Can We Predict Eat-Out Preference of a Person from Tweets?

Md. Mahabur Rahman[1], Md Taksir Hasan Majumder[2], Md Saddam Hossain Mukta[3],
Mohammed Eunus Ali[4], Jalal Mahmud[5]
[1,2,3,4]Bangladesh University of Engineering & Technology
[5]IBM Research-Almaden
{mmr2512[1],mythoss1092019[2],saddam944[3]}@gmail.com,
eunus@cse.buet.ac.bd[4], jumahmud@us.ibm.com[5]

ABSTRACT

Twitter, a microblogging site, has become a major platform of communication of users on the web. Recently, location based social networking sites such as Foursquare have become popular which enable users to publish their visited places through check-ins. In this paper, we present a study of users' eat-out preferences who share their Foursquare restaurant check-ins through Twitter. Our study reveals a strong correlation of a user's eat-out preference and the linguistic features of her tweets, i.e., word use. Hence, our proposed model enables one to predict a user's eat-out preference from her word-use in Twitter.

CCS Concepts

•**Human-centered computing → Social network analysis;** *Social recommendation;* •**Computing methodologies → Classification and regression trees;**

Keywords

Twitter; Foursquare; Check-ins; Eat-out preferences

1. INTRODUCTION

Twitter allows a user to express herself or share information with others using tweets. In recent studies, researchers have been able to find many interesting insights such as personality, value and preferences of users by analyzing the texts of these tweets [2, 5, 1]. Location based social networking sites such as Foursquare have also become increasingly popular in recent years with the proliferation of location-aware technologies and smartphones. By using these sites, users can share their location information of visiting different venues/places through check-ins. A Foursquare check-in consists of latitude, longitude, the name and the category of the venue, and the time of the check-in.

In this paper, we exploit users' check-ins of different categories of restaurants, i.e., cheap, moderate, expensive, and very expensive through the Foursquare links of users' tweets, and combine these check-ins information with the users' pattern of word-use in tweets to build a model that enables us to predict a user eat-out preferences from her tweets. A large number of applications can be benefited from our eat-out preference prediction model that include recommendation systems, business location identification, and financial prediction of users.

Though a plethora of works have been produced to identify user behaviors from a single social media usage, an emerging research trend is to combine the data from multiple social media usages and finding interesting insights of user behaviors from that is gaining attention in recent years. In this paper, we collect tweets and Foursquare check-ins of 731 Twitter users. We first find the user patterns of visiting four different categories of restaurants. Then, we build a model that correlates between user's word use in her tweets and visiting frequency in different categories of restaurants. Finally, we can predict a person's eat-out preference, i.e., frequency of visiting different categories of restaurants, from her tweets.

2. METHODOLOGY

In this paper, we identify eat-out preference of a person from her pattern of word use in Twitter and check-ins shared via Foursquare. To this end, we first crawl only English tweets of a user containing the Foursquare check-ins. Later, we collect all the tweets of a user in a separate file having the Foursquare check-ins. The more Foursquare check-ins a user has, our system builds more accurate model to predict the result. Then, we use an *html* parser to get the price information (restaurant category) from the links of the foursquare check-ins. Later, we compute visiting frequency of a user to a specific restaurant type. Next, we compute LIWC category of words in each user's tweets. Then, we perform Pearson's correlation analysis between LIWC category of words and her visiting frequency to a restaurant type based on the price. We build a linear regression model with the statistically significant LIWC category of words and visiting frequency in a restaurant type. We also compute the strength (i.e., the R^2) of our linear regression model.

3. DATA COLLECTION AND ANALYSIS

We used Twitter advance search technique to find users whose English tweets contain Foursquare links. If a user uses Foursquare links, the check-ins of her tweets usually contain

WebSci '16 May 22-25, 2016, Hannover, Germany

ACM ISBN 978-1-4503-4208-7/16/05.
DOI: http://dx.doi.org/10.1145/2908131.2908199

keywords such as “4sq”, and “Foursquare”. To find out required Twitter users, we search with these two keywords through advance search technique. We use English words for LIWC analysis. After selecting Twitter id of users who regularly tweets using Foursquare links, we collected their tweets through http://greptweet.com/. We collect tweets of 731 users. Among the tweets of the users, we discover a total of 72662 Foursquare links of restaurants that are categorized into four categories by Foursquare based on the food price. We have found 23986, 36187, 10335 and 2154 links for cheap, moderate, expensive and very expensive categories of restaurants, respectively. We used *html* parsing to gather information about the place or restaurant found in the link. Thus, we calculated how many times a user visited each type of restaurant. As a ground truth data, we use frequency of visits in a restaurant type by a user. After finding out the frequency, we have conducted LIWC based analysis of users' tweets. Then, we applied Pearson's correlation analysis to find significant correlations between LIWC category of words in user's tweet and her frequency of visit to each type of restaurant. Table 1 shows the correlations between few LIWC category of words and restaurant type. To save space, we only show some significant correlations.

