Title: Destination Image through Social Media Analytics and Survey Method

Abstract

Purpose

Recent tourism research has adopted social media analytics to examine tourism destination image (TDI) and gain timely insights for marketing purposes. Comparing the methodologies of social media analytics and intercept surveys would provide a more in-depth understanding of both methodologies and a more holistic understanding of TDI than each method on their own. This study aims to investigate the unique merits and biases of social media analytics and a traditional visitor intercept survey.

Design/methodology/approach

This study collected and compared data for the same tourism destination from two sources: responses from a visitor intercept survey (n=1,336) and Flickr social media photos and metadata (n=11,775). Content analysis, machine learning, and text analysis techniques were used to analyze and compare the destination image represented from both methods.

Findings

The results indicated that the survey data and social media data shared major similarities in the identified key image phrases. Social media data revealed more diverse and more specific aspects of the destination, whereas survey data provided more insights in specific local landmarks. Survey data also included additional subjective judgment and attachment towards the destination.

Together, the data suggested that social media data should serve as an additional and complementary source of information to traditional survey data.

Originality

This study fills a research gap by comparing two methodologies in obtaining TDI: social media analytics and a traditional visitor intercept survey. Furthermore, within social media analytics, photo and metadata are compared to offer additional awareness of social media data's underlying complexity. The results showed the limitations of text-based image questions in surveys. The findings provide meaningful insights for tourism marketers by having a more holistic understanding of TDI through multiple data sources.

Keywords: Tourism destination image (TDI), survey, social media analytics, textual analysis, image analysis, machine learning

1. Introduction

Big data includes user-generated content (UGC) (e.g., online reviews and Twitter data), device data (e.g., real-time mobile location data), and transaction data (e.g., online shopping data). The volume of these data sources has grown exponentially in the past decade (Li *et al*., 2018), leading to big data analytics, which has generated critical insights for researchers and marketers and affected business practices. Big data analytics requires a diverse set of statistical or analytical tools to generate the understanding of human behaviors. For example, Marriott launched a social platform focused on using UGC to capture real customer experiences and striving for better performance based on the analysis (Marriott, 2015).

The growth in UGC volume has transformed the hospitality and tourism industry due to consumers' trust in UGC channels during their decision-making process (Mariani *et al*., 2018). However, Mariani *et al*. (2018)'s systematic literature review revealed a lack of research using business intelligence and big data to assess and enhance consumer satisfaction and the quality and memorability of tourist experiences. They recommended that researchers use UGC to supplement findings from traditional survey-based research methods and gain additional insights. Compared to traditional surveys and focus group methods with small and fixed-scale data sets (Li *et al*., 2017), social media analytics (SMA), as part of big data analytics, collects and analyzes unstructured and complex data with unmatched breadth, depth, and scale (e.g., Du *et al*., 2018; Park *et al*., 2020; Zhang *et al*., 2020). Additionally, SMA can be used to solve real-life hospitality and tourism industry issues in near real-time.

Previous research investigating tourist experiences using SMA has mainly focused on analyzing UGC data (i.e., online metadata/photo data) to address issues of guest satisfaction in hospitality settings and has compared guest satisfaction across multiple platforms (e.g., Xiang *et*

al., 2015; Xiang *et al*., 2017). Few studies have holistically investigated the tourist experience at a destination using multiple methodologies (Jiang *et al*., 2021). Tourist experiences in a destination are complex and associated with various aspects of the experience, including transportation, events, attractions, lodging, and restaurants (Chang, 2018). In addition, assessing tourism destination image (TDI) through different methodologies can help researchers form a more comprehensive understanding and help destination marketers develop more effective marketing strategies. While survey methodology has traditionally been used to evaluate TDI (Stepchenkova and Morrison, 2008; Wang and Hsu, 2010; Yeh *et al*., 2012), recent studies have used SMA to study this concept (Deng and Li, 2018; Jiang *et al*., 2021; Stepchenkova and Zhan, 2013). However, studies have not yet compared and supplemented traditional survey methods with SMA. In particular, within SMA, the comparison between metadata and photo data is not conducted in the extant literature.

Moreover, by using data collected through two different methods, hospitality and tourism researchers could better understand the comparative strengths and weaknesses of each method. For example, Ruths and Pfeffer (2014) suggested using one type of social media data may exhibit many biases, such as population bias. Such biases could be evaluated by comparing results between data collected using different methods, such as visitor intercept surveys and both metadata and photo data from social media.

Accordingly, this study aims to compare and supplement traditional survey data with social media analytics to measure tourists' image of a destination. Specifically, this research addresses three questions:

1. What aspects of TDI can be revealed from social media analytics, from both metadata and photo data?

2. Compared to visitor intercept surveys, what are the unique characteristics and biases of social media analytics?
3. How can we incorporate TDI from different sources to provide insights to tourism marketers?

The findings will contribute to the literature by filling a gap in using multiple methodologies to provide a more comprehensive understanding of destination image and identifying the pros and cons of social media analytics and survey methods.

2. Literature Review

The following literature review will focus on research related to tourism destination image (TDI), big data analytics, and more specifically social media analytics (SMA) in the hospitality and tourism. The literature review will also deliberate on the traditional survey methods and SMA used to study TDI. Missing gaps will be identified to explain why this study aims to use data collected through the two methods to provide a more comprehensive understanding of this concept and compare the unique characteristics of both methods in assessing TDI.

2.1 Tourism destination image (TDI)

Since the 1970s, the topic of tourism destination image (TDI) has been extensively studied because creating a positive and memorable image is vital for understanding tourists' destination selection process, as well as destination differentiation and positioning (e.g., Afshardoost and Eshaghi, 2020; Sahin and Baloglu, 2011; Stylidis *et al*., 2020). Researchers

have investigated both the static structure and dynamic nature of TDI, including its conceptualization (Baloglu and McCleary, 1999), components (MacKay and Fesenmaier, 1997), measurements (Echtner and Ritchie, 1993), influencing factors (Hu and Ritchie, 1993; Milman and Pizam, 1995; Stepchenkova and Zhan, 2013), and outcomes (e.g., perception-behavior link; Wang and Hsu, 2010). Though the concept of TDI possesses internal, external, and foundational vagueness (Lai and Li, 2016), it can be defined as "the perception of a person or a group of people regarding a place" (Sahin and Baloglu, 2011, pp. 71). One's beliefs, ideas, and impressions sum to form such a perception (Crompton, 1979).

TDI, as a multidimensional construct, has been examined through different lenses. Researchers have categorized TDI based on the temporal dimensions of pre-visit or post-visit image (Beerli and Martin, 2004); tourism attributions of the functional/psychological axes, the common/unique axes, and the holistic/attribute-based axes (Gallarza *et al*., 2002); tourist responses in terms of cognitive evaluations and affective evaluations (Baloglu and McCleary, 1999). The last perspective is adopted as it fits the scope of our research to explore tourist perceptions and evaluations.

The cognitive/perceptual evaluation refers to "the beliefs or knowledge about a destination's attributes" (Baloglu and McCleary, 1999). In a cognitive evaluation, a tourist evaluates multiple objective attributes to form a perception about that place. According to Wang and Hsu (2010), such elements of a destination include attractions to be seen (e.g., sand and beach), the environment to be perceived (e.g., weather, public hygiene), and experiences to remember (e.g., surfing, swimming). The affective evaluation refers to "feelings toward, or attachment to that place" (Baloglu and McCleary, 1999). In such a mental construct, a tourist evaluates the affective qualities of environments (Hanyu, 1993). For example, Stylidis *et al*.

