Analyzing Instagram Media Over Time for Zoo Business Insights

Noa Magrisso, Shaked Sapir, and Dr. Michael Fire *Department of Software and Information Systems Engineering Ben-Gurion University of the Negev* Beer-Sheva, Israel
{noamagri, shaksa, mickyfi}@post.bgu.ac.il

***Abstract*— When making insights on business attractions, such as placement of booth locations in entertainment sites and locating areas requiring frequent maintenance, a preliminary acquaintance with the different areas of the place which are observed by visitors over different periods in time, may contribute to the selection of areas to focus on in the future, so that the booth owners can maximize their financial profit or making logistics much easier. As a case study, by analyzing Instagram media taken at the "Woodland Park Zoo" in Seattle (Washington, United States), we will locate the "hot" regions (most visited) in the zoo over time - per season/per month/per hour, so we can bring some business and financial insights to managers' knowledge.**

We provide an easy-to-use tool for displaying those changes over time on a heatmap intuitively, allowing food stall owners to deduce where to place their stalls, where to concentrate focus in cleaning and placing trash cans, etc.

Keywords: Geographic Data Mining, Instagram, Tagging, Zoo, Entertainment Sites, Social Media, Image Classification

I. INTRODUCTION

The data revolution leads to an era in which data is made, stored, and analyzed for many trends such as forecasting, business analytics etc. Such vast amounts of data enable pattern recognition in many organizations and provides the ability of detecting insights and trends we can then use to improve these organizations.

One type of such knowledgebase is social media – Instagram, Twitter, Facebook, and many more are the 'city square' for many discussions, news and trends reporting, recommending where we should go and when, sharing opinions and ideas and many more. Social media contains many kinds of data: Texts, Images, and videos (also in posts and comments), geographic locations, histograms, statistics, etc. In this paper we focused on Instagram, specifically in business analytics out of posts related to organizations so these organizations may deduce some insights about their business and act accordingly. As a case study, we chose 'Woodland Park Zoo' in Seattle (Washington, US) for deducing where should they place their food stands or food carts, where to place many other mobile attractions in the park during the day, such as jugglers, magicians, so they face most visitors along the day, thus raising the potential income from these attractions.

Our method utilizes Instagram posts related to the zoo, such as posts in which the zoo's user was tagged (aka 'tagged posts'), posts in which the zoo's location was tagged (aka 'location posts') and posts which contain *hashtag* (a marking sign to mention a subject in a post, for bringing the post to attention of people interested in this subject) related to the zoo. We extracted animals appearing in those posts, that way we build heatmaps of areas in the zoo visited by people over time in the zoo's opening hours (in different seasons over the year), so the zoo can see where most visitors pay visit in different times of the month/season, thus planning on food stands and attractions' locations during the day, as well as garbage collecting which is required in those regions. We also provide a user-intuitive tool for visualizing those heatmaps, with the option to focus on area of interest on the zoo map (integrated over a Google-Maps map), using several layers of geographical data.

II. RELATED WORKS

A. Image Processing and Classification

Image processing is a large-growing concept, used in many fields of data analysis. A very common algorithm in this field is *image classification*, in which the algorithm can determine which object/concept lies in that image. In order to train and perform such algorithms, we may need to collect a lot of data and train a model to learn how to distinguish between different images, especially for distinguishing between some objects somehow related to each other.

B. Geographical Data Mining

In a work based on geographic data, there is a reference to the coordinates as well as the geometry of the entities. Using Google Maps allows you to place objects according to the coordinates. Adding an image map as a raster, can be anchored on top of Google Maps so that we can get the coordinates of different regions in the park, so we can build a vector layer of polygons on top of it, from which a heatmap can be created.

C. Social Media Data Mining

Nowadays Social media creation and management is very easy, therefore a lot of data is created by the simple user. Because of that, social media mining is used vastly to analyze

various phenomena these days. As their popularity grows over time, many social media frameworks are offering some free-of-charge tools for fetching data from them, such as public API's. The downside of these public data-fetching tools is that most of them are limited in fetching some interesting and important details about that data, so there are many code libraries and packages for dealing with this problem.

