Study of People's Eat-out Behavior using Natural Language Processing (NLP) on Tweets for Targeted Marketing

Abul H. Fahad (20764979), Syed Anjoom Iqbal (20640625), Mohamed H. Abdalla (20800994)

***Abstract*— Social networking sites (SNS) such as Twitter can capture wide varieties of information of every user ranging from psychological attributes to life-style preferences. To extract valuable knowledge from this ocean of information was the prime motivation for this project. Efforts were made to extract user's eat-out preference from his/her tweets using a dataset of tweets from 671 social networkers who were active both in Twitter and Foursquare. These tweets were processed through Natural Language Processing (NLP) and Machine Learning model (Classification-Regression) was built to predict eat-out preference in different categories of restaurants. Both training and testing scores of all applied algorithms were presented. The described workflow and tools can be utilized in targeted marketing applications.**

***Index Terms*—Human-centered computing; Twitter; Foursquare; Check-ins; Eat-out preferences; Predictive models, Supervised learning**

I. Introduction

Twitter is a major social-media platform of communication among users on the web. This platform allows users to express their opinions and share information with others via tweets (micro-blogs). In addition, location-based social networking sites such as Foursquare have become popular, enabling users to publish their visited places through check-ins [1]. A Foursquare check-in consists of latitude, longitude, the name and the category of the venue, and the time of the check-in. Accordingly, Social Media researchers are being able to find many interesting insights, such as personality, value, and preferences of users by analyzing the texts of their tweets and check-ins [2] [3] [4]. Drawing inspiration from these works, authors in [5] collected tweets with Foursquare check-ins of 671 Twitter users. They first figured out each user's pattern of visiting four different categories of restaurants, and then built a model that correlates between user's word use in tweets and visiting frequency in different categories of restaurants. Finally, they proposed a prediction model, which can predict eat-out preference of a person by analyzing the tweets with Foursquare check-ins.

The current study utilized dataset from [5]. In the project, Natural Language Processing (NLP) and Machine Learning (ML) were applied to extract knowledge and make prediction on people's eat-out behavior by analyzing their tweets. For Natural Language processing, Linguistic analysis was carried out using tools like LIWC and Empath [6] and [7], which provide linguistic features for Machine learning workflow. Webscraping was performed for URL mining, where restaurant category was extracted from Foursquare webpage of the restaurant. All features and restaurant visiting numbers were normalized for machine learning workflow. Machine Learning workflow consists of formulating regression and classification problems.

The rest of the paper is organized as follows: Section II contains review of linguistic analysis tools, Section III is Methods and Approach and Section IV contains the Results. Finally, Section V concludes the paper.

II. Review of Linguistic Analysis Process

In this study, Natural Language Processing (NLP) or Linguistic Analysis was performed using two very similar tools- LIWC [6] and Empath [7]. LIWC or Lingusitic Inquiry and Word Counting was originally developed in order to provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples. The LIWC tool relies on an internal default dictionary that defines which words should be counted in the target text files. With each text file, approximately 80 output variables are written as one line of data to a designated output file. This data record includes the file name, 4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (periods, commas, etc).

But like other popular lexicons, LIWC is small: it has only 40 topical and emotional categories, many of which contain fewer than 100 words. Further, many potentially useful categories like violence or social media don't exist in current lexicons, requiring the ad hoc curation and validation of new gold standard word lists. Other categories may benefit from updating with modern terms like "paypal" for money or "selfie" for leisure. To address these problems, came Empath: a living lexicon mined from modern text on the web. Empath allows researchers to generate and validate new lexical categories on demand, using a combination of deep learning and crowdsourcing. Empath also analyzes text across 200 built-in, pre-validated categories drawn from existing knowledge bases and literature on human emotions, like neglect (deprive,

refusal), government (embassy, democrat), strength (tough, forceful), and technology (ipad, android). Empath combines modern NLP techniques with the benefits of handmade lexicons: its categories are transparent word lists, easily extended and fast. And like LIWC (but unlike other machine learning models), Empath's contents are validated by humans.