(2017) suggest that four semantic differential scales (i.e., sleepy–arousing, unpleasant–pleasant, boring–exciting, and distressing–relaxing) could be used to measure affective components. The relationship between cognitive evaluations and affective evaluations is "distinct but hierarchically related" (Gartner, 1993). Diverse researchers agreed that the affective response serves as the cognitive response's descendant within the cognitive-affective-overall image tradition (Baloglu and McCleary, 1999; Stepchenkova and Morrison, 2008; Wang and Hsu, 2010). In addition, tourists' perceptions of both cognitive and affective attributes form an overall image, leading to a favorable or unfavorable attitude about the destination (Beerli and Martin, 2004). However, some studies suggest that the overall TDI should be treated as a third component and measured separately since it may be similar to or different from the simple summation of both parts (Fakeye and Crompton, 1991; Phelps, 1986). The important issue of delineating the relationship between the overall image and sub-components is whether structured or unstructured approaches should be adopted to investigate TDI.

Overall, both cognitive and affective evaluations positively and directly influence the overall image (Baloglu and McCleary, 1999; Wang and Hsu, 2010; Woosnam *et al*., 2020). Once the holistic TDI is constructed, it further impacts a tourist's future behavioral intentions (Wang and Hsu, 2010). A particular research stream has examined the relationship between TDI and behavioral intentions (e.g., Bigné *et al*., 2001; Fakeye and Crompton, 1991; Lee *et al*., 2005).

2.2 Social media analytics vs. visitor intercept survey on TDI

Since TDI is constructed through a complex learning process and information sharing, the multiple, complex, and dynamic nature of TDI leads to the debate over whether structured or unstructured approaches could better serve the investigation (Gallarza *et al*., 2002; Wenger,

2008). For structured methods such as Likert and Semantic Differential scales, participants were asked to rate pre-determined attributes (e.g., scenarios, activities, buildings, quality of service). However, predefined, standardized attributes can only reflect the objective reality. Since such a fixed set of items cannot incorporate all functional or psychological traits of TDI, they may not be "relevant or descriptive enough" for participants to reflect or express their specific and unique views about a place, thus neglecting potentially critical and non-fitting responses (Wang *et al*., 2020). Over-reliance on structured approaches may fail to capture TDI's holistic and unique components, as noted by many researchers (Echtner and Ritchie, 1993; Gallarza *et al*., 2002; Tasci and Holecek, 2007). Some researchers claim that unstructured approaches, such as open-ended questions, interviews, online media might reduce the structured approach's inherent bias and irrelevance, as those data allow informants or content creators to take initiatives regarding what they wish to express (Pan and Li, 2011; Stepchenkova and Morrison, 2008). Unstructured methods could bring breakthroughs and innovations to TDI research. Hence, this study analyzes the unstructured data from social media (e.g., online text and photos) and visitor intercept surveys (e.g., open-ended questions) and, subsequently, compares the TDIs formed from these two approaches.

With the remarkable growth of technology in the era of Web 2.0, the Internet, big data, and related technologies have brought forth new data sources and caused a paradigm shift in scientific research, including in the field of hospitality and tourism (Li *et al*., 2018; Mariné-Roig, 2019; Park *et al*., 2020; Zhang *et al*., 2020). Soical media analytics (SMA), as a subset of big data analytics, utilizes any form of content available via social media platforms such as blogs, discussion forums, posts, chats, tweets, podcasting, pins, digital images, video, audio files, or others (Choi *et al*., 2007). Therefore, social media data tend to be massive and complex. In terms

of data quality, the large scale of social media data could effectively mitigate sample size limitations and sampling bias issues (Kirilenko *et al*., 2021; Park *et al*., 2020). Social media data is also more informative and complex, thus identifying the underlying behavior patterns that can offer meaningful insights (Kirilenko *et al*., 2021). SMA could be used to approach new research questions with diverse analytical tools to observe patterns and provide insights via analyzing large amounts of data (Aiden and Michel, 2014).

Currently, tourists own dual identities- as consumers, and as "the efficient, active, and effective destination promoters" (Gurung and Goswami, 2017). Therefore, social media content has been increasingly impacting destination awareness and TDI formation (Tussyadiah and Fesenmaier, 2009). A large amount of SMA research utilizes data from community-based online review platforms (e.g., Tripadvisor, Yelp, and LonelyPlanet) and transaction-based sites (e.g., Expedia, Bookings.com) to explore, assess, and categorize dimensions of TDIs (Liu *et al*., 2020; Jiang *et al*., 2021). For example, Mariné-Roig (2017) examined 387,414 TripAdvisor tourist reviews on 'Things to Do' in France and found that online tourist reviews contribute to the construction of perceived TDI on five dimensions (cognitive, spatial, temporal, evaluative, and affective attributes). Later the same approach was applied to capture three major aspects of Attica's (in Greece) TDI: designative, appraisive, and prescriptive (Mariné-Roig, 2019).

When comparing images of the same destination formed from different data sources, both textual material and image data appear in SMA. For instance, Stepchenkova and Zhan (2013) used a comparative analysis of DMO (destination marketing organizations) and Flickr images of Peru. Results have shown that the DMO is more likely to present a well-rounded destination image and emphasizes natural tourism resources. In contrast, Flickr's images reflect tourists' interests in local lifestyles and cultural attractions. Moreover, Deng and Li (2018) also

constructed a machine-learning-based model to select photo elements from the viewers' perspective and assist destination marketing organizations (DMO) in photo selection process. Though researchers compared images of the same destination through different lenses, most research only focuses on a single approach. No study has explored similarities or dissimilarities between different methodological approaches.

As suggested above, one important unstructured data to understand TDI is online photo data, a subset of UGC data that is posted on photo-sharing websites like Flickr, Panoramio, and Instagram (Li *et al*., 2018; Wang *et al*., 2020; Zhang *et al*., 2019). It contains a rich set of information including metadata and the photo itself (e.g., content and composition of the photo) (Albers and James, 1988). Metadata refers those textual data associated with the photos, including user ID, date, title, descriptions, and tags entered by the users (Albers and James, 1988). Since tourists take photos to capture the most salient destination attributes they perceive (Albers and James, 1988; Day *et al*., 2002), online photo data could convey tourists' experiences and perspectives. Therefore, it could help study tourist behavior, tourism recommendations, and TDI (Li *et al*., 2018).

Overall, in discussing the preferred methodology to capture TDI, the strengths and challenges of SMA and visitor intercept surveys should be considered. On the one hand, the traditional visitor intercept survey cannot match SMA in terms of sample sizes, frequencies, and details (Whitaker, 2014). Due to the selection of samples (Echtner and Ritchie, 1993), traditional surveys may encounter spatial and temporal limitations and reduce the generalizability of the findings. Some surveys also struggle with precision issues, recall biases, and nonresponse biases since answers heavily rely on people's memory (Biemer, 2010). Conversely, amongst tourists, the message extracted from social media data tends to be perceived as highly credible and

independent (Gitelson and Kerstetter, 1995). A large data volume makes constructing TDI from online photos much more convenient, as social media data's enormous sample sizes "support more detailed analysis regarding space, time and other subgroups" (Callegaro and Yang, 2018, p.183). Online photo posts on social media allow researchers to capture any communicable attribute of TDI since such pertinent images reflect both tangible and intangible aspects of the destination (Echtner and Ritchie, 2003). They can be used to trace image changes across different periods and, as a result, could even be viewed as a "pseudo-longitudinal" study and help marketers make better decisions (Tasci and Holecek, 2007).

