Aspect Sentiment and Cultural Bias Mining from User Reviews of the Top Free-to-Play Games on the App Store

Anonymous ACL submission

Abstract

Game players generate different in-game behaviors patterns, social interaction habits and commentary feedback driven by various internal or external motivations (e.g. What do players find fun? Why do they find it fun? Does their feeling of fun change over time? What happens when they stop having fun? Does the sense of fun vary according to the demographic? is there an association between user’s gender and game they play?). In this paper, we analyze the top 50 free-to-download games on the Apple’s App Store. We utilize Aspect Based Sentiment Analysis (ABSA) approach to extract fine-grained information related to each game. We also utilize the name of a reviewer as a gender indicator. We then use this information to investigate cultural differences of the same games in three other English speaking countries: United Kingdom, Australia, and Canada.

1 Introduction

Why did we pick only mobile game? Gaming community takes pride in its enormous amount of users which bestows us with sufficient training data for domain specific analysis (Yannakakis and Togelius, 2018). Gaming industry is fast evolving, trending games are constantly changing, compared to other online communities, the gaming community is faster to accumulate reviews, faster to win or lose users, and faster to appear in multiple social media sites. In addition, games typically share same gamification techniques and designing principles (Deterding et al., 2011), such as visual or audio effects, leaderboards, badges, avatars, collecting gadgets, social media connections and so on (Hunicke et al., 2004). Such uniformity in turns provides uniformity in aspect mining. Moreover, gamification techniques has been adopted in many other human interacting systems (Werbach and Hunter, 2012). As a result, the knowledge learned from the gaming community (e.g., gamers) has the potential to be transferred to more real world scenarios and usages, which is why AI research labs are training AI agents on complex games (Silver et al., 2016; Mnih et al., 2015, 2013).

Different software domains can exhibit different phenomena. Industry has started mining reviews to better understand the its demand and supply relationship. In fact, mining app stores reviews has gain a lot of attention from researchers over the years (Genc-Nayebi and Abran, 2017). Many studies have found that user reviews contain a wealth of information that can be used to help requirement engineers to better meet user needs, notify developers of unexpected behavior, and inform other users about their experiences from using the application (Pagano and Maalej, 2013; Fu et al., 2013; Panichella et al., 2015; Khalid et al., 2015). Pagano et al. (Pagano and Maalej, 2013) found that user feedback (reviews and ratings) on the app store has a direct impact on the number of downloads that app has. Interestingly, they also found that the app store serves as a communication channel among users and with developers. In an effort to help reduce the manual effort in investigating these reviews, many machine learning approaches have been proposed. Neural attention-based aspect extraction techniques has been proven useful to automatically extract aspects from reviews. In addition, the extracted aspects can largely capture notions of user preferences (Mitcheltree et al., 2018).

Moreover, several researches that studied cultural differences have always utilize “active
crowd-sourced” approach (i.e., directly survey participants) to collect data for this type of study, such as in (Hofstede, 2001; Trompenaars and Hampden-Turner, 2011; Ayed et al., 2017; Hofstede, 2011). In our case, we used data that is readily available online (”passive crowd-sourced approach”) - user reviews - to study cultural differences between gamers from different countries.

In this paper, we analyzed reviews of the top 50 free-to-play apps on the App Store. We utilized multiple information extraction and machine learning approaches to extract fine-grained information from these reviews. We then used this information to study the cultural differences of the same games in different countries.

The rest of the paper is organized as follows. In Section 2, we describe our data collection and annotation process and result. In Section 3, we discuss our models and experiment setup. In Section 4, we present downstream application - cross-cultural study. We highlight related work in Section 5. Finally, in Section 6, we conclude the paper and outline plans for future research.

2 Datasets and Annotation

In this section, we describe our approach for collecting the reviews of games used in this study.