III. Methods

In this section we describe the complete data processing pipeline.

First, we fetched posts from Instagram using a dedicated code library. A *media* is a type of resource in Instagram: image, video, reel, story, etc. As an Instagram's post might contain more than a single media, we extracted all the media out of the posts we got. For the scope of this project, we decided to focus on images alone, so in case a post contained a video – we cut a representing image out of it (using the video's metadata brought using the library). Thus, we got a total of 15,000 images to analyze, together with their timestamp (the time at which they were posted). Next, we have to be able to detect which animal appears in each image so we can construct a visualization over time of which animals were seen and when, over the zoo's map with its different animal's regions.

So, we decided to examine some pre-trained image classification models, such as VGG16, VGG19 and Google's InceptionV3. As we saw that Inception got the best classifying results – we chose it to be our model. But then, after examining all 58 kinds of animals in the zoo, we realized that only 33 animals species were classified properly by the model (and some of them under a different name than their name in the zoo – e.g. a 'wolf' in the zoo was always detected as a 'white wolf' by the Inception model, as all the wolves in the zoo are white), thus we had 25 more animal species in the zoo which couldn't be identified properly. Therefore, we made two adaptations:

1. We built a name-alias map between an animal's name in the zoo and its potential name/class declared by the model, so we can detect the right animal even if the Inception model gives us a 'right' prediction but with a different name than the name we expected, like the 'wolf' example we mentioned above.
2. For each animal which couldn't be recognized by the Inception, we gathered a dataset from the iNaturalist website [1] which contains a lot of images: plants, animals and many more, and over the combined dataset of all those animals' pictures we trained a Catboost model (aka *cm*) for covering up upon the Inception's lack of classification. For this manner, we fixed a threshold *th* determining that a *cm* classification decision will be accepted if its probability is equal or higher than *th* .In particular, we added a class named 'UNKNOWN' for images which *cm* wasn't 'sure enough' about (its probability was lower than *th*), thus we could ignore images which contained only plants, humans and other stuff which were irrelevant to our ability to infer where the picture was taken.

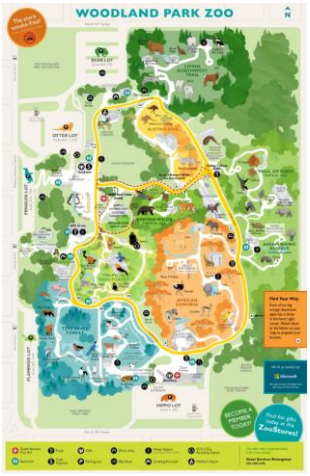

Figure 1: *The zoo official map*

For training *cm* we used about 1,000 images for each animal and got 78.635% accuracy. Further examining this revealed that most of the time *cm* did great, but there were some classes which were closely-related to each other – such classes are '*river otter'* and '*asian small-clawed otter'* as different kinds of otters which look a lot like each other, and another example is the different kinds of pigs: '*warty pig', 'pudu', 'kunekune pig' and* '*tapir'.* As most of the animals which were close to each other were close in their regions at the zoo, we were forgiving for these misclassifications (One is able to see it in a confusion matrix in the project's notebook).

Here we present a t-SNE reduction of those animals which demonstrates the stated above:

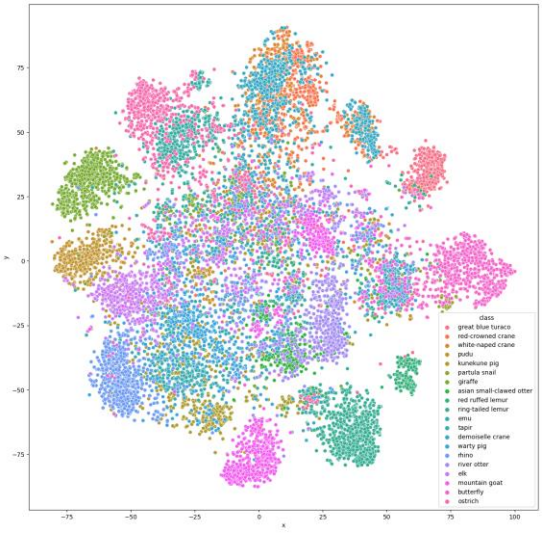

Figure 2: *A t-SNE reduction of the animals' vectors*

As mentioned above, we mapped each animal to its region in the zoo, so we can then see which regions are most and least popular by visitors.