Table 1: Pearson's Correlations between LIWC category of words and visiting frequency in different types of restaurants. [for significance level: *p<0.05, **p<0.01]

category	**Cheap**	**Moderate**	**Expensive**	**V. exp.**
family	**0.116****	-0.029	**-0.112****	-0.035
friend	**-0.145****	-0.002	**0.211****	-0.0006
leisure	**-0.114****	**0.107****	0.023	-0.0008
money	**-0.075***	-0.014	**0.086***	**0.093***

We have performed linear regression analysis using WEKA machine learning toolkit to predict the eat-out preference of a person from her social media word use [3]. To eliminate collinearity among independent LIWC category words and visiting frequency of a restaurant type, we compute lasso penalized linear regression using *glmnet* R package [1] [4]. Finally, we perform the linear regression analysis with a 10-fold cross-validation with 10 iterations. Table 2 presents that R^2 (9.12%-17%) and adj-R^2 (7.12%-14%) were small but substantial across all restaurant categories. Motivated by the work [1], we also conduct prediction potential with major classifiers using WEKA machine learning toolkit.

Table 2: Strength of linear regression models and correlation coefficient with different categories of restaurant.

Restaurant category	**R^2 of linear regression**	**Adjusted R^2 of linear regression**	**Correlation coefficient**
Cheap	14%	11%	0.278
Moderate	10.8%	7.24%	0.1139
Expensive	17%	14%	0.319
V. expensive	9.12%	7.12%	0.106

Table 3 presents the best classifier, it's TPR, FPR and AUC scores for computing visiting frequency of each of the restaurant category. We observe that our classifiers achieved moderate improvement over random chances.

4. DISCUSSION

Our work identifies a number of significant correlations between the users' psycholinguistic categories in their tweets and frequency of visiting a restaurant type. These psycholinguistic categories can be useful to predict the preference of a user to a restaurant type. We observe that *very expensive* category has positive correlation with *money* and *certain* LIWC category of words. For instance, it is likely that the users who usually write about money in their tweets and express their opinions with certainty, they tend to visit very expensive restaurants. Similarly, expensive restaurants are positively correlated with *friend* category words, in contrast to being negatively correlated with *family* category words. Similar to users of very expensive restaurants, expensive restaurants also use money related words frequently. The result of the analysis also focuses that visiting a *moderate* type restaurant correlates with the usage of *health*, and *leisure*, words of LIWC categories. Moderate restaurants are strongly correlated only with *leisure* category words. It can be assumed that users are more prone to visit restaurants at leisure time. Again, we find for cheap type restaurants that their users differ from expensive restaurant users in several ways. *Swear*, *family*, *friend*, *ingest*, and *money* category of words have opposite correlations with cheap and expensive restaurant users. It shows that cheap and expensive restaurant category users are psychologically different. However, we also find few correlations that are not intuitively explainable.

Table 3: Best performing classifier to predict different restaurant categories.

Restaurant category	**Highest AUC achieving classifier**	**AUC**	**TPR**	**TNR**
Cheap	SVM	0.624	0.624	0.376
Moderate	Logistic Reg.	0.608	0.59	0.41
Expensive	Logistic Reg.	0.67	0.624	0.376
V. expensive	Logistic Reg.	0.564	0.551	0.449

5. CONCLUSIONS

In this paper, we demonstrate how can we predict a person's eat-out preference from her tweets and Foursquare check-ins. The main advantage of our approach is that we can successfully predict eat-out preference of a person in spite of not using Foursquare check-ins in her tweets.

6. ACKNOWLEDGMENT

This research is funded by ICT Division, Ministry of Posts, Telecommunications and Information Technology, Government of the People's Republic of Bangladesh.

7. REFERENCES

[1] J. Chen, G. Hsieh, J. U. Mahmud, and J. Nichols. Understanding individuals' personal values from social media word use. In *CSCW*, pages 405–414. ACM, 2014.

[2] H. Cramer, M. Rost, and L. E. Holmquist. Performing a check-in: emerging practices, norms and'conflicts' in location-sharing using foursquare. In *MobileHCI*, pages 57–66. ACM, 2011.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *SIGKDD*, 11(1):10–18, 2009.

[4] T. Hastie and J. Qian. Glmnet vignette., 2014.

[5] K. Joseph, C. H. Tan, and K. M. Carley. Beyond local, categories and friends: clustering foursquare users with latent topics. In *Ubicomp*, pages 919–926. ACM, 2012.