TWEET SENTIMENT ANALYSIS ON GREEN SPACES

Lino Garda Denaro[*], Yudianto Sujana[*], Adindha Surya Anugraha[**], Hafsah Fatihul Ilmy[**]

[*]National Cheng Kung University/Department of Computer Science and Information Engineering, Tainan, Taiwan

[**] National Cheng Kung University/Department of Geomatics, Tainan, Taiwan

ABSTRACT

Twitter has become one of the most significant resources for text mining. Twitter can provide information about human activities, mobility, and emotional patterns along with location data. Many types of text research can be made with these data, one of which is sentiment analysis. This study evaluates the potential of deriving emotional responses of individuals from tweets while they experience and interact with urban green space. A machine learning model using Support Vector Machine (SVM) and corpus from over 2000 movie reviews has been made. This model is used to classify incoming tweets into positive and negative sentiments. Then the web-based recommender system has been built to provide suggestions for green spaces based on users' preferred activities.

Keywords—sentiment analysis, text mining, twitter data, SVM, machine learning

1. Introduction

Green spaces provide a place for human activity. There are a variety of green spaces, such as a park, which encourages a lot of people to use their social media in finding an interesting park to conduct their activity. People will be looking for good park recommendations to do their activities, but there is not much information about the park recommendations for a specific activity.

The growth in the field of recommender systems (RS) has enabled people to receive personalized suggestions and services. Recommender systems try to identify the need and preferences of users, filter the huge collection of data accordingly and present the best-suited option before the users by using some well-defined mechanism [12].

The recommender system provides recommendations using users' opinions, item features, ratings, or contextual information. Nowadays, contextual information is easy to get through Twitter because it has location data for users who activate it, and Twitter users usually tweet (give their opinions via Twitter) their feelings, so these tweets can be analyzed to make a recommender system.

In this paper, a recommender system has been developed for various activities such as workouts, socializing, and relaxation at the park based on Twitter data. The system analyzes geotagged tweets within the park to find out whether the park is suitable or not for users to conduct their activities (workout, socializing, and relaxation). In addition, the system is capable of updating the latest recommendations based on the incoming tweets each day (real-time data).

1.1 Related Work

A. Sentiment in New York City: A High-Resolution Spatial and Temporal View

The research in this area mainly focused on analyzing the sentiments with regard to a certain location. The author [2] developed a classier specially tuned for 140-character Twitter messages or tweets, using keywords, phrases and emoticons to determine the tweet's mood. This method, combined with geotagging provided by users, enables the author to gauge public sentiment on extremely re-grained spatial and temporal scales.

B. Recommendation System and Urban Planning Using Twitter Data.

The authors in [11] 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.

2. Methodology

There are several steps that have been done in order to create a recommender system (RS) as shown in the flowchart method in Figure 3.1.
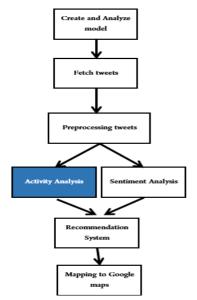

Figure 3.1. Flowchart of Method

A. Data Collection

The tweets in this research have been collected from Twitter Streaming Application Programming Interface (API), filtered by location to those geotagged within a bounding box of four New York City Parks. A total of 1,100 tweets about Central Park, Bronx Park, Pelham Bay Park, and Van Cortland Park have been retrieved in the first week of July 2018.

B. Create and Analyze Model

The model in this research has been created using a movie review dataset (1,000 positive and 1,000 negatives) [7]. The non-words, extra spaces, stop words have been removed and all sentences have been changed into lower case. Then, a classifier based upon Text-Frequency- Inverse Document Frequency (TF-IDF) has been constructed with 2,000 features, three minimum document frequencies, and 60% maximum document frequencies. Afterward, the data has been split into 80% training and 20% testing.

Three classification methods have been used to try the model, including: logistic regression, Naïve Bayes, and Support Vector Machine (SVM). The cross-validation result shows that SVM is the suitable method for the model.

Table 1 Model Accuracy

Model	Accuracy
Logistic Regression	0.84
Naïve Bayes	0.79
Support Vector Machine	0.84

C. Fetch Tweets

The tweets have been collected from Twitter Streaming Application Programming Interface (API), filtering by keyword and location to those geotagged within a bounding box of 4 New York City Park. A total of 1,100 of these tweets about Central Park, Bronx Park, Pelham Bay Park, and Van Cortland Park have been retrieved in the first week of July 2018.