On the other hand, SMA may exhibit some reliability concerns since online data is normally messy and unstructured (Pan *et al*., 2012). Ruths and Pfeffer (2014) suggest that SMA might encounter a series of validity problems, including platform biases, data availability biases, and data authenticity issues. Also, unlike open-ended questionnaires designed to address questions researchers aim to investigate, big data comes with big noise (Waldherr *et al*., 2017). The meanings embedded in online photos are subjective to researchers' interpretation, thus making it harder to collect tourists' attitudes, opinions, and emotions about the destination (Callegaro and Yang, 2018; Stepchenkova and Zhan, 2013).

While both SMA and visitor intercept surveys have a lot to offer, to our surprise, no studies have empirically measured TDI using both approaches. We argue that these two approaches are comparative. In cultural studies, pictorial materials are also considered unstructured and a form of "text," enabling comparison with an unstructured, open-ended questionnaire (Stepchenkova and Zhan, 2013). Comparing SMA and survey data could provide a more comprehensive understanding of TDI and address the shortcomings of each methodology. To fill this gap in the literature, this study will address the unique characteristics, strengths, and

biases among online texts and photos from social media platforms and open-ended questions from the visitor intercept survey when analyzing one destination's overall TDI.

3. Methodology

3.1 Destination location

In order to address the comparative strength and weakenss of two methodologies, researchers picked a college town in the United States. Centre County is located in the Commonwealth of Pennsylvania, with a population of 162,385 (United States Census Bureau, 2019). Centre County consists of several cities and townships, including Bellefonte, Centre Hall, Port Matilda, Snow Shoe, and State College. The University Park campus of The Pennsylvania State University is located in State College. The university's athletic teams are known as the Penn State Nittany Lions. Penn State football games are the most famous and among the most popular events in Centre County (https://gopsusports.com/).

3.2 Social media data collection

Two data sources—social media data and survey data—were collected. The social media data, including photographs and related metadata, was collected through Flickr's API (https://www.flickr.com/services/api/) in March 2020. Flickr is a popular social media platform employed in TDI research (Stepchenkova and Zhan, 2013; Kim and Stepchenkova, 2015; Deng and Li, 2018). The keywords used for searching photos included the names of seven cities and townships within Centre County (Bellefonte, Boalsburg, Centre Hall, Philipsburg, Port Matilda, State College, and Snow Shoe) plus "P.A./Pennsylvania." In total, 14 keywords were employed in searching for photos.

The types of metadata returned by Flickr API included username, user I.D., the name of users, origins of residence, number of views, uploaded date, titles of photos, tags of photos, descriptions of photos, and photos' URLs, covering the years 2004-2020. Since the purpose of this study was to explore the TDI of tourists, after the initial data collection, users accounts reported with a residence in Centre County, PA, were removed. In addition, official accounts of local organizations, such as Penn State World Campus and the Department of Education of Pennsylvania, were removed. Also, duplicate photos were removed based on the URLs (Universal Resource Locators) of photos. A total of 11,775 photos and their related information were included. Table 1 shows the number of photos taken in each city within the county.

[Insert Table 1 Here]

3.3 Image Recognition

3.3.1 Machine Learning Methods

To analyze each photo's content, Google's Cloud Vision API (Google Cloud Vision, 2017), an image recognition cloud platform, was employed to label all collected photos. The API has powerful machine learning models to detect objects and faces and extract labels from images (Richards and Tunçer, 2018). Up to ten labels were extracted from one photo through the API.

3.3.2 Data pre-processing

To include all information from the photos, photo metadata—including titles, tags, and descriptions—was combined with the photo labels returned by Google Cloud Vision API before data analysis. Next, all collected metadata were pre-processed following these established steps (Xiang *et al*., 2017; Ma and Kirilenko, 2020):

1. Transfering of all words to lower case;
2. Tokenization with the removal of short words (below 3 letters);
3. Removal of English stop words (e.g. “the”) by nltk.corpus.stopwords;
4. Conducting of bigrams and trigrams;
5. Lemmatization;
6. Filtering tokens (only keeping nouns, adjectives, and adverbs);
7. Removing infrequent words (those encountered in less than 1.5% of words) and extremely frequent words (those encountered in over 80% of words).

3.3.3 Latent Dirichlet Allocation (LDA)

After data pre-processing, Latent Dirichlet Allocation (LDA) (Blei *et al*., 2003), a topic modeling approach, was employed to reduce text materials’ dimensions and extract potential main TDI. Gensim, a Python package for topic modeling, was applied to conduct LDA (Rehurek and Sojka, 2010). The results of LDA return lists of words by topic and the weight value of each word associated with those topics. In addition, the proportion of each topic was calculated to identify the popularity of each topic.

The number of topics, K, will affect the results since the LDA algorithm allocates words to each topic based on the number of topics. Therefore, in this study, the authors tried K= 5, 6, ……14, 15. The final decision of the number of topics, K, is determined by the degree of interpretability of those topics.

3.4 Survey approach

3.4.1 Data collection

The survey was part of a larger project for the Happy Valley Adventure Bureau (the DMO for Centre County). The project investigated the demographic background, behavioral profile, and satisfaction of visitors. The survey was conducted at over 20 sites, which represented different geographic locations in Centre County. The sites represented accommodations, attractions, and special events. The data collection period ranged from May 2019 to January 2020. In total, 1,336 completed onsite responses were returned. To examine TDI from visitors, the authors focused on two open-ended survey questions: (1) *What comes to your mind when you think of this area?;* and (2) *What is the most iconic image of Centre County?*

3.4.2 Data cleaning and coding

The answers to the two open-ended questions from the intercept survey went through a data cleaning process. First, data with clear typos were corrected (e.g. revising "pdu" to "psu", the acronym of The Penn State University), stopwords and non-target language were deleted, and sentences were split into words and phrases in preparation for further analysis (e.g. splitting "Rural but an intellectual center" into "rural" and "intellectual center"). Second, words and phrases with identical meanings were merged (e.g., merging "psu", "penn state", and "penn state university").

The researchers then conducted content analysis on the cleaned data. Each word or phrase was regarded as a single unit of content (Krippendorff, 2004; Neuendorf, 2002). Kim and Stepchenkova's (2015) paper provided the theoretical grounding for category generation. To assess inter-rater reliability (Stepchenkova and Zhan, 2013; Stepchenkova and Li, 2014), two

researchers separately categorized the TDI described in the content units and then held discussions to reach an acceptable agreement on all categories. A third coder further reviewed the results. These categories were clarified and refined by the authors through multiple iterations of discussions.

The researchers calculated the frequencies of different words/phrases in each category (Pan and Li, 2011). The frequencies of different words/phrases were ranked from most to least used, and the top and bottom phrases were identified. The goal of this step was to illustrate the TDI of Centre County indicated in the survey, to prepare for further comparison between survey data and the results of SMA analytics.

