Predicting Restaurant Consumption Level through Social Media Footprints

Yang Xiao, Yuan Wang, Hangyu Mao, and Zhen Xiao
School of Electronics Engineering and Computer Science, Peking University, China
{xiaoyang, wangyuan, mhy, xiaozhen}@net.pku.edu.cn

Abstract

Accurate prediction of user attributes from social media is valuable for both social science analysis and consumer targeting. In this paper, we propose a systematic method to leverage user online social media content for predicting offline restaurant consumption level. We utilize the social login as a bridge and construct a dataset of 8,844 users who have been linked across Dianping (similar to Yelp) and Sina Weibo. More specifically, we construct consumption level ground truth based on user self-report spending. We build predictive models using both raw features and, especially, latent features, such as topic distributions and celebrities clusters. The employed methods demonstrate that online social media content has strong predictive power for offline spending. Finally, combined with qualitative feature analysis, we present the differences in words usage, topic interests and following behavior between different consumption level groups.

1 Introduction

Over the past decade, microblogging services like Twitter and Sina Weibo have built up a huge user base. For example, by the end of 2014, Sina Weibo has accumulated more than 500 million users, out of which 167 million are monthly active users. With the growing popularity of microblogging service, businesses are also looking for new opportunities on social media, e.g., identifying target consumers and marketing their products. Compared with traditional consumer targeting scenarios, social media has several factors in its favour. Firstly, traditional consumer targeting techniques are mainly based on users' query logs or web access histories, and it is generally limited by session length (Dasgupta et al., 2012). While on social media, users have accumulated much more abundant traces, such as tweets, relationships and profiles. Secondly, according to a recent study (Nielson, 2012), 92% of consumers believe recommendations from friends over all other forms of advertising and 64% of salesperson believe word of mouth is the most effective way of marketing. Since social network links both friends and families, it becomes a natural platform for business to take advantage of word of mouth utility (Trusov et al., 2009).

Demographic profile is the starting point of defining target consumers for marketing. Demographic includes multiple aspects such as simple attributes like gender and age, and more complicated attributes like income, personality and consumption level. Since a large number of users provide their profiles on social media, inferring user attributes such as gender (Ciot et al., 2013; Liu and Ruths, 2013; Rao et al., 2011), age (Al Zamal et al., 2012; Nguyen et al., 2013), political polarity (Volkova et al., 2014), or occupation(Preoţiuc-Pietro et al., 2015a) from social media has already been widely studied. In this paper, we focus on the consumption level attribute prediction.

Consumption behavior reflects one's economic capacity and living standards (Brewer et al., 2012). An accurate consumption level prediction model can help business dealers identify their target consumers and recommend suitable products. Moreover, since consumption level is an important factor of economic status (Stutzer, 2004), effective prediction of consumption level will facilitate social economic researches on social media. However, unlike the gender attribute that is displayed on one's page, or political orientation that is unequivocally stated in one's tweets, consumption behavior related attributes are hard to acquire automatically from microblogging service.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3328–3338, Osaka, Japan, December 11-17 2016.

In this paper, we take advantage of the "socialization" of third party websites to build the ground truth collection. Specifically, we map one's Weibo (the largest microblogging service in China) account to his or her corresponding Dianping (the largest consumer review site in China) account. We carefully construct a dataset of 8,844 users who have been linked across Dianping and Weibo, and use one's self-report spending history to build the consumption level ground truth.

We propose a systematic method that takes social media features for consumption level prediction. We hypothesize that consumption level is correlated with various features, mainly including three aspects. The first aspect is user profiles including gender, age and education. Secondly, we hypothesize that language use in social media is a predictive factor for consumption level. We take textual features as the second aspect for prediction. Finally, we hypothesize that people of different economic statuses have their own unique tastes and interests. We take the following links which reflect one's interests and tastes as the third aspect for prediction. Owing to the noisy tweet content and sparse following relationships, this problem is technically challenging. We make use of topic modeling and LIWC categories to generate dense representations of text features. For graph features, due to the large quantities of users on Weibo, using raw following relationships directly has sparsity problem. To address it, we propose a matrix factorization algorithm on following matrix and naturally generate dense representations of user following preferences.