2.1 App Selection

We collected reviews from the top 50 free-to-download games in the US Apple’s App Store on September 4, 2018. We selected top apps because we believed that the developers of these apps want to continue maintaining and improving the app and that these apps often contain a large quantity of reviews. We retrieved a maximum of 1.5 years worth of data, starting March 27th, 2017 to September 4th, 2018, for each app through a web crawler tool we developed. In total, we obtained 618,691 reviews. This data was used in our preliminary analysis and model generation. To conduct a cross-cultural study, we also collected reviews of the same games from the App Store of three other English speaking countries: Canada, Australia, and United Kingdom. We chose to study reviews available only on Apples App Store for two reasons. First, it has a clear separation of stores for different countries. This makes the identification of a reviews origin easier. Second, the store prohibits users to submit a review anonymously on it. Although earlier this year, Google announced that they would remove anonymous reviews and disallow users to leave a review anonymously on its store, these reviews are still visible and displayed under the name A Google user. Such a review makes the task of identifying the gender of a reviewer unattainable. We retrieved reviews from the last 90 days before Nov 24, 2018 on all four countries. In the end, we acquired a total of 193,452 reviews. This dataset was used in our cross-cultural study.

2.2 Manual Annotation

Although there are many datasets for aspect-based sentiment analysis available to use, for example, SemEval-2016 Task 5 (Pontiki et al., 2016) and SemEval-2015 Task 12 (Pontiki et al., 2015), these datasets are domain-specific (Restaurant and Consumer Electronic reviews). Due to the absence of annotated data in our domain (i.e., mobile game), we have to build our own oracle manually.

To create the ground-truth set, we randomly selected 481 reviews from our dataset through stratified random sampling and ensuring that there is a review for each star rating of each game. Several meetings were held in order to determine the categories we would use in the study. In addition, one annotator annotated all the reviews, while another annotator annotated 30% of the reviews to assess the reliability of the annotation. We resolved disagreements by discussion.

2.3 Annotation Guidelines

Table 1 shows the list of categories and their corresponding description we used to annotate each review. This list is based on the list of categories proposed by Yee (Yee, 2006). We applied similar annotation template as SemEval-2015 Task 12 dataset, namely, each review consists of one or many tuples that contain information regarding the category, sentiment, and aspect (Pontiki et al., 2015). In addition, we assigned gender of the reviewer based on the name the reviewer used. Each name was classified as either female, male, or unidentifiable. unidentifiable name includes names that can be applied to both sexes (e.g., Ohgoditsjordan) and that cannot be identified as male or female (e.g., SixOTHree, CupcakeFyte, AlchemicKnowledge, etc.)

Table 2 shows an example of manual annotation result. Each reviewer name will have a gender associated with and each review will have one or more category tuples associated with.
Table 1: List of categories and descriptions

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Game</td>
<td>Regarding the overall or general assessment of the game or app (e.g., gameplay and cost).</td>
</tr>
<tr>
<td>Performance</td>
<td>Regarding the performance aspect of the app or game (e.g., crash, freeze, network connection, etc.).</td>
</tr>
<tr>
<td>Advertisement</td>
<td>Regarding the advertisement aspect of the game or app.</td>
</tr>
<tr>
<td>Achievement</td>
<td>Advancement, Mechanics, and Competition aspect of the game.</td>
</tr>
<tr>
<td>Social</td>
<td>The social aspect of the game (relationship, teamwork, community, etc.).</td>
</tr>
<tr>
<td>Immersion</td>
<td>Aspect that makes user feel deeply involved in the game (discovery, customization, escapism, roleplay, etc.).</td>
</tr>
</tbody>
</table>

3 Experiment

3.1 Model - Gender

As aforementioned, analyzing each review manually is not ideal and a laborious task. For example, it is almost impossible to analyze all Facebook reviews as Facebook can receive over 10,000 reviews a day. To help study the cultural differences in terms of the bias in user demographics (i.e., Gender) to a greater extent, we experimented with two supervised machine learning approaches: Character-level n-gram Naive Bayes and Character-level Recurrent Neural Network (LSTM).

Initially, we found that Character-level n-gram Naive Bayes model outperformed Character-level LSTM model in terms of the F1 score (see Table 3). Hence, in this study, we used Character-level n-gram Naive Bayes model for our name-gender analysis. Figure 1 depicts the entire pipeline of the name-gender model. Note that the training sets we used (Kaggle’s twitter data¹ and SSN administration’s popular baby names with gender²) do not contain “U - unidentifiable” category.

