Online Product Review with Country Information

Ruiqi Zhu, Zexing Song, Haoliang Jiang
Georgia Institute of Technology
Atlanta, Georgia 30332

Abstract

We study the country information present in online product reviews and we approach the study from two angles. First, we want to examine whether stereotypes occur when reviewers mention country in their reviews. Second, we study whether country information in reviews affect review helpfulness. Furthermore, we explore the motivations behind why reviewers mention countries. We fine-tune a pretrained BERT model based on 3700 manually annotated reviews to classify the reviews and predict the motivation when reviewers mention country. We utilize the Amazon review data and use econometric approach, named entity recognition, bag of words with TF-IDF, word2vec and BERT for processing data and analyzing the results. We identify some preliminary evidence for country stereotypes by looking at word frequency and also the distance between words in the reviews of Amazon U.S. before 2014. We also find that country information in reviews affect review helpfulness. In addition, we demonstrate the probability for using the BERT model to classify reviewer's motivation in mentioning the country-related key words. Our code and our manually annotated dataset are available at $https://github.com/haoliangjiang/CS6471\_project$.

Introduction

E-commerce has fundamentally transformed how consumers purchase products and one of the most unique features of E-commerce is the availability of online reviews. When you purchase something on Amazon or Taobao, you can browse detailed reviews by other consumers who have bought the same product. The reviews serve as essential information on the product content and quality and plays an integral role in the purchase decision of the consumer. When looking at the reviews, country information shows up frequently. For example, reviewers would mention that the product is made in China, is of U.S. quality, or that it is made with Japanese standard. The number of reviews that include information on country is staggering, and we are want to understand the role that country information plays in the reviews.

This project approach the question from two directions. First, we want to examine if stereotypes are present when people mention countries in their shopping reviews. In this project, we adopt the standard definition for stereotypes that are "beliefs about the characteristics, attributes and behaviors of certain groups" (Chattalas, Kramer, and Takada 2008). Stereotype for a country is a oversimplified view for the country and its products during purchasing(Hinner 2010). Stereotype for countries can be explained as a result from a need to reduce the world's complexity to a manageable level whose effect are also known as country-of-origin effect(Chattalas, Kramer, and Takada 2008). Stereotypes can be positive or negative. On the one hand, stereotypes can help people make easy decision on purchasing. On the other hand, it leads to irrationality in the purchase decision. For instance, reviews that mention China might signal that the product is a knockoff or of poor quality while reviews that mention U.S.A. might signal superior product quality. Several studies have studied the variables for the country of origin effect (Godey et al. 2012; Yang, Ramsaran, and Wibowo 2016; Chattalas, Kramer, and Takada 2008). They establish hypothesis and then utilize regression analysis to analyze the variables' influence. To the best of our knowledge, there is no existing research about extracting stereotypes for country of origin in online shopping review. We hope that our work from the lens of country information in online reviews could shed some light on the problem of identifying stereotypes for country of origins.

Second, we would like to study if including country information would affect the review helpfulness and this could help us answer if country information is useful for online shoppers who look at reviews. There is a rich literature on online product reviews and researchers have studied it from many different directions. One work that is related to our question studies the factors that make an online review helpful (Mudambi and Schuff 2010) and the authors also used the Amazon review dataset. They divided products into search goods and experience goods and explored how review extremity, review depth, and product type affect the perceived helpfulness of the reviews. Our work also studies the factors that impact perceived helpfulness of the reviews and we use the content in the reviews, most specifically the country information included in the review. Another interesting work studies the strategies that online reviewers adopt to compete for attention. They found that online reviewers do behave strategically in order to compete for scare attention and enhance reputation, and their decisions are determined by the dynamic review environment at the time they review the product (Shen, Hu, and Ulmer 2015). Even though our research does not directly study online reviewer strategy,

mentioning country in reviews could also be a strategy that is adopted by online reviewers. Works by Mayzlin et al. (Mayzlin, Dover, and Chevalier 2014) investigates promotional reviews, where firms' incentives to manufacture biased reviews impede review usefulness. The researchers study how hotels create promotional reviews with the incentive of getting a higher rating and compare the rating difference for the same hotel on Expedia.com and TripAdvisor.com. This work is a great example of biased reviews and gives us some insights on why biased reviews are taking place, which is helpful in understanding country stereotypes in online reviews. However, we find that there is few research that directly analyze how information of product related to country present in the review text such as country of origin influences the review helpfulness.

To answer the above two research questions, we tap into the existing amazon review dataset, amazon review 2014 for review text analysis. To extract the country of origins information, we utilize the following strategy: we first use the named entity recognition (NER) tool (Manning et al. 2014) to extract country candidates in each review text; then we utilize a BERT model (Devlin et al. 2018) to indicate the purpose of reviewers putting the candidate word in their reviews. We identify the purpose candidates by examining 1500 samples and annotating 3700 reviews. The pre-trained BERT model is fine-tuned based on the annotated data. With the refined review dataset for country of origins, we apply bag of words with term frequency–inverse document frequency (TF-IDF) and word2vec (Mikolov et al. 2013) to qualitatively identify the stereotypes for countries in text. Furthermore, the country information are utilized for regression analysis on the review helpfulness and rating. In summary, our work is three-fold:

- We extract reviews with country of origin information in review text by NER and BERT
- We qualitatively identify strong stereotypes for country of origins in review text
- We analyze the relationship between country information in review text and review helpfulness

In the following sections, we will discuss the amazon review 2014 dataset and our proposed methods. After that, we will describe the experiment configuration, our annotation process, and final results. Finally, we will conclude our work and discuss the limitations.