III. Methods & Approach

In order to solve the problems described in previous section, the following workflow, outlined in Figure 1, was followed.

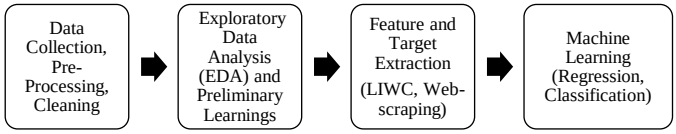

Fig. 1. Data Mining and Machine Learning Workflow for our project

A. Data Collection, Pre-Processing, Cleaning

The dataset was collected from authors of paper [5]. It consists of 671 *.txt files. File names were the Twitter handles of each user. Each file had hundreds of Tweets with a timestamp. All the 671 *.txt files were read into a master file with 23M+ tweets. Each tweet was split into reasonable section i.e. `tweetid`, `day_of_week`, `month`, `day`, `year`, `hour`, `minute`, `second`, `tweet` etc. There were many inconsistencies encountered in reading the text files and those exceptions were handled appropriately. A CSV file was prepared. In addition to the mentioned columns, few more columns were added, i.e. `tweetcount_by_all_users`,`userid`, `total_tweet_by_a_user`. These columns were created to for better-tracking on the master CSV file and not losing information when transforming the data from text files to the CSV.

B. Exploratory Data Analysis and Preliminary Knowledge Extraction

In this step, the goal was to analyze the `day_of_week` and hour columns of every tweet in the CSV file from previous step. For each user 15 bins were created, 7 bins for 7 days of the week, and 8 bins for every 3 hour window of each day. Table I presents the bins.

TABLE I
BINS FOR ANALYZING WHEN USERS ACTIVE ON TWITTER PORTAL

Day_of_week bins	Mon, Tue, Wed, Thu, Fri, Sat, Sun
Time_of_use bins	Night_1= 00:00- 02:59 Night_2= 03:00- 05:59 Morning_1= 06:00- 08:59 Morning_1= 09:00- 11:59 Afternoon_1= 12:00-14:59 Afternoon_2= 15:00- 17:59 Evening_1= 18:00- 20:59 Evening_2 = 21:00- 23:59

The number of times a user tweeted matching with each bin was counted. This raw data was exported to a CSV file. In addition, this number was normalized with respective `total_tweet_by_a_user` column value and the normalized data is exported to another CSV file. visually represent the data of one user by plotting the percentage tweet from the user on 7 days of the week.

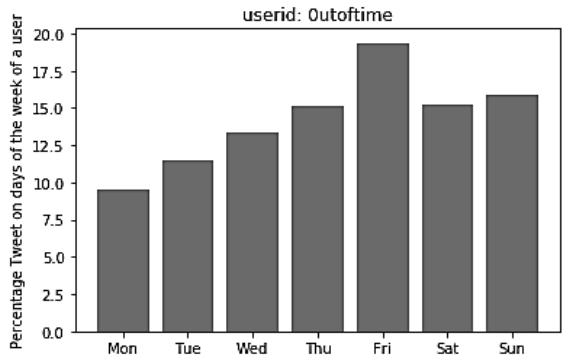

Fig. 5. Tweets (%) of user '0utoftime' throughout different days of week

From fig. 5, it could be interpreted that this user tweets more and more as it approaches towards the weekend and tweets the most on Friday. Therefore, this person might be a good candidate for ads and promotions for eating out on Friday. We also plot the percentage tweet from the same user on every 3 hours interval of the day.