To account for this category, once we trained the model and tested it on the subset of our annotated data (only names that were manually classified as either “M” or “F” category) to evaluate the accuracy and f1 score, we used the trained model to output the probability of each name falling into either “M” or “F” of the entire testing dataset, including the “U - unidentifiable” names. The output consists of a list of one dimensional array with 2 columns where the first column depicts the probability of the name being “M - male” and the second column depicts the probability of the name being “F - female”. We used Matthews Correlation Coefficient (MCC) to find the threshold that would give us the best F1 score for “M - male” and “F - female” category while also including “U - unidentifiable” category (Note: the default threshold for binary classification is 0.5). In other words, we converted uncertain predictions to “unidentifiable” category. The overall accuracy after accounting for the “U - unidentifiable” category is 0.58. The best threshold for male and female is 0.65 and 0.7, respectively. At the end of the pipeline, we saved the trained model and the best threshold and later used them in the analysis part of our study.

Figure 1: The pipeline of name-gender model - the two outcomes of the model which were used in the gender study are depicted in red

3.2 Model - Aspect and Sentiment Extraction

Sentiment analysis tries to automatically extract the subjective information from a user written textual content and classifies it into one of the predefined set of classes, e.g. positive, negative, neu-

¹https://www.kaggle.com/crowdflower/twitter-user-gender-classification
²https://www.ssa.gov/oact/babynames/limits.html
Table 2: Example of manual annotation result

<table>
<thead>
<tr>
<th>Name</th>
<th>Review</th>
<th>Gender</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>mark’em x</td>
<td>“New updates and bonuses keep the game enjoyable”</td>
<td>M</td>
<td>(Game, updates, pos)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Achievement, bonuses, pos)</td>
</tr>
<tr>
<td>Keepsakerose</td>
<td>“The game is fun. But an ad after EVERY turn? It’s ruins the game. And the price to remove ads is way too much.”</td>
<td>F</td>
<td>(Game, null, positive)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Advertisement, ad, negative)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Game, price, negative)</td>
</tr>
<tr>
<td>Lisamontoya</td>
<td>“I love this game it keeps me busy and the graphics are wonderful.”</td>
<td>F</td>
<td>(Game, null, positive)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Immersion, null, positive)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Immersion, graphics, positive)</td>
</tr>
<tr>
<td>SeanKU08</td>
<td>“Give us the option to play without bots!”</td>
<td>M</td>
<td>(Achievement, bots, negative)</td>
</tr>
</tbody>
</table>

Table 3: F1 score of the two name-gender models we experimented

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Char-level n-gram Naive Bayes</td>
<td>0.843</td>
</tr>
<tr>
<td>Char-level LSTM</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Our task of ”Aspect and Sentiment Extraction” is a complete pipeline of generalized task on aspect-based sentiment analysis (ABSA), which focus on extracting aspect and the relating sentiment from unstructured reviews at the same time. Several approaches of aspect extraction and aspect level sentiment classification in product reviews have been proposed independently. In this section, we focus on building the state of art baseline of these two tasks. First, aspect extraction. Second, aspect level sentiment classification.

3.2.1 Aspect Level Sentiment Classification

The main idea behind aspect level sentiment classification is to develop neural architectures that are able to learn continuous features and capture the intricate relation between a target and context words. However, to sufficiently train these models, substantial aspect level annotated data is required, which is expensive to obtain in practice. We explore both pre-training and multitask learning for transferring knowledge from document level to aspect level. For our problem, we hypothesize that aspect-level sentiment classification can be improved by employing knowledge gained from document-level sentiment classification, as these two tasks are highly related semantically (He et al., 2018).

3.2.2 Models

Attention-based LSTM

We first describe a conventional implementation of an attention-based LSTM model for this task. We use it as a baseline model and extend it with pretraining and multitask learning approaches for incorporating document-level knowledge. The inputs are a sentence consisting of n words, and an opinion target occurring in the sentence consisting of a sub-sequence of m words from s. Each word is associated with a continuous word embedding. LSTM is used to capture sequential information, and outputs a sequence of hidden vectors. An attention layer assigns a weight i to each word in the sentence. The final target-specific representation of the sentence s is notated as z. The sentence representation z is fed into an output layer to predict the probability distribution p over sentiment labels on the target. We refer to this baseline model as LSTM+ATT. It is trained via cross entropy minimization.

