Happiness is a Choice: Sentiment and Activity-Aware Location Recommendation

Jia Wang
The University of Melbourne
Australia
jiaw8@student.unimelb.edu.au

Yungang Feng
The University of Melbourne
Australia
yungangf@student.unimelb.edu.au

Elham Naghizade
The University of Melbourne
Australia
e.naghi@unimelb.edu.au

Lida Rashidi
The University of Melbourne
Australia
rashidi.l@unimelb.edu.au

Kwan Hui Lim
The University of Melbourne
Australia
kwan.lim@unimelb.edu.au

Kate Lee
The University of Melbourne
Australia
kate.lee@unimelb.edu.au

ABSTRACT

Studying large, widely spread Twitter data has laid the foundation for many novel applications from predicting natural disasters and epidemics to understanding urban dynamics. Recent studies have focused on exploring people's emotional response to their urban environment, e.g., green spaces versus built up areas, through analysing the sentiment of tweets within that area. Since green spaces have the capacity to improve citizen's well-being, we developed a system that is capable of recommending green spaces to users. Our system is unique in the sense that the recommendations are tailored with regard to users' preferred activity as well as the degree of positive sentiments in each green space. We show that the incoming flow of tweets can be used to refine the recommendations over time. Furthermore, We implemented a web-based, user-friendly interface to solicit user inputs and display recommendation results.

CCS CONCEPTS

• **Information systems → Web applications**; **Social networks**;

KEYWORDS

Social Networks, Sentiment Analysis, Location Recommendation

ACM Reference Format:
Jia Wang, Yungang Feng, Elham Naghizade, Lida Rashidi, Kwan Hui Lim, and Kate Lee. 2018. Happiness is a Choice: Sentiment and Activity-Aware Location Recommendation. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3184558.3191583

1 INTRODUCTION

Green spaces in urban areas bring positive effects to citizens both mentally and physically [5, 8]. This positive effect can be associated with a variety of activities that people can do in green spaces. Online reviews and ratings as well as comments in social media on these green spaces can be utilised as a reference for people to choose which space they would prefer to go to. However, users may wish to search for the parks that are suitable for a specific type of activity, while these reviews and ratings contain no or little information about that specific activity.

The growth in the field of recommender systems (RS) has enabled people to receive personalised suggestions and services. There are various types of RS, such as collaborative filtering based RS, knowledge based RS and context-aware RS [14]. These techniques can provide recommendations using information such as users' ratings, items' features or the contextual information. Recently, there has been an increasing trend to use Twitter data to provide location recommendations since tweets have the unique property capturing time, location and textual information [13].

The research in this area has mainly focused on either analysing the sentiments with regard to a certain location [3] or detecting the most popular topics related to a given location [12]. However, a purely sentiment-aware system is not capable of tailoring the recommendations based on users' preferences, especially with regard to green spaces that have the capacity to offer a variety of activities (as opposed to a restaurant for instance). An activity-aware recommender system on the other hand may fail to detect the underlying sentiment around a location. Imagine a user is looking for green spaces that have "running" as one of their most prevalent topics. There might be a certain green space that is actually unpopular for running, e.g., many users have complained about the large number of potholes on the running track of that park, but the system would recommend it due to the large amount of running related tweets.

To address this issue, we build a novel system that provides highly tailored suggestions of green spaces based on users' preferred activities, e.g. workout, relaxing or socializing as well as the general sentiment around green spaces. It is worth noting that these three groups of activities were selected based on the popular facilities that are available in green spaces. Our system analyses geo-tagged tweets within green spaces to learn whether or not they refer to our pre-specified set of activities and determines the *popularity* of the green spaces with respect to those activities. We also use natural language processing techniques to determine the sentiment of the labelled tweets at each location. To this end, we use the NRC Word-Emotion Association Lexicon dictionary to map the tweet content to different emotions. The frequency of words in each group is later used to compute the *polarity*, i.e., the extent of positive sentiments, of the tweets.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191583

Our system is capable of dynamically updating the recommendations based on the incoming tweet stream. It also provides users with a fine-grained searching capability as well as an interactive, easy-to-use interface.

1.1 Related Work

1.1.1 Topic modelling and sentiment analysis in tweets. For analysing tweets, an important task is to extract the topic and activity from tweets. The authors in [7] applied Latent Dirichlet Allocation (LDA), which is an unsupervised algorithm, to analyse tweets that were posted in London. It abstracted 20 topics in the urban area of London, such as "Photography and Tourism" and "Sport and Games". However, the tweets have a word limitation of 140 characters, which also limits the amount of information in one tweet. The autors in [2] argue that LDA is not an effective method for modeling the topics of tweets due to the limited number of words contained in one tweet [2]. In [1], Akbari et al. utilise hashtags in tweets to study and identify specific topics and events.

