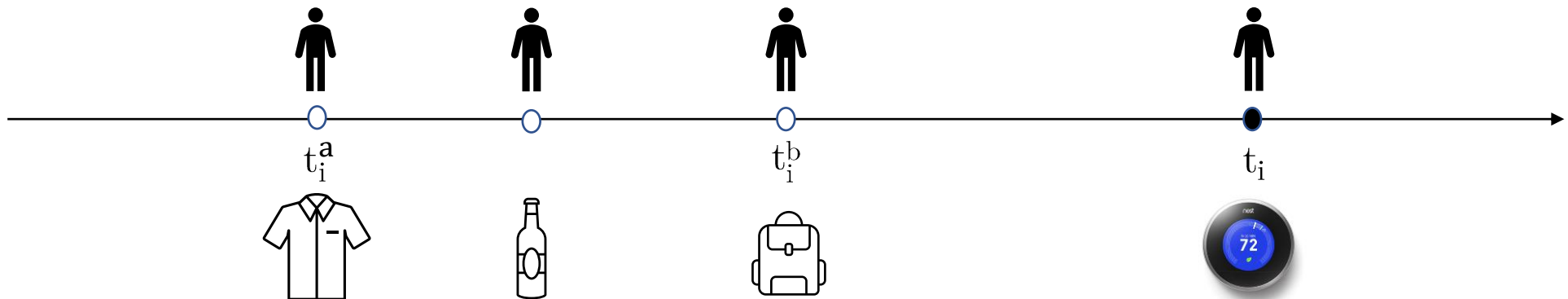


- What are digital footprints (DFs)?

- **User DFs:** reviewer i 's DFs before writing a review of thermostat p on day t_i .

$$\sum_{t_i^a}^{t_i^b} df_{ipt_i}(\cdot), \text{ where } t_i^a = \operatorname{argmax}_{t_i^a} |t_i - t_i^a| \text{ and } t_i > t_i^a$$

$$t_i^b = \operatorname{argmin}_{t_i^b} |t_i - t_i^b| \text{ and } t_i > t_i^b \geq t_i^a$$



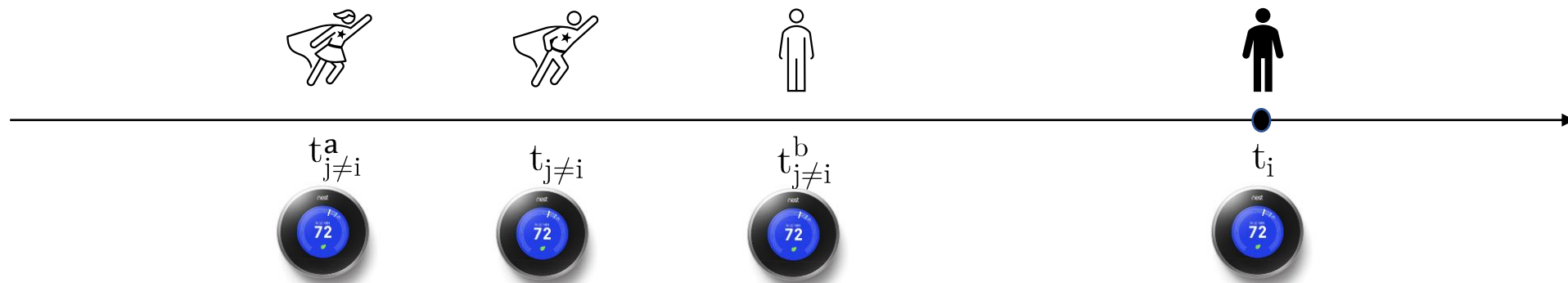
- What are DFs?

- **Crowd DFs:** other prior reviewers' DFs for thermostat p before i writes a review of p on day t_i .

$$\sum_{j \neq i}^J \sum_{t_j^a}^{t_j^b} df_{j|pt_j}(\cdot), \text{ where } \{\forall J \in \mathbb{R} \text{ and } 1 \leq j \leq J < \infty \mid i, t_i, p\}$$

$$t_j^a = \operatorname{argmax}_{t_j^a} |t_i - t_j^a| \text{ and } t_i > t_j^a$$

$$t_j^b = \operatorname{argmin}_{t_j^b} |t_i - t_j^b| \text{ and } t_i > t_j^b \geq t_j^a$$



Research Design

Step. 1: Data cleaning

Step. 2: Topic modeling and annotation

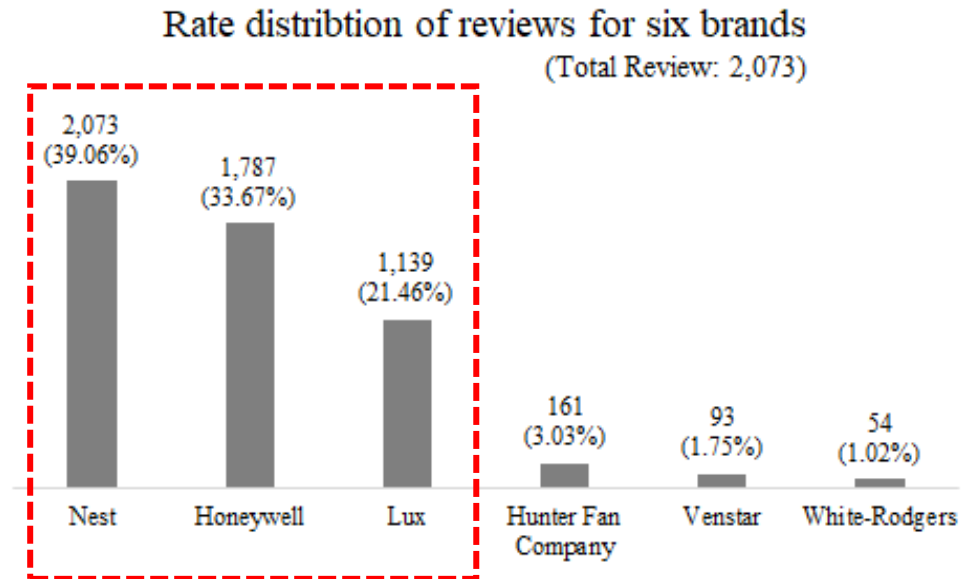
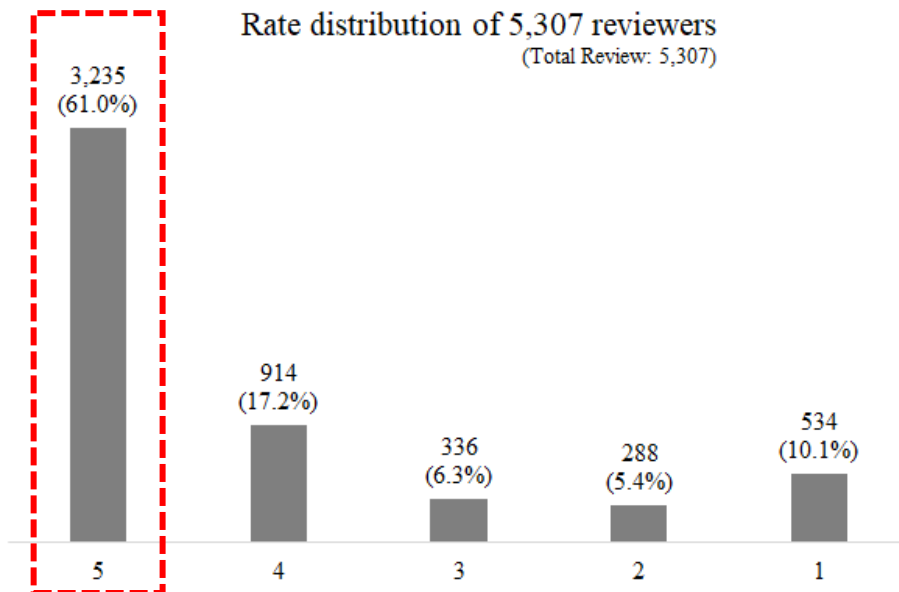
Step. 3: Econometric analysis

Step. 4: Predicting potential individual consumers' ratings

Step. 5: Review sentiment classification

STEP 1. Data Cleaning

- Cleaning 'suspicious one-time reviewers' and 'always-the-same-rating reviewers' from sample
- Deleting reviews with no DFs
- Selecting the top 6 brands based on # of reviews



STEP 2. Topic modeling (LDA) and annotation

- Identifying five latent product content dimensions (PCDs) in the review text using LDA*

* Latent Dirichlet allocation

Topic dimensions	Interpretation	Top 15 keywords in each topic
1. Connectivity	The review describes WiFi, wireless connection issues with software (e.g., App) and hardware (e.g., HVAC)	wire, WiFi, power, device, connected, connect, wireless, Issue, common, update, app, router, software, hvac, connection,
2. Easiness	The review mentions ease of use, including simplicity of installation, programming, and use.	easy, work, install, program, installation, instruction, installed, simple, programming, nice, programmable, well, took, product, set
3. Saving	The review talks about energy savings, including money savings by reducing energy consumption.	energy, control, save, away, money, saving, heater, month, app, bill, iphone, electric, temperature, feature, best
4. Setting	The review contains content related to setting and control, and information related to temperature, time, scheduling, heating, and other devices.	temperature, time, set, heat, turn, day, back, go, temp, setting, system, need, want, work, change
5. Support	The review focuses on consumer support services before, during, and after they make a purchase.	support, customer, call, product, service, called, tech, told, said, company, hvac, issue, worked, working, customer_service

STEP 2. Topic modeling (LDA) and annotation

Modifying PCDs by leveraging the domain expert's knowledge.

- The expert extends **five product content** dimensions from the LDA model to **nine dimensions** based on domain knowledge and the purpose of the research design.
- The nine dimensions are **1. Smart-connectivity, 2. Easiness, 3. Energy Savings, 4. Functionality, 5. Support, 6. Price value 7. Privacy, 8. The Amazon effect, and 9. Environment friendliness.**
- **The domain expert manually annotates 47,763** labeling (two months) tasks for the reviewers' sentiment toward each product content dimension to transfer domain knowledge to the models.


STEP 3: Research Question 1

1. Can consumers' preferences be identified?


STEP 3. Econometric analysis for ratings

- **Ratings** could censor the strength of reviewers' latent utility therefore, consumers' observable ratings indicate the range of their unobservable continuous preferences
(Green 2012)

Create Review

 Nest Learning Thermostat, 2nd Generation, Works with Amazon Alexa

Overall rating



Add a headline

What's most important to know?

Write your review

What did you like or dislike? What did you use this product for?

$$R_{ipt} = 1, \text{ if } -\infty < U_{ipt}^* \leq c_1$$

$$R_{ipt} = 2, \text{ if } c_1 < U_{ipt}^* \leq c_2,$$

$$R_{ipt} = 3, \text{ if } c_2 < U_{ipt}^* \leq c_3,$$

$$R_{ipt} = 4, \text{ if } c_3 < U_{ipt}^* \leq c_4,$$

$$R_{ipt} = 5, \text{ if } c_4 < U_{ipt}^* < \infty.$$

STEP 3. Econometric analysis for ratings

- Heteroskedastic ordered probit model (HEPTO)

- U_{ipt}^* denotes the unobservable continuous utility of reviewer i for product p on day t as follows:

$$U_{ipt}^* = x'_{ipt}\beta + \rho\varepsilon_{it}, \quad \varepsilon_{it} \sim \text{i. i. d Normal } (0, 1)$$

- $\rho = 1$ in an ordered probit model (OP)
- $\rho_i = \exp(\mathbf{Z}'_{it}\boldsymbol{\gamma})$ in a heteroskedastic ordered probit model (HETOP)

STEP 3. Econometric analysis for ratings

- At time variables (observable variables)

Variable	Description
rating (dependent)	i (the reviewer)' five-scale star-rating for a PT at t_i^*
sum_len	i 's length of review summary (headline) at t_i
rev_len	i 's length of review body at t_i
title_len	The length of tittle for the PT reviewed by i at t_i
desc_len	The length of description for the PT reviewed by i at t_i
nest	Brand dummy for the Nest (base group is White Roger)
honey	Brand dummy for the Honeywell
hunter	Brand dummy for the Hunter Fan
lux	Brand dummy for the Lux
venstar	Brand dummy for the Venstar