Methods

Dataset

We leverage the amazon dataset that contains reviews on Amazon from 1996 to 2014. In the project, we will utilize a subset of the dataset due to the limitation of our computational resources. Compared to the entire dataset, the subset of the data filters out products and reviewers with less than 5 reviews. The information contained in each review includes: review ID, product ID, product category, reviewer name, helpfulness vote, review text, overall rating of the product, summary of the review, time of the review, and the image. The main information we leverage is the summary, the overall rating, the review text, and the helpfulness ratio as shown in figure 1. A review with its data format is shown in figure 12 in Appendix. We pick 10 categories from the entire dataset. Following the definition given in (Mudambi and Schuff 2010), we choose 4 categories in experience goods and search goods respectively as shown in table 1. With these 10 categories, we have 1,118,849 reviews. The number of review for each category is shown in figure 2. In addition, figure 2 also show the statistics of data in terms of review time and histograms of helpfulness and overall rating.

Figure 1: An example of the information used in our project

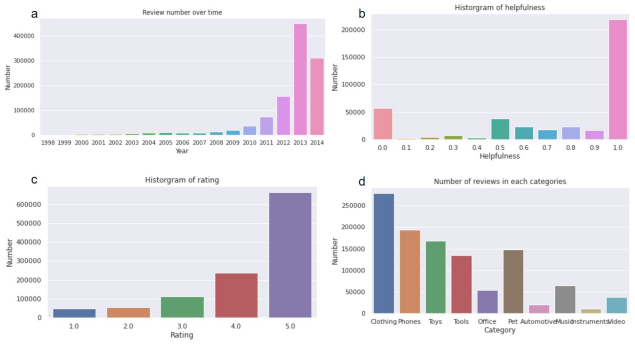

Figure 2: Summary statistics of the dataset. (a). How number of reviews changes over time. (b). The histogram of helpfulness ratio. (c). Histogram of rating. (d). Number of reviews in each categories.

Table 1: The categories from amazon review 2014 used.

Search Goods	Experience goods
cell phone	digital music
office products	instant video
automotive	games and toys
tools	
clothing	
musical instruments	
pet supplies	

Furthermore, we annotate 3700 samples in the above dataset with labels on why reviewers mention country in their reviews. The dataset statistics and analysis are discussion in the annotation section.

Proposed Methodology

In this subsection, we introduce our methods for data extraction, purpose classfication, and analysis methods for RQ1 and RQ2.

Data Extraction and Post-process To identify the words referring to countries in the reviews, we utilize Stanford CoreNLP (Manning et al. 2014) library to complete the task of named entity extraction. Figure 3 gives an example when the review sentence is "There is a different between Galaxy II that is carried by USA and what's carried by Asia." We extract mainly two types of entities from the results, country and nationality. Then, we manually find the words or phrases that refer to the same country into the same category. For example, 'us', 'usa', 'the states' and 'the united states' related reviews are all treated as reviews related to 'usa'. We furthermore filter out countries and nationalities with less than 200 reviews in the dataset. As a result, we have 15 countries and 17 nationalities left in the dataset.

Word	NER
[Sentence 1]	
There	O
is	O
a	O
different	O
between	O
Galaxy	O
II	O
that	O
is	O
carried	O
by	O
USA	COUNTRY
and	O
what	O
's	O
carried	O
by	O
Asia	CITY
.	O

Figure 3: An example of a review after NER using Stanford CoreNLP Toolkit.

Purpose Classification We utilize BERT to identify reviewers' purposes in mentioning a country in the review. The pretrained BERT is fine-tuned on our annotated data. As shown in figure. 4, BERT takes a review text as input and generates the embedding for each token. We take the average embeddings for the tokens representing the given country-related words by masking out the representations of other tokens and apply an extra linear layer for classification. After verifying the results of the classifier, we apply it to the data for analysis in RQ1. Both the analysis results with or without the classifier are conducted for comparison between the results.

RQ1 To investigate the stereotype people might have for countries in their shopping reviews, we utilize two word-level methods, TF-IDF and word2vec. We use TF-IDF to see the most frequent words in the reviews related to specific countries to see the differences in the usage of words. Word2vec helps us see the most representative words for

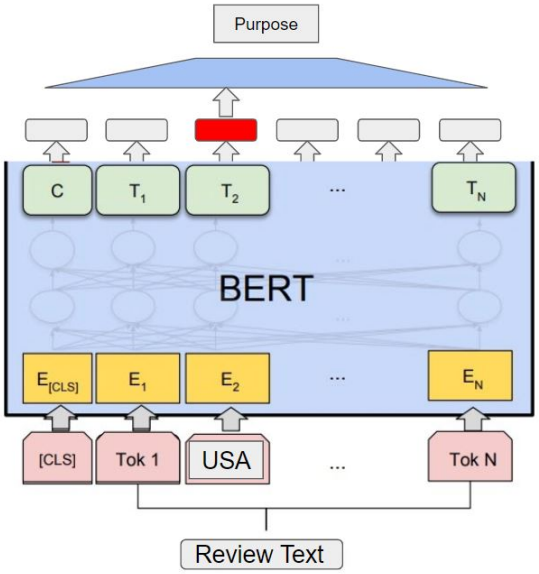

Figure 4: The architecture of our BERT model. The masked embeddings are shown as grey blocks and the activated embeddings are shown as a red block. After averaging the embeddings, a linear layer and a logsoftmax layer are applied upon vanilla BERT to get the scores for classification.

countries by looking at the words around the country-related word in a small window.