We take the raw features, and especially latent features as input features. We adopt the gradient boosted decision tree that can generate nonlinear combinations of input features, to predict user's consumption level. Empirical experiments demonstrate that rich features on social media have strong predictive power for consumption level, and latent features have best prediction performance. We conduct Spearman correlation test between topic preference and restaurant spending. We have found that the topic preference of a user is significantly correlated with the consumption level. We also report several new findings about consumption level and behavior on social media, e.g., users who follow topics such as luxury brands and politics, or talk more about money (e.g., audit, cash, owe) tend to be of higher consumption level, while users who follow topics about popular stars, use more character expressions and more assent words (e.g., agree, OK, yes) tend to be of lower consumption level. These findings will facilitate future attempts to consumer targeting, and may suggest extension application to spending prediction in other domains. The flexibility of our approach lies in that we identify important and general types of correlations that are easy to leverage from external social media sites.

2 Dataset Description

We focus on Sina Weibo, the largest Chinese microblogging service, as the studied microblogging service. We select a popular crowd-sourced review site, Dianping as the external website to help construct user consumption level ground truth. We do not adopt the automatic user linking methods but use "self-disclosure" to identify the same user across these two medias: some Weibo users provide their Dianping links in their tweets. This approach generates an accurate linking of across-website users.

Weibo Dataset: Sina Weibo is the largest Chinese microblogging service. We crawl the Weibo search page[1] to find those who declare their Dianping pages in their tweets. In this way, we get 62,015 users and then we crawl all the detailed information of these linked users, including profiles, tweets, followers and following links.

Dianping dataset: Dianping[2] is the largest social based crowd-sourced review site in China, which is similar to Yelp in terms of overall design and service. Dianping has two major components: users and local businesses. Users on Dianping can comment on and grade for local businesses, e.g., restaurants and hotels. Besides score and comment, user can also provide how much an average person spend for a meal in the restaurant. Each restaurant is usually assigned a small set of taste categories, price, score and starts a thread of reviews. We crawl all the reviews of 62,015 linked users mentioned above. The linked users have posted 286,069 reviews for 35,650 restaurants. We also crawl the 35,650 restaurants pages. The 35,650 restaurants are located in 45 cities in China, covering 98 different taste categories.

[1] http://s.weibo.com/

[2] http://www.dianping.com/

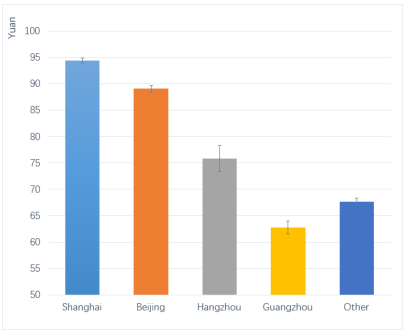

Figure 1: Average spending in different cities

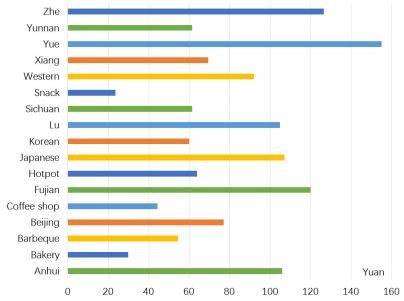

Figure 2: Average spending of different categories

We calculate the average restaurant spending in different cities and the result is shown in Figure 1. As shown in the figure, big cities such as Beijing and Shanghai are more expensive than other cities such as Guangzhou and Hangzhou. Moreover, users in Beijing and Shanghai take a share of 78.8% of our users. To make the population in our dataset more homogeneous, we only keep those who are located in Beijing or Shanghai. Dianping contains many categories of restaurants and different categories have different prices. The prices of different categories are shown in Figure 2. Since categories like cafe shop mainly provide drinks and can not be taken as real restaurants, we filter out restaurants in such categories. In order to gain a reliable estimation of consumption level, we only keep those who have post more than ten reviews. Finally, we obtain a total of 8,844 users. We find that these users are active on Weibo. They have posted 13,026,078 tweets and have 3,078,497 following links in total. We summarize the data statistics in Table 2.

Users	Tweets	Followings	Restaurants	Reviews
8,844	13,026,078	3,078,497	35,650	286,069

Table 1: Data statistics of our dataset for linked across-media users.

3 Problem Definition

If we can estimate one's consumption level according to his or her microblogging account, we can help the businesses design suitable marketing strategies, e.g., sending coupons to those who are more sensitive to price. We formulate this task as a typical prediction task: it aims to estimate the user consumption level. We assume the following general definition of the task: given a user's accumulated spending history $h^{(i)}$ and corresponding social media data s, we would like to test with varying type of s how accurate introduction of s can predict the label of $h^{(i)}$.