4. Results

4.1 Social media analytics

Social media analytics results contain three sections: 1) topics identified from image labels generated from Google Cloud Vision API; 2) topics identified from metadata associated with images; 3) topics identified from both image labels and metadata. LDA identified seven topics from image labels, including Event & Performance, Buildings, Transportation, Natural Landscape, Residential & Food, Leisure, and Sports (Table 2).

[Insert Table 2 here]

LDA extracted eight topics from images' metadata, namely Bellefonte Historic District, Railroad, University, Centre County, Event, Reunion, Memory, and Infrastructure (Table 3). Significantly, unlike topics extracted from image labels, topics from metadata involve more specific attractions in Centre County. For example, the Bellefonte Historic District depicts historic buildings built in 1795. Railroad indicates Bellefonte Central Railroad, constructed in the

late 19th century, connecting Bellefonte to State College. Topics extracted from metadata do not include Sports and Natural Landscape extracted from image labels.

[Insert Table 3 here]

LDA identified eight topics from the combined image labels & metadata: Transportation, Buildings, University, Natural landscape, Food, Residential, Sports, and Event and Performance (Table 4). The topics retrieved from combined metadata and image labels are similar to those from image labels. Figure 1 shows the topic proportion for three data representation. The most popular topic in image labels is Event and performance. The most popular topic in metadata is University that greatly exceeds other topics. Sports is the most discussed topic in image labels & metadata.

[Insert Table 4 Here]

[Insert Figure 1 Here]

4.2 Survey results

The results of the survey are listed in Table 5 and Table 6. Table 5 lists the responses to the question, "*What comes to your mind when you think of this area?"* After the cleaning and coding the survey data, 116 unique and 1,128 total phrases were summarized. Finally, eleven categories were developed to represent all essential characteristics of the destination: Penn State, nature and scenery, activities, subjective judgment, people. The top five categories were Penn State (n=562), Nature and scenery (n=151), Activities (n=83), Subjective judgment (n=79), and People (n=79). Penn State was listed as the top category with keywords related to its sports, campus, memories of college, and education. Other than common TDI categories (e.g., Nature and scenery, Activities, and People), Subjective judgment was listed as one of the top categories.

It included descriptions of how the respondents perceived Centre County as a destination. The rest of the categories are Rural, Attraction, Memory and Attachment, Location, Accommodation and Food, and Community. Unlike the keywords identified in the social media data, these keywords in the survey represented the respondents' feelings, attitudes, and emotions toward the destination. Such abstract, higher-order keywords cannot be easily assessed by the photo and text social media data.

[Insert Table 5 Here]

Table 6 lists the responses to the question, "*In your opinion, what is the most iconic image of Centre County*?" After the cleaning and coding the survey data, 71 unique and 818 total phrases were summarized. The top five categories were Penn State (n=470), Attraction (n=194), Nature and scenery (n=74), People (n=30), and Subjective judgment (n=16). The results of this question were consistent with the other TDI-related survey question. Nevertheless, within the Penn State and attraction categories, many of the identified keywords were more specific than the previous question. For example, in Table 5, Penn State's top five keywords were Penn State, football, campus, college memory, and education. The top five keywords of Penn State in Table 6 were Penn State, Nittany Lion, Nittany Lion Shrine, campus, and college town. The differences indicated that the keywords identified in the survey method were related to how the questions were framed, indicating what type of information researchers intend to obtain from the respondents. On the other hand, compared to the social media data, the survey respondents were able to identify specific attractions (e.g., Nittany Lion Shrine, Beaver Stadium, and Mount Nittany). Given the sample and the analysis technique, the social media data cannot identify specific attractions and locations, as garnered by the survey respondents.

[Insert Table 6 Here]

4.3 Comparing Social Media and Survey Methodology

Six categories were generated from social media analytics: Natural landscape, Sports, Residential, Activities, Buildings, and Transportation. Eleven categories emerged from the survey data: Penn State, Nature and scenery, Activities, Subjective judgment, People, Rural, Attraction, Memory and attachment, Location, Accommodation and food, and Community. Among the categories, the survey data and social media data shared similarities in Nature and scenery, Activities, Attraction, and Location, which have been major components of TDI (Pan and Li, 2011; Stepchenkova and Zhan, 2013). Therefore, the findings suggest that both survey data and social media data can generate certain consistent TDI components for a destination.

There were also substantial differences between the results. The survey data did not include two categories—Buildings and Transportation—whereas the social media data did include these. The reason could be that the information in the Flickr photos contained buildings, architecture, and vehicles, offering supplementary information other than metadata. In other words, social media data offered more specific information from the photos. For example, the social media photo data could identify specific keywords from the photo content such as “sky” and “grass” from the natural landscape category, “room” from the activities category, and “metal” from the transportation category. These specific words were not recalled in the survey questions or the UGC text, but they were available in the photos, which were generally too trivial to be recalled in the TDI survey questions. This result verified that photos were the source of detailed image items when UGC was used to assess TDI for a destination (Stepchenkova and Zhan, 2013).

Moreover, the survey method was limited due to the time and space of data collection. The Flickr UGC (i.e., photos and metadata) within the social media data does not suffer from

such limitations. The social media user can upload their information at their convenience, which could alleviate these biases related to data collection. Therefore, social media data results could represent more diversified locations and respondents for the sample location and mitigate biases from survey data collection confined to a certain time and space.

Compared to the social media data, the survey data showed more landmark recognition in Centre County as the iconic components of the destination (e.g., Penn State, Old Main, Nittany Lion Shrine, and Mount Nittany). On the other hand, the social media LDA results did not reflect such specific landmarks and attractions, but identified them as "buildings." One potential reason could be that the Google Vision API only generated general labels for photos without identifying the actual landmark's name. In addition, Flickr users were not only attracted by State College but were also interested in other adjacent cities. The landmark name did not get enough weight in such a diverse and massive sample as social media.

Moreover, survey respondents expressed subjective judgments and attachment with State College and Centre County by using adjectives such as "beautiful," "nice," and "nostalgia." These adjectives reflected respondents' personal feelings and emotions toward the destination, which is an essential part of TDI (Yeh *et al*., 2012). On the other hand, social media data analysis had difficulty identifying the intangible TDI of Centre County. Such information was not easily decipherable in the photos or related metadata.

5. Discussion and Conclusions

5.1 Conclusions

This study compares social media and tourist survey data and assessed the TDI of Centre County, Pennsylvania. . Previous literature has only assessed TDI by one method, and research

on each perspective's comparative strengths and weaknesses is scant. Comparing categories and frequencies of image items from two data sources suggests that social media data should serve as an additional and complementary source of information to traditional survey data, given the major similarities and differences in TDI's major categories and particularly in identified keywords. At the same time, social media data could alleviate the restrictions of location and space with survey collection and offer more diverse and specific content from UGC. In contrast, a survey could provide more recognition of specific local landmarks and contain subjective judgment and attachment toward the destination. Therefore, this study contributes to the field of destination image by providing a potential analytical approach to achieve a more comprehensive TDI. With the potential technology advancement, social media photo data may provide insightful information related to ambience and sentiments. Such a comprehensive TDI can combine the local knowledge from survey data and extensive viewpoints from SMA.