STEP 3. Econometric analysis for ratings

- User DFs variables

Variable	Description
u_help_dfs	The number of helpfulness upvote for i in all categories by t_i^b
u_no_help_dfs	The number of helpfulness downvote for i in all categories by t_i^b
u_avg_len_sum	i 's average length of summary in all categories by t_i^b
u_sd_len_sum	i 's SD of length of summary in all categories by t_i^b
u_avg_len_rev	i 's average length of review body in all categories by t_i^b
u_sd_len_rev	i 's SD of length of review body in all categories by t_i^b
u_cum_reviews	i 's number of reviews in all categories by t_i^b
u_cate_diversity	Shanon index for i 's category diversity of reviews posted by t_i^b
u_avg_rating	i 's average star-rating in all categories by t_i^b
u_sd_rating	i 's SD of star-rating in all categories by t_i^b

STEP 3. Econometric analysis for ratings

- Target consumers' volume of prior reviews in each sub-category

Variable	Description
sum_amz_video	i's number of reviews in the amazon instant video category by t_i^b
sum_appliance	i's number of reviews in the appliance category by t_i^b
sum_apps	i's number of reviews in the apps for android category by t_i^b
sum_arts_crafts	i's number of reviews in the art crafts category by t_i^b
sum_automotive	i's number of reviews in the automotive category by t_i^b
sum_baby	i's number of reviews in the baby category by t_i^b
sum_beauty	i's number of reviews in the beauty category by t_i^b
sum_books	i's number of reviews in the book category by t_i^b
sum_buyakindle	i's number of reviews in the kindle category by t_i^b
sum_cdsvinyl	i's number of reviews in the cds and vinyl category by t_i^b
sum_cellphone	i's number of reviews in the cell phones category by t_i^b
sum_clothes	i's number of reviews in the clothes, shoes, jewelry category by t_i^b
sum_computers	i's number of reviews in the computer category by t_i^b
sum_digit_music	i's number of reviews in the digital music category by t_i^b
sum_electronics	i's number of reviews in the electronics category by t_i^b
sum_giftcards	i's number of reviews in the gift cards category by t_i^b
sum_grocery	i's number of reviews in the grocery gourmet food category by t_i^b
sum_healthcare	i's number of reviews in the health personal care category by t_i^b
sum_home_kitch	i's number of reviews in the home kitchen category by t_i^b
sum_industry_spe	i's number of reviews in the industry specific category by t_i^b
sum_kindle_store	i's number of reviews in the kindle store category by t_i^b
sum_magazine	i's number of reviews in the magazine subscription category by t_i^b
sum_movies_tv	i's number of reviews in the move and tv category by t_i^b
sum_musical_ins	i's number of reviews in the musical instrument category by t_i^b
sum_office_prod	i's number of reviews in the office products category by t_i^b
sum_patio_lawn	i's number of reviews in the patio, lawn, and garden category by t_i^b
sum_pet_supp	i's number of reviews in the pet supplies category by t_i^b
sum_software	i's number of reviews in the software category by t_i^b
sum_sports_out	i's number of reviews in the spots and outdoors category by t_i^b
sum_tools_home	i's number of reviews in the tools & home category by t_i^b
sum_toys_games	i's number of reviews in the tops and games category by t_i^b
sum_video_games	i's number of reviews in the video games category by t_i^b

STEP 3. Econometric analysis for ratings

- Crowd DFs variables

Variable	Description
c_cum_reviews	p's number of crowd's reviews by t_i^b
c_avg_rating	p's average rating of crowd by $t_{j \neq i}^b$ *
c_sd_rating	p's SD of crowd's rating by $t_{j \neq i}^b$
c_avg_len_sum	p's average length of review summary written by crowd until $t_{j \neq i}^b$
c_sd_len_sum	p's SD of review summary written by crowd until $t_{j \neq i}^b$
c_avg_len_rev	p's average length of review body written by crowd until $t_{j \neq i}^b$
c_sd_len_rev	p's SD for the length of review body written by crowd until $t_{j \neq i}^b$
c_rating_rec	p's average rating of crowd at $t_{j \neq i}^b$
c_len_sum_rec	p's the length of review summary written by a crowd at $t_{j \neq i}^b$
c_len_rev_rec	p's the length of review body written by a crowd at $t_{j \neq i}^b$

STEP 3. Econometric analysis for ratings

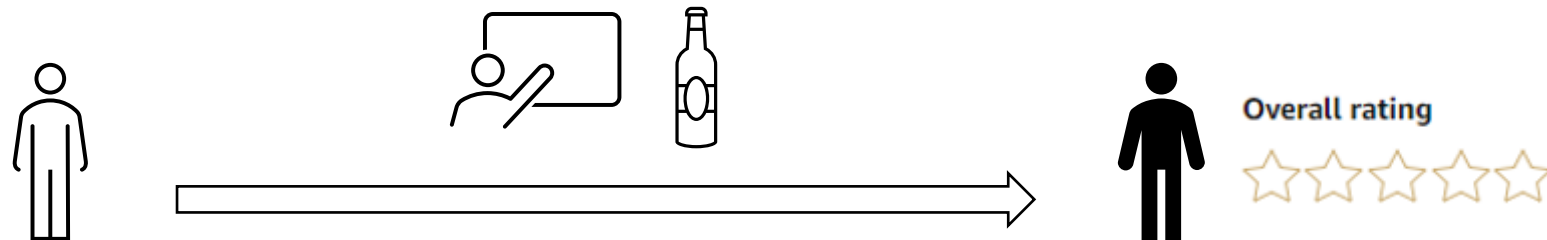
- Time variables

Variable	Description
day	Day dummies for t_i and base day is Monday (0)
month	Month dummies for t_i and base month is January (1)
year	Year dummies for t_i and base year is 2005
holiday	US holiday dummies and base is not holiday (0)
interval	The time interval between p 's the day of first review and t_i
nest_avail	Dummy for the first day of the Nest's PT on Amazon (Dec 15, 2011)

STEP 3. Econometric analysis for ratings

- Sentiment variables toward product content dimensions (PCDs)

Variable	Description
smart_con	i's sentiment of p's smart connectivity in i's review at t_i
easy	i's sentiment of p's easiness in i's review at t_i
save	i's sentiment of p's energy saving in i's review at t_i
func	i's sentiment of p's functionality in i's review at t_i
support	i's sentiment of p's support in i's review at t_i
price value	i's sentiment of p's perceived price value in i's review at t_i
privacy	i's sentiment of p's privacy issues in i's review at t_i
amazon	i's sentiment of p's Amazon effect in i's review at t_i
env	i's sentiment of p's environmental friendliness in i's review at t_i



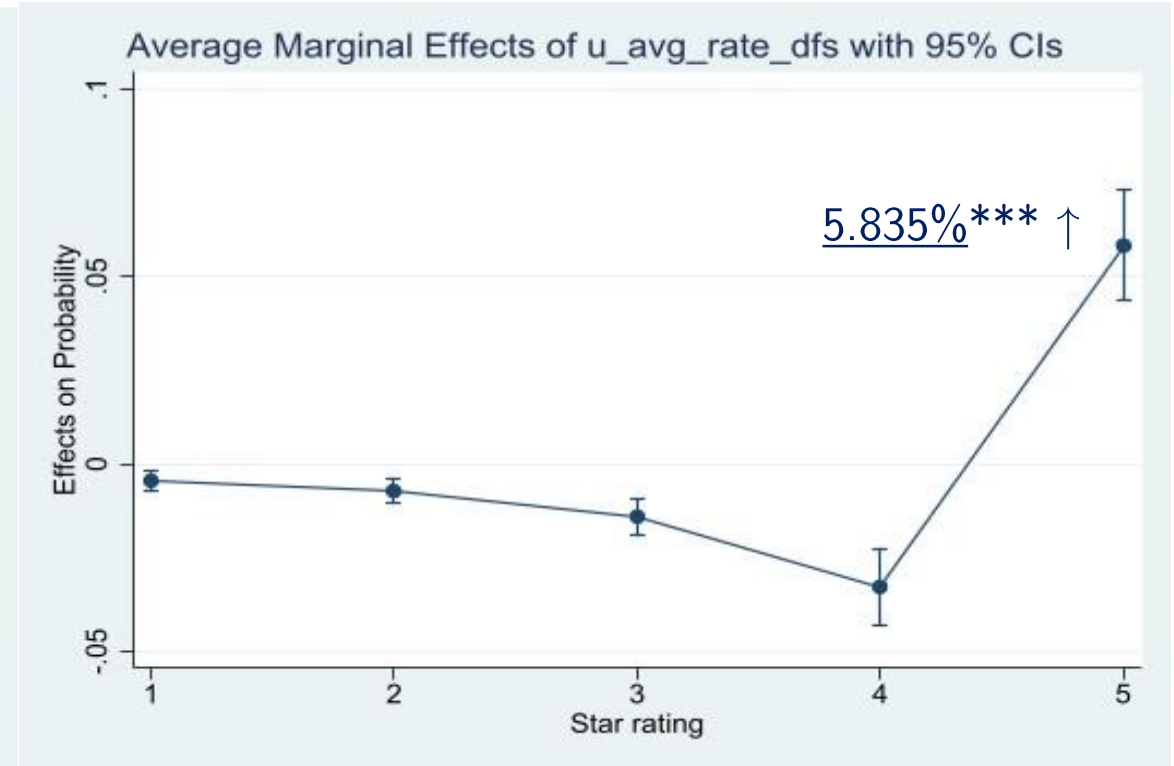
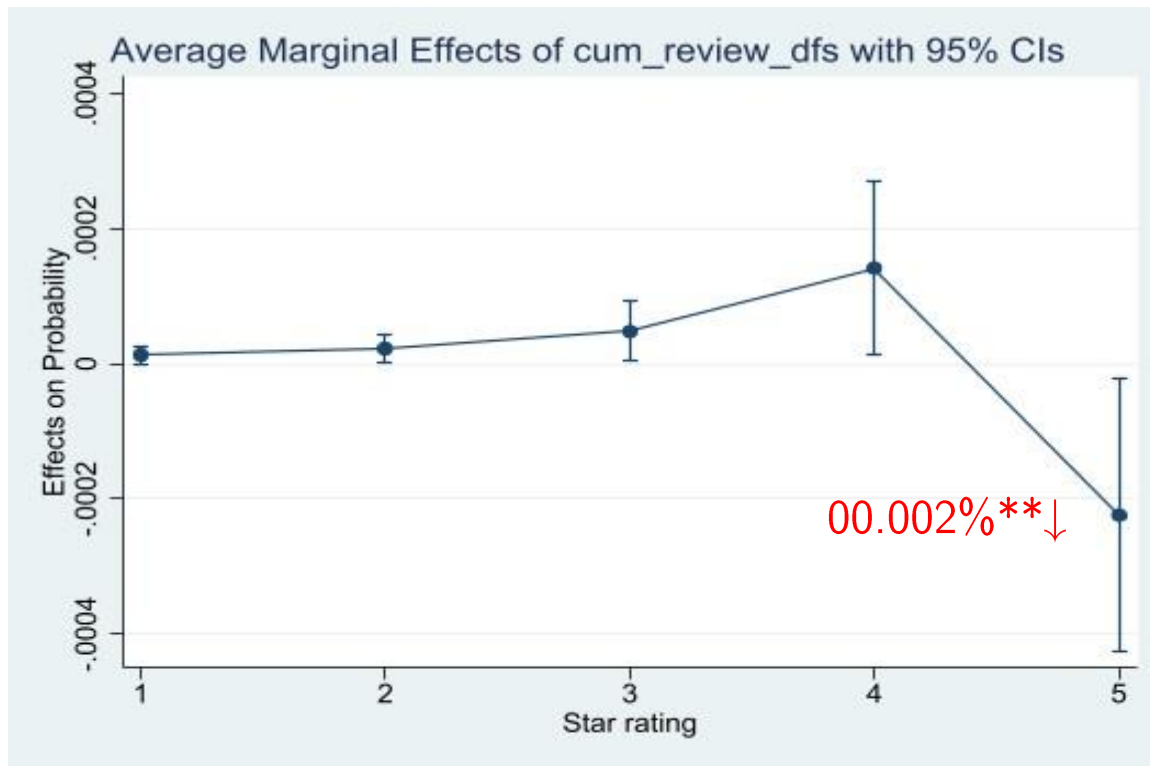
STEP 3. Econometric analysis for ratings

- Price variables (at the time of web scrapping and users' price DFs)

Variable	Description
price	p (the PT reviewed by i at t_i)'s price (at the time of web scrapping)
u_avg_p_dfs	i's average price for reviewed products in all category by t_i^{b*}
u_sd_p_dfs	i's SD of price for reviewed products in all category by t_i^b
u_max_p_dfs	i's the highest price among reviewed products in all category by t_i^b