Many studies focus on analysing sentiments in tweets using various methodologies, such as lexicon-based analysis [4]. Emojis have become an important element of a tweet since about 19.6% of tweets contain emojis [10]. However, there are limited studies that consider using emojis in their topic and sentiment analysis tasks.

1.1.2 Recommendation System and Urban Planning Using Twitter Data. The authors in [13] introduced a context-aware system that can recommend tourist attractions based on tweets. This system analyses tweets to extract the sentiment on a specific tourist attraction, and make recommendations based on the extracted information as well as location. Our work, on the other hand can be used for location recommendations to both citizens and tourists with a focus on their desired activity as well as the overall sentiment around a location.

2 SYSTEM OVERVIEW

The system is a web-based application, which consists of a back-end server and a front-end user interface. The back-end server is responsible for crawling tweets, data pre-processing, classifying tweets into different activities and analysing their sentiment. The front-end component is utilised for interaction with users, handling users' requests, communication with the back-end server and presenting the recommendation results.

Back-end server and front-end user interface are de-coupled, which enables the system to update or optimise its recommendation algorithms without affecting the front-end part of the system. This is important since the recommendations can potentially change over time, e.g., due to the change of season, or occasional events.

2.1 Data collection

Our dataset is composed of 210,000 geo-tagged tweets within 159 parks within the city centre from July 2014 to September 2017. In addition, the boundary of the parks was extended by 100 meters to account for GPS noise. A thorough cleaning of tweets was required to ensure the quality of lebeling and sentiment analysis, however, the detail of our pre-processing is not discussed for brevity.

2.2 Back-end Component: Mining Tweets

The purpose of our back-end component is to provide the recommendation based on popularity and polarity of each park. Tourists may be more likely to choose those green spaces which are popular while citizens may prefer green spaces that can help to keep positive moods. Therefore, these two aspects are both taken into consideration for recommending green spaces.

2.2.1 Popularity. The popularity of an activity is measured by the number of occurrences of tweets related to that activity over the total number of tweets in a park. The key problem of popularity is how to filter tweets with a specific topic. In this section, we discuss the method and process of filtering relevant tweets.

Topic Modeling. We address the task of filtering specific topical tweets as a text categorization problem. One simple way to achieve that is by using keyword searching, which is suitable for an activity like "BBQ" or "picnic". Nevertheless, due to the coexistence of many possible meanings for a word in English (polysemy), this is not a suitable solution for a keyword like "run". For example, "How many Carradines are there left? Will we ever run out of them?", which contains the word "run", however it is not related to workout.

Another widely used unsupervised approach is Latent Dirichlet Allocation(LDA) [7]. LDA is a topic model which is capable of demonstrating the topics in documents in terms of probability models. The advantage of using LDA is that it does not require manually constructed training data. However, LDA is not efficient in handling fine-grained topic problems. Moreover, the problem of polysemy of "run" and "walk" might not be solved without supervision.

As a result, we adopted a supervised method to infer topics of tweets. Similar to [6], we used linear Support Vector Machine (SVM) with Text Frequency-Inverse Document Frequency (TF-IDF) to retrieve our topics. However, as a supervised learning method, there is the requirement of assigning the tweets to their ground-truth labels. To this end, we manually labelled a small sample of our dataset denoted as T_M (around 1% of the data). We use hashtags to semi-

Algorithm 1 Incremental training

function ActivityLabelling(T_U, T_L)
  $frequentWordsVector \leftarrow$ TF-IDF(T_L)
  $model \leftarrow$ SVM.train($frequentWordsVector$)
  $T_L^{new} \leftarrow model$.predict($T_U$)
  return T_L^{new}
procedure Training
  $T_L \leftarrow T_M$ + *hashtags*
  while T_U is available **do**
    $T_L^{new} \leftarrow$ ActivityLabelling(T_U, T_L)
    $T_L \leftarrow T_L \cup T_L^{new}$

automatically label the training data (ActivityLabeling function in Algorithm 1). Twitter allows users to use hashtags to classify their posts. For example, "Good morning! Beautiful day for some hill sprints @alicek66 #run #running". Using the categorised hashtags (Table 1), the processing of labelling tweets can be achieved in a semi-automatic manner.