To qualitatively investigate the results, we compare the frequency of words and the similarity of words between specific countries in the results for TF-IDF and word2vec. For TF-IDF, we mainly look into words related to product type, concerns about products and sentiment-related words. For word2vec, we mainly look into sentiment-related words and also the t-SNE visualization for similar words to countries in word-embedding space. For sentiment-related words, we utilize the positive and negative dictionaries from (Hu and Liu 2004)

Empirical Model For our second research question, we want to examine whether country information embedded in online product reviews are correlated with the helpfulness of review and the product rating. We address our question using the same Amazon review dataset and we extracted new variables in the dataset that includes the review length, number of unique words, and various class variables that indicate whether certain keywords on country and nationality are mentioned in the review. We ignore the information on when the review was made, making our data cross sectional and then perform linear regression on the data. The specific description of the control and dependent variables are presented in figure 5.

In addition, since there is a number of product categories on Amazon and the review can be drastically different for diverse product categories, we adopt paradigm from information economics and divide the product categories into two types: search goods and experience goods (Mudambi and

Variable	Description
Control Variables	
review_length	number of words in the review
number_of_unique_word	number of unique words in the review
summary_length	number of words in review summary
number_of_unique_word_in_summary	number of unique words in review summary
no_country	Binary variable that equals 1 if no country-related keywords are mentioned in the review and 0 otherwise
no_nationality	Binary variable that equals 1 if no nationality-related keywords are mentioned in the review and 0 otherwise
china	Number of times keyword "china" is mentioned in the review
chinese	Number of times keyword "chinese" is mentioned in the review
american	Number of times keyword "american" is mentioned in the review
us	Number of times keyword "us" is mentioned in the review
*review_length*china*	Interaction of the two variables
*number_of_unique_word*china*	Interaction of the two variables
*review_length*us*	Interaction of the two variables
*number_of_unique_word*us*	Interaction of the two variables
*review_length*chinese*	Interaction of the two variables
*number_of_unique_word*chinese*	Interaction of the two variables
*review_length*american*	Interaction of the two variables
*number_of_unique_word*american*	Interaction of the two variables
Dependent Variables	
rate	Product rating that the review gives and is a number from 1-5
helpfulness_ratio	Number of up-votes that consider the review as helpful and the ratio ranges from 0 to 1. Value becomes -1 when there are only down-votes.

Figure 5: Variable description for the empirical model

Schuff 2010). The specific division is presented in Table 1.

Results

Experiment Setup

Configurations and Implementations In the training of BERT, we utilize the vanilla BERT model with pre-trained parameters. We utilize 500 samples as evaluation data and fine-tune our model on the remaining 3200 samples. We train the model with a learning rate of $1e^{-6}$, batch size of 4, and the adam optimizer. The maximum sequence length of model input is 512. For the vocabulary of TF-IDF, words appear greater than 10 times are considered. For vocabulary of word2vec, words or phrases from unigram and bigram that appear more than 10 times and 30 times respectively are considered. The dataset we apply TF-IDF to is the country dataset filtered by BERT. The dataset we apply word2vec to is the country and nationailty dataset filtered by BERT.

Annotation We design the labels for purposes after examining around 1500 reviews and summarizing the motivation of the reviewer. We propose five categories of purposes according to the context of reviews as shown in table 2.

As was stated above, we group the reviews that mention countries into five main categories. First, for phrases like "...made in...", we could classify those reviews into *origin*. Second, if some key words such as "...Shoes US 10...", we would classify those reviews into *Attributes*; Next, if the usage of certain products or their function description is related to the country, those should be in *Place of usage*. In cases other than the three mentioned, we classify the review into *Others*. Finally, some key words do not refer to specific region or city such as "Captain America", so those groups of reviews are assigned to *Ignore*.

Table 2: The purposes for a reviewer to mention a country

Name	Description
Origin	Place of production or place of origin
Attributes	Attributes of product
Place of usage	Where the product is used and consumed
Others	Other purposes
Ignore	The word is not referring to a country

Table 3: The summary statistics of the annotate dataset.

	Origin	Attribute	Place of usage	Others	Ignore
USA	705	388	209	326	77
China	829	13	10	11	35
UK	70	84	89	109	13
Japan	44	23	13	4	0
Germany	46	8	13	5	0
Canada	22	11	20	7	0
Korea	14	1	1	5	0
Mexico	39	13	18	14	0
All	1915	597	497	600	128

The figure 6 gives an example of our custom-designed annotation interface. Each review was read from the *.Json* files and related information in Jupyter Notebook. We manually assign those reviews into the above five categories and save their types into a new *.Json* file for further model training and analysis.