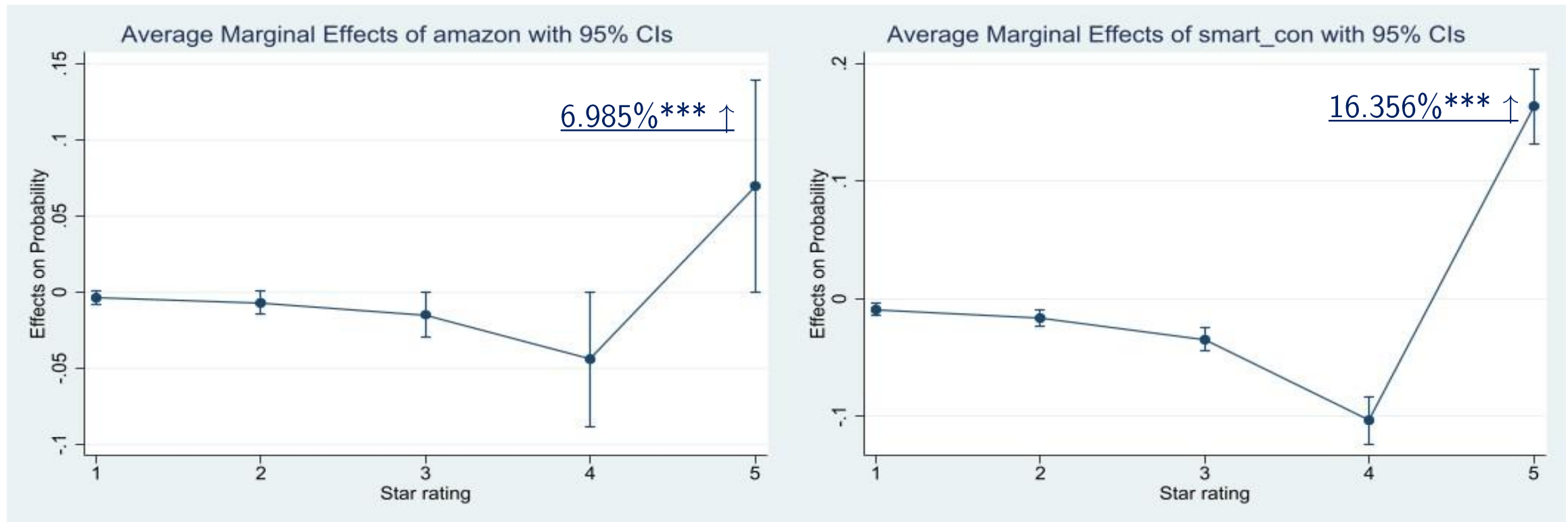
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The target consumer' volume of prior reviews and average rating



STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The target consumer's sentiment toward the amazon's service quality and smart connectivity.



STEP 3. Econometric analysis for ratings

Management Implications for the Nest (the smart thermostat company)

- Longer review summary length could be a signal for a lower probability of 5-star rating
- A consumer with a lower volume of prior reviews and a higher average rating in the prior reviews may have a higher probability of a 5-star rating
- Marginal effect for probability of a 5-star rating is **privacy > functionality > support > easiness > energy saving > smart connectivity > price value > amazon service effect***.

* This study is the first study to measure the effect of online service platform on star ratings

STEP 3. Econometric analysis for ratings

The effect of volume of prior reviews on each subcategory on ratings.

- The results show that a reviewer is **more likely to give a five-star rating** for the reviewed PT who
- (1) **writes a smaller volume of prior reviews** in the specific eight product categories (“Amazon instant video”, “apps for Android”, “cell phones”, “clothes, shoes, and jewelry”, “grocery gourmet food”, “health and personal care”, “magazine subscriptions”, and “software”)
- (2) and **writes a larger volume of reviews** in the “appliance” category.

STEP 4: Research Question 2

2. Can potential individual consumers' preferences be predicted?

STEP 4. Predicting potential consumers' ratings

- This study defines two different counterfactual scenarios as “full ex ante” and “partial ex ante” predictions.
- The designation “ex ante” indicates a firm’s prediction of consumers’ preferences before they make a purchase (full ex ante) or write a review of the purchased product (partial ex ante.)

STEP 4. Predicting potential consumers' ratings

- **Six popular supervised machine learning** models are applied, including
 1. **two base models** (kernel support vector machine and decision tree),
 2. **tree ensemble models** (random forest and extreme gradient boosting),
 3. **deep learning** (artificial neural net and long- short- term memory).

STEP 4. Predicting potential consumers' ratings

- Each machine learning model predicts potential consumers' star ratings **with six different ex ante variable sets** to identify the effect of adding
 1. digital footprint variables,
 2. the volume of prior reviews in each category,
 3. product dummies (partial ex ante, firms know the type of product purchased)
 4. potentially biased price variables.

STEP 4. Predicting potential consumers' ratings

- The star ratings (label) are skewed to five-star ratings (majority classes); therefore, the Amazon review dataset is an **imbalanced dataset**.

Star rating	Total Set		Total Train Set		Train Set		Valid Set		Test Set	
	Count	Shares	Count	Shares	Count	Share	Count	Share	Count	Share
5	3,235	60.96%	3,039	60.73%	2,841	60.41%	198	65.78%	196	64.69%
4	914	17.22%	872	17.43%	829	17.63%	43	14.29%	42	13.86%
3	336	6.33%	322	6.43%	308	6.55%	14	4.65%	14	4.62%
2	288	5.43%	268	5.36%	258	5.49%	10	3.32%	20	6.60%
1	534	10.06%	503	10.05%	467	9.93%	36	11.96%	31	10.23%
Total	5,307	100.00%	5,004	100.00%	4,703	100.00%	301	100.00%	303	100.00%
Period	Oct 12, 2005 – July 17, 2014		Oct 12, 2005 – May 17, 2014		Oct 12, 2005 – Mar 16, 2014		Mar 17, 2014 – May 17, 2014		May 18, 2014 – July 17, 2014	

STEP 4. Predicting potential consumers' ratings

- The prediction performance criteria for classification are:
 - 1) “accuracy,”
 - 2) “precision,”
 - 3) “recall,”
 - and 4) “F1 score,”
- **Accuracy:** the ratio of the total number of correctly classified reviews over the total number of reviews
- **Precision:** the fraction of reviews correctly classified for a rating over the total number of reviews classified as the rating.
- **Recall:** the fraction of reviews correctly classified for a given rating over the true number of reviews belong to the rating
- **F-measure:** the weighted average of precision and recall in the following format: **F1 score** = $\frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}}$

STEP 4. Predicting potential consumers' ratings

- Cohn and De La Torre (2013), the F1 score is a better evaluation criterion for an imbalanced dataset.
- The weighted average macro F1 score (WA F1) is the evaluation criterion:

$$\text{Weighted average macro F1 score (WA F1)} = \sum_{k=1}^K \frac{N_k}{N} \times \text{k class's F1 - score}$$

STEP 4. Predicting potential consumers' ratings

- Extreme grading boosting (XGBoost) shows the stable and best prediction performance.

Five star rating prediction performance (weighted average F1 Score)

Models	At-time	Ex ante base	Ex ante sub	Ex ante sub price	Partial ex ante	Partial ex ante price
Variables	Obs (37)	Obs + DFs (59)	Obs + DFs +Sub (90)	Obs + DFs +Sub + P (94)	Obs + DFs +Sub + Item (161)	Obs + DFs +Sub + Item +P (165)
Heteroprobit	0.51	0.51	0.52	0.51	N/A	N/A
Kernel SVM	0.50	0.51	0.51	0.51	0.51	0.51
Decision Tree	0.51	0.51	0.51	0.51	0.51	0.51
Random Forest	0.51	0.53	0.53	0.54	0.52	0.53
XGBoost	0.51	0.55	0.56	0.54	0.57	0.57
ANN	0.51	0.52	0.53	0.52	0.53	0.52
LSTM	0.51	0.52	0.53	0.53	0.52	0.54

Obs: observable variables, DFs: digital footprint variables, Sub: volume of prior reviews in each subcategory, P: price variables, Item: product dummies. A number in the parentheses means the number of variables

STEP 4. Predicting potential consumers' ratings

- Performance in three- and binary classification is **higher** than five-star rating classification
- Extreme grading boosting (XGBoost) is the best and stable prediction performance.

Three- and binary classification (Weighted average F1 score)

Class range Model	Three-class classification		Binary classification	
	Ex ante base	Partial ex ante sub	Ex ante base	Partial ex ante sub
Heteroprobit	0.72	N/A	0.72	N/A
Kernel SVM	0.69	0.69	0.71	0.69
Decision Tree	0.69	0.69	0.73	0.73
Random Forest	0.74	0.71	0.74	0.71
XGBoost	0.74	0.74	0.74	0.73
ANN	0.71	0.71	0.72	0.71
LSTM	0.70	0.7	0.73	0.71

STEP 4. Predicting potential consumers' ratings

- Minority class (3star rating) prediction accuracy is very zero.
- The lower performance of minority class could cause unfairness problem.

Models	Hyperparameter	Accuracy	Precision	Recall	F1-score	Confusion matrix																
Xgboost	Tree number: 100 Depth: 4 Learning rate:0.2	0.802	1: 0.78 2: 0.00 3: 0.80 WA: 0.76	1: 0.14 2: 0.00 3: 0.99 WA: 0.80	1: 0.23 2: 0.00 3: 0.89 WA: 0.74	<table border="1"><thead><tr><th></th><th>1</th><th>2</th><th>3</th></tr></thead><tbody><tr><th>1</th><td>7</td><td>0</td><td>44</td></tr><tr><th>2</th><td>0</td><td>0</td><td>14</td></tr><tr><th>3</th><td>2</td><td>0</td><td>236</td></tr></tbody></table>		1	2	3	1	7	0	44	2	0	0	14	3	2	0	236
	1	2	3																			
1	7	0	44																			
2	0	0	14																			
3	2	0	236																			

STEP 3. Econometric analysis for ratings

Takeaways:

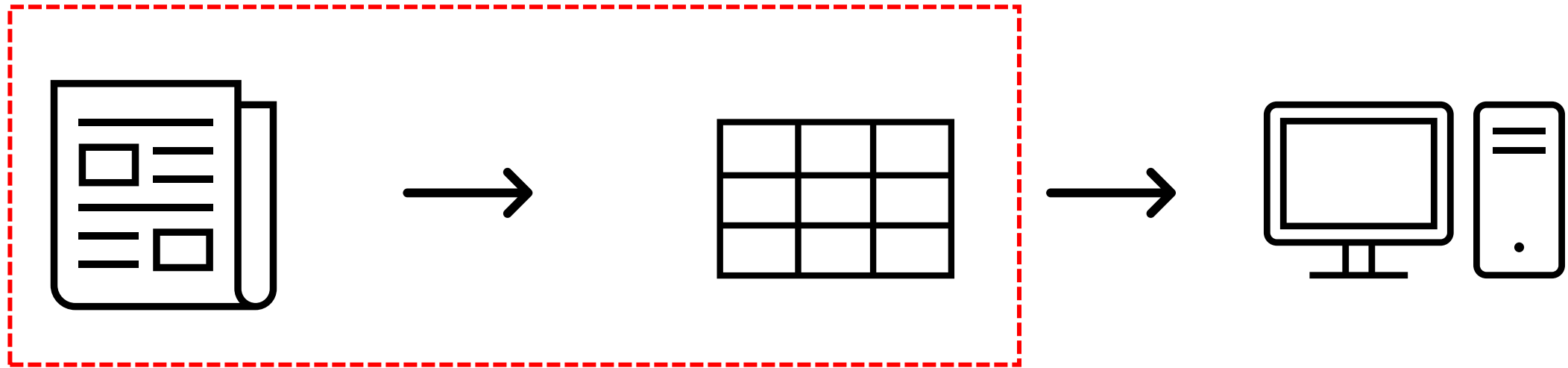
- XGBoost is the best, and stable prediction machine.
- Reducing class ranges improves performance.
- Prediction is skewed toward the majority class.
- Prediction performance for the minority class is low (imbalance problem).

STEP 5: Research Question 3

3. Can consumers' sentiments be classified?

STEP 5. Review sentiment classification

- Word embedding is a way of mapping text data into numerical vectors



STEP 5. Review sentiment classification

Three popular word embedding models are applied.

- 1. Frequency-based word embedding: **TF-IDF** (high dimensional)
- 2. Word-distribution based word embedding: **Word2Vector** (similarity, dense)

Context-free embedding methods : a word has the same embedding vector.