For relaxing or socializing activities, we only used the hashtags to label the tweets since polysemy is not an issue in those categories.

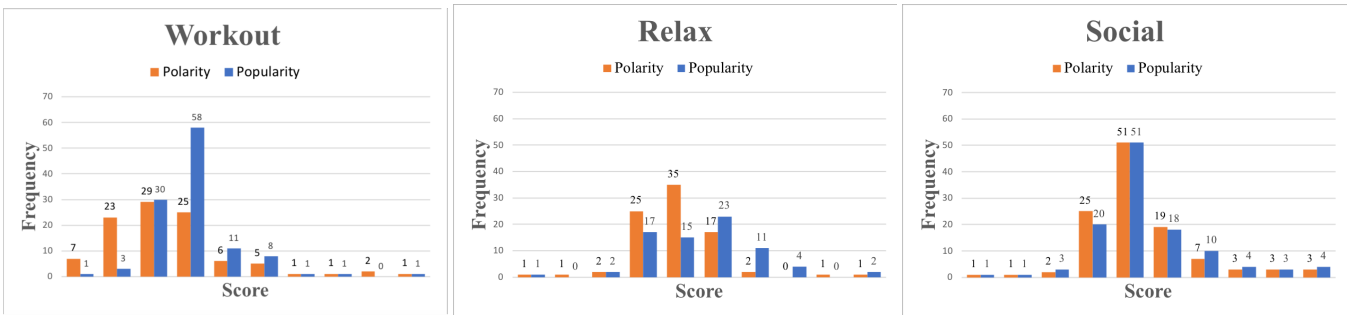

Figure 1: Histogram of popularity and polarity scores in parks.

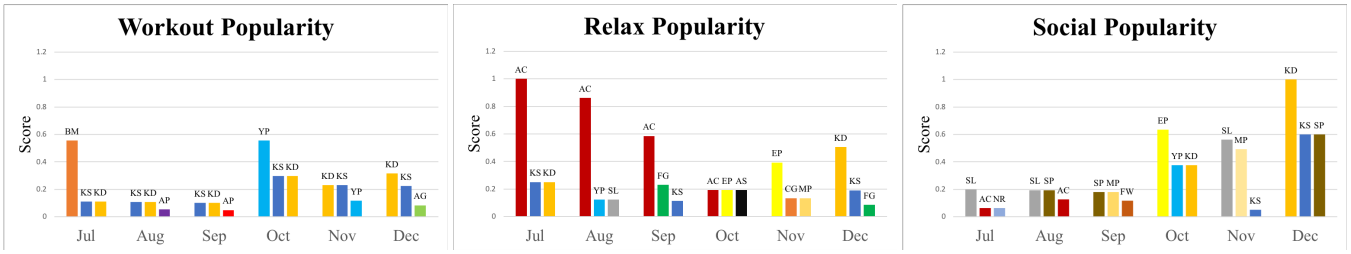

Figure 2: The effect of time in the top 3 most popular park query. Each color represents a different park.

Activity	Hashtags
Workout	#Endomondo, #endorphins, #run, #running, #parkrun, #morningrun, #jog, #jogging, #walking, #walk, #walkies, #ride, #cycling
Socializing	#relax, #relaxing, #meditation, #reading, #lunchbreak, #chill, #mindfulness, #yoga, #yogainpark
Relaxation	#meetup, #wedding, #bbq, #picnic, #catchup, #friends, #festival, #hangout, #party, #birthday

Table 1: Hashtags for the three types of activities.

However, finding the labels for the workout activity is treated as a binary classification problem, and hence the tweets were labelled in "workout" or "others" classes using a combination of manually labeled tweets and hashtags. Our classification model achieves a high

	Precision	Recall	F1-score	Support
Others	0.98	1.0	0.99	1716
Workout	0.99	0.91	0.95	433

Table 2: Classification results for workout and non-workout (others) classes.

accuracy, as detailed in Table 2. After labelling the tweets, we use the frequency of our three topics within parks to estimate the popularity p_a^g of each of the activities $a \in \{workout, socializing, relaxation\}$ for each park $g \in \{g_1, g_2, ..., g_{159}\}$ using Equation 1. Figure 1 shows the popularity distribution (blue histogram) for each activity.

$$p_a^g = \frac{\#tweets_a^g}{Max_{g' \in g_{1..159}}(\#tweets_{g'}) - Min_{g' \in g_{1..159}}(\#tweets_{g'})} \quad (1)$$

Moreover, to make sure that the system can efficiently update the recommendations based on the incoming tweet stream, we modeled the labeling process as shown in Algorithm 1. The Training procedure uses the previously labeled tweets as well as the new incoming tweets with hashtags to label the new unlabeled tweets. This process saves the time of manually labeling tweets and is essential since the popularity of green spaces changes over time, e.g., following a seasonal trend (Fig 2).