In addition, future research on TDI should consider using both photo and metadata when analyzing social media data. Within social media analytics, photo and metadata are compared to offer additional awareness of social media data's underlying complexity. Recent studies have used either social media photo data (e.g., Taecharungroj and Mathayomchan, 2020; Wang *et al*.,, 2020) or metadata (e.g., Liu *et al*.,2020; Mariné-Roig, 2019). However, to our best knowledge, none of the studies compared social media photo and metadata by using machine learning algorithm. Using Google's Cloud Vision API, SMA photo data can only identify broad categories such as "event" and "building" without local knowledge about the detailed information about the events and buildings. The social media metadata provides local knowledge to supplement the photo data with additional information such as names, descriptions, and

emotions. Therefore, SMA can offer a complete representation of the destination image with both photos and metadata.

5.2 Theoretical implications

Theoretically, TDI are merely a tourist's perception and experience of a place (Echtner and Ritchie, 1993). It has been expressed as words representing a set of attributes. However, our perceptions and experiences are beyond those communicated words and may not be articulated (Scarles, 2010). Think of the impression of a place where you had a first date, or a memorable scene from a classic movie. Those feelings may not be expressed clearly through languages. The text data from surveys may suffer from such limitations. The photos taken in a destination by tourists may represent a more subconscious perception of the TDI. Such photos may contain more inherent information on how the TDI is perceived. For example, when thinking of State College, one may think of brick university buildings, which may not have a distinct name in tourists' recollections. Nonetheless, they may form a significant part of one's impression of the place.

Similarly, even the information contained in photo labels derived from machine learning tools and the photos' metadata created by the tourists gave different aspects of TDI: machine learning tools tell about the objects contained in the photo; the metadata represents the authors' local knowledge and interpretation. A brick building in a label became Penn State or Classrooms in the metadata. In the survey, the same building became college life and nostalgia. Thus, one may argue that TDI contains three layers in different abstraction levels: the lower layer contains specific colors, sounds, and objects; the middle layer has places, names, and ambiance; the upper layer is filled with emotions and intentions. These layers are overlapping and not clearly

separated from each other. They nonetheless form the complex TDI of a destination. Different research methodologies may reveal information from these different layers.

5.3 Practical implications

This study proposes a new methodological approach of combining photos and metadata in social media and comparing to and complement survey data. Such methodological contribution entails a more comprehensive and less biased destination composition than previous methods that only examine the information contained in photos. Practically, this study's results could help a local DMO evaluate local TDI using SMA tools and a visitor intercept survey together to promote a more holistic understanding of TDI. From the perspective of SMA, given the accessibility of social media data, the DMO could consider monitoring up-to-date destination image-related attributes via SMA tools. In particular, compared to surveys that capture TDI during a period of time at specific locations, social media data can provide a source of continuous and timely monitoring of how tourists perceive a destination. For example, the local DMO can set up a platform to automatically obtain and analyze SMA metadata and photo data by using machine learning techniques. Destination marketers have the flexibility to choose the time range, specific destination, and even social media platforms. Such information may uncover tourists' changing perceptions and reveal new insights to help the DMO construct marketing and management plans.

Another practical approach is automatically identifying representative colors, shapes, locations, and objects of a destination on social media through machine learning techniques. The advancements of image analysis and machine learning have made this possible. Destination image could be represented by images directly, instead of through articulation of words and

phrases. Through statistical analysis by combining images with metadata and survey results, researchers could identify those images that can promote repeat visits and loyalty. Those images can be directly and automatically promoted and communicated to tourists. For example, an image on the DMO's homepage could be auto-generated and rotated frequently.

However, survey tools are still fundamental and desirable to capture subjective judgment information from tourists and especially emotions that cannot be easily assessed via photos or text. Sophisticated survey instruments are necessary to capture emotions and attachments from tourists' minds and help the local DMO carry out an effective marketing plan using this information. Moreover, survey tools can include structural measurements (e.g., dichotomous and Likert scales) that can systematically assess visitors' emotions, attachments, and satisfactions and monitor the subjectivity through time. The results from structural measurement can also be used to conduct additional advanced statistical analyses (e.g., regression) to examine the measurements' relationship.

5.4 Limitations and future research

The current study is subject to some limitations. First, this study only uses one social media platform—Flickr—to obtain photos and related metadata. We did not investigate social media data from multiple platforms to enhance robustness. The major reason why is that the difference in UGC between local residents and tourists, and the differences between organizational photos and personal photos, are not clearly distinguished on other social media platforms such as TripAdvisor and Instagram. Such differences are essential for the accuracy of the visitors' TDI generation. Future research can aim to overcome such challenges and compare data from different sources. Second, this study utilized Google Vision API to analyze Flickr

photo data. However, such a machine learning technique cannot recognize the actual names of local landmarks in photos. Given the limited number of landmark photos and the fact that training in landmark recognition requires a large number of cases, machine learning techniques for local landmark recognition have not yet been successfully developed. Future research can investigate major tourist locations that have massive photo data to train a sophisticated model and compare it with traditional survey data. Third, future research can move from structured questions to unstructured questions to capture additional information from both the survey and social media data. In addition, even surveys and tourists' photos may not be able to capture the entire spectrum of a tourist's experience and impression. How about the sound and smell of the place which might be too sutle to express in words? Tourists' videos or the emotions expressed during interviews may also offer additional information. If the layers of TDI are valid in the previous discussion, multiple methodolgoies, including surveys and interviews, and crowd-sourced data such as photos, audios, and videos, should be adopted in researching a more comprehensive TDI of a destination.

References

Afshardoost, M. and Eshaghi, M.S. 2020. Destination image and tourist behavioural intentions: a meta-analysis. *Tourism Management*, 81, p. 104154.

Aiden, E. and Michel, J.B., 2014. *Uncharted: big data as a lens on human culture*. Penguin.

Albers, P.C. and James, W.R., 1988. Travel photography: a methodological approach. *Annals of tourism research*, 15(1), pp.134-158.

Baloglu, S. and McCleary, K.W., 1999. A model of destination image formation. *Annals of tourism research*, *26*(4), pp.868-897.

Beerli, A. and Martin, J.D., 2004. Factors influencing destination image. *Annals of tourism research*, *31*(3), pp.657-681.

Biemer, P.P., 2010. Total survey error: design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), pp.817-848.

Bigné, J.E., Sanchez, M.I. and Sanchez, J., 2001. Tourism image, evaluation variables and after purchase behaviour: inter-relationship. *Tourism management*, 22(6), pp.607-616.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), pp.993-1022.

Callegaro, M. and Yang, Y., 2018. The role of surveys in the era of "big data". In *The Palgrave handbook of survey research* (pp. 175-192). Palgrave Macmillan, Cham.

Chang, S., 2018. Experience economy in hospitality and tourism: gain and loss values for service and experience. *Tourism Management*, *64*, pp.55-63.

Choi, S., Lehto, X.Y. and Morrison, A.M., 2007. Destination image representation on the web: Content analysis of Macau travel related websites. *Tourism management*, *28*(1), pp.118-129.

Crompton, J.L., 1979. Motivations for pleasure vacation. *Annals of tourism research*, 6(4), pp.408-424.