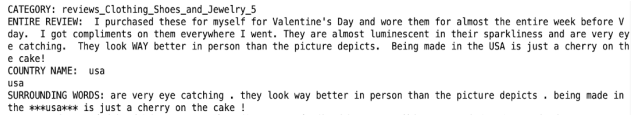

Figure 6: Annotation Interface

From the sample annotation interface, the category, entire review text, keywords and the surrounding words are clearly shown. The annotator is able to label this review to one of the purposes based on the give information.

With the efforts from two our our team members, we annotate around 3700 reviews before the deadline of the final report. In table 3, we give the statistics for the annotation results in terms of each class. The histogram of reviews in terms of countries and their number in each class is given in the table as well. We conduct the *Welch Two Sample t-test* based on 200 samples for the annotation results from two annotators in order to check the consistency of our result from different annotators. Based on the test below, we can see that the $t - Statistics$ is 0.40693 and the $P - value$ is 0.6845 which is way larger than $\alpha = 0.05$. Therefore, we can not reject the Null hypothesis and conclude that there is no statistically significant difference of labeled results from different annotators.

Results

In this section, we our results for data pre-processing - including NER and BERT - and analysis for RQ1 and RQ2.

Preprocess When annotating the data, we introduce the class 'Ignore' since we found that in some cases, CoreNLP might identify false positive entities. Words or phrases such as 'us' or 'english breakfast tea' might cause difficulties to the system during inference. However, during our annotation, we verify that such cases are not as much as we expected as table 3. In the first row of table 4, we show the statistics about the number of reviews each country is related to.

After annotating the data with country information, we use the data to train our BERT model. Table 5 demonstrate the evaluation results of the best model. The overall micro-f1 score is 58.95% and the overall macro-f1 score is 70.43%. Based on the performance, we can see the model is doing a good job in identifying the class of product origin. This is mainly because of two reasons. First of all, the amount of data with this label is greater than half of the annotated data. Secondly, by calculating the word frequency in reviews related to this class, we found that 'made' is a very common word in these reviews. This might make the classification of this class easier than other classes. Based on the situation that the f1-score for other classes is not ideal enough to bring out convincing labels when applied to the whole dataset, we decide to use the fine-tuned BERT as a binary classifier to identify whether the review is to intended to discuss the products' origin when mentioning the country.

After applying the model to the entire dataset, table 4 demonstrates the statistics for the top countries in the 'products' origin' dataset which will be used in the analysis of our RQ1. For countries like UK and Japan, their experience goods such as music and videos are popular. These goods typically bring in more reviews about the attributes of the product. Thus, they lost many of their reviews after classification. However, countries like China and Korea are more concentrated in search goods such as cell phones. This could probably explain why a majority of reviews on those two countries are about products' origin.

RQ1 We apply TF-IDF to capture frequent words reviewers use in country-related reviews. The results for 8 most common countries is shown in table 6. We mainly consider words related to product types, people's concerns about the product and sentiment words by dictionary proposed by (Hu and Liu 2004). We find that for product type, it mainly infers what kind of product are popular on Amazon for that country. For concerns on product and sentiment-related words, people are using similar words for different countries.

The frequent words for product type mainly reflects the most popular product type in U.S. for that country. For example, phone made in Korea, jeans made in Mexico and album from the UK. Countries share similar words of concerns including, quality, price and size. Since the results are filtered out by BERT and for countries like Korea or Canada, a small number of reviews are collected. It is hard to indicate the focus of people's concern by the order. Also, the order might vary between products. For the sentiment-related words, the most frequent words people use in reviews are very similar, including words like 'great', 'like' or 'good'. Some differences exists between the three Asian countries and others. We see word 'cheap' for China, 'classic' for Japan and 'thicker' and 'issue' for Korea.

For word2vec, we care more about the words which are most representative (similar) to the country. First, we utilize the same sentiment dictionary and get the top words for countries as shown in table 7. We also calculate the ratio of positive sentiment-related words in the top 50 sentiment-related words. The similarity score of the words involved is greater than 0.35. The results show that for countries focusing on manufacturing such as China, Korea and Mexico, they are more often mentioned with negative words. Especially for China, extreme words like 'junk' and 'trash' are highly related. For countries like 'UK' and 'Germany', very high positive ratios are observed. Thus, we conclude that a large difference in word patterns exists when reviewer mention different countries as the product's origins. For countries like 'China', 'Mexico' and 'Korea', the comments tend to infer connection between the poor product quality and the country more. While mentioning countries like 'Germany' or 'UK', people tend to use positive words. Possible explanations for the phenomenon might be consumer ethnocentrism, country image, and cultural difference (Yang, Ramsaran, and Wibowo 2016). The t-SNE visualizations are shown in figure 7 below and figure 13 in Appendix. The results demonstrate similar conclusions to the sentiment-related word analysis we have here. From the word space visualization, it is obvious to see there are strong stereotypes associated with different countries' products. For the visualization of Germany, we can see many words such as experienced, engineer, precision and professional. However, in the word space for China, many negative works like waste money, negative review, trash, and piece junk are used.