2.2.2 Polarity. Polarity in this work is an important aspect that measures the sentiments of users towards a specific activity in the park. One of the commonly used approaches for sentiment analysis is to calculate the number of positive and negative words in tweets to determine the sentiment. The authors in [9] use the Emotion Lexicon listed on NRC Word-Emotion Association Lexicon (EmoLex) [11] to analyse the polarity of tweets. We improve upon this approach by considering emojis in our sentiment analysis. Figure 1 shows the polarity score (orange histogram) of each activity in the parks. Also, typical emojis that indicate positive or negative sentiment are shown in the Figure 3.

2.3 Front-end Component: User Interface

An overview of the front-end interface is shown in Figure 4. The front-end interface consists of three main components:

(1) **Preference stepper** (red zone 1 in Figure 4): the selector with which user can choose their preferred options about the green spaces.
(2) **Main map** (red zone 2 in Figure 4): the map component for demonstration, which is based on OpenStreetMap™.
(3) **Search results** (red zone 3 in Figure 4): the list of the search results based on users' search query, which includes the

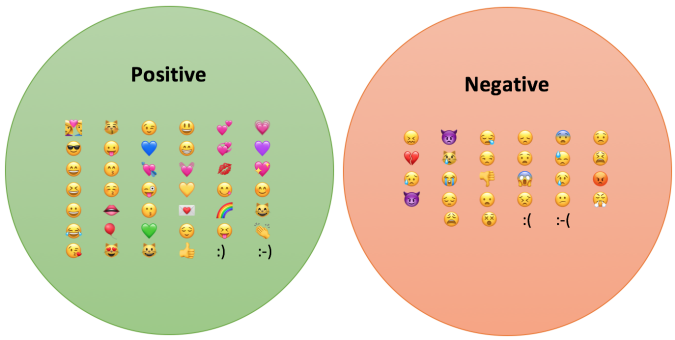

Figure 3: Positive and negative emojis used for the sentiment analysis.

polarity and popularity scores of the recommended green space.

(4) **Highlighted search result region** (red zone 4 in Figure 4): the region that belongs to the search results. It includes the highlighted park region and a number of interactive icons.

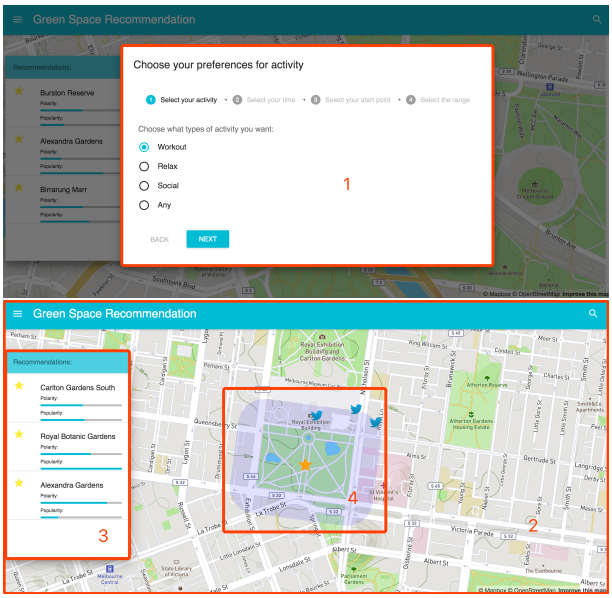

Figure 4: Front-end UI screenshots

Use Case Scenario. Users can start a new search through the front-end UI. Apart from the option of looking for green spaces with the highest scores, a user can ask for a customised search with options for activity, time, start point and distance (shown as the red region numbered "1" in Figure 4). The options and the corresponding choices are illustrated in Table 3.

The application sends users' query to the back-end server, and the server processes the query to provide recommendations. The recommended green spaces will be shown on the main map. We also incorporate interactive components in the region of the recommended green space such as a set of sample tweets in the green space and a word cloud (Figure 5) to further suggest activities, events or available facilities that might be of interest to a user based on their choice of activity. For instance, the word cloud shown in Figure 5 (right) includes keywords such as tan, bike path,and cycle that are indicative of popular types of workout facilities in the park.