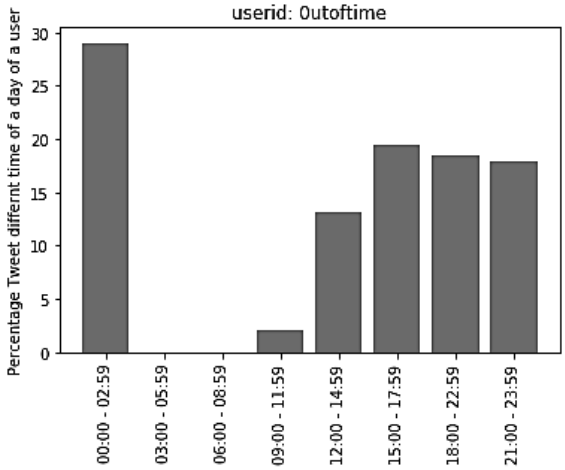

Fig. 6. Tweets (%) of user '0utoftime' throughout different time of day

From fig. 6, we can understand that, this user stays up late as most of the tweets are after midnight. In addition, we see that before noon there is very minimum activity compared to afternoons and evenings. Therefore, most likely this person will be a good candidate for ads and promotions of a dinner place compared to a place for lunch. We see that, this analysis gives us good insight about the person's habits and schedules. Therefore, we group the users with similar habits and schedules and make separate CSV lists of each group. For simplicity, we find the max value out of the 7 bins of days of the week and the max value of the 8 bins of time windows of the day for every user, and add that user to the list of both that day and that time window. This way we generate 15 lists of users with similarity. Each user is in 1 day list and 1 time list only.

C. Feature and Target Extraction

1) Feature Matrix Creation

First, we conduct LIWC based analysis of users' tweets. For this purpose, we use LIWClite7- a student version of LIWC tool [6]. LIWC analyzes 70 different features of text in different categories. The categories are linguistic processes (word count, words longer than 6 letters, total pronouns, common verbs etc.),

psychological processes, personal concerns (work, leisure etc. related words) and spoken categories (assent, noninfluencies etc.). The psychological processes is divided into five categories. It includes social process (words related to family, friends etc.), affective process (words related to positive emotion, negative emotion, anger etc.), cognitive process (insight, discrepancy, inhibition etc. related words), perceptual process (see, hear etc. related words) and biological process (body, health etc. related words). But LIWC provides limited features and many potentially useful categories like violence or social media don't exist in current lexicons, requiring the ad hoc curation and validation of new gold standard word lists. To address these problems, Empath was proposed [7]. Empath is a living lexicon mined from modern text on the web. Empath allows researchers to generate and validate new lexical categories on demand, using a combination of deep learning and crowdsourcing. Empath can generate and validate a category for social media. Empath also analyzes text across 200 built-in, pre-validated categories drawn from existing knowledge bases and literature on human emotions, like neglect (deprive, refusal), government (embassy, democrat), strength (tough, forceful), and technology (ipad, android). Empath combines modern NLP techniques with the benefits of handmade lexi-cons: its categories are transparent word lists, easily extended and fast. And like LIWC (but unlike other machine learning models), Empath's contents are validated by humans. Hence, our Emapth workflow is outlined in Figure 2:

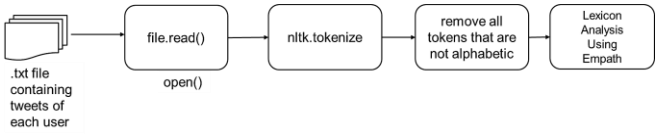

Fig. 8. Linguistic Analysis workflow for feature extraction

2) Computation of visiting different categories of restaurants (Target Matrix creation)

The computation of number of visit to different categories of restaurants for each user was a challenging one. The process was carried out into two steps. First, each text file was analyzed and URLs were extracted. This process was repeated for all 671 users and all URLs for each user was gathered in a file. In the next step, for every user we analyzed each single URL to identify Foursquare-listed Restaurants and then used web-scraping to identify the price category of the restaurant. The details of these two workflow is outlined in the Figure 3 below. Thus, we obtain the frequency of visit for each category of restaurant.