(Example) I disliked **the device**. I love **the device** now. → **the device** have the same vector.

STEP 5. Review sentiment classification

Three popular word embedding models are applied.

- 1. Frequency-based word embedding: TF-IDF (high dimensional)
- 2. Word-distribution based word embedding: Word2Vector (similarity, dense)
- 3. **Context-based word embedding**: BERT (similarity, ambiguity, context, dense)

(Example) I disliked **the device**. I love **the device** now. → **the device** have different vectors

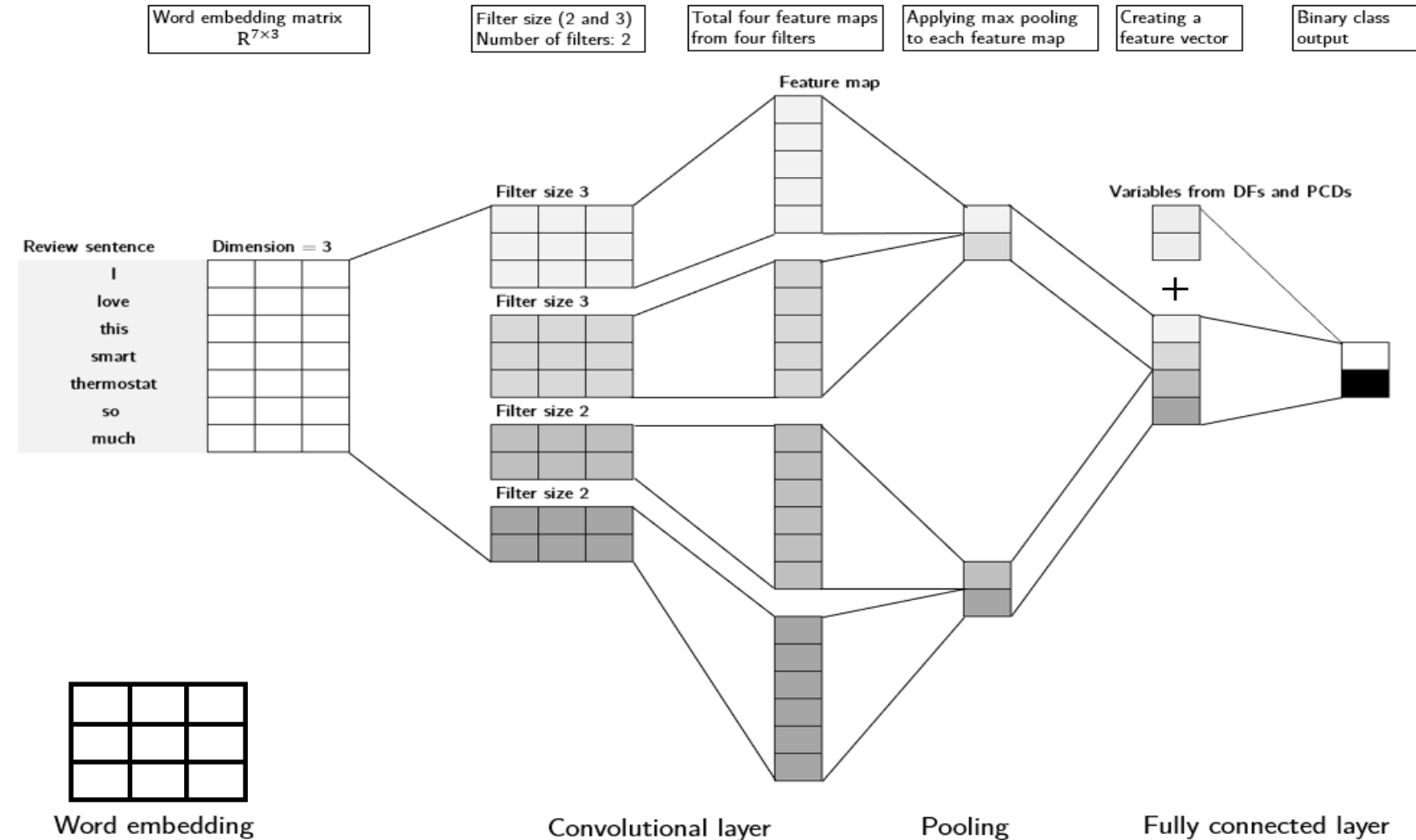
STEP 5. Review sentiment classification

- Convolutional Neural Network (CNN) on embedding vectors (W2V or BERT)

- Text + Structured Data

- W2V and BERT embedding

- CNN is a classifier



STEP 5. Review sentiment classification

- Class distribution is more balanced dataset so that accuracy is a suitable for evaluation.

Sentiment distribution in the functionality dimension

	Total Set		Total Train Set		Sub Train Set		Valid Set		Test Set	
-1	1,355	25.53%	1,281	25.60%	1,211	25.75%	70	23.26%	74	24.42%
0	1,739	32.77%	1,625	32.47%	1,523	32.38%	102	33.89%	114	37.62%
1	2,213	41.70%	2,098	41.93%	1,969	41.87%	129	42.86%	115	37.95%
Total	5,307	100%	5,004	100%	4,703	100%	301	100%	303	100%

STEP 5. Review sentiment classification

- Two different feature sets are applied:
 - 1. Partial model : Text only
 - 2. Full model : Text + Structured Variables

STEP 5. Review sentiment classification

- CNN with fine-tuned BERT embedding shows the best prediction performance.

The results of the sentiment classification of reviews about functionality dimension

Models	Partial model (Text only)		Full model (Text + partial ex ante-sub model)	
	WA F1	Accuracy	WA F1	Accuracy
RF + IDF	0.62	0.637	0.64	0.644
XGB + IDF	0.65	0.650	0.73	0.723
CNN + W2V	0.67	0.673	0.69	0.686
CNN (BERT)	0.72	0.719	0.73	0.729
CNN(BERT_S)	0.71	0.710	0.71	0.713
CNN(BERT_L)	0.72	0.719	0.72	0.719

Notes: Two different online product review datasets are applied for further pre-training: (1) BERT S (N = 169,809), containing all reviews of the target reviewers across all categories over the entire sample period; and (2) BERT L (N = 1,926,047), consisting of all reviews in the “tool and home improvement category.”

STEP 5. Review sentiment classification

Takeaways:

- Embedding is a way to transform text into vectors.
- Deep learning is better than ensemble models.
- CNN on fine-tuned BERT embedding is the best sentiment classification method.

Conclusion

- These approaches are **interpretable, scalable, and applicable** for **different goods** in a **specific industry**.
- These approaches can be used by **industry** to **design customer-oriented marketing strategies**.
- **Policy makers** can use these approaches to **identify public needs and opinions** using UGC.

Thank you



Pre-Processing

Step. 1: Selecting reviews with no missing values, which resulted in a set of 110 PTs.

- The PT set without missing values in either brand or price variables will be called “programmable thermostats” in the rest of this paper; there are 110 thermostats in this set.
- This study considers only the inexperienced consumers' first review on the PTs, because inexperienced consumers may become experienced consumers after they write their first review.

Pre-Processing

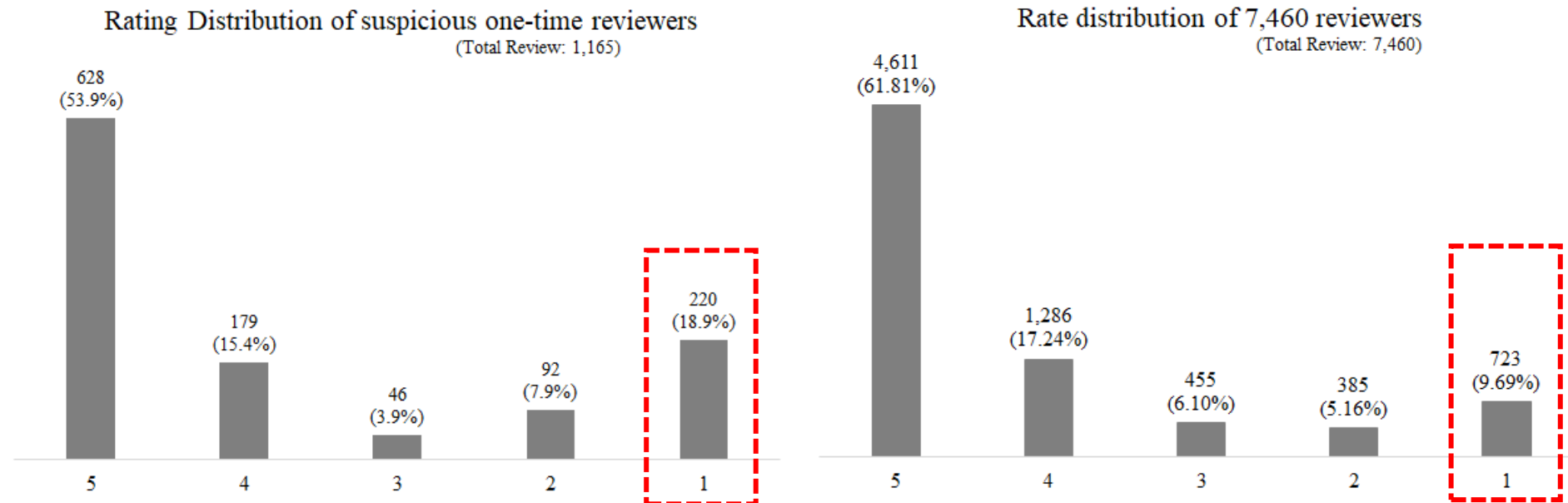
Step. 2: Cleaning 'suspicious one-time reviewers' and 'always-the-same-rating reviewers.'

- Mayzlin et al. (2014) defined the “suspicious reviewer” as one who writes only a review for a hotel for the first time only during the sample period (October 2011)
- This study defines “suspicious one-time reviewers” as those who write only a review for a PT as a first review and do not write reviews for any other products over the entire sample period (May 1996 – July 2014)

Pre-Processing

Step. 2: Cleaning 'suspicious one-time reviewers' and 'always the same rating reviewers.'

- Rating distributions of suspicious one-time reviewers and reviewers after cleanings.



Pre-Processing

Step. 2: Cleaning 'suspicious one-time reviewers' and 'always the same rating reviewers.'

- The '**Always the Same Rating Reviewers (ASR)**' is a reviewer who wrote more than **eight** times with the same rating level.
- A five-star rating showed the highest probability as of 0.595 in the "Tool and home improvement" category. The probability of nine consecutive five-star ratings is 0.00934, which is **less than 0.01**.
- **Only 69 reviewers** wrote more than eight reviews at the same star rate level (5 stars), surprisingly designating them as **Always happy raters (ASRs)**.

Pre-Processing

Step. 3: Deleting reviewers and reviews for products with no DFs

- Without DFs, it is impossible to measure the effect of DFs on a reviewer's rating

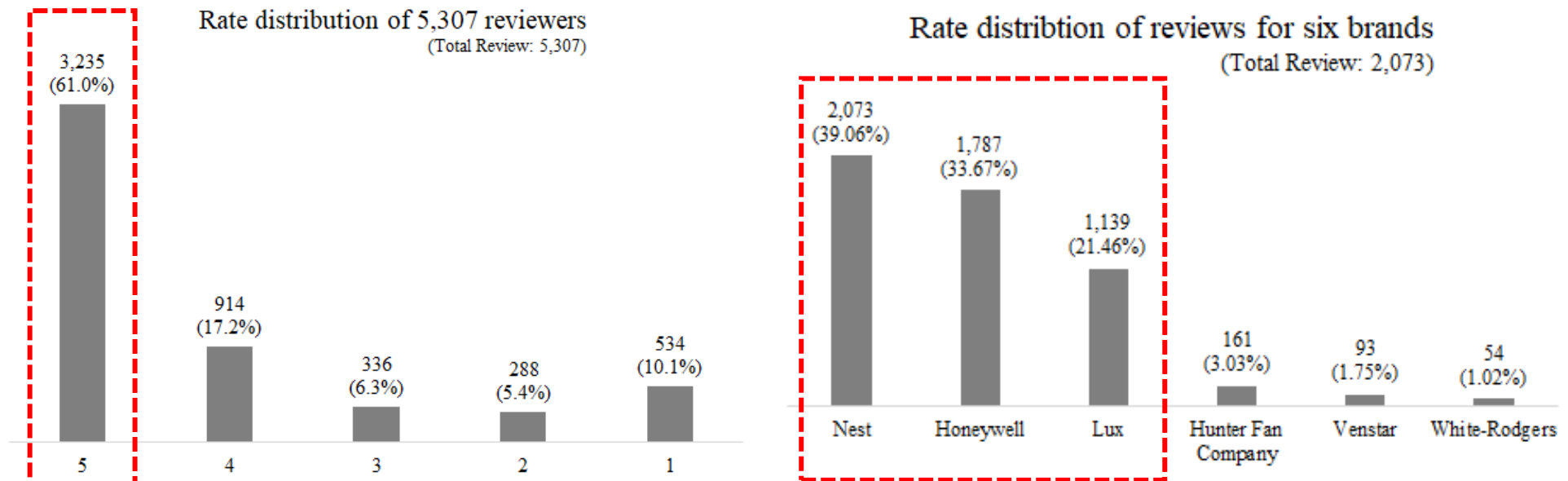
Accordingly, this procedure removed:

- (1) 1,965 reviewers do not have any previous reviews on other products
- (2) 91 reviewers write a review for the PTs that do not have any previous reviews

Pre-Processing

Step. 4: Selecting the top 6 brands among 26 brands.