To formulate the consumption level prediction task, we assume that there are a set of m users $U = \{u^{(1)}, u^{(2)}, u^{(3)}, ...u^{(m)}\}$. For each user, let $y^{(i)}$ denotes the consumption level of user $u^{(i)}$. Higher value of $y^{(i)}$ means higher consumption level. A feature vector $x^{(i)}$ can be constructed for each user $u^{(i)}$. The aim of the learning task is to derive a prediction function f such that, for each feature vector $x^{(i)}$, it outputs a consumption level $f(x^{(i)})$.

4 Features

In this section, we discuss features extracted from Weibo service. In particular, we study how to derive effective latent features for the task.

4.1 Raw Features

Raw features are features that can be directly extracted from one's Weibo homepage. In this category, we consider the following three aspects:

I. METADATA Demographic fields of users. Fields such as gender and education level are binary features. Education level with value one indicates the user has accessed to university education. Tags

are a list of self reported words that users describe themselves. A separate binary feature is included for each unique tag. A user has 4 tags on average. For example, a user use tags such as music and shopping to represent one's interest.

II. RAWWORDS Unigrams in one's tweets and retweets. We use binary features for each unique unigram.

III. RAWFOLLOW All users that one follows. We use binary features for each unique followee.

4.2 Latent Features

In addition to raw features, we consider leveraging the latent features to improve predictions. We attempt to find semantics from the sparse user word matrix and user following matrix.

LIWC of Tweets (LIWCT)

Linguistic Inquiry and Word Count is a dictionary that classifies English words into psychological meaningful categories (Tausczik and Pennebaker, 2010). LIWC demonstrates its ability to detect psychological meaning in a wide range of applications such as emotionality investigation (Alpers et al., 2005) and personality prediction (Pennebaker and King, 1999). Previous study on social media shows that LIWC provides effective psychological features to determine the relationship between users (Adali et al., 2012) and the credibility of comments (Ott et al., 2011). Chinese LIWC (Huang et al., 2012) is a Chinese version LIWC that classifies 7444 Chinese words into 71 dimensions. Chinese LIWC have some slight differences from English one due to the difference in language, e.g., stemming is not needed for Chinese and verbs in Chinese do not have tense form. In general, the function of Chinese LIWC is similar to the English one.

We hypothesize that people at different consumption levels have different scores on LIWC features. For each user, we calculate the count of words that fall into each LIWC category, and get a vector of 71 dimensions. The vector is then normalized by the sum of the all values in it. Each value in the vector represents the user's usage preference on the linguistic category.

Topic Modeling of Tweets (LDAT)

In addition to employing raw text features, i.e., RAWWORD, we also use a high level representation of words to discover latent semantics. In order to distill topics from tweets, we adopt Latent Dirichlet Allocation method (Blei et al., 2003), which is an unsupervised learning method to discover latent topic distribution using large amounts of documents.

The model has two parameters, i.e., the document topic distribution θ_i and the topic word distribution φ_i. By learning θ_i and φ_i, document's topic distribution can be obtained and hence we get user's preferences on each topic. Since we care about user's topic distribution instead of a single tweet's topic distribution, we aggregate the user's tweets and retweets into documents and take the documents as the input of LDA model.

We take the resulted vector θ_i to describe user u_i's topic distribution. In this study, we set topic number k to 200 and run LDA with 500 iterations using Gibbs[3] sampling.

SVD Following Matrix (SVDF)

Because economic status is the central concern of hierarchical society, people of different economic statuses have their own unique tastes and interests (Wong and Ahuvia, 1998). Intuitively, if one is interested in some areas, it is natural that he or she will follow some related celebrities on social media. Therefore, celebrities that one follows can reflect the interest of the user (Lim and Datta, 2012). Since celebrities have large quantities of followers, in this paper, we take the account that has more than 30,000 followers as a celebrity account. The most direct way of utilizing celebrity features is to take each individual celebrity as a distinct feature aspect, i.e., RAWFOLLOW. However, since the celebrity accounts are of a huge number, using the celebrity features directly can result in sparsity problem. To tackle this problem, we use matrix decomposition to capture the hidden interest of one user. The row of the matrix denotes users in our dataset, the column of the matrix denotes the celebrity accounts and the element