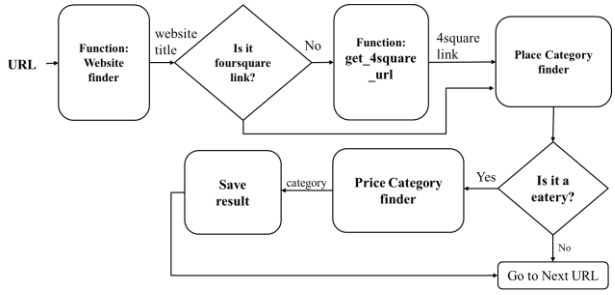

Fig. 9. Restaurant Category Estimation from URL using web-scraping

D. Prediction Model Creation (Machine Learning)

This stage was conducted in three steps:

- Stage One (No Feature Selection)- Keeping all features from linguistic analysis
- Stage Two (Implementing Feature Selection)- Feature Reduction application
- Stage Three (Hyper-parameter Tuning)

At each of the above steps, the prediction problems were formulated as regression problem and classification problem for each price category (cheap, moderate, expensive and very expensive).

The workflow formulation is outlined in the Figure 4 below:

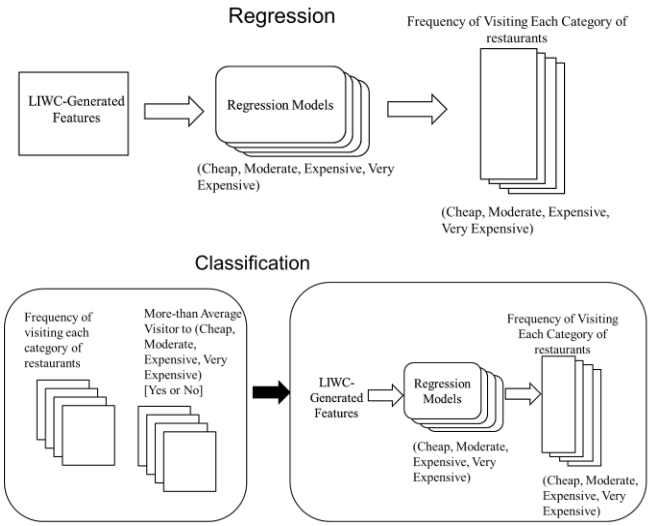

Fig. 10. Restaurant Category Estimation from URL using web-scraping

IV. Results

To improve our models, our data frame was inspected for outliers by using the Tukey outlier labeling. If a value lies below the first quartile or above the third quartile by more than 1.5 times the interquartile range (IQR), it will be considered an outlier. After applying this method, 1% of the data were considered outliers. Next, outliers will be capped to either the first quartile-1.5 * IQR or the third quartile+1.5 * IQR (since the outliers here are legitimate values not an error value, therefore replacing them with mean or dropping them is not preferred). In addition to that the features are normalized using Z-Score normalization.

1) Stage One (No Feature Selection)

In this stage, the Regression problem of finding the exact percentage of visiting each class of restaurant, relative to visiting all the classes, for each user using the default settings of each regression model was attempted. The results are outlined in Table I-V:

TABLE I
KNN Regression R2 Scores

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.321	0.248	0.30	0.247
Testing Score	-0.034	-0.134	-0.158	-0.292

TABLE II
SVM REGRESSION R^2 SCORES

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.051	0.045	0.029	-0.0007
Testing Score	-0.021	0.006	0.017	-0.063

TABLE III
LASSO (LEAST ABSOLUTE SELECTION AND SHRINKAGE OPERATOR) REGRESSION R^2 SCORES

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.18	0.045	0.167	0
Testing Score	0.092	0.017	-0.009	-0.008

TABLE IV
RANDOM FOREST REGRESSION R^2 SCORES

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.833	0.821	0.834	0.708
Testing Score	-0.021	-0.137	-0.171	-0.479

TABLE V
NEURAL NETWORK REGRESSION R2 SCORES

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.999	0.999	0.999	0.999
Testing Score	-0.389	-0.865	-0.727	-0.848

Next the Classification problem of finding if the user tends to visit the specific class of restaurants (cheap, moderate, expensive, and very expensive) more than the average users or not was formulated. Using the default settings of each classification model, the following results were obtained:

TABLE VI
KNN CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	57.7%	52.3%	55.3%	61.9%
AUC	0.54	0.50	0.55	0.45
f1	0.56	0.52	0.54	0.48

TABLE VII
SVM CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	60.1%	55.3%	66.6%	73.8%
AUC	0.65	0.54	0.65	0.62
f1	0.58	0.55	0.62	0.50