IV. Result Analysis

We analyzed the different regions in the zoo in the manner of how much popular they are, using grouping on the animals'

counts resulted by our algorithm, when we divided the results into two groups regarding the zoo's opening hours: The first season is from May to August (months 5-8) and the second is from September to April (months 9-12, 1-4). Then, for each season, we used Google Maps on which we anchored a map of the zoo (as detailed in the zoo's website [2]), so we can build heatmaps upon these counts and project them over the map, to see the interesting sites in the zoo. The polygons representing the zoo's different regions were sampled by hand over the zoo's map, exported as GeoJSON objects and were shown in the heatmap with some transparency so we can see through the zoo's trails, meadows, toilets, and some other facilities of it, resulting in where exactly we should place some attractions for the visitors.

A. Heatmaps per season

We built two heatmaps on top of the zoo's park map, each belong to seasonal opening hours according to viator.com [3] – the first is of months 5-8 (May-August, aka *summer*) and the second one is of months 1-4 and 9-12 (September to April, aka *winter*). We used an image of the official park map from 2023, and pasted it onto Google Maps as raster, so that on the one hand we can clearly see the division of the park into the regions, entrances, paths, parks, and other information, and on the other hand we can get the coordinates of any point we want in the park. We created the polygons and exported them to a GeoJson file, so we then converted it to the Geo Data Frame we were working on.

The heat maps allow support for three things:

1) Moving on the scroll bar at the top of the map allows dragging and moving between months (forward and backward) in sequence.
2) Changing the number in the text box allows jumping directly to the requested month.
3) Choosing from the layers icon in the right corner of the map the different heat maps by month.

A change in the scroll bar is automatically reflected in the selected layer. For convenience, the summer season is in shades of red, while the winter season is in shades of blue.

Figure 3: *Summer season heatmaps per months*

Figure 4: *Winter season heatmaps per months*

B. Showing optimal locations for some purposes

We show the concept of "optimal recommended locations" in the zoo: we place markers representing optimal locations to place the food stand near "hot" regions and even inside a park with benches and places to sit, so that people can sit comfortably to eat and drink, taking a break between trips in the different regions. By adding a marker, you can click on the marker and the coordinates will pop up.

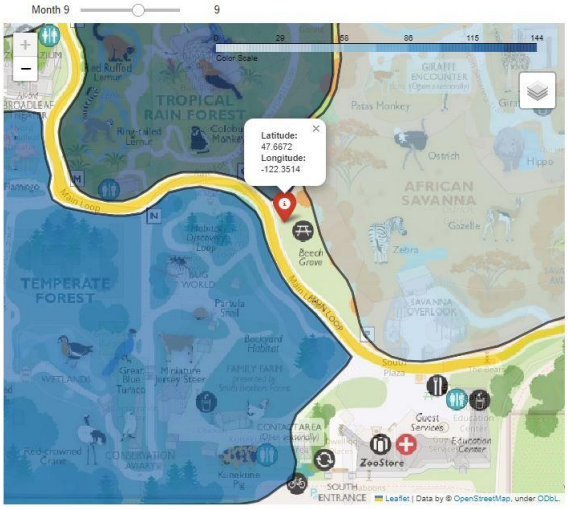

Figure 5: *An optimal location for a stand in September*

C. media distribution over months per region

We show graphs of media amounts per month in each region, so we can visualize the difference between visitors' numbers in each region over several months, thus having a complete sight on months which are high or low populated. For example, we can see that there is a significant jump in the "Tropical Rain Forest" region, as well as in the "Living Northwest Trail" region, and these findings do agree with the temperature in which the animals in these regions are

comfortable with. In addition, the "Temperate Forest" region, which was also the most visited region in the summer season, is also significant in the winter season, and this indeed makes sense since this is a region which is adapted to all 4 existing seasons of the year. The "African Savanna" region is also prominent, especially in the month of November.