[3]http://gibbslda.sourceforge.net/

f_{ij} is set to 1 if user u_i follows celebrity c_j. Different from traditional matrix decomposition task that has elements ranging from 1 to 5, in our case, the element of matrix only have two values representing whether user follows the celebrity. Hence, we choose the following logistic loss as our optimization goal.

$$U(i) = \arg\min_{w} \sum_{j} \log(1 + \exp(-f_{ij} \cdot w^T C(j))) + \lambda \|w\|^2$$

where f_{ij} denotes whether user u_i follows celebrity c_j, $U(i)$ denotes the latent vector of u_i, $C(j)$ denotes the latent vector of c_j and λ is the regularization coefficient.

We take the resulted $U(i)$ with varying length to describe one's interest. In this study, we employ stochastic gradient descent for matrix factorization parameter inference (Rendle, 2012) .

5 Clustering and Labeling

In this section, we introduce how we derive user consumption level label $y^{(i)}$ from the consumption history $h^{(i)}$. We use Gaussian mixture model to cluster users over their price space. The motivation for clustering is to find the natural structure of consumption prices and avoid manual threshold settings. This makes the labels of users more reliable and applicable to other dataset. The following part introduces the Gaussian mixture model and how we apply it to our dataset.

Gaussian mixture model is a probabilistic model which assumes that data is generated from finite number of Gaussian distributions. Given n data points $\mathbf{x} = \{x_1, x_2, x_3, ..., x_n\}$, the probability of generating the data x_i is as follows:

$$p(x_i|\pi, \Theta) = \sum_{z=1}^{k} p(z|\pi) p(x_i|\theta_z)$$

where π denotes the distribution over components, $p(x_i|\theta_z)$ is normal distribution where $\theta_z = (\mu_z, \sigma_z)$. $p(x_i|\theta_z)$ can be formulated as $p(x_i|\theta_z) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{(x_i-\mu_z)^2}{2\sigma_z^2}}$

In our task, we calculate average of the user u_i's spending history $h^{(i)}$ for each user, and in this way we get a list containing average spending of all users, i.e., $L = \{avg(h^{(1)}), ...avg(h^{(m)})\}$. $L^{(i)}$ ranges from 33 Yuan to 436 Yuan. We calculate average spending irrespectively of restaurant category because we find that most users visit diverse categories of restaurants. Since we focus on the relative consumption level of users instead of the real spending of users, we formalize the task as classification task. We apply the Gaussian mixture model to the spending list L. We assume that users come from k (k=2) different consumption levels, and the user's label $y^{(i)}$ is the cluster number the user belongs to.

6 Experiment

In the above section, we have shown how to extract features from social media and how to derive labels from spending history. We are going a step further to figure out the feasibility of using these social media features to predict the user's consumption level. We conduct experiments on the collected Weibo and Dianping dataset as described in Section 2. We set the component number of Gaussian mixture model to 2, i.e., users are labeled either one or zero indicating whether they are of high consumption level. We take 60% of users as training portion, 20% as validation portion and the remaining 20% as the test portion. For the evaluation, we take the accuracy, precision, recall and F1 measure as the evaluation metrics.

6.1 Quantitative Evaluation

We employ GBDT[4] and logistic regression[5] as the prediction models. Since traditional survey-based consumption behavior analysis task mainly focus on demographic attributes (Jang et al., 2004; Fodness, 1994), in this paper, we refer to the results obtained with feature METADATA as *baseline*. Specifically,

[4]http://github.com/dmlc/xgboost

[5]https://www.csie.ntu.edu.tw/ cjlin/liblinear/

Category	Name	Accuracy	Precision	Recall	F1
BASELINE	Age	0.5471	0.5547	0.5108	0.5318
	EDU	0.5507	0.5564	0.5324	0.5441
	TAG	0.5655	0.5629	0.6408	0.5993
	ALL	0.5889	0.5775	0.6715	0.6210
RAW	RAWWORD	0.6574	0.6544	0.6715	0.6628
	RAWFOLLOW	0.6945	0.6783	0.7529	0.7137
	ALL	0.7118	0.6969	0.7610	0.7276
LATENT	LIWCT	0.6066	0.5908	0.6982	0.6400
	LDAT	0.7451	0.7303	0.7863	0.7573
	SVDF	0.7673	0.7760	0.7635	0.7697
	ALL	0.8012	0.7821	0.8413	0.8106