LSTM+ATT is used as our aspect-level model. We also build a standard LSTM-based classifier.
based on document-level training examples. This network is the same as the LSTM+ATT apart from
the lack of the attention layer. The training objective is also cross entropy minimization. Pre-
training (PRET): In this setting, we rst train on document-level examples. The last hidden vec-
tor from the LSTM outputs is used as the doc-
ument representation. We initialize the relevant
parameters of LSTM+ATT with the pretrained
weights, and train it on aspect-level examples to ne-
tune those weights. Multitask Learning (MULT):
This approach simultaneously trains two tasks
document-level and aspect-level classiication. In
this setting, the embedding layer and the LSTM
layer are shared by both tasks, and a document is
represented as the mean vector over LSTM out-
puts. The other parameters are task-specic.

Combined (PRET+MULT): In this setting, we
rst perform PRET on document-level examples.
We use the pretrained weights for parameter
initialization for both aspect-level model and
document-level model, and then perform MULT
as discussed above.

3.2.3 Experiments

We run experiments on four benchmark aspect-
level datasets, taken from SemEval 2014, SemEval
2015 and SemEval 2016. Statistics of the result-
ing datasets are presented in Table 1. In all ex-
periments, 300-dimension GloVe vectors are used
to initialize E, and E when pretraining is not con-
ducted for weight initialization. These vectors are
also used for initializing E in the pretraining phase.
Values for hyperparameters are obtained from ex-
periments on development sets. We randomly
sample 20% of the original training data from the
aspect-level dataset as the development set and
only use the remaining 80% for training. For all
experiments, the dimension of LSTM hidden vec-
tors is set to 300, is set to 0.1, and we use dropout
with probability 0.5 on sentence/document repre-
sentations before the output layer. We use RM-
SProp as the optimizer with the decay rate set to
0.9 and the base learning rate set to 0.001. The
mini-batch size is set to 32. Table 2 shows the Av-
erage accuracies and Macro-F1 scores over 5 runs
with random initialization. The best results are in
bold.

4 Cross-cultural Study Results

4.1 Gender

4.1.1 By Country

Figure 2: Normalized distribution of gender by
country

Figure 2 shows the distribution of genders be-
tween each studied country. We applied Chi-
square test of independence ($\chi^2$) to determine
whether there is a signicant association between
the two variables (gender vs country). Chi-square
test ($\chi^2$) doesn’t show that there is no association
between genders at the country-level ($p > 0.05$).
This is contradict to the study done by (Guzman
et al., 2018). However, the author of the paper in-
cluded India, Hong Kong, and Singapore in their
gender study, which could signicantly aect the
outcome of the chi-square test. We plan to incor-
porate more countries in our future work. How-
ever, the name of people in these countries (e.g.,
India, Hong Kong, Singapore, etc.) are different
from the name of people in the Canada, United
Kingdom, and Australia. This requires an addi-
tional gender-name ground truth set that is curated
speciically for these countries as English names
are not applicable to them.

4.1.2 By Game

We applied Chi-square test of independence ($\chi^2$)
to test the association between gender and coun-
try at the game-level. Chi-square test of inde-
pendence shows an association between genders
and country in 10 out of 50 games (Color Road!,
Roblox, Toon Blast, Pokémon Go, Twisty Arrow!,
Hungry Dragon™, Fortnite, TENKYU, and 8 Ball
Pool™) with statistical signicance ($p < 0.05$).
This provides an evidence that the distribution of
gender of gamers on these games are different at
the country-level.
4.2 Aspect Level Sentiment

4.2.1 By Country

4.2.2 By Game

5 Related Work

To our knowledge, our novelty appear in two dimensions. First, we incorporate approaches concerning aspect level sentiment analysis and emotion dynamics with the state of the art techniques to understand the cognitive process of users. Second, we conduct cross-cultural study and uncover hidden biases in the data. We focus the related work discussion in two areas: aspect level sentiment extraction and cross-cultural bias mining.

5.1 Aspect and Sentiment Extraction

5.1.1 Aspect extraction

One of the challenging tasks under information extraction is aspect extraction. The task aims to extract product aspects (i.e., feature or attribute) on which opinions, expressions, or sentiments are expressed. For example, when applying a typical sentence-level sentiment analysis to the review "I love the touchscreen of my phone but the battery life is underwhelming," it will result in the overall sentiment closer to neutral. However, aspect extraction will deconstruct the sentence into two set of features (i.e., touchscreen and battery life) and link the associate polarity to those features (Poria et al., 2014). Compared to document-level sentiment analysis or sentence-level sentiment analysis, aspect extraction is more fine-grained (Zhang et al., 2018). There are mainly two types of study done in this area: non deep-learning based and deep-learning based. Note that the tasks can also be divided into two sub-tasks: extracting aspects and linking sentiment value to the identified aspects (i.e., sentiment analysis).