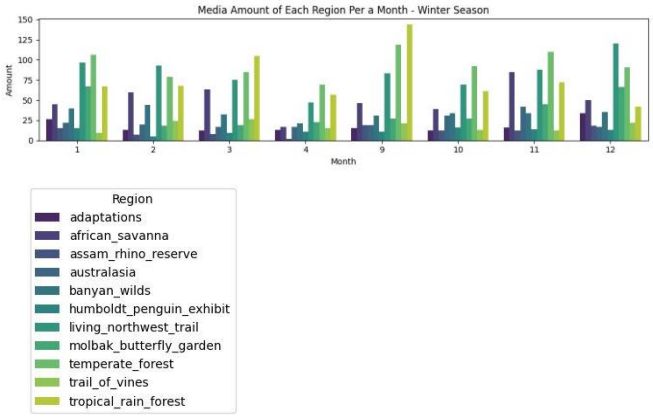

Figure 6: *A bar plot of the media amount of each region per a month in the winter season*

D. media distribution over months by hours

We show the distribution of pictures per hour each month, so we can see the best hours to visit the zoo in each month / which hours are the most popular among the zoo's visitors, so the zoo's facilities can be logistically prepared for it.

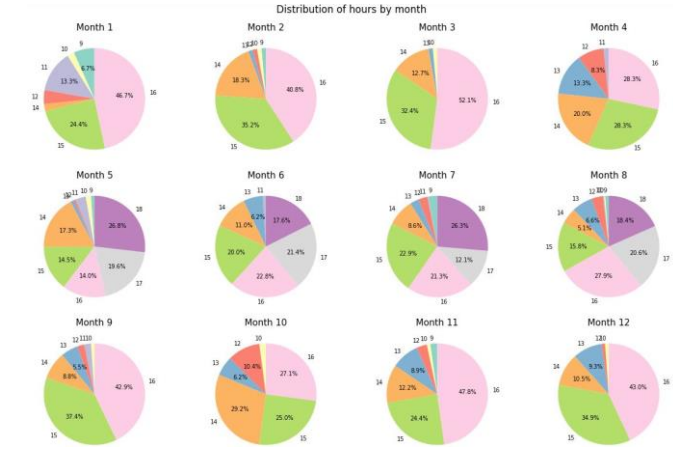

Figure 7: *A pie-chart of the distribution of hours per month*

Overall, the hours when most posts are uploaded start at noon, and especially in the afternoon (when the visit is over). It is possible that there is an increase of posts also at noon, when the visitors are resting, because most of the animals in the park are more active in the morning or afternoon [3].

In the summer season, i.e., in the months of May to August, you can see that posts are also uploaded at 18:00 onwards, this is because the opening hours end later.

Further analyze code can be viewed in the following GitHub repository: https://github.com/NoaMagrisso/Instagram_Photos_Analysis_Over_Time

V. Conclusions and Future Directions

We analyzed the popularity of different regions in the zoo using geographic and image analysis, which could help the zoo's managers to gather insights and ideas which areas should be treated more carefully, where to put some attractions for money making, etc. This method could be applied to many locations and organizations and help them with their own needs. In future work, we will shrink the granularity of regions to capture a single animal per polygon so we can be more accurate and make better insights about the park, we could use data from other social media like Facebook and add some vector layers such as plants, trails, benches and facilities, gathering visitors opinions regarding seasons/ months/ hours per regions in which there was less interest (using NLP) so we can perform better geo-layer compounds and have much more interesting insights.

Acknowledgments

Throughout the work we used the ChatGPT tool for debugging purposes (exploring error messages raised during the coding phase).

References

[1] https://www.inaturalist.org/taxa/1-Animalia - iNatuarlist website

[2] Woodland Park Zoo: All for Wildlife - Woodland Park Zoo Seattle WA - Woodland park zoo's website

[3] "https://www.viator.com/Seattle-attractions/Woodland-Park-Zoo/d704-a1315#experiences", a website containing details and opening hours of many attractions over the US