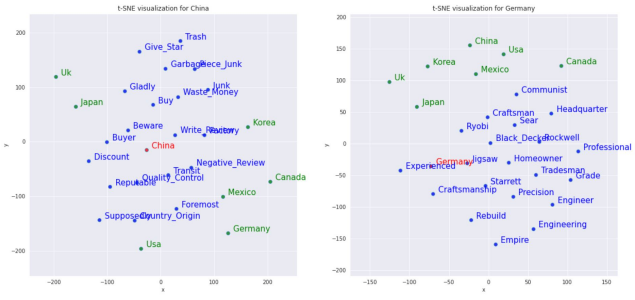

Figure 7: The t-SNE visualizations of word space for China (low sentiment positive rate) and Germany (high sentiment positive rate).

RQ2 First, we use helpfulness ratio as the dependent variable and run the regression on all the data (shown in figure

Table 4: The countries with most reviews and the number of reviews in different stages.

	US	China	UK	Japan	Germany	Mexico	Canada	Korea	All
all country-related	14333	7495	3328	807	579	708	553	177	28960
products' origin	4600	5799	335	299	311	288	168	108	14751

Table 5: Evaluation results of BERT on classifying reviewers' purposes on mentioning countries.

	Origin	Attributes	Place of usage	Others	Ignore
Recall (%)	89.05	54.84	64.15	28.13	53.33
Precision (%)	89.4	55.74	55.73	30	72.73
F1-Score (%)	89.26	55.28	59.65	29.03	61.54

8) using control variables listed in the figure 5. We observe that the coefficient estimate for review length is negative and the estimate for number of unique words in the review is positive, both of which are statistically significant. This supports findings in the literature and also further justifies the validness of our data and result. What is interesting from our result is that mentioning country in the review appears to contribute to the helpfulness of the review because the estimation result for variables related to country information are all positive and statistically significant. In addition, we also find that country information has a significant positive moderating effect on review length and a significant negative moderating effect on number of unique words in the review. It is possible that since country information is helpful, having a longer review would increase the helpfulness of the review. The effect of country information on number of unique words seems rather perplexing and requires further exploration. One potential explanation is that when country information is present in the review, there is a common theme or reason for mentioning the country, like where the product is manufactured, and discussing something irrelevant to that theme could jeopardize the helpfulness of the review.

We performed the same analysis by dividing our data into search goods and experience goods, and presented our results in figure 9. One important observation is that country information in reviews are considered useful only for search goods but does not make much difference for experience goods. This can be seen from the estimate of the variable no_country and no_nationality, which is negative and statistically significant for search goods but not significant for experience goods. One potential explanation is that when country information is mentioned for search goods, it is usually related to where the product is made or where the company is from, and that information could be an important quality signal. On the other hand, when country information is mentioned for experience goods like videos or music, it is related to the content of the good itself. For instance, U.S. is mentioned a lot in videos because the movie is about a story that took place in the United States. In that case, the country information does not play such a critical role in product quality information.

Furthermore, we break down our sample further by performing analysis on the sub-sample divided by product cat-

	helpfulness_ratio		
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	-0.73	0.01	**<0.001**
review_length	-0.00	0.00	**<0.001**
number_of_unique_word	0.01	0.00	**<0.001**
summary_length	0.00	0.00	0.846
number_of_unique_word_in_summary	0.00	0.00	0.286
no_country	-0.03	0.01	**<0.001**
no_nationality	-0.01	0.01	0.090
china	0.16	0.02	**<0.001**
chinese	0.19	0.03	**<0.001**
american	0.23	0.02	**<0.001**
us	0.19	0.01	**<0.001**
review_length * china	0.00	0.00	**<0.001**
number_of_unique_word * china	-0.00	0.00	**<0.001**
review_length * us	0.00	0.00	**<0.001**
number_of_unique_word * us	-0.00	0.00	**<0.001**
review_length * chinese	0.00	0.00	**0.047**
number_of_unique_word * chinese	-0.00	0.00	**0.004**
review_length * american	0.00	0.00	**<0.001**
number_of_unique_word * american	-0.00	0.00	**<0.001**
Observations	1118849		
R^2 / R^2 adjusted	0.106 / 0.106		

Figure 8: Overall result with helpfulness ratio as DV

	Search Good	**Experience Good**
Predictors	*Estimates*	*Estimates*
(Intercept)	-0.71 ***	-0.71 ***
review_length	-0.00 ***	-0.00 ***
number_of_unique_word	0.01 ***	0.01 ***
summary_length	-0.01 **	0.02 ***
number_of_unique_word_in_summary	0.01 **	-0.01
no_country	-0.05 ***	-0.01
no_nationality	-0.02 *	-0.00
china	0.17 ***	0.15 **
chinese	0.14 ***	0.37 ***
american	0.18 ***	0.22 ***
us	0.14 ***	0.26 ***
review_length:china	0.00 ***	0.00
number_of_unique_word:china	-0.00 ***	-0.00
review_length:us	0.00 ***	0.00 ***
number_of_unique_word:us	-0.00 ***	-0.00 ***
review_length:chinese	-0.00	0.00 **
number_of_unique_word:chinese	-0.00	-0.01 ***
review_length:american	0.00 **	0.00 ***
number_of_unique_word:american	-0.00 ***	-0.00 ***
Observations	849420	269429
R^2 / R^2 adjusted	0.085 / 0.085	0.147 / 0.147

** p<0.05 ** p<0.01 *** p<0.001*

Figure 9: Result by product type with helpfulness ratio as DV

Table 6: Frequent words in country-related reviews for 8 most common countries by TF-IDF.