Non Deep learning based: Earlier works in the area of opinion mining have been relying on feature engineering to construct rules to extract aspects and sentiments from a sentence (Poria et al., 2014; Popescu and Etzioni, 2007; Kiritchenko et al., 2014). Aspect extraction can be seen as a general information extraction problem, for which techniques based on sequential labeling are generally used. The most popular methods in this context, in particular, are Hidden Markov Models (HMM) and Conditional Random Fields (CRF) (Lafferty et al., 2001). Jin and Ho (Jin et al., 2009) used a lexicalized HMM for joint extraction of opinions along with their explicit aspects. Niklas and Gurevych (Jakob and Gurevych, 2010) used CRF to extract explicit aspects in a custom corpus with data of different domains. Li et al. (Li et al., 2010), Choi and Cardie (Choi and Cardie, 2010), and Huang et al. (Huang et al., 2012) also used CRF for extraction of explicit aspects. Traditionally, researchers utilize NLTK pipeline (Guzman and Maalej, 2014) or modified LDAs (Park et al., 2015) for opinion mining.

Deep learning based: Poria et al. (Poria et al., 2016) presented the first supervised aspect extraction technique based on deep learning. Their model consists of 7-deep convolutional neural network (CNN). Their experimental results show that CNN based approach can perform significantly better than the state-of-the-art approach that is based on feature engineering (rule-based) described in (Popescu and Etzioni, 2007). He et al. (He et al., 2017) proposed an attention-based neural model for unsupervised aspect extraction. They used the attention mechanism to de-emphasize words that are not part of aspects with in a sentence, which in turn allows model to focus on aspect words. In other words, they utilized attention model to help construct aspect embedding. Their entire process is similar to autoencoder approach.

5.1.2 Aspect sentiment classification

Sentiment analysis or opinion mining is the computational study of peoples opinions, sentiments, emotions, appraisals, and attitudes towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes (Liu, 2015). The inception and rapid growth of the field coincide with those of the social media on the Web, for example, reviews, forum discussions, blogs, microblogs, Twitter, and social networks, because for the first time in human history, we have a huge volume of opinionated data recorded in digital forms. Existing research has produced numerous techniques for various tasks of sentiment analysis, which include both supervised and unsupervised methods (Zhang et al., 2018; Ravi and Ravi, 2015). In the supervised setting, early papers used all types of supervised machine learning methods (such as support vector machines, maximum entropy, nave Bayes, etc.) and feature combinations. Unsupervised methods include various methods that exploit sentiment lexicons, grammatical analysis, and syntactic patterns. Sev-
eral survey books and papers have been published, which cover those early methods and applications extensively (Brody and Elhadad, 2010). Since about a decade ago, deep learning has emerged as a powerful machine learning technique (Goodfellow et al., 2016) and produced state-of-the-art results in many application domains, ranging from computer vision and speech recognition to NLP. Applying deep learning to sentiment analysis has also become very popular recently. (Zhang et al., 2018) The target of our paper is entity or an entity aspect. For simplicity, both entity and aspect are usually just called aspect. Different from the document-level and the sentence-level sentiment classification, aspect-level sentiment classification considers both the sentiment and the target information, as a sentiment always has a target. Aspect-level sentiment classification is challenging because modeling the semantic relatedness of a target with its surrounding context words is difficult. Different context words have different influences on the sentiment polarity of a sentence toward the target. Therefore, it is necessary to capture semantic connections between the target word and the context words when building learning models using neural networks.

Ruder et al. (Ruder et al., 2016) proposed to use a hierarchical and bidirectional LSTM model for aspect-level sentiment classification, which is able to leverage both intra- and inter-sentence relations. The sole dependence on sentences and their structures within a review renders the proposed model language independent. Word embeddings are fed into a sentence-level bidirectional LSTM. Final states of the forward and backward LSTM are concatenated together with the target embedding and fed into a bidirectional review-level LSTM. At every time step, the output of the forward and backward LSTM is concatenated and fed into a final layer, which outputs a probability distribution over sentiments.