Table 2 Tweet Location

Central Park	-73.9817503577, 40.7649522867, -73.9496211749, 40.8008585051
Bronx Park	-73.8837856795, 40.8417750476, -73.8705677534, 40.8691708974
Pelham Bay Park	-73.8267176937,40.8459270955, -73.776764233,40.8887011171
Van Cortlandt Park	-73.9057641969, 40.8814179494, -73.8689885307, 40.9112746603

Table 3 Keywords

workout	'workout', 'running', 'walking', 'run', 'parkrun', 'jog', 'jogging', 'walk', 'walking', 'ride', 'cycling'
socializing	'relax', 'relaxing', 'meditation', 'reading', 'lunch', 'chill', 'mindfulness', 'yoga'
relaxation	'meetup', 'wedding', 'bbq', 'picnic', 'catchup', 'friends', 'festival', 'hangout', 'party', 'birthday'

D. Pre-processing Tweets

The pre-processing process have been applied to every tweet: retrieving English-only tweets, removing the hyperlink, removing non-words, changing tweets to lowercase, removing stop words, modifying some words e.g., "that's" to "that is", etc.

E. Sentiment Analysis

Sentiment Analysis is used to measure the sentiment of the tweet. The tweets have been classified into positive and negative sentiment. In more detail, sentiment analysis calculated the number of positive and negative words in tweets to determine the sentiment. The number of positive and negative words is strongly correlated with the condition of the parks. Users tend to give positive words for their tweet about good parks and tend to give negative words in their tweet about bad parks.

F. Activity Analysis

Activity Analysis is used to measure the percentage of positive tweets in the parks for several activities. The threshold of the activity type which used in this paper is a workout, socializing, and relaxation. The result was used to determine the popular park.

G. Recommendation System

A recommendation system is used to give a recommendation to users. The recommendation depends on the type of activity. The system will give a different recommendation for each activity type. The recommendation will show the popular parks to users based on the activity type.

H. Mapping to Google Maps

Mapping to Google Maps is used as an interface. The motivation is to give the easiest way for users to understand the result from the recommendation system. It will show an emoticon-based on the recommendation system in each park and in each activity type. For example, the happy emoticon shows that the park is very popular for certain activity types, and the sad emoticon shows the park is not popular for certain activity types.

3. Result and Discussion
3.1. EXPERIMENTAL EVALUATION

A. Classification Result

Support Vector Machine (SVM) is used to classify the tweets from users to give a good recommendation for users to go to parks to do certain activity types. The recommendation is obtained from the measurement in sentiment analysis and activity analysis. The example result of classification can be seen in figure 4.1

There are eight columns in figure 4.1 showing Id, Tweets, Text, Date, Sentiment, Parkname, Location, and Keyword. The classification result shows the information from the users, such as the tweets, the location where users were tweeting, and the time while users were tweeting. The important thing from the classification result is that the system defines whether the tweets have positive sentiment or negative sentiment. Furthermore, the systems can also define the tweets belonging to relaxation activity, workout activity, or socializing activity.

Id	tweet	text	date	sentiment	parkname	location	keyword
1	Omg so happy that Louis the vampire and Rachel fro...	omg so happy that louis the vampire and rachel fro...	2018-06-22 15:51:00	1	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	relaxation
2	I thought her friends were okay or whatever, the b...	thought her friends were okay or whatever the boy...	2018-06-22 15:51:51	0	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	relaxation
3	Another walk in the books! #teamprh @ 79th Street...	another walk in the books teamprh th street boat b...	2018-06-22 15:53:30	0	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	workout
4	Me when I dub all my friends for my future goddess...	me when dub all my friends for my future goddess...	2018-06-22 15:58:35	1	Pelham Bay Park	[-73.8267176937, 40.8459270955, -73.776764233, 40...	relaxation
6	I want off this ride please.	want off this ride please	2018-06-22 16:02:07	0	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	workout
7	I need the On The Run Tour remix of Run This Town	need the on the run tour remix of run this town	2018-06-22 16:07:00	0	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	workout
8	@donald_faison Happy birthday	donald_faison happy birthday	2018-06-22 16:08:39	1	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	relaxation
9	When your daughter drops a cookie on the floor and...	when your daughter drops cookie on the floor and y...	2018-06-22 16:10:46	1	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	workout
10	@divyaspandana @AmitShah @BJP4India Piteous rant b...	divyaspandana amitshah bjp india piteous rant by ...	2018-06-22 16:22:38	0	Pelham Bay Park	[-73.8267176937, 40.8459270955, -73.776764233, 40...	relaxation
12	Wait,McLaren is running an INDYCAR next year?	wait mclaren is running an indycar next year	2018-06-22 16:22:53	0	Pelham Bay Park	[-73.8267176937, 40.8459270955, -73.776764233, 40...	workout
14	Anyone running pickups	anyone running pickups	2018-06-22 16:27:22	0	Central Park	[-73.9817503577, 40.7649522867, -73.9496211749, 40...	workout