	Product type	Concerns of product	Sentiment
USA	dog, food, album, chicken	quality, price, size	like, good, love, great
China	phone, battery, dog, food	quality, price, size	like, good, cheap, well
UK	album, song, music, guitar	material	best, great, love, like
Japan	album, battery, song, guitar	quality, price, material	great, like, classic, best
Germany	puzzle, game, tool, pencils	quality, color, price	well, like, good, best
Canada	food, album, dog, song	price, quality, size	good, like, better, well
Korea	battery, phone, charger, screen	material, quality, price	thicker, well, issue, like
Mexico	jeans, tool, pencils, battery	size, quality, price	good, like, well, better

Table 7: The more similar sentiment-related words for each country of word2vec and the corresponding rate of positive sentiment words in the top 50 words.

	Sentiment-related words	Positive rate
USA	taint, strict, assure, trustworthy	0.4
China	beware, junk, trash, foremost	0.22
UK	sexy, joy, upbeat, heaven	0.74
Japan	fidelity, humble, integral, inaccurate	0.5
Germany	unacceptable, skill, versatility, rapid	0.66
Canada	pride, trustworthy, wary, strict	0.46
Korea	sleek, genuine, shocked, fake	0.38
Mexico	inferior, inconsistent, shoddy, rant	0.28

egory in figure 11. From the result, we see that country information is extremely helpful for product categories like phones, clothes, and office products, but are not helpful for product categories like videos. Moreover, when keywords on China or U.S. are mentioned, the review becomes more helpful overall.

The same analysis were performed using product rating as the dependent variable. Results are calculated using all the sample and also dividing the sample by product categories, which is presented in figure 10, 14, and 15. We see that not mentioning country information is correlated with a lower product rating. Furthermore, keywords on China leads to a lower product rating than mentioning U.S. We think that the R^2 result for the regression is too low to give any meaningful interpretation of the coefficient estimate.

Work Divisions

Haoliang Jiang: Literature review, data process, data analysis, NER, RQ1, variable generation for RQ2, and annotation preparation, BERT, and result analysis for purpose classification.

Zexing Song: Regression analysis for RQ2, data annotation for purpose classification.

Ruiqi Zhu: Literature review, entire RQ2, data annotation for purpose classification.

Conclusion

In this project, we first utilize NER and BERT to extract reviews with country of origin mentioned in the review text. Then we conduct bag of words with TF-IDF to check frequent words and word2vec to find the most representative

	rate		
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	4.62	0.01	**<0.001**
review_length	0.00	0.00	**<0.001**
number_of_unique_word	-0.00	0.00	**<0.001**
summary_length	-0.02	0.00	**<0.001**
number_of_unique_word_in_summary	-0.01	0.00	**0.046**
no_country	-0.09	0.01	**<0.001**
no_nationality	-0.10	0.01	**<0.001**
china	-0.71	0.02	**<0.001**
chinese	-0.69	0.05	**<0.001**
american	-0.06	0.02	**0.016**
us	-0.06	0.02	**<0.001**
review_length * china	-0.00	0.00	**0.008**
number_of_unique_word * china	0.00	0.00	**<0.001**
review_length * us	-0.00	0.00	**<0.001**
number_of_unique_word * us	0.00	0.00	**<0.001**
review_length * chinese	-0.00	0.00	0.051
number_of_unique_word * chinese	0.00	0.00	**0.001**
review_length * american	-0.00	0.00	0.178
number_of_unique_word * american	0.00	0.00	0.096
Observations	1118849		
R^2 / R^2 adjusted	0.010 / 0.010		

Figure 10: Overall result with product rating as DV

words. We find that strong stereotype exists among the review text. For countries such as 'China', 'Mexico' and 'Korea', people are more leaning toward using negative words when mentioning the country as the products' origin. On the other hand, when countries like 'UK' and 'Germany' are mentions as origins, the ratio of positive words in the text is much higher. The visualization of word space demonstrates a similar tend.

By performing linear regression on the review data, we find that country information in the review has a positive and statistically significant relationship with the helpfulness of the review. Even though review length has a negative relationship with review helpfulness, when country information is mentioned, a longer review becomes more helpful. In addition, country information is more helpful for search goods (e.g. clothes, office products) compared to experience goods like videos.

There are several limitations we would like to highlight. First of all, we cannot infer whether people mentioning the origin of the product intend to provide information on products' quality. Though rare, people might mention the product's origin as a neutral statement and do not indicate any connection between the origin and the product quality. Secondly, we cannot infer the appropriateness of the review that builds a connection between the country and the product because difference in consumer preference is not an objective measure of product quality. For example, the tastes of the same type of fruit from different countries might be different because of the weather. This is part of the reason why we do not include the category, food, and grocery in our analysis. However, there is no guarantee that similar cases do not exist in other categories as demonstrated in previous work that product type is important in terms of the country-of-origin effect (Chattalas, Kramer, and Takada 2008). Thirdly, to simplify the classification task, we do not include some main purposes such as comparison or emphasis. Ignoring these classes can also bring noise into the analysis. In future work, an important task will be to refine our BERT model in term of its data amount and functionality. Studying more quantitative and stable metric for stereotype identification could also be a promising direction. Furthermore, it will be interesting to explore how the word patterns and the influence of country information change over time and tie it back with the findings in our project.