Option	Choices
Activity	{Workout, Relax, Social, Any}
Time	{Day, Night}
Start Point	Drop a pin within the map area
Range	1 km to 5+ km

Table 3: Available options for a search query.

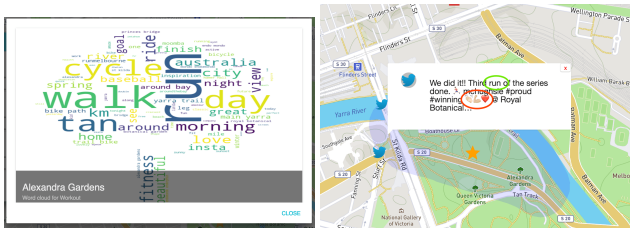

Figure 5: The word cloud (left) shows an overview of the workout-related topics around a certain park. The sample tweet (right) provides a testimonial for the park.

3 CONCLUSION

This paper proposes a novel recommendation system for green spaces, which considers both the sentiment and user's preferred activity. This system identifies the activities in tweets to analyse and provide a fine-grained, highly personalised recommendation. One of the advantages of our system is its incremental training process that does not require the entire data to be available at once. An interactive front-end UI is also implemented to recommended green spaces based on user preferences and provide related information through a word cloud and users' testimonials.

ACKNOWLEDGEMENT

This work was funded by the Melbourne Networked Society Institute. Kate Lee is supported by the Clean Air and Urban Landscapes hub of the National Environmental Science Program. Kwan Hui Lim is supported by a Defence Science and Technology Group Postdoctoral Fellowship.

The authors also thank Andrew MacKinlay, Noel Faux, Gail Hall, and research partners from IBM-Research - Australia and the City of Melbourne for their inputs.

REFERENCES

[1] Mohammad Akbari, Xia Hu, Liqiang Nie, and Tat-Seng Chua. 2016. From Tweets to wellness: wellness event detection from Twitter streams.. In *AAAI*. 87–93.

[2] Gennady Andrienko, Natalia Andrienko, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, , and Dennis Thom. 2013. Thematic patterns in georeferenced Tweets through space-time visual analytics. *Computing in Science & Engineering* 15, 3 (2013), 72–82.

[3] Karla Z. Bertrand, Maya Bialik, Kawandeep Virdee, Andreas Gros, and Yaneer Bar-Yam. 2013. Sentiment in New York city: A high resolution spatial and temporal view. (2013).

[4] Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*. 450–453.

[5] Terry Hartig, Richard Mitchell, Sjerp De Vries, and Howard Frumkin. 2014. Nature and health. *Annual Review of Public Health* 35 (2014), 207–228.

[6] Thorsten Joachims. 1998. Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning* (1998), 137–142.

[7] Guy Lansley and Paul A. Longley. 2016. The geography of Twitter topics in London. *Computers, Environment and Urban Systems* 58 (2016), 85–96.

[8] Andrew CK Lee and Ravi Maheswaran. 2011. The health benefits of urban green spaces: a review of the evidence. *Journal of Public Health* 33, 2 (2011), 212–222.

[9] Kwan Hui Lim, Kate E. Lee, Dave Kendal, Lida Rashidi, Elham Naghizade, Stephan Winter, and Maria Vasardani. 2018. The Grass is greener on the other Side: Understanding the effects of green spaces on Twitter user sentiments. *2018 Web Conference (WWW'18)* (2018).

[10] Nikola Ljubešić and Darja Fišer. 2016. A Global Analysis of Emoji Usage. In *Proceedings of the 10th Web as Corpus Workshop*. 82–89.

[11] Saif M Mohammad and Peter D Turney. 2013. *Nrc emotion lexicon*. Technical Report. NRC Technical Report.

[12] Tatsuhiro Sakai, Keiichi Tamura, and Hajime Kitakami. 2015. Identifying main topics in density-based spatial clusters using network-based representative document extraction. *IEEE 8th International Workshop on Computational Intelligence and Applications (IWCIA)* (2015), 77–82.

[13] Yosuke Sakamoto and Yasufumi Takama. 2017. Proposal of sentiment-based tourist spot recommendation system using RDF database. *IEEE 10th International Workshop on Computational Intelligence and Applications (IWCIA)* (2017), 61–66.

[14] Shahab Saquib Sohail, Jamshed Siddiqui, and Rashid Ali. 2017. Classifications of recommender systems: A review. *Journal of Engineering Science & Technology Review* 10, 4 (2017), 132 – 153.