Figure 4.1. Classification Result

B. Mapping to Google Maps

Mapping to Google Maps is the latest step from the method flowchart in this paper. It shows the emotion based on the calculation of the number of positive tweets divided by the total number of tweets.

$$Score = \frac{\sum positive\ tweets}{N\ tweets}$$

(1)

In this paper, the score threshold is 0 to 25%, 25% to 50%, 50% to 75%, and 75% to 100%. More detail can be found in figure 4.2.

Score	Emoticon
>= 0.75	😃
>= 0.5	🙂
>= 0.25	😐
>= 0	🙁

Figure 4.2. Scores and Emoticons

After defining the threshold of the score and the emoticon, the last step is to give the marker to google maps in each park based on the sentiment from tweets. The result of mapping to google maps can be seen in figure 4.3
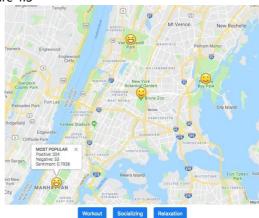

Figure 4.3. Mapping to Google Maps

Figure 4.3 shows the result of the workout activity. It can be seen that Central Park and Van Cortland Park have the highest recommendation for users in doing the workout. The marker in Manhattan Central Park shows the number of positive sentiment, negative sentiment, and the score of the sentiment. The information of each park can be seen by clicking the marker in that park.

Users can see the recommendation of socializing activity and relaxation activity by clicking the emoticon. The result of each activity in each park may be different.

4. Conclusion

The result of this research shows that Twitter data can be used for developing a recommendation system on certain locations based on tweets sentiment analysis. This recommendation system can be improved with a good dataset and use the deep learning method.

REFERENCES

[1] Ali, A., Stratmann, T., Park, S., Schöning, J., Heuten, W., & Boll, S. (2018). Measuring, Understanding, and Classifying News Media. CHI 2018 (pp. 1-13). Montreal, QC, Canada: ACM.

[2] Bertrand, K., Bialik, M., Virdee, K., Gros, A., & Baryam, Y. (2013). Sentiment in New York City: A High Resolution Spatial and Temporal View.

[3] Chen, X., Sykora, M., Jackson, T., & Elayan, S. (2018). What about Mood Swings? Identifying Depression on Twitter. International World Wide Web Conference Committee (pp. 1653- 1660). Lyon, France: Creative Commons CC.

[4] Chiu , S.-I., & Hsu, K.-W. (2018). Predicting Political Tendency of Posts on Facebook. ICSCA 2018 (pp. 110-114). Kuantan, Malaysia: Association for Computing Machinery.

[5] Das, S., Sun, X., & Dutta, A. (2015). Investigating User Ridership Sentiments for. Journal of Transportation Technologies, 69-75.

[6] Lim, K., Lee, K., Kendal, D., Rashidi, L., Naghizade, E., Winter, S., & Vasardani, M. (2018). The Grass is Greener on the Other Side: Understanding the. WWW 2018 (pp. 275-282). Lyon, France: Creative Commons CC.

[7] Pang, B., & Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. Proceedings of the ACL. Baltimore, Maryland, USA.

[8] Roberts, H., Resch, B., & Sadler, J. (2018). Investigating the Emotional Responses of Individuals to Urban Green. Urban planning, 21-33.

[9] Roberts, H., Sadler, J., & Chapman, L. (2017). Using Twitter to investigate seasonal variation inphysical activity in urban green space. Geo: Geography and Environment, 1-14.

[10] Roberts, H., Sadler, J., & Chapman, L. (2018). The value of Twitter data for. Urban studies journal, 1-18.

[11] Sakamoto, Y., & Takama, Y. (2017). Proposal of sentiment-based tourist spot recommendation system using RDF database. IEEE 10th International Workshop on Computational Intelligence and Applications (IWCIA) (2017), (pp. 61-66).

[12] Sohail, S., Siddiqui, J., & Ali, R. (2017). Classifications of Recommender Systems: A review. Journal of Engineering Science and Technology Review, 132-153.

[13] Wang, J., Feng, Y., Naghizade, E., Rashidi, L., Lim, K. H., & Lee, K. (2018). Happiness is a Choice: Sentiment and Activity-Aware Location. WWW '18 Companion Proceedings of the The Web Conference 2018 (pp. 1401-1405). Lyon, France: Creative Commons CC.