Day, J., Skidmore, S. and Koller, T., 2002. Image selection in destination positioning: A new approach. *Journal of vacation marketing*, 8(2), pp.177-186.

Deng, N. and Li, X.R., 2018. Feeling a destination through the "right" photos: A machine learning model for DMOs' photo selection. *Tourism Management*, *65*, pp.267-278.

Du, X., Bian, J. and Prosperi, M., 2018, September. An operational deep learning pipeline for classifying life events from individual tweets. In *Annual International Symposium on Information Management and Big Data* (pp. 54-66). Springer, Cham.

Echtner, C.M. and Ritchie, J.B., 1993. The measurement of destination image: An empirical assessment. *Journal of travel research*, *31*(4), pp.3-13.

Fakeye, P.C. and Crompton, J.L., 1991. Image differences between prospective, first-time, and repeat visitors to the Lower Rio Grande Valley. *Journal of travel research*, *30*(2), pp.10-16.

Gallarza, M.G., Saura, I.G. and García, H.C., 2002. Destination image: Towards a conceptual framework. *Annals of tourism research*, *29*(1), pp.56-78.

Gartner, W.B., 1993. Words lead to deeds: Towards an organizational emergence vocabulary. *Journal of business venturing*, *8*(3), pp.231-239.

Gitelson, R. and Kerstetter, D., 1995. The influence of friends and relatives in travel decision-making. *Journal of Travel & Tourism Marketing*, *3*(3), pp.59-68.

Google Cloud Vision, 2017. Release Notes. available at: https://cloud.google.com/feeds/vision-release-notes.xml (accessed 6 March 2021)

Gurung, D.J. and Goswami, C., 2017. User Generated Content on Sikkim as an Image Formation Agent: A Content Analysis of Travel Blogs. *International Journal of Hospitality & Tourism Systems*, *10*(2).

Hanyu, K., 1993. The affective meaning of Tokyo: Verbal and non-verbal approaches. *Journal of Environmental psychology*, 13(2), pp.161-172.

Hu, Y. and Ritchie, J.B., 1993. Measuring destination attractiveness: A contextual approach. *Journal of travel research*, *32*(2), pp.25-34.

Jiang, Q., Chan, C.-S., Eichelberger, S., Ma, H. and Pikkemaat, B. 2021. Sentiment analysis of online destination image of Hong Kong held by mainland Chinese tourists. *Current Issues in Tourism*, 0(0), pp. 1–22.

Kim, H. and Stepchenkova, S., 2015. Effect of tourist photographs on attitudes towards destination: Manifest and latent content. *Tourism Management*, 49, pp.29-41.

Kirilenko, A.P., Stepchenkova, S.O. and Dai, X. 2021. Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management*, 83, p. 104241. doi: 10.1016/j.tourman.2020.104241.

Krippendorff, K., 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3), pp.411-433.

Lai, K. and Li, X., 2016. Tourism destination image: Conceptual problems and definitional solutions. *Journal of Travel Research*, *55*(8), pp.1065-1080.

Lee, C.K., Lee, Y.K. and Lee, B., 2005. Korea's destination image formed by the 2002 World Cup. *Annals of tourism research*, *32*(4), pp.839-858.

Li, J., Xu, L., Tang, L., Wang, S. and Li, L., 2018. Big data in tourism research: A literature review. *Tourism Management*, *68*, pp.301-323.

Li, X., Pan, B., Law, R. and Huang, X., 2017. Forecasting tourism demand with composite search index. *Tourism management*, *59*, pp.57-66.

Liu, M.T., Liu, Y., Mo, Z. and Ng, K.L., 2020. Using text mining to track changes in travel destination image: the case of Macau. *Asia Pacific Journal of Marketing and Logistics*.

Ma, S. and Kirilenko, A.P., 2020. Climate change and tourism in English-language newspaper publications. *Journal of Travel Research*, 59(2), pp.352-366.

MacKay, K.J. and Fesenmaier, D.R., 1997. Pictorial element of destination in image formation. *Annals of tourism research*, *24*(3), pp.537-565.

Mariani, M., Baggio, R., Fuchs, M. and Höepken, W., 2018. Business intelligence and big data in hospitality and tourism: a systematic literature review. *International Journal of Contemporary Hospitality Management*.

Mariné-Roig, E., 2017. Measuring destination image through travel reviews in search engines. *Sustainability*, *9*(8), p.1425.

Mariné-Roig, E., 2019. Destination image analytics through traveller-generated content. *Sustainability*, *11*(12), p.3392.

Marriott., 2015., Meet the People at the Heart of Marriott. available at: https://news.marriott.com/news/2015/01/20/meet-the-people-at-the-heart-of-marriott (accessed 6 March 2021)

Milman, A. and Pizam, A., 1995. The role of awareness and familiarity with a destination: The central Florida case. *Journal of travel research*, *33*(3), pp.21-27.

Neuendorf, K.A., 2002. Defining content analysis. *Content analysis guidebook*. Thousand Oaks, CA: Sage.

Pan, B. and Li, X.R., 2011. The long tail of destination image and online marketing. *Annals of Tourism Research*, 38(1), pp.132-152.

Pan, B., Wu, D.C. and Song, H., 2012. Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*.

Park, S.B., Kim, J., Lee, Y.K. and Ok, C.M. 2020. Visualizing theme park visitors' emotions using social media analytics and geospatial analytics. *Tourism Management*, 80, p. 104127.

Penn State University Athletics. Retrieved from: https://gopsusports.com/

Phelps, A., 1986. Holiday destination image—the problem of assessment: An example developed in Menorca. *Tourism management*, *7*(3), pp.168-180.

Rehurek, R. and Sojka, P., 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Richards, D.R. and Tunçer, B., 2018. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosyst. Serv*. 31, 318–325.

Ruths, D. and Pfeffer, J., 2014. Social media for large studies of behavior. *Science*, *346*(6213), pp.1063-1064.

Sahin, S. and Baloglu, S., 2011. Brand personality and destination image of Istanbul. *Anatolia–An International Journal of Tourism and Hospitality Research*, *22*(01), pp.69-88.

Scarles, C., 2010. Where words fail, visuals ignite: Opportunities for visual autoethnography in tourism research. *Annals of tourism research*, 37(4), pp.905-926.

Stepchenkova, S. and Li, X.R., 2014. Destination image: Do top-of-mind associations say it all?. *Annals of Tourism Research*, 45, pp.46-62.

Stepchenkova, S. and Morrison, A.M., 2008. Russia's destination image among American pleasure travelers: Revisiting Echtner and Ritchie. *Tourism management*, *29*(3), pp.548-560.

Stepchenkova, S. and Zhan, F., 2013. Visual destination images of Peru: Comparative content analysis of DMO and user-generated photography. *Tourism management*, *36*, pp.590-601.

Stylidis, D., Shani, A. and Belhassen, Y., 2017. Testing an integrated destination image model across residents and tourists. *Tourism Management*, 58, pp.184-195.

Stylidis, D., Woosnam, K.M., Ivkov, M. and Kim, S.S. 2020. Destination loyalty explained through place attachment, destination familiarity and destination image. International *Journal of Tourism Research*, 22(5), pp. 604–616.

Taecharungroj, V. and Mathayomchan, B., 2020. Traveller-generated destination image: Analysing Flickr photos of 193 countries worldwide. *International Journal of Tourism Research*.