Wang et al. (Wang et al., 2016) proposed an attention-based LSTM method with target embedding (ATAELSTM), which was proven to be an effective way to enforce the neural model to attend to the related part of a sentence. The attention mechanism is used to enforce the model to attend to the important part of a sentence, in response to a specific aspect. Likewise, Yang, Tu, Wang, Xu, and Chen (Yang et al., 2017) proposed two attention-based bidirectional LSTMs to improve the classification performance. Liu and Zhang (2017) extended the attention modelling by differentiating the attention obtained from the left context and the right context of a given target/aspect. They further controlled their attention contribution by adding multiple gates.

Tang et al. (Liu and Zhang, 2017) introduced an end-to-end memory network for aspect-level sentiment classification, which employs an attention mechanism with an external memory to capture the importance of each context word with respect to the given target aspect. This approach explicitly captures the importance of each context word when inferring the sentiment polarity of the aspect. Such importance degree and text representation are calculated with multiple computational layers, each of which is a neural attention model over an external memory.

Lei et al. (Lei et al., 2016) proposed to use a neural network approach to extracting pieces of input text as rationales (reasons) for review ratings. The model consists of a generator and a decoder. The generator specifies a distribution over possible rationales (extracted text) and the encoder maps any such text to a task-specific target vector. For multi-aspect sentiment analysis, each coordinate of the target vector represents the response or rating pertaining to the associated aspect.

Li et al. (Li et al., 2017) integrated the target identification task into sentiment classification task to better model aspect-sentiment interaction. They showed that sentiment identification can be solved with an end-to-end machine learning architecture, in which the two subtasks are interleaved by a deep memory network. In this way, signals produced in target detection provide clues for polarity classification, and reversely, the predicted polarity provides feedback to the identification of targets.

5.2 Cross-cultural study and cultural bias mining

To best of our knowledge, there is a dearth in research on cross-cultural study in app store reviews. However, many studies, not necessarily from the computer science domain, have shown that the cultural background of survey respondents can affect their responses (Trompenaars and Hampden-Turner, 2011; Ayed et al., 2017; Hofstede, 2001, 2011; Guzman et al., 2018). Motivated by the findings of these works and the lack of knowledge in
this area, we investigate whether or not user perceptions on application aspects vary across users from different cultural backgrounds.

Hofstede (Hofstede, 2001) investigated how values in workplace differ across individuals from different cultures. He collected over 116,000 responses from IBM employees from 72 countries and later additional, unrelated to IBM, data was added which can be matched by countries to the IBM dataset. Through theoretical reasoning and statistical analysis of the entire dataset, he uncovered five dimensions on which country cultures differ.

- Power Distance - an extent to which a culture programs a less powerful members of an organization or institution to accept and expect power to be distributed unequally.
- Uncertainty Avoidance - an extent to which a culture programs an individual to feel comfortable or uncomfortable in an unstructured situation.
- Individualism vs Collectivism - an extent to which a culture programs an individual to feel the needs to look after themselves or remain dependent.
- Masculinity vs Femininity - a perception of emotion associated with genders.
- Long-term vs Short-term orientation - an extent to which a culture programs an individual to accept delayed gratification.

Ayed et al. (Ayed et al., 2017) studied the impact of cultural differences on the adoption of agile practices. They interviewed 19 practitioners from three countries: Belgium, Singapore, and Malaysia on practices, challenges and impediments encountered by software development teams. Their qualitative analysis suggested a potential relationship between cultural background factors (i.e., from Hafstede model (Hofstede, 2011)) and practice adoptions.

The work that is closely related to ours is Guzman et al.’s work (Guzman et al., 2018). In that work, they investigated cultural differences in user reviews of 7 mobile applications from three categories on Google App Store. A representative review sample size of 2560 reviews written by users from eight countries with diverse national cultures was manually analyzed. They found that many aspects of user reviews (e.g., sentiment, length, and rating) can differ at the country level. Their work is related to ours because they investigated culture differences using user reviews as a data source. However, their study was not done in a fine grained level of user reviews (i.e., not through aspect extraction analysis), but rather on a sentence-level. In addition, their study requires high manual efforts. In particular, we are interested in whether aspect extraction can reveal more insightful information regarding culture differences in app store reviews of game applications.

6 Conclusion

References


Christopher Mitchelltree, Veronica Wharton, and Avneesh Saluja. 2018. Using aspect extraction approaches to generate review summaries and user profiles. NAACL HLT 2018 pages 68–75.