References

Chattalas, M.; Kramer, T.; and Takada, H. 2008. The impact of national stereotypes on the country of origin effect. *International Marketing Review* .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Godey, B.; Pederzoli, D.; Aiello, G.; Donvito, R.; Chan, P.; Oh, H.; Singh, R.; Skorobogatykh, I. I.; Tsuchiya, J.; and Weitz, B. 2012. Brand and country-of-origin effect on consumers' decision to purchase luxury products. *Journal of Business research* 65(10): 1461–1470.

Hinner, M. B. 2010. Stereotyping and the Country-of-Origin Effect. *China Media Research* 6(1).

Hu, M.; and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Mayzlin, D.; Dover, Y.; and Chevalier, J. 2014. Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review* 104(8): 2421–55.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Mudambi, S. M.; and Schuff, D. 2010. Research note: What makes a helpful online review? A study of customer reviews on Amazon. com. *MIS quarterly* 185–200.

Shen, W.; Hu, Y. J.; and Ulmer, J. R. 2015. Competing for Attention. *Mis Quarterly* 39(3): 683–696.

Yang, R.; Ramsaran, R.; and Wibowo, S. 2016. A conceptual model for country-of-origin effects. *Asia Pacific Journal of Advanced Business and Social Studies* 2(1): 96–116.

Appendix

	Video	Automotive	Phone	Clothing	Music	Instrument	Office Product	Pet Supply	Tool	Toys
Predictors	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*
(Intercept)	-0.87 ***	-0.60 ***	-0.68 ***	-0.71 ***	-0.15 ***	-0.74 ***	-0.53 ***	-0.78 ***	-0.59 ***	-0.79 ***
review_length	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***
number_of_unique_word	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***	0.01 ***
summary_length	0.01	-0.01	-0.01	0.01	0.01	-0.01	0.01	-0.02 **	-0.01	-0.00
number_of_unique_word_in_summary	-0.00	0.01	0.02 **	-0.01	-0.00	0.01	-0.01	0.03 **	0.01	0.01
no_country	0.06 *	-0.11	-0.05	-0.01	-0.00	-0.14	-0.15 ***	-0.03 *	-0.03	0.02
no_nationality	0.02	-0.02	-0.15 ***	-0.05 **	-0.02	0.02	-0.05	0.04 ***	0.01	-0.01
china	0.18	0.18	0.00	0.21 ***	-0.13	0.28	0.13	0.27 ***	0.17 ***	0.17 **
chinese	0.19	0.31	-0.00	-0.11	0.13	0.10	0.01	0.29 **	0.14 *	0.61 ***
american	0.10 *	0.25	0.24 *	0.22 **	0.22 ***	0.63 ***	0.03	0.09	0.15 *	0.08 *
us	0.11 *	0.12	0.17 ***	0.11 ***	0.20 ***	0.08	-0.15	0.05	0.25 ***	0.23 ***
review_length:china	-0.00	0.00	0.00	0.00 ***	-0.00	0.00	0.00	0.00	0.00 **	0.00
number_of_unique_word:china	0.00	-0.01	-0.00	-0.00 ***	0.00	-0.01	-0.00	-0.00 *	-0.00 ***	-0.00
review_length:us	0.00	0.00	0.00 **	0.00 *	0.00 ***	-0.00	-0.00 *	0.00	0.00 ***	0.00 **
number_of_unique_word:us	-0.00	-0.01	-0.00 ***	-0.00 **	-0.00 ***	0.00	0.00 *	-0.00	-0.00 ***	-0.00 ***
review_length:chinese	0.00	-0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.00	0.00	0.00
number_of_unique_word:chinese	-0.00	0.00	0.00	0.00	-0.00	0.00	0.00	-0.00	-0.00	-0.01 *
review_length:american	0.00	0.00	0.00	0.00	0.00 ***	0.01 ***	0.00	-0.00	0.00	0.00 ***
number_of_unique_word:american	-0.00	-0.01	-0.00	-0.00 *	-0.00 ***	-0.02 ***	-0.00	-0.00	-0.00	-0.00 ***
Observations	37126	20473	194439	278677	64706	10261	53258	157836	134476	167597
R^2 / R^2 adjusted	0.122 / 0.121	0.082 / 0.081	0.102 / 0.102	0.056 / 0.056	0.091 / 0.090	0.126 / 0.125	0.092 / 0.092	0.076 / 0.076	0.099 / 0.099	0.098 / 0.098

** p<0.05 ** p<0.01 *** p<0.001*

Figure 11: Result by product category with helpfulness ratio as DV

Table 8: The distribution of overall rating for the dataset after filtered by BERT for the most common countries.