Table 2: Prediction results with different feature sets.

for *baseline* features, we use logistic regression to combine all features together, which is a one layer classifier; for the RAW features and LATENT features, we construct a two layer classifier to combine heterogeneous features, i.e., in the first layer, we construct base classifiers using single feature sources, then we build a second layer classifier on top of the first layer classifier which takes the output of base classifiers as input. In our study, for the second layer classifier, we employ logistic regression to combine heterogeneous features. Since features such as RAWWORD and RAWFOLLOW are of high dimensions, we use L1 regularized logistic regression to construct base classifier; for the LATENT features, we use GBDT as the prediction model to construct non-linear base classifier. For LATENT feature base classifier selection, we have compared logistic regression with GBDT and found that GBDT preforms better than logistic regression. Therefore, we choose GBDT as the base classifier. The prediction results with different set of features are listed in Table 2.

For *baseline* features, we conduct experiments on age, gender, education and tags. As shown in Table 2, tags are the most predictive features and perform better than education and age. This is reasonable because tags contain much richer information than age and education and are related to user's profession and interests. Furthermore, the gender feature performs the same as random guess, hence we do not incorporate the gender feature into baselines. This suggest that the man and the woman do not have significant difference in restaurant consumption.

For RAW feature source, the amount of feature candidates is very large, e.g., hundreds of thousands of words. By borrowing the idea of feature selection in previous text classification task (Forman, 2003), we use χ^2 test to select representative features for classification. We select 4715 features for RAWWORD and 7820 features for RAWFOLLOW, which achieves best performance for prediction. Results in Table 2 suggest that RAW features are competitive for consumption level prediction, e.g., while baseline features can achieve 58.89% accuracy and 62.10% F1, RAWWORD alone can achieve 65.74% accuracy and 66.28% F1, and RAWFOLLOW alone can achieve 69.45% accuracy and 71.37% F1.

For LDAT feature source, we set the vocabulary size to 31514 and the number of topics to 200. By distilling topic semantics from tweets of users, the prediction accuracy can achieve 74.51%, which has been improved by 13.34% in contrast to RAWWORD. For LIWCT feature source, we would have expected LIWCT performs better than RAWWORD, since it categorizes the words into psychological and linguistic meaningful categories. A closer analysis of LIWCT features reveal that the vocabulary of LIWCT has relatively small overlap with the most distinguishing words in RAWWORD. For SVDF feature source, we conduct experiments with varying length of SVDF features. We find that when the length is more than 25, the performance does not increase, hence we set the length to 25 in our experiments. As presented in Table 2, accuracy of SVDF method achieves 76.73% and is much better than the result of RAWFOLLOW feature. This is also because SVDF features can capture high level interest of users.

Topic ID	Label	**Topic** (most frequent words, *translations*)	ρ	p value
13	Seafood	三文鱼,刺身,生蚝,日料,海胆,金枪鱼,鲍鱼,大闸蟹,鲜美,米其林*(salmon, sashimi, oyster, Japanese cooking, urchins, tuna, abalone, steamed crab, tasty, Michelin)*	0.85	0.0001
32	Politics	反腐,受贿,公职,公安局长,批捕,缓刑,查清,名下,收受*(anti-corruption, accept bribes, public employment, public security bureau chief, ratify the arrest, probation, investigation, name, take)*	0.82	3.81E-05
71	Luxury brands	vogue, victoria, miranda, chanel, kerr, alexander, dior, collection, louis, mcqueen *(vogue, victoria, miranda, chanel, kerr, alexander, dior, collection, louis, mcqueen)*	0.75	0.0017
198	Driving	牌照,高架,成品油,中环,远光,私车,93号,车友会,立交,油门*(vehicle license, elevated highway, product oil, median cycle, high beam, private car, No. 93 gasoline, car club, Interchange, gas)*	0.74	0.0014
120	Tennis	roger,莎拉波娃,罗杰,马卡洛娃,彭帅,阿扎伦卡,彭帅,郑洁,oba *(Roger, Sharapova, Roger, Makarova, Peng Shuai, Azarenka, Peng Shuai, Azarenka, Zheng Jie, oba)*	0.71	0.0001
45	Shanghai dialect	哪能,阿拉,今朝,老早,腔调,模子,白相,事体,闲话,辰光,喔唷*(how, I, today, previously, cool, personal loyalty, play, thing, talk, time, ugh)*	0.69	0.0260
192	Auto	车展,发动机,suv,保时捷,别克,沃尔沃,引擎,凯迪拉克,雷克萨斯,比亚迪*(auto show, engine, suv, Porsche, Buick, Volvo, engine, Cadillac, Lexus, BYD)*	0.61	0.0180
135	Mass brands	美宝莲,宝洁,阿芙,origins,美优,olay,多芬,spa,玉兰油,梦妆*(Maybelline, P&G, AFU, origins, beaubeau.com, olay, dove, spa, olay, mamonde)*	-0.77	0.0054
19	Cooking	关火,八角,豆瓣酱,土豆丝,豆角,切末,桂皮,鸡丁,炸酱面,葱油*(take off heat, aniseed, thick broad-bean sauce, shredded potato, French bean, mince, cinnamon, chicken cubes, Noodles)*	-0.81	0.0008
112	Stars	吴亦凡,朴灿烈,张艺兴,吴世勋,exo-m,金钟仁,边伯贤,黄子韬,exo-k, 泰妍*(exo Kris, Park Chan Yeol, exo Lay, Oh Se-hoon, exo-m, exo-k Kai, Baekyun, exo-m Tao, exo-k, Taeyeon)*	-0.81	9.19E-06
142	Character expression	2333, wwww, hhhh, OwO, hhhhh, 233333, QvQ, QuQ, wwwww, 0v0 *(2333, wwww, hhhh, OwO, hhhhh, 233333, QvQ, QuQ, wwwww, 0v0)*	-0.57	0.0322