- Major 6 brands that had received more than fifty reviews.



Step. 5: Identifying latent product content dimensions in the review text by using LDA.

- **Latent Dirichlet Allocation** (LDA, Blei, Ng, and Jordan 2003) is an unsupervised learning model; it was used to identify latent topics in each review and the distribution of these topics in each review.

Pre-Processing

Step. 5: Identifying latent product content dimensions in the review text by using LDA.

LDA assumes that \mathbf{w}_R (words in reviews) is generated from the joint distribution of θ_R (the review's topic distribution) and φ_K (the topic's word distribution). $z_{i,n}$ is a vector in \mathbb{R}^K that maps the n th word in the i th review to topic k . The joint distribution indicates the word generation process in reviews as follows:

$$p(\varphi_K, \theta_R, z_R, \mathbf{w}_R | \alpha, \beta) = \prod_{k=1}^K p(\varphi_k | \beta) \prod_{i=1}^R p(\theta_i | \alpha) \sum_{n=1}^N p(z_{i,n} | \theta_i) p(w_{i,n} | \varphi_k, z_{i,n} | \theta_i)$$

Excluding \mathbf{w}_R , the other variables are latent variables. During the training process of LDA, the optimal values of the latent variables maximize the posterior probability. The posterior probability is estimated by variational inference or Gibbs sampling.

$$p(\varphi_K, \theta_R, z_R | \mathbf{w}_R) = \frac{p(\varphi_K, \theta_R, z_R, \mathbf{w}_R)}{p(\mathbf{w}_R)}$$

Topic 1

1. Can consumers' preferences be identified through the analysis of DFs?

- The marginal effect of the HETOP model (continuous variables case):

	(1) case of $x_{it}^a \in x_{it} \cap Z_{it}^c$	(1) case of $x_{it}^b \in x_{it} \cap Z_{it}$
The marginal effect of x_{it} at $R_{ipt} = 1$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_1 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$
The marginal effect of x_{it} at $R_{ipt} = j$ where $j \in \{2, 3, 4\}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) - \emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \right] \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_j - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$ $- \left[\emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_{j-1} - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$
The marginal effect of x_{it} at $R_{ipt} = 5$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{\beta_{x_{it}^b} + (c_4 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$

Topic 1

1. Can consumers' preferences be identified through the analysis of DFs?

- The marginal effect of the HETOP model (continuous variables case):

	(1) case of $x_{it}^a \in x_{it} \cap Z_{it}^c$	(1) case of $x_{it}^b \in x_{it} \cap Z_{it}$
The marginal effect of x_{it} at $R_{ipt} = 1$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_1 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$
The marginal effect of x_{it} at $R_{ipt} = j$ where $j \in \{2, 3, 4\}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) - \emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \right] \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_j - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$ $- \left[\emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_{j-1} - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$
The marginal effect of x_{it} at $R_{ipt} = 5$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{\beta_{x_{it}^b} + (c_4 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$

Topic 1

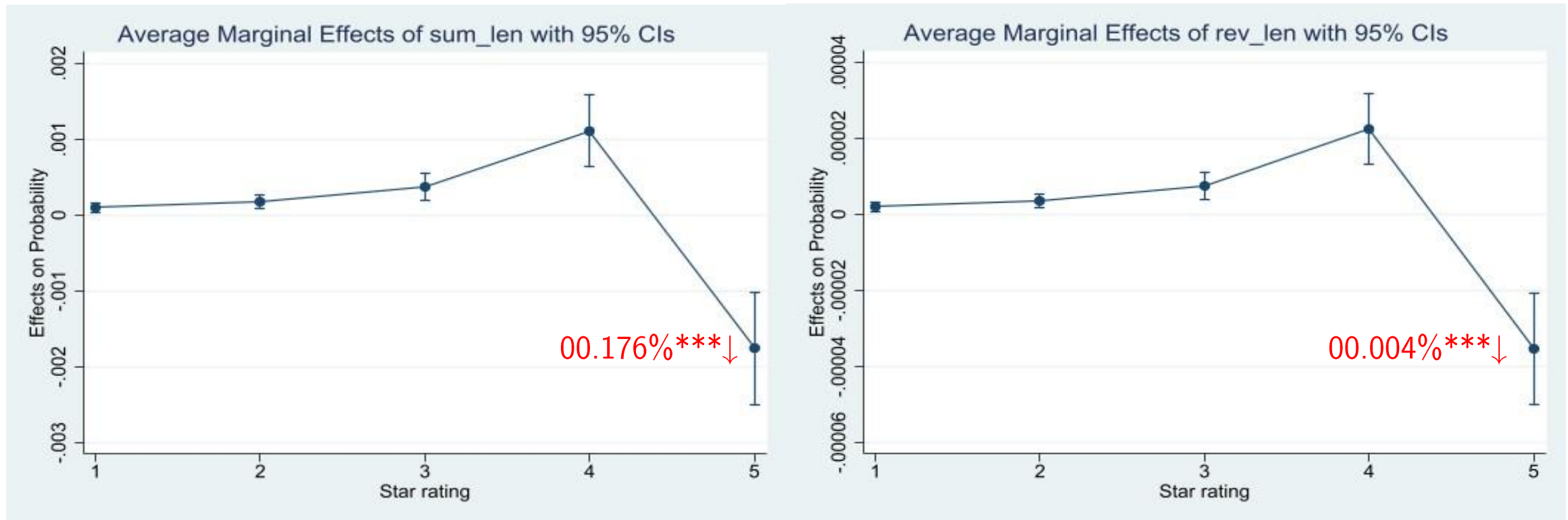
1. Can consumers' preferences be identified through the analysis of DFs?

- The marginal effect of the HETOP model (continuous variables case):

	(1) case of $x_{it}^a \in x_{it} \cap Z_{it}^c$	(1) case of $x_{it}^b \in x_{it} \cap Z_{it}$
The marginal effect of x_{it} at $R_{ipt} = 1$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_1 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_1 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$
The marginal effect of x_{it} at $R_{ipt} = j$ where $j \in \{2, 3, 4\}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) - \emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \right] \frac{-\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\left[\emptyset \left(\frac{c_j - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_j - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$ $- \left[\emptyset \left(\frac{c_{j-1} - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{-\beta_{x_{it}^b} - (c_{j-1} - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right) \right]$
The marginal effect of x_{it} at $R_{ipt} = 5$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \frac{\beta_{x_{it}^a}}{\exp(Z'_{it}\gamma)}$	$\emptyset \left(\frac{c_4 - x'_{it}\beta}{\exp(Z'_{it}\gamma)} \right) \left(\frac{\beta_{x_{it}^b} + (c_4 - x'_{it}\beta)\gamma_{x_{it}^b}}{\exp(Z'_{it}\gamma)} \right)$

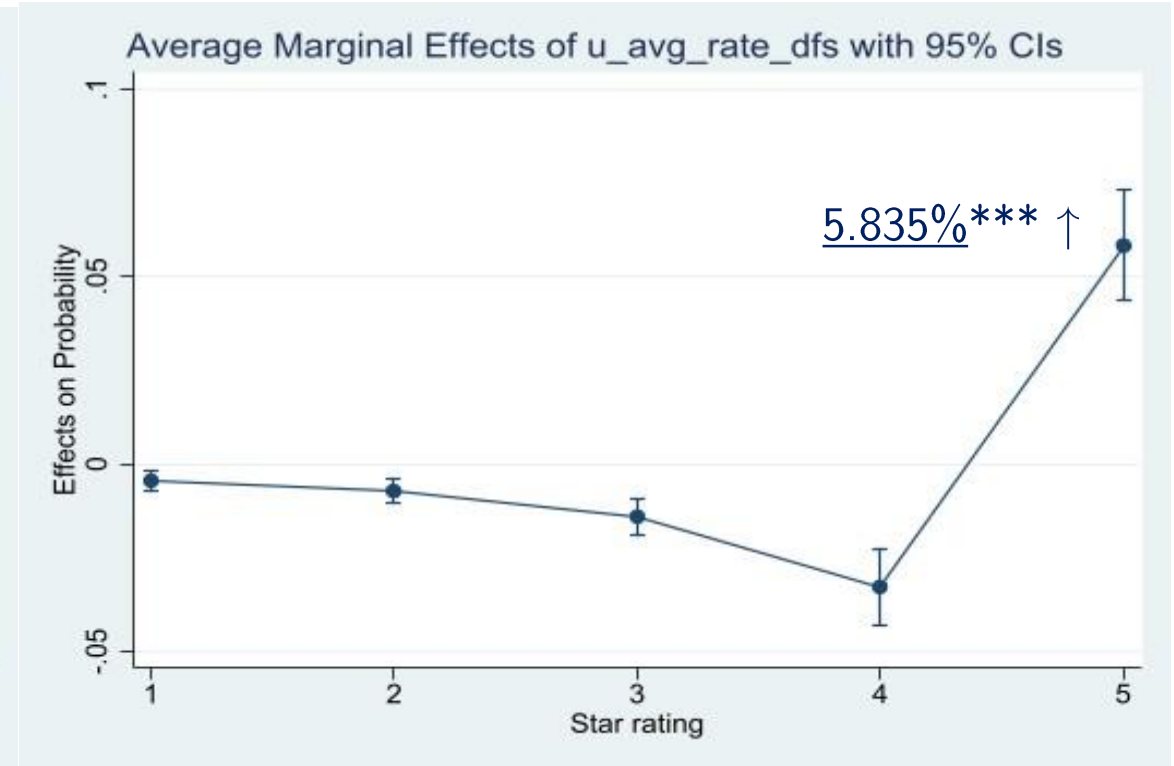
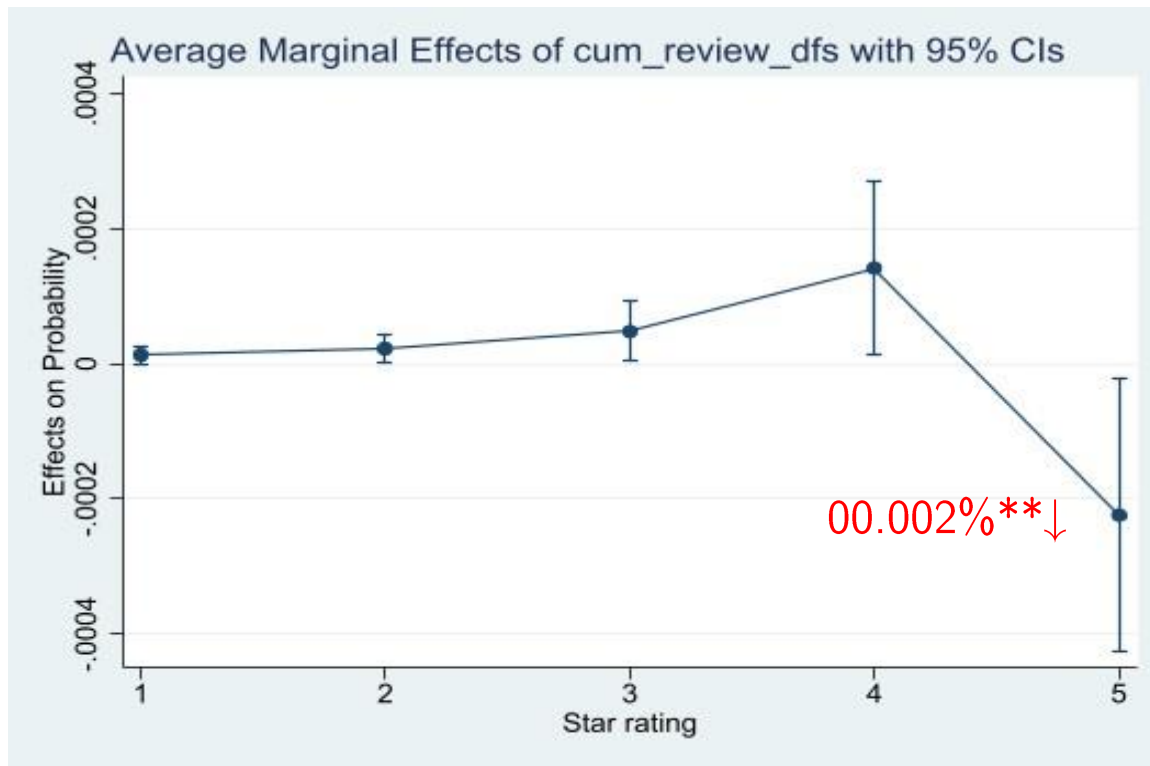
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The target consumer's length of review summary, length of review body