Tasci, A.D. and Holecek, D.F., 2007. Assessment of image change over time: The case of Michigan. *Journal of Vacation Marketing*, 13(4), pp.359-369.

Tussyadiah, I.P. and Fesenmaier, D.R., 2009. Mediating tourist experiences: Access to places via shared videos. *Annals of tourism research*, *36*(1), pp.24-40.

United States Census Bureau. 2019. Population estimates, July 1, 2019. available at: https://www.census.gov/quickfacts/fact/table/centrecountypennsylvania# (accessed 6 March 2021)

Waldherr, A., Maier, D., Miltner, P. and Günther, E., 2017. Big data, big noise: The challenge of finding issue networks on the web. *Social Science Computer Review*, 35(4), pp.427-443.

Wang, C.Y. and Hsu, M.K., 2010. The relationships of destination image, satisfaction, and behavioral intentions: An integrated model. *Journal of Travel & Tourism Marketing*, *27*(8), pp.829-843.

Wang, R., Luo, J. and Huang, S.S., 2020. Developing an artificial intelligence framework for online destination image photos identification. *Journal of Destination Marketing & Management*, *18*, p.100512.

Wenger, A., 2008. Analysis of travel bloggers' characteristics and their communication about Austria as a tourism destination. *Journal of Vacation Marketing*, *14*(2), pp.169-176.

Whitaker, S., 2014. Big data versus a survey. FRB of Cleveland Working Paper No. 14-40, 29 December.

Woosnam, K.M., Stylidis, D. and Ivkov, M. 2020. Explaining conative destination image through cognitive and affective destination image and emotional solidarity with residents. *Journal of Sustainable Tourism*, 28(6), pp. 917–935.

Xiang, Z., Du, Q., Ma, Y. and Fan, W., 2017. A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, pp.51-65.

Xiang, Z., Schwartz, Z., Gerdes Jr, J.H. and Uysal, M., 2015. What can big data and text analytics tell us about hotel guest experience and satisfaction?. *International Journal of Hospitality Management*, *44*, pp.120-130.

Yeh, S.S., Chen, C. and Liu, Y.C., 2012. Nostalgic emotion, experiential value, destination image, and place attachment of cultural tourists. In *Advances in hospitality and leisure*. Emerald Group Publishing Limited.

Zhang, K., Chen, Y. and Li, C. 2019. Discovering the tourists' behaviors and perceptions in a tourism destination by analyzing photos' visual content with a computer deep learning model: The case of Beijing. *Tourism Management*, 75, pp. 595–608.

Zhang, X., Yang, Y., Zhang, Y. and Zhang, Z. 2020. Designing tourist experiences amidst air pollution: A spatial analytical approach using social media. *Annals of Tourism Research*, 84, p. 102999.

Table 1: Photo counts by six cities/townships in Centre County, PA

Locations	Photo Count
State College	8,454
Centre Hall	1,086
Bellefonte	912
Boalsburg	521
Philipsburg	515
Snow Shoe	167
Port Matilda	120
Total	11,775

Table 2: Keywords & weights in each topic extracted from image labels

Events and performance		Buildings		Transportation	
Word	Weight	Word	Weight	Word	Weight
event	0.120	property	0.106	vehicle	0.148
sky	0.041	building	0.097	black	0.081
performance	0.039	architecture	0.093	car	0.075
team	0.036	home	0.057	transport	0.062
floor	0.034	facade	0.047	stadium	0.057
music	0.030	build	0.035	aircraft	0.040
light	0.026	tree	0.033	metal	0.035
night	0.023	neighborhood	0.032	white	0.026
community	0.022	town	0.031	tourism	0.026
crowd	0.020	house	0.023	flight	0.024
Natural landscape		Residential and food		Leisure	
Word	Weight	Word	Weight	Word	Weight
plant	0.143	room	0.092	photography	0.123
tree	0.091	art	0.054	recreation	0.087
grass	0.053	furniture	0.052	technology	0.058
landscape	0.029	product	0.052	asphalt	0.052
font	0.023	table	0.049	commercial	0.029
ceremony	0.023	wood	0.034	family	0.029
leaf	0.023	game	0.032	pink	0.026
glass	0.023	branch	0.024	shirt	0.024
stage	0.021	lunch	0.013	parking	0.021
door	0.017	food	0.013	line	0.020

Sports	
Word	Weight
sport	0.145
player	0.059
equipment	0.055
tournament	0.037
job	0.031
muscle	0.031
football	0.029
wrestling	0.026
basketball	0.023
uniform	0.021

Table 3: Keywords & weights in each topic extracted from metadata

Bellefonte historic district		Railroad		University	
Word	Weight	Word	Weight	Word	Weight
building	0.044	image	0.102	college	0.102
show	0.039	railroad	0.088	park	0.080
home	0.035	page	0.024	first	0.051
view	0.028	early	0.023	stream	0.020
detail	0.025	track	0.016	barn	0.019
blogspot	0.024	center	0.015	night	0.019
tree	0.024	archive	0.014	snowfall	0.018
way	0.024	version	0.014	blackie	0.016
bellefonte_historical	0.021	appearance	0.012	target	0.015
company	0.018	wood	0.012	april	0.014
Centre County		Event		Reunion	
Word	Weight	Word	Weight	Word	Weight
bellefonte	0.048	time	0.054	family	0.041
art	0.048	title	0.035	year	0.037
photo	0.043	design	0.035	site	0.035
downtown	0.038	illustration	0.035	water	0.029
car	0.038	people	0.028	foot	0.021
centre_hall	0.025	line	0.028	long	0.021
back	0.023	group	0.024	collection	0.019
famous	0.021	school	0.023	also	0.019
sculpture	0.020	central	0.019	courthouse	0.019
pennsylvania	0.019	spring	0.018	display	0.017
Memory		Infrastructure			
Word	Weight	Word	Weight		
old	0.048	train	0.060		
day	0.036	book	0.035		
man	0.036	city	0.029		
picture	0.030	store	0.026		
historic	0.027	tank	0.022		
second	0.026	former	0.020		
new	0.022	abuse	0.019		
history	0.022	sign	0.016		
area	0.018	boalsburg	0.014		
great	0.016	good	0.013		

Table 4: Keywords & weights in each topic extract from both image labels and metadata

Transportation		Buildings		University	
Word	Weight	Word	Weight	Word	Weight
helicopter	0.093	building	0.099	annual	0.070
vehicle	0.056	architecture	0.078	black	0.042
alecbuck	0.041	property	0.066	football	0.038
visit_alec	0.041	facade	0.037	university	0.029
medical_center	0.040	game	0.036	penn	0.026
transport	0.033	store	0.031	photography	0.025
center	0.030	door	0.026	town	0.024
aircraft	0.024	picture	0.026	nittany	0.023
flight	0.013	neighborhood	0.025	night	0.022
light	0.013	window	0.023	photo	0.022
Natural landscape		Food		Residential	
Word	Weight	Word	Weight	Word	Weight
tree	0.117	restaurant	0.073	room	0.093
plant	0.088	lunch	0.065	floor	0.054
home	0.061	crowd	0.064	seat	0.035
sky	0.054	tourism	0.044	target	0.033
car	0.047	meal	0.039	technology	0.030
grass	0.045	infrastructure	0.030	ceiling	0.029
nature	0.025	supper	0.025	asphalt	0.025
branch	0.022	uniform	0.025	family	0.025
fast	0.021	road	0.021	adaptation	0.021
park	0.021	photographer	0.021	wildlife	0.020
Sports		Event and performance			
Word	Weight	Word	Weight		
college	0.350	event	0.119		
stadium	0.035	statecollege	0.067		
sport	0.035	team	0.063		
table	0.034	player	0.038		
basketball	0.028	performance	0.037		
furniture	0.023	recreation	0.034		
product	0.022	community	0.026		
art	0.021	conference	0.023		
metal	0.014	ceremony	0.020		
wood	0.013	tournament	0.019		

Table 5: Keywords in each category in the question "*What comes to your mind when you think of this area*?"