Table 3: Topics sorted by absolute of Spearman correlation coefficient ρ. Topic labels are manually created.

6.2 Qualitative Analysis

In the above, we demonstrated that latent features, such as LDAT and SVDF, are the most effective features for predicting user consumption level. In this section, we conduct further hypothesis test to analyze the language divergence and the interest divergence between users at different consumption levels.

To select topics that are most correlated with consumption level, we present the formal Spearman correlation coefficient[6] test. The Spearman's coefficient ρ lies in the interval $[-1, 1]$, and a value of "+1" or "-1" indicates a perfect, positive or negative Spearman correlation. Intuitively, it is straightforward to generate two rankings of users, either by average spending or by topic preference. However, it is noted that ρ is sensitive to small value differences of both measures, and it will be difficult to obtain robust correlation values in this case. To capture the general trend, therefore, we group users according to their average spending. We sort users according to their average spending in descending order, split users equally into 100 buckets and examine the correlation at the group level.

Table 3 demonstrates the most correlated topics sorted by absolute of ρ. It is worth noting that topics with positive correlation coefficients reflect interests of high consumption level users, while topics with negative correlation coefficients reflect interests of low consumption level users. As shown in the table, users of higher consumption level prefer topics such as "Luxury Brands", while in contrast users of lower consumption level care more about "Mass Brands". This is consist with previous study on consumer behavior (Wong and Ahuvia, 1998), which shows that people buy luxury brands to take it as publicly visible markers of their economic status. Interestingly, we also find that users who speak Shanghai dialect are more likely to be of higher consumption level. We also conduct the correlation test within only Shanghai users, and we find that Shanghai dialect is also significantly correlated with high spending.

[6]http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient

Topic ID	Label	t (Age)	t (Gender)
13	Seafood	0.8837	$2.2599^{\ddagger}$
32	Politics	$10.1372^{\S}$	$-30.1144^{\S}$
71	Luxury brands	$-1.8778^{\dagger}$	$9.5550^{\S}$
198	Driving	$7.8684^{\S}$	$-7.2142^{\S}$
120	Tennis	$-2.5192^{\ddagger}$	-0.8891
45	Shanghai dialect	$4.7150^{\S}$	$5.9072^{\S}$
192	Auto	$2.8303^{\S}$	$-13.6531^{\S}$
135	Mass brands	-0.7032	$7.0779^{\S}$
19	Cooking	$-2.8099^{\S}$	$7.3084^{\S}$
112	Stars	$-2.7430^{\S}$	$2.7556^{\S}$
142	Character expression	$-7.4935^{\S}$	1.0283

Table 4: Topic preference t-test between different age and gender groups. "†","‡", "§" indicate the t test is significant at the level of 0.1, 0.05 and 0.01 respectively.