	1	2	3	4	5
US	70	53	94	201	413
China	1115	639	916	1517	2656
UK	6	5	8	35	64
Japan	25	18	34	68	210
Germany	10	11	28	65	208
Canada	9	14	12	38	102
Korea	9	8	16	24	56
Mexico	35	35	43	65	132


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the piano.
He is having a wonderful time playing these old hymns.  The music  is
at times hard to read because we think the book was published for
singing from more than playing from.  Great purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Figure 12: The data example of Amazon review 2014

Figure 13: The t-SNE visualizations for USA, Canada, Mexico, Japan, Korea and UK.

	Search Good	**Experience Good**
Predictors	*Estimates*	*Estimates*
(Intercept)	4.59 ***	4.75 ***
review_length	0.00 ***	0.00 ***
number_of_unique_word	-0.00 ***	-0.01 ***
summary_length	-0.02 ***	-0.01 *
number_of_unique_word_in_summary	-0.01	-0.01
no_country	-0.07 ***	-0.11 ***
no_nationality	-0.11 ***	-0.08 ***
china	-0.70 ***	-0.64 ***
chinese	-0.68 ***	-0.55 ***
american	-0.15 ***	-0.03
us	-0.02	-0.07 *
review_length:china	-0.00	-0.00 *
number_of_unique_word:china	0.00 ***	0.00 ***
review_length:us	-0.00	-0.00 **
number_of_unique_word:us	-0.00	0.00 ***
review_length:chinese	-0.00	-0.00
number_of_unique_word:chinese	0.00 *	0.00
review_length:american	0.00	-0.00 ***
number_of_unique_word:american	-0.00	0.00 ***
Observations	849420	269429
R^2 / R^2 adjusted	0.009 / 0.009	0.016 / 0.016

** p<0.05 ** p<0.01 *** p<0.001*

Figure 14: Result by product type with product rating as DV

	Video	**Automotive**	**Phone**	**Clothing**	**Music**	**Instrument**	**Office Product**	**Pet Supply**	**Tool**	**Toys**
Predictors	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*	*Estimates*
(Intercept)	4.59 ***	4.80 ***	4.31 ***	4.69 ***	4.40 ***	4.71 ***	4.64 ***	4.66 ***	4.74 ***	5.00 ***
review_length	0.00 ***	0.00 ***	0.00 **	0.00 ***	0.00 ***	0.00 *	0.00 ***	0.00	0.00 ***	0.00 ***
number_of_unique_word	-0.01 ***	-0.00 ***	-0.00 *	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.00 ***	-0.01 ***
summary_length	-0.02	-0.07 **	-0.02 *	-0.01	0.00	-0.03	0.00	-0.01	-0.04 ***	-0.02 *
number_of_unique_word_in_summary	0.01	0.05 *	-0.01	-0.03 ***	-0.00	0.01	-0.02	-0.02 *	0.02 *	-0.01
no_country	-0.08 *	0.10	-0.04	-0.11 ***	-0.10 ***	-0.06	-0.05	-0.09 ***	-0.08 **	-0.14 ***
no_nationality	-0.01	-0.19 ***	-0.01	-0.12 ***	-0.08 ***	0.02	0.02	-0.13 ***	-0.08 **	-0.17 ***
china	-0.09	-0.81 ***	-0.56 ***	-0.91 ***	-0.02	0.01	-0.33 *	-0.82 ***	-0.71 ***	-0.82 ***
chinese	-0.23	-0.65 *	-0.71 ***	-0.58 ***	-0.20	-0.77 **	-0.70 *	-0.52 ***	-0.83 ***	-0.75 ***
american	-0.08	-0.62	-0.22	-0.53 ***	-0.02	0.31	-0.40	-0.01	-0.06	0.08
us	-0.08	0.14	-0.13 *	-0.20 ***	0.03	-0.07	0.13	0.17 ***	-0.05	-0.03
review_length:china	0.00	0.00	-0.00	-0.00 **	0.00	0.00	0.00	-0.00	-0.00	-0.00 *
number_of_unique_word:china	0.00	0.00	0.00	0.01 ***	-0.00	-0.01	-0.00	0.00 **	0.00 *	0.01 ***
review_length:us	-0.00	-0.00	-0.00	-0.00 **	0.00	-0.00	0.00	0.00	-0.00	-0.00 *
number_of_unique_word:us	0.00 *	0.01	0.00	0.00 **	-0.00	0.00	-0.00	-0.00 *	0.00	0.00 *
review_length:chinese	-0.00	0.00	-0.00	0.00	-0.00	-0.00 *	0.00	-0.00	-0.00	0.00
number_of_unique_word:chinese	0.00	-0.00	0.00	-0.00	0.00	0.01 *	-0.00	0.01	0.00	-0.00
review_length:american	-0.00	-0.01	-0.00	-0.00	-0.00	0.00	-0.00 *	0.00	0.00	-0.00
number_of_unique_word:american	0.00 *	0.03	0.00	0.00	0.00	-0.01	0.01 *	-0.00	-0.00	0.00
Observations	37126	20473	194439	278677	64706	10261	53258	157836	134476	167597
R^2 / R^2 adjusted	0.031 / 0.031	0.023 / 0.022	0.006 / 0.006	0.013 / 0.013	0.003 / 0.003	0.015 / 0.013	0.021 / 0.021	0.013 / 0.013	0.018 / 0.018	0.031 / 0.031

** p<0.05 ** p<0.01 *** p<0.001*

Figure 15: Result by product category with product rating as DV