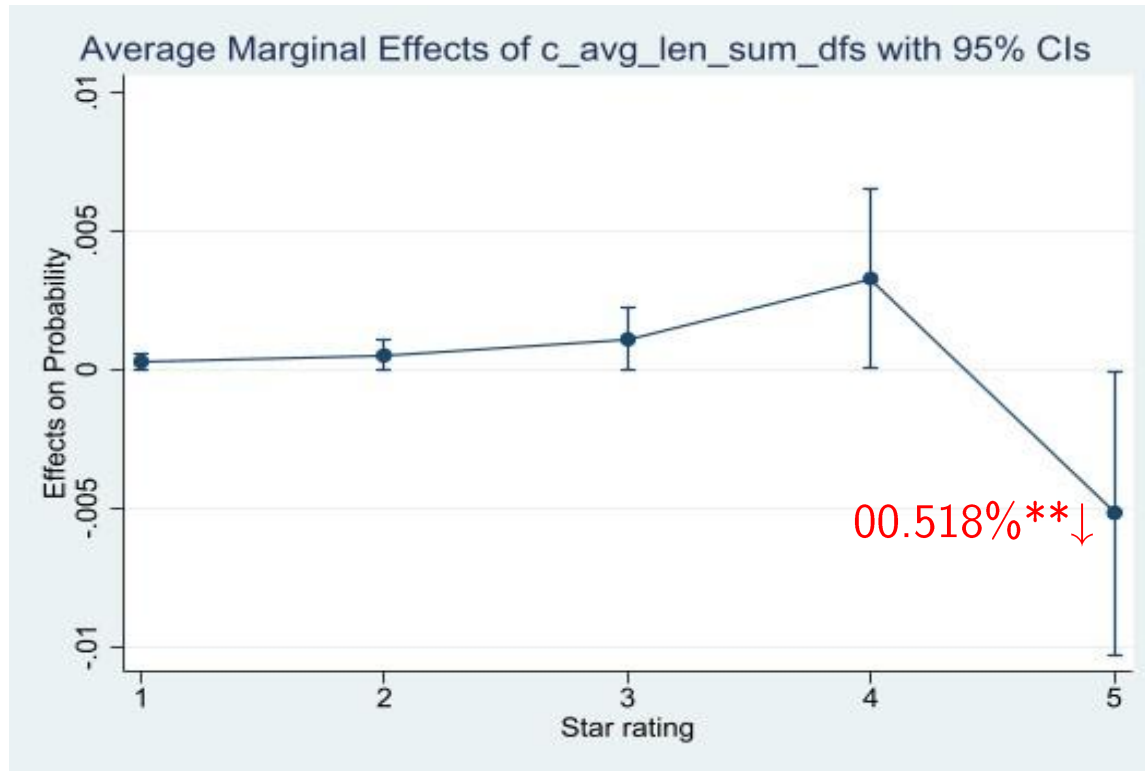
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The target consumer' volume of prior reviews and average rating



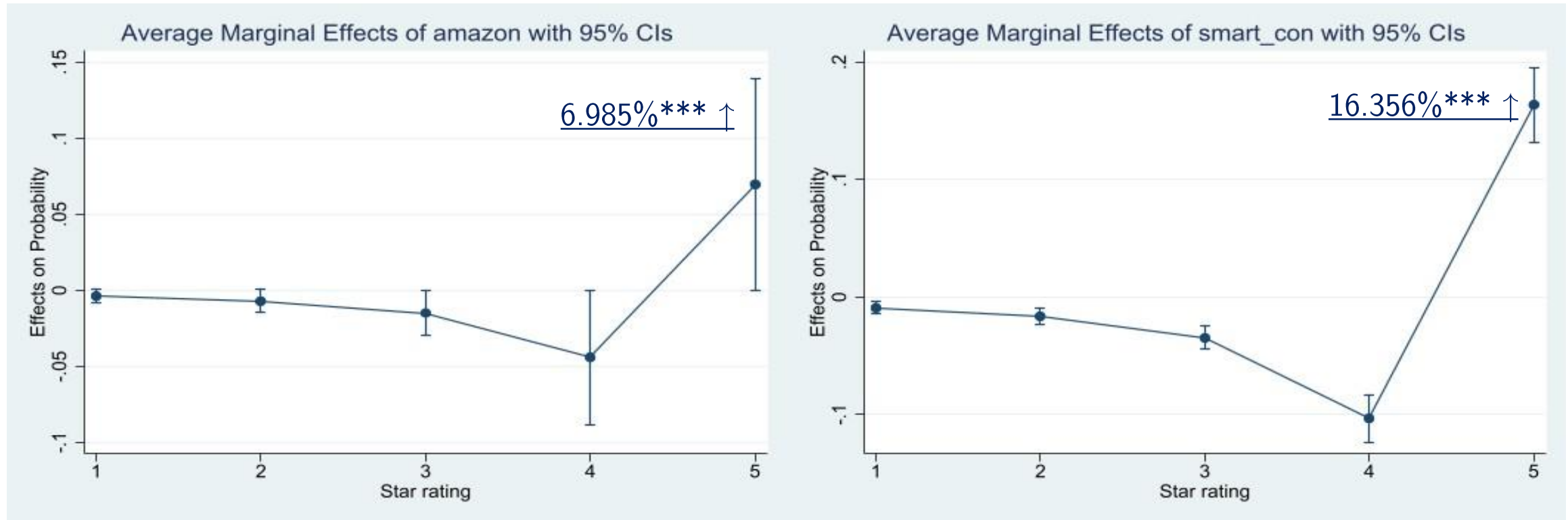
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The crowd's average summary length on the PT (that reviewed by the target consumer)



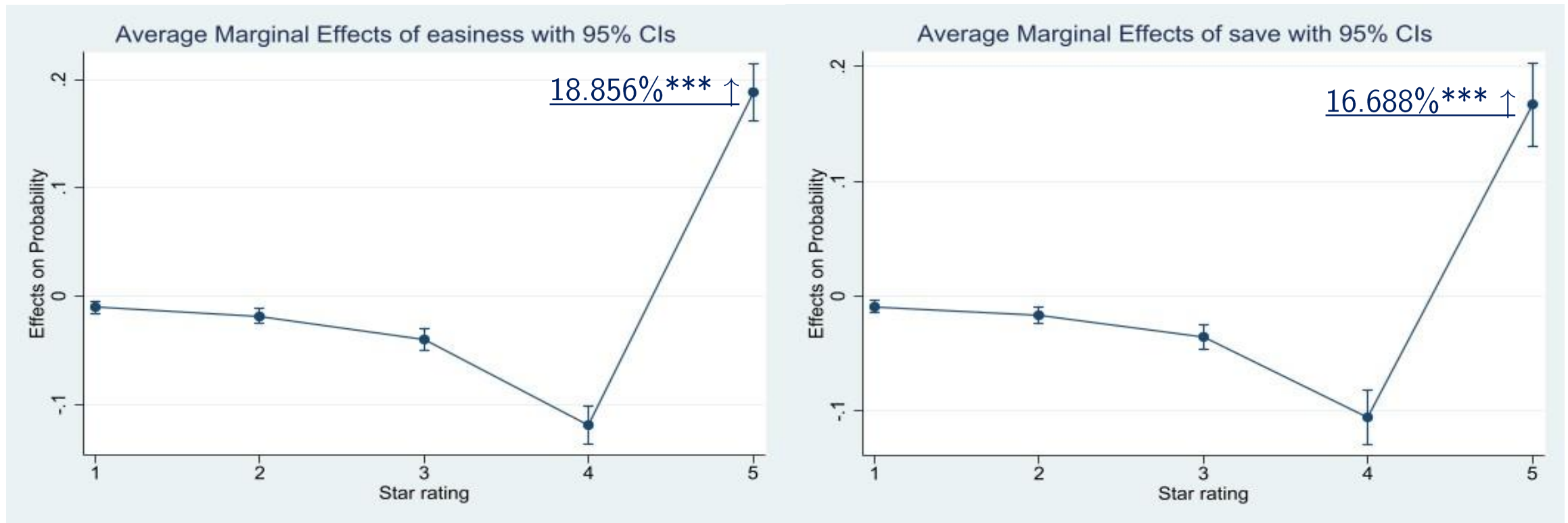
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- The target consumer's sentiment toward **the amazon's service quality and smart connectivity.**



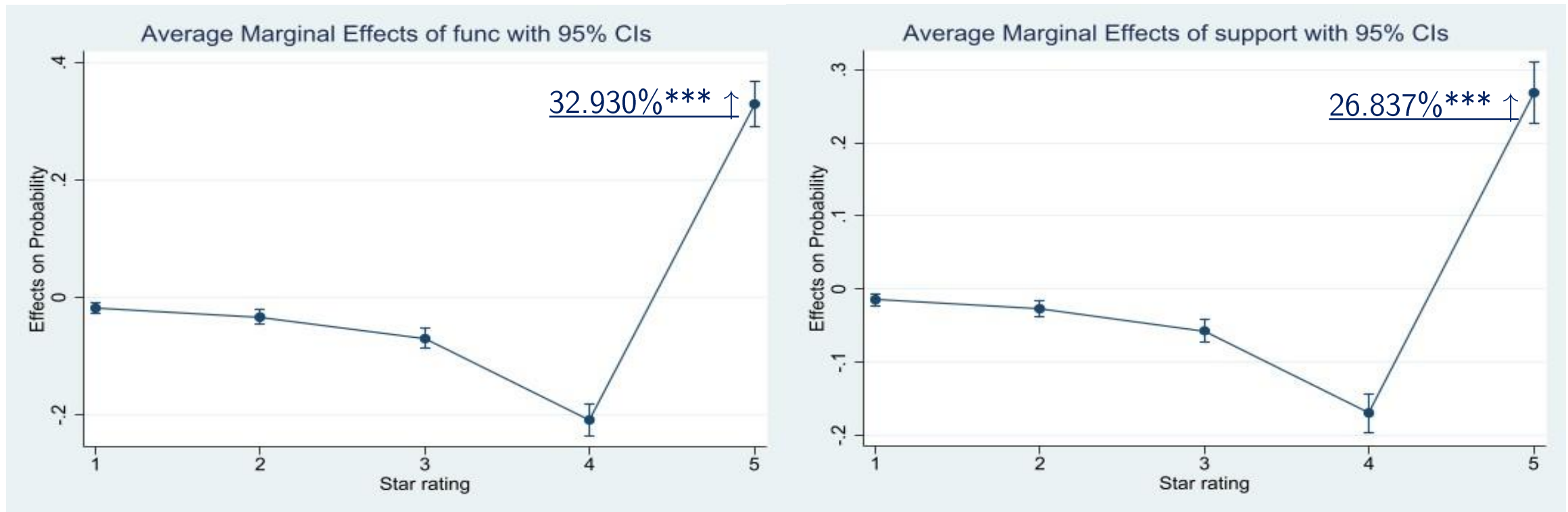
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- Target consumer's sentiment toward **easiness** and **energy saving**.