While on the other hand, users who use more web trending "Character expressions" tend to be of lower consumption level. Since our ground truth dataset is based on spending in restaurants, the top topics also cover food related topics, i.e., "Seafood" and "Cooking". "Seafood" is generally more expensive and hence it is an indicator of higher consumption level, while users who love "Cooking" may prefer dine in and consume less in restaurants. Moreover, high consumption level users talk more about "Politics", "Driving" and "Auto".

We conduct t test to analyze the interaction between topic preference and profile factors. Table 4 demonstrates the t test results. According to user self report age, we divide users into two groups, i.e., older than 30 years old or younger than 30 years old. A positive t indicates that elder group has higher preference on the topic. As shown in the table, elder users prefer topics such as "Politics", "Driving", "Shanghai dialect" and "Auto", and younger users prefer topics such as "Character expression", "Stars", "Luxury brands", "Cooking" and "Tennis". Generally speaking, elder people have higher spending power. Therefore, topics that elder users prefer are positively correlated with spending, while most topics that younger users prefer are negatively correlated with spending. Similarly, we conduct t test of topic preference between female users and male users. Topics such as "Character expression" and "Tennis" have no significant difference between females and males. Female users have significantly higher preference on topics such as "Luxury brands", "Cooking" and "Seafood", and male users have higher preference on topics such as "Driving" and "Auto".

Celebrity		Celebrity		Celebrity	
Dianping Coupon Shanghai	-	Beijing subway	-	Tourism related company	+
Beijing TV cusine programme	-	Reciting words app	-	International radio anchor	+
Comic dialogue player	-	Beijing SKP	+	Waldorf astoria	+
UK shopping	+	Wine related magazine	+	Charity fund	+

Table 5: Top celebrity features selected by χ^2. '+' or '-' indicates the sign of correlation.

For the celebrity features, we select the celebrities with highest χ^2 scores and present them in Table 5. As listed in the table, users of low consumption level follow celebrity such as coupon, subway, TV cuisine programme and reciting words app. On the contrary, high consumption level users follow celebrity such as traveling, wine, high grade hotels and international radio host.

For the LIWCT features, we also select the categories that have strongest correlation coefficient ρ with consumption level. Interestingly, we found that high consumption level users talk more about money (e.g., audit, cash, owe). On the contrary, low consumption level users talk more about time (e.g, end, until, season) and assent words (e.g., agree, OK, yes).

7 Related Work

User profiling aims to infer attributes of users from massive online data. Demographic attributes are widely used for ad targeting (Cheng and Cantú-Paz, 2010) and product recommendation (Wang et al., 2015). Traditional user profiling is mainly based on users' search logs or web access histories (Weber and Castillo, 2010; Hu et al., 2007). Recently, more and more researchers focus on user profiling on social media (Fink et al., 2012; Goswami et al., 2009; Tu et al., 2015).

In addition to simple demographic attributes such as gender or age, recently, researchers focus on complicated attributes such as political orientation (Pennacchiotti and Popescu, 2011), tags (Feng and Wang, 2012), locations (Backstrom et al., 2010; Pavalanathan and Eisenstein, 2015), occupation (Preoţiuc-Pietro et al., 2015a) and personal interests (Yang et al., 2011). However, the economic status related attributes have not been fully explored, which is partially due the difficulty in ground truth collection. Preoţiuc-Pietro et al. (2015b)'s work on income prediction from social media is the most relevant work to ours. Though consumption and income are related, as previous work on social economics (Brewer et al., 2012) has pointed out, " the amount of consumption in any period is not constrained to be equal to income in that period". Recent work on social economic classification (Lampos et al., 2016) is also related to our work. Their work focus on behavior features while we focus on language features and latent features. Furthermore, besides prediction task, we conduct an exploratory data analysis of language use patterns between users of different consumption levels. This work is also related to the study of food consumption on twitter (Abbar et al., 2015), and the work showed that foods mentioned in tweets are correlated with national obesity and diabetes statistics. The authors (Abbar et al., 2015) conduct experiments mainly from nutrition and health aspects, while we conduct experiments from the social economic aspect.

Our work is also related to mining heterogeneous social networks (Deng et al., 2012; Wang et al., 2011; Deng et al., 2011). Recently, many researchers focus on mapping accounts on different sites to one person (Zafarani and Liu, 2009; Liu et al., 2013) in real world. By utilizing these studies, we can link more users from Dianping and Weibo, and hence scale our task to larger dataset. Moreover, there are also works that utilize user linking feature to leverage social media knowledge for solving the cold start problem on third party website (Xiao et al., 2014; Zhang and Pennacchiotti, 2013). Since we estimate user consumption level accurately, it can also be used to solve the cold start problem in recommendation scenario.