Top 5	Freq.	Last 5	Freq.
Penn State (Total frequencies: 562; Unique phrases: 12)			
Penn State	388	Graduate school	2
Football	93	College library	1
Campus	21	Football scandal	1
College memory	18	Fraternities	1
Education	14	Tradition	1
Nature and scenery (Total frequencies: 151; Unique phrases: 21)			
Beautiful	39	Ice and snow	1
Mountains	32	Lakes	1
Nature	29	Long winter	1
Trees	15	Outdoors	1
Scenery	9	Sunny	1
Activities (Total frequencies: 83; Unique phrases: 33)			
Art festival	11	Soccer	1
Shopping	11	Thon	1
Lots to do	7	Trout fishing	1
Sports	5	Vacation	1
Art	4	Wrestling	1
Subjective judgment (Total frequencies: 79; Unique phrases: 32)			
Peaceful	11	Smart people	1
Friendly	8	Spacious	1
Nice	7	Surroundings	1
Small	6	Versatile	1
Small town	5	Well-organized	1
People (Total frequencies: 79; Unique phrases: 17)			
Family	30	People	1
Friends	8	Rednecks	1
Son	8	Relatives	1
Daughter	6	Sibling	1
Husband	6	Wife	1
Rural (Total frequencies: 44; Unique phrases: 2)			
Rural	27		
Agriculture	17		
Attraction (Total frequencies: 42; Unique phrases: 19)			
Creamery	18	Park	1
Arboretum	2	Penn's Creek	1
Bald Eagle State Park	2	Poe Valley	1
Beaver Stadium	2	State park	1
Mount Nittany	2	Talleyrand Park	1
Memory and attachment (Total frequencies: 29; Unique phrases: 8)			
Home	11	Pride	2
Memory	7	Nostalgia	1
Youth	3	Second home	1

I love it	2		
Life	2		
Location (Total frequencies: 26; Unique phrases: 7)			
Happy Valley	10	Downtown	1
State College	7	Location	1
MONW (Middle of nowhere)	4		
Bellefonte	2		
Center	1		
Accommodation and food (Total frequencies: 20; Unique phrases: 9)			
Food	11	Hilton Garden Inn	1
Nittany Lion Inn	2	Hotel	1
Apples	1	Waffle shop	1
Beer	1	Waffles	1
Cabins	1		
Community (Total frequencies: 11; Unique phrases: 6)			
Community	5	Traffic	1
Economy	2		
Construction	1		
Lee Metal Products	1		
Rail trail	1		

Table 6: Keywords in each category in the question "*In your opinion, what is the most iconic image of Centre County?*"

Top 5	Freq.	Last 5	Freq.
Penn State (Total frequencies: 470; Unique phrases: 9)			
Penn State	211	Football	3
Nittany Lion	164	College life	1
Nittany Lion shrine	61	Education	1
Campus	23	Engineering	1
College town	5		1
Attraction (Total frequencies: 194; Unique phrases: 18)			
Beaver Stadium	90	Park	1
Old Main	53	Penn's Caves	1
Mount Nittany	26	Prison	1
Creamery	9	Shaver's Creek	1
Arboretum	2	State park	1
Nature and scenery (Total frequencies: 74; Unique phrases: 12)			
Mountains	31	Fall trees	1
Trees	13	Land	1
Nature	7	Landscape	1
Fall foliage	2	Weather	1
Scenery	2	Wild life	1
People (Total frequencies: 30; Unique phrases: 5)			
Joe Paterno	20		
Students	6		
Smiles	2		
Amish	1		
Spirited people	1		
Subjective judgment (Total frequencies: 16; Unique phrases: 7)			
Don't know	8	Space	1
N/A	3	Vintage	1
Culture	1		
Historical	1		
Nice	1		
Rural (Total frequencies: 14; Unique phrases: 4)			
Agriculture	7		
Rural	4		
Barn	2		
Dairy farms	1		
Location (Total frequencies: 13; Unique phrases: 3)			
Happy Valley	6		
State College	6		
Downtown Bellefonte	1		
Activities (Total frequencies: 9; Unique phrases: 5)			
Grange Fair	4		
Sport	2		
Conference	1		

Hiking trails	1
Recreation	1
Community (Total frequencies: 5; Unique phrases: 4)	
Community	2
Blue Loop	1
Construction	1
Truck	1
Accommodation and food (Total frequencies: 3; Unique phrases: 3	
Allen Street Grill	1
Nittany Lion Inn	1
Wine	1

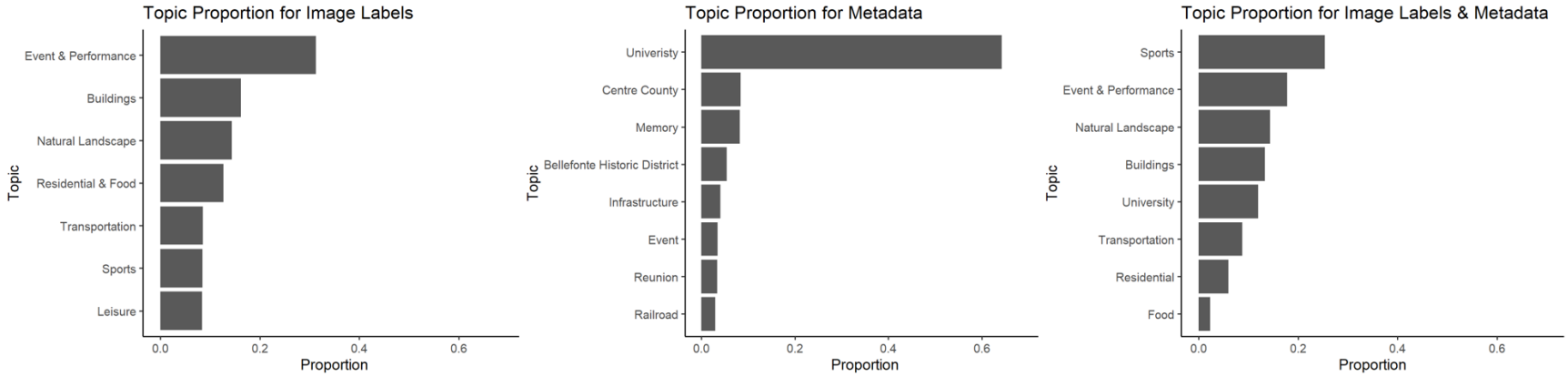

Figure 1 Topic Proportion for Three Datasets