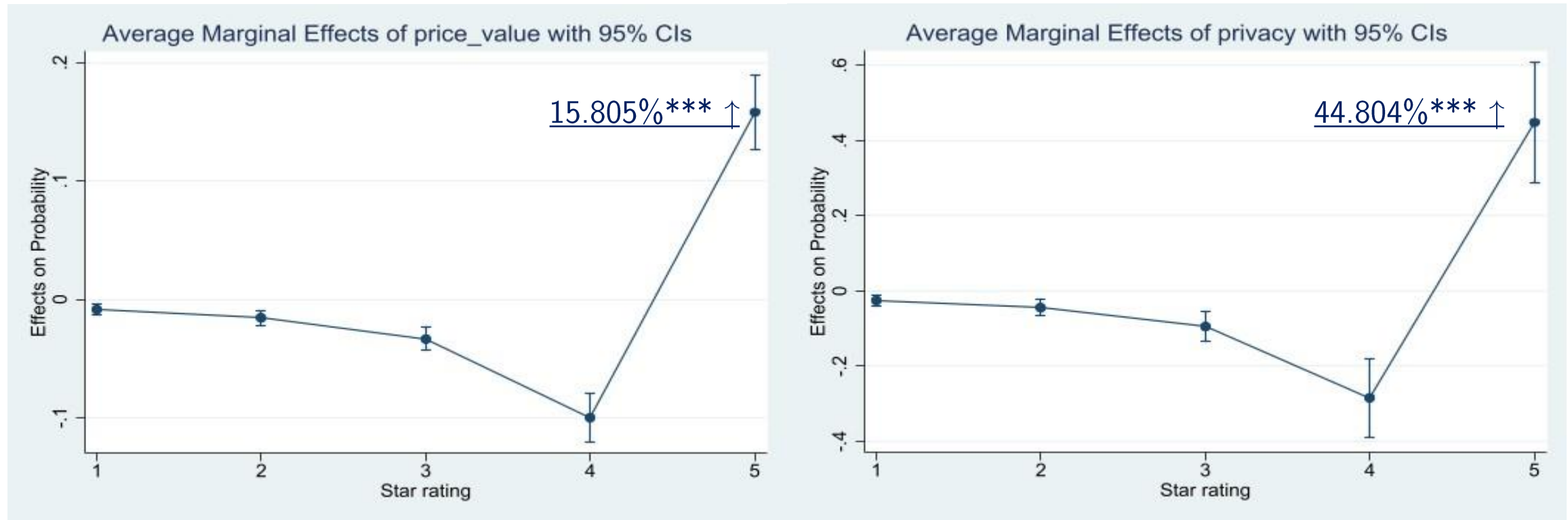
STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- Target consumer's sentiment toward **functionality** and **support**.



STEP 3. Econometric analysis for ratings

- Marginal Effect for the Nest
- Target consumer's sentiment toward **perceived price value** and **privacy**.



Topic 1

- Model_h4 shows the effect of volume of prior reviews on each subcategory on ratings.

Variable	model_h4
sum_len	-0.008***
rev_len	-0.000***
desc_len	-0.000**
u_sd_len_sum	0.010
cum_reviews	
u_avg_rating	0.254***
c_avg_len_sum	-0.028**
c_sd_len_sum	0.034**
smart_con	0.826***
easy	0.937***
save	0.858***
func	1.621***
support	1.326***
price_value	0.765***
privacy	2.337***
amazon	0.331*
sum_amz_video	-0.667*
sum_appliance	0.227*
sum_apps	-1.995*
sum_cellphone	-0.051*
sum_clothes	-0.091*
sum_grocery	-0.043**
sum_healthcare	0.055**
sum_magazine	-0.367*
sum_pet_supp	-0.053*
sum_software	-0.052

- The results show that a reviewer is less likely to give a five-star rating for the reviewed PT who
- (1) writes a larger volume of prior reviews in the specific product categories (“Amazon instant video”, “apps for Android”, “cell phones”, “clothes, shoes, and jewelry”, “grocery gourmet food”, “health and personal care”, “magazine subscriptions”, and “software”)
- (2) and writes a smaller volume of reviews in the “appliance” category.

Topic 1

- Robust check : unobservable variables' effect on the coefficients are limited.

Variable	Base (47 variables)	model_h with control (66 variables)
sum_len	-0.008*** (0.002)	-0.008*** (0.002)
rev_len	-0.000*** (0.000)	-0.000*** (0.000)
desc_len	-0.000** (0.000)	-0.000** (0.000)
nest	0.286 (0.199)	0.170 (0.248)
honey	0.425** (0.206)	0.431** (0.213)
hunter	-0.104 (0.237)	-0.008 (0.265)
lux	0.498** (0.220)	0.555** (0.234)
venstar	0.491* (0.271)	0.428 (0.278)
u_sd_len_sum	0.009** (0.004)	0.009* (0.005)
cum_reviews	-0.001** (0.000)	-0.001** (0.000)
u_avg_rating	0.230*** (0.052)	0.225*** (0.053)
c_avg_len_sum	-0.025** (0.011)	-0.022* (0.012)
u_sd_len_sum	0.032** (0.013)	0.030** (0.013)
smart_con	0.708*** (0.149)	0.699*** (0.147)
easy	0.808*** (0.156)	0.806*** (0.154)
save	0.700*** (0.152)	0.713*** (0.154)
func	1.408*** (0.264)	1.407*** (0.263)
support	1.148*** (0.224)	1.147*** (0.223)
price value	0.665*** (0.138)	0.675*** (0.139)
privacy	1.938*** (0.500)	1.915*** (0.496)
amazon	0.291* (0.162)	0.298* (0.162)
env	0.022 (0.686)	0.058 (0.698)
Z.u_avg_rating	-0.033* (0.019)	-0.032* (0.019)
Z.nest	0.544*** (0.174)	0.544*** (0.174)
Z.honey	0.401** (0.174)	0.398** (0.174)
Z.lux	0.382** (0.174)	0.386** (0.174)
Z.hunter	0.594*** (0.196)	0.582*** (0.196)
Z.venstar	0.175 (0.224)	0.229 (0.224)

Research Topic 2

- (Base model 1) Kernel support vector machine (Nonlinear SVM)
- The support vector machine (SVM) model finds the linear separable hyperplane in the feature space to classify labels
- To deal with non-linearly separable, noisy, and outlier data, Cortes and Vapnik (1995) introduced a slack variable.
- A kernel SVM is applied in this study to consider the non-linearity of data.
- A kernel function K implicitly maps original data to a high-dimensional functional feature space $\Phi: \mathbf{x} \rightarrow \varphi(\mathbf{x})$, such that $K(\mathbf{x}, \mathbf{x}') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$ for two samples \mathbf{x} and \mathbf{x}' .
- The Gaussian radial basis function (RBF) is the kernel function, as follows:

$$K_{\text{rbf}}(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2)$$

Research Topic 2

- (Base model 1) Kernel support vector machine (Nonlinear SVM)

- Overall, the dual problem of kernel SVM can be expressed as follows:

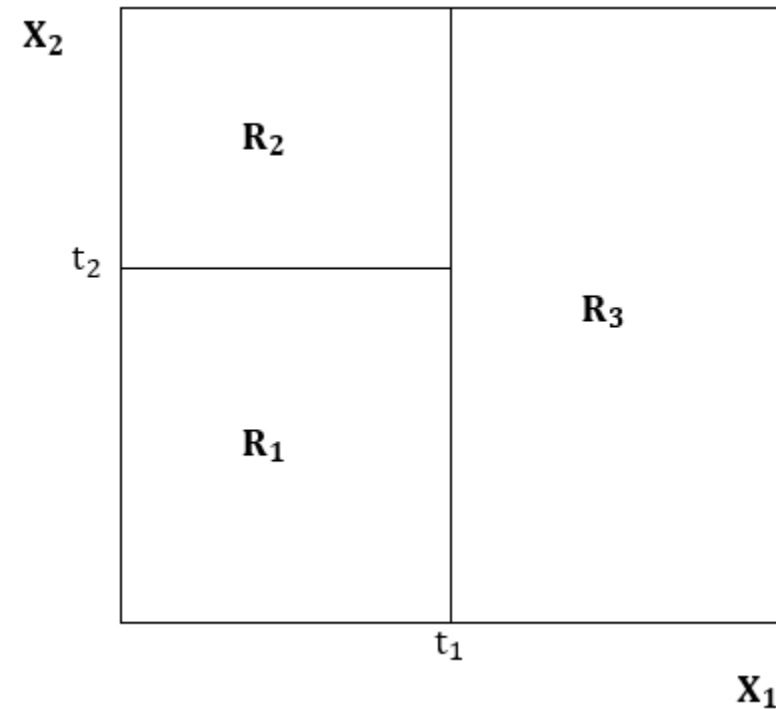
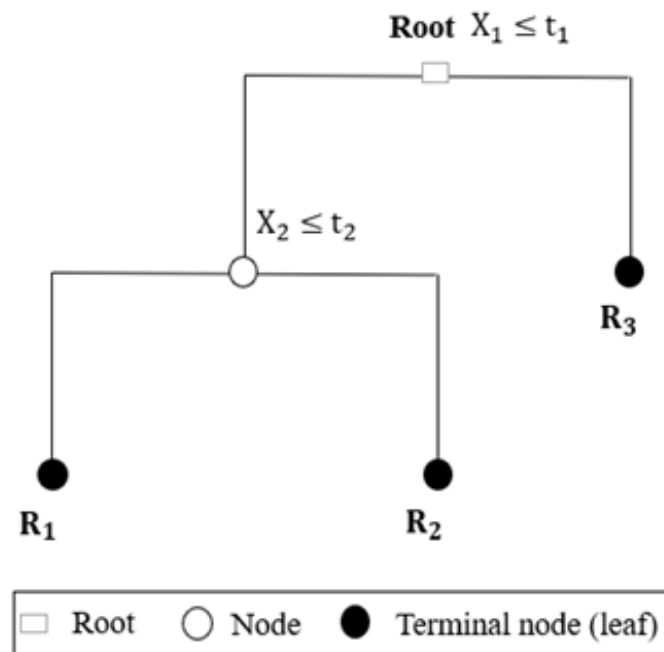
- $$\max_{\alpha_i} \sum_{i=1}^N \xi_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K_{\text{rbf}}(\mathbf{x}_i, \mathbf{x}_j),$$

- where $C \geq \alpha_i \geq 0$ and $\sum_{i=1}^N \alpha_i y_i = 0$

- α_i denotes the Lagrange multipliers, and $\{\mathbf{x}_i \mid C > \alpha_i > 0, \forall i\}$ are the support vectors deciding the decision boundary.
- C is an upper bound of ξ_i in this kernel SVM optimization setting.
- In addition, C and γ are two hyperparameters of SVM.

Research Topic 2

- (Base model 2) Decision tree (DT)
- The decision tree (DT) model recursively partitions the feature space into a disjointed set of rectangle regions such that each region contains the same classes



Research Topic 2

- (Base model 2) Decision tree (DT)
- The decision tree (DT) model recursively partitions the feature space into a disjointed set of rectangle regions such that each region contains the same classes
- (Pros) The decision tree is simple, interpretable, applicable for regression and classification with continuous and/or categorical variables, and acceptable for a dataset containing missing values.
- (Cons) the decision tree has high variance due to its hierarchical structure so that a small change of features can cause different split results. Further, the classification of the DT on imbalanced data could be biased toward the majority class.
- Therefore, the tree ensemble models (random forest and extreme gradient boosting) are applied to mitigate these problems

(Tree ensemble model 1) Random forest (RF)

- Ensemble methods use a set of base classifiers. The random forest (RF) is a tree ensemble model called bootstrap aggregating.
- The RF is able not only to improve the prediction performance by reducing variation but also to maintain robust prediction performance with an increasing number of noisy variables (Friedman, Hastie, and Tibshirani 2001.)
- Breiman (2001) argued that the RF's prediction performance depends on individual DTs' performance and the correlation between DTs.