8 Conclusion

In this paper, we focus on understanding the relationship between user's online social media behavior and offline restaurant spending. We link user's social media account and corresponding review site account, and then build consumption level ground truth based on user self report spending in their reviews. We propose the topic modeling methods and the matrix factorization methods to tackle the feature sparsity problem. We demonstrate that raw features and latent features on social media can predict consumption level with strong accuracy. The empirical analysis measures the correlation between social media features and consumption levels, and sheds light on language use differences across users at different consumption levels.

Acknowledgements

The authors would like to thank the anonymous reviewers for their comments. This work was supported by the National Grand Fundamental Research 973 Program of China under Grant No.2014CB340405 and the National Natural Science Foundation of China under Grant No.61572044. The contact author is Zhen Xiao.

References

Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Sibel Adali, Fred Sisenda, and Malik Magdon-Ismail. 2012. Actions speak as loud as words: Predicting relationships from social behavior data. In *Proceedings of the 21st International Conference on World Wide Web*.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *International AAAI Conference on Weblogs and Social Media*.

G. Alpers, A. Winzelberg, C. Classen, H. Roberts, P. Dev, C. Koopman, and C. Barr Taylor. 2005. Evaluation of computerized text analysis in an internet breast cancer support group. *Computers in Human Behavior*, 21(2).

Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan).

Mike Brewer, Cormac O'Dea, et al. 2012. *Measuring living standards with income and consumption: evidence from the UK*. Institute for Social and Economic Research, University of Essex.

Haibin Cheng and Erick Cantú-Paz. 2010. Personalized click prediction in sponsored search. In *WSDM*.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Empirical Methods in Natural Language Processing*.

Anirban Dasgupta, Maxim Gurevich, Liang Zhang, Belle Tseng, and Achint O Thomas. 2012. Overcoming browser cookie churn with clustering. In *WSDM*.

Hongbo Deng, Jiawei Han, Bo Zhao, Yintao Yu, and Cindy Xide Lin. 2011. Probabilistic topic models with biased propagation on heterogeneous information networks. In *KDD*.

Hongbo Deng, Jiawei Han, Michael R Lyu, and Irwin King. 2012. Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *JCDL*.

Wei Feng and Jianyong Wang. 2012. Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In *KDD*, pages 1276–1284. ACM.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *Sixth International AAAI Conference on Weblogs and Social Media*.

Dale Fodness. 1994. Measuring tourist motivation. *Annals of tourism research*, 21(3).

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. 2007. Demographic prediction based on user's browsing behavior. In *Proceedings of the 16th International Conference on World Wide Web*.

CL Huang, CK Chung, N Hui, YC Lin, YT Seih, WC Chen, and JW Pennebaker. 2012. The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, 54(2).

SooCheong Shawn Jang, Billy Bai, Gong-Soog Hong, and Joseph T O Leary. 2004. Understanding travel expenditure patterns: a study of japanese pleasure travelers to the united states by income level. *Tourism Management*.

Vasileios Lampos, Nikolaos Aletras, Jens K Geyti, Bin Zou, and Ingemar J Cox. 2016. Inferring the socioeconomic status of social media users based on behaviour and language. In *ECIR*. Springer.

Kwan Hui Lim and Amitava Datta. 2012. Following the follower: Detecting communities with common interests on twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*.

Wendy Liu and Derek Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium: Analyzing Microtext*.

Jing Liu, Fan Zhang, Xinying Song, Young-In Song, Chin-Yew Lin, and Hsiao-Wuen Hon. 2013. What's in a name?: an unsupervised approach to link users across communities. In *WSDM*.

Dong-Phuong Nguyen, Rilana Gravel, RB Trieschnigg, and Theo Meder. 2013. How old do you think i am? a study of language and age in twitter. In *International AAAI Conference on Weblogs and Social Media*.

AC Nielson. 2012. Global trust in advertising. *NY: USA, Nielsen Media Research, ACNielsen*.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *ACL*, pages 309–319.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. *arXiv preprint arXiv:1506.02275*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. Democrats, republicans and starbucks afficionados: user classification in twitter. In *KDD*. ACM.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6):1296.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An analysis of the user occupational