TABLE VIII
RANDOM FOREST CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	53.5%	49.4%	62.5%	70.8%
AUC	0.60	0.47	0.63	0.46
f1	0.45	0.58	0.49	0.51

TABLE IX
NEURAL NET CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	55.3%	54.1%	56.5%	67.2%
AUC	0.60	0.54	0.61	0.62
f1	0.54	0.54	0.54	0.59

2) Stage Two (Feature Selection):

Next, dimensionality reduction using feature ranking with mutual information between the features and each target as a metric has been implemented, as shown in Figure 4 below. The best 20 features for each specific target have been used in building the specific target prediction model.

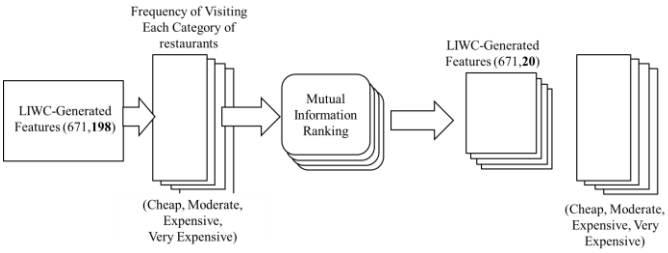

Fig. 11. Restaurant Category Estimation from URL using web-scraping

The following Tables shows the top 3 features ranked for each target and their mutual information score.

TABLE X
FREQUENCY OF VISITING CHEAP RESTAURANT

air_travel	achievement	swimming
0.13093052	0.0870093	0.08630168

TABLE XI
FREQUENCY OF VISITING MODERATE RESTAURANT

office	alcohol	family
0.09203442	0.08195523	0.08058522

TABLE XII
FREQUENCY OF VISITING EXPENSIVE RESTAURANT

philosophy	toy	occupation
0.0901525	0.07765963	0.0771492

TABLE XII
FREQUENCY OF VISITING VERY EXPENSIVE RESTAURANT

affection	tool	meeting
0.08395045	0.07485812	0.07333734

Next, the regression and classification problems results are presented in Tables XIII- X

TABLE XIII
KNN REGRESSOR

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.297	0.186	0.282	0.205
Testing Score	0.046	-0.116	-0.113	-0.311

TABLE XIV
SVM REGRESSOR

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.048	0.031	0.008	-0.031
Testing Score	-0.006	-0.002	0.04	-0.081

TABLE XV
LASSO REGRESSOR

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.058	0.004	0.106	0
Testing Score	0.054	-0.002	-0.008	-0.008

TABLE XVI
RANDOM FOREST REGRESSOR

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.812	0.816	0.833	0.737
Testing Score	0.042	-0.103	-0.211	-0.141

TABLE XVII
NEURAL NET REGRESSOR

	Cheap	Moderate	Expensive	Very Expensive
Training Score	0.968	0.817	0.997	0.994
Testing Score	-0.422	-0.801	-0.976	-2.277

TABLE XVIII
KNN CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	54.1%	52.3%	55.3%	64.8%
AUC	0.55	0.53	0.57	0.45
f1	0.53	0.56	0.57	0.49

TABLE XIX
SVM CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	55.9%	47.0%	63%	75.5%
AUC	0.64	0.47	0.65	0.51
f1	0.53	0.50	0.57	0.45

TABLE XX
RANDOM FOREST CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	54.1%	50.5%	65.4%	72.6%
AUC	0.61	0.52	0.63	0.62
f1	0.56	0.47	0.53	0.53

TABLE XXI
NN CLASSIFIER

	Cheap	Moderate	Expensive	Very Expensive
Accuracy	57.1%	48.2%	58.3%	61.9%
AUC	0.59	0.48	0.60	0.55
f1	0.55	0.51	0.52	0.51

3) Stage Three:

In this stage, Hyper parameters of each model is tuned using Cross-Validation. To answer the question: did feature selection result in getting rid of useless information or did it loss beneficial information? Both stage one and two were compared. After comparing the results from stage one and two, it is clear that using feature selection enhanced our regression models but not our classification models. Therefore, the regression models were tuned with the implementation of feature selection, whereas, the classification models were tuned without implementing feature selection. In addition, one can notice a clear overfitting problem in the Random Forest and Neural Network models; therefore, to address this problem regularization terms are tuned in this stage as well. The following Hyper parameters were tuned: kNN (Number of neighbors, Wight function used in prediction, SVM (C, gamma), LASSO (regularization parameter Alpha), Random Forest (Number of trees , Max depth of the trees, Class Weight) and Neural Network (L2 regularization parameter, Hidden layers structure, Solver, Max number of iteration).

TABLE XXII
REGRESSION R^2 SCORES AFTER HYPER-PARAMETER TUNING

Restaurant Type	Regression	Testing R^2 Score
Cheap restaurant	Random Forest [Number of trees=100, Max depth of the trees=10]	0.1
Moderate restaurants	SVM [C=1, gamma=1]	0.017
Expensive restaurants	LASSO [Alpha=1]	0.05
Very Expensive restaurants	LASSO [Alpha=1]	0.003

TABLE XXIII
CLASSIFICATION AUC SCORES AFTER HYPER-PARAMETER TUNING

Restaurant Type	Classification	F1	Obtained AUC	Original Paper's AUC
Cheap restaurant	SVM [C=2, gamma=0.01]	0.43	0.66	0.624
Moderate restaurants	Random Forest [Number of trees=100, Max depth of the trees=20]	0.57	0.64	0.608
Expensive restaurants	SVM [C=0.5, gamma=0.01]	0.59	0.67	0.67
Very Expensive restaurants	Neural Network [Hidden layers structure= 5000, L2 regularization parameter= 0.0001, Max number of iteration= 50, Solver= Adam]	0.59	0.64	0.564

It can be seen that after stage three, our results improved and in all categories, in addition our classification models achieved better results than [5] except in the Expensive restaurants target for which the results were equal. These classification models resulted in moderate improvement over random chance and, therefore, can target more potentially interested users.

V. CONCLUSION

In this work, we have predicted users' eat-out preferences after linguistic analysis of their tweets. The Dataset allowed us to study the data fusion of Twitter and Foursquare. We demonstrated LIWC categories importance in predicting users' preference to a particular price-category of restaurants. We also improved our models by hyper-parameter tuning and regularization. The main advantage of our approach is that Restaurant owners of specific price category can use our prediction models to target the potentially interested twitter users on the time and day of week which the user is most-likely to be online in twitter portal.

REFERENCES

[1] Henriette Cramer , Mattias Rost , Lars Erik Holmquist, Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare, Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, August 30-September 02, 2011, Stockholm, Sweden

[2] Jilin Chen , Gary Hsieh , Jalal U. Mahmud , Jeffrey Nichols, Understanding individuals' personal values from social media word use, Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, February 15-19, 2014, Baltimore, Maryland, USA

[3] Kenneth Joseph , Chun How Tan , Kathleen M. Carley, Beyond "local", "categories" and "friends": clustering foursquare users with latent "topics", Proceedings of the 2012 ACM Conference on Ubiquitous Computing, September 05-08, 2012, Pittsburgh, Pennsylvania

[4] Hamshaw RJT, Barnett J, Lucas JS. Tweeting and Eating: The Effect of Links and Likes on Food-Hypersensitive Consumers' Perceptions of Tweets. Front Public Health. 2018;6:118. Published 2018 Apr 23.

[5] Md. Mahabur Rahman , Md Taksir Hasan Majumder , Md Saddam Hossain Mukta , Mohammed Eunus Ali , Jalal Mahmud, Can we predict eat-out preference of a person from tweets?, Proceedings of the 8th ACM Conference on Web Science, May 22-25, 2016, Hannover, Germany

[6] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates 71 (2001), 2001

[7] Ethan Fast , Binbin Chen , Michael S. Bernstein, Empath: Understanding Topic Signals in Large-Scale Text, Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, May 07-12, 2016, Santa Clara, California, USA