(Tree ensemble model 1) Random forest (RF)

- The RF's procedure is: (1) generating an independent training set s_i by selecting a subset of the sample from training set S with replacement;
- (2) creating de-correlated RF rf_i , by selecting a subset of features;
- (3) training rf_i with s_i and using fitted rf_i to classify new data x ; and
- (4) repeating the above steps B times and classifying new data by using majority voting as follows:

$$\hat{y} = \frac{1}{B} \sum_{i=1}^B rf_i(x; \theta_i)$$

(Tree ensemble model 2) Extreme gradient boosting (XGB)

- Boosting combines multiple *weak classifiers* to build a *strong classifier*.
- Extreme gradient boosting (XGB; Che, and Guestrin 2016) implements gradient boosting (Friedman 2001) by regularizing the complexity of the tree structure.
- The prediction of a tree ensemble model is the sum of K DTs:

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F$$

- where $F = \{f(x) = w_{q(x)} \mid q(x) \in \{1, \dots, T\} \text{ and } w \in \mathbb{R}^T\}$

(Tree ensemble model 2) Extreme gradient boosting (XGB)

- Each DT has an objective function (OF). A smaller value of the OF means a better tree structure.
- The OF (= training loss + regularization term):

$$= \sum_i^N L(y_i, \hat{y}_i) + \sum_{k=1}^K \left[\gamma T + \frac{1}{2} \lambda \|w\|^2 \right]$$

Topic 2

(Tree ensemble model 2) Extreme gradient boosting (XGB)

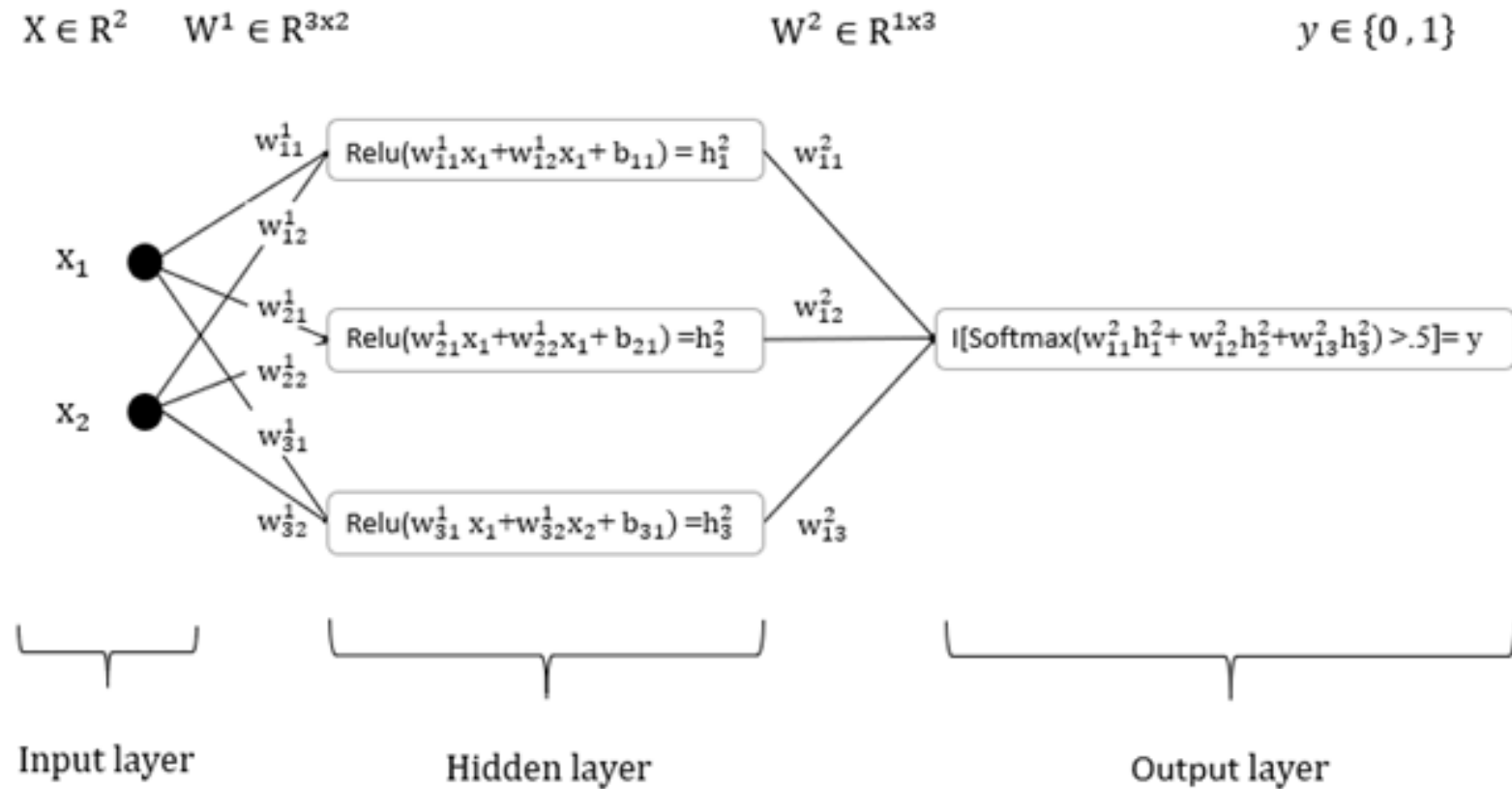
- additive training is applied for the optimization by adding a new function $f_t(\mathbf{x}_i)$ in each iteration t and using second-order Taylor approximation:

- $$OF^{(t)} \approx \sum_i^N L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) + \sum_{k=1}^K \left[\gamma T + \frac{1}{2} \lambda \|w\|^2 \right]$$

- where
$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \text{ and } h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)^2}}$$

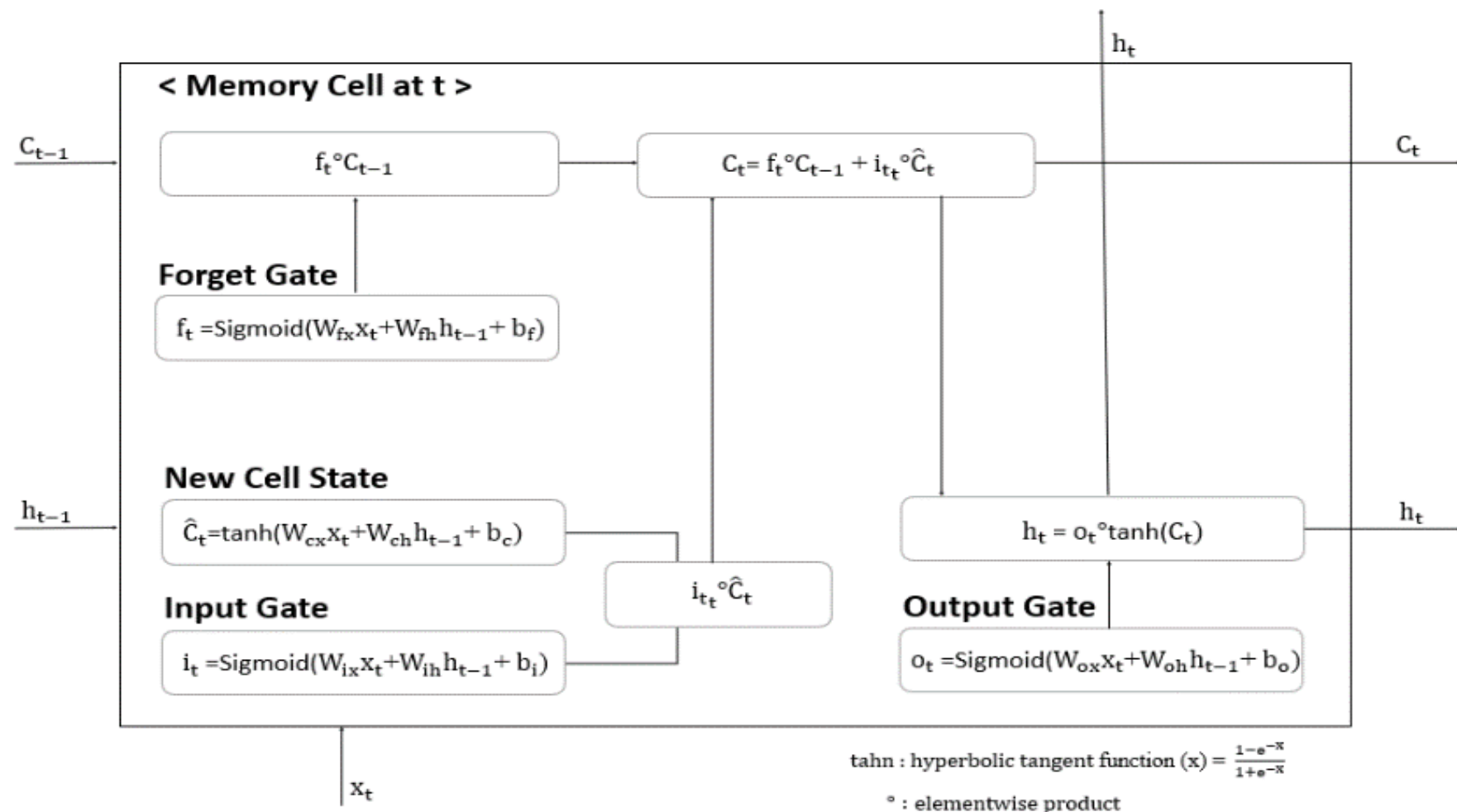
Topic 2

- (Deep learning model 1) Artificial neural network (ANN) example



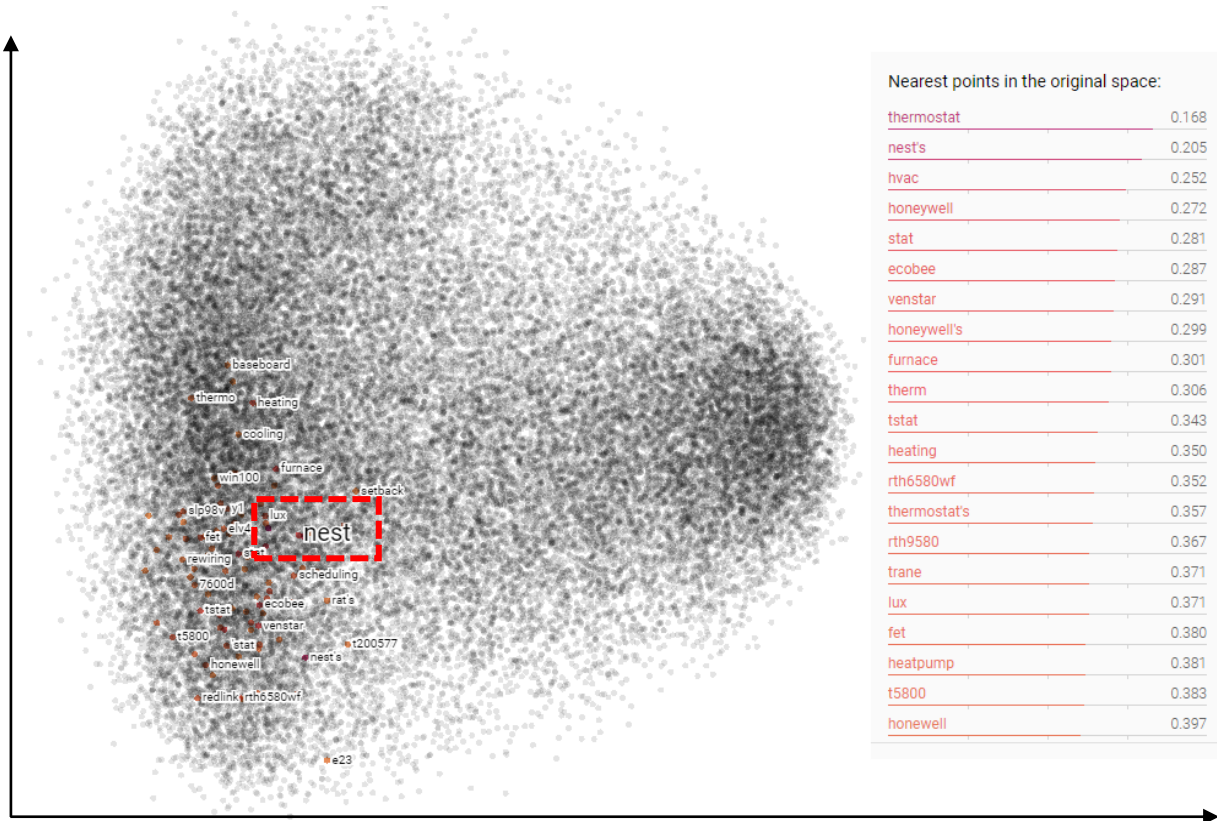
Topic 2

- (Deep learning model 2) LSTM



TOPIC 3

- Example: W2V trained on all reviews in the home category (D=100, W=5)



- PCA

- Reduce dimension 100 to 2 (PC1 and 2)

- Cosine similarity = $\frac{A \cdot B}{\|A\| \|B\|}$

TOPIC 3

- **Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018)**
- Using **WordPiece** tokenizer (relax out-of-vocabulary problem) e.g., embedding → em# bed #ding
- 800 million words using a **book corpus** (Zhu et al. 2015) and 2,500 million words from **Wikipedia**
- Using **12 layers of transformer encoders** and **12 multi-head attention** (self-attention)
- **masked language modeling (MLM)** and **next sentence prediction (NSP)**
- Using "the BERT-base model" (30,522 unique tokens with 768 embedding dim) → Roberta or Longformer (future)
- Using **fine-tuned BERT** and **further-pre-trained BERT** (idea from ACL 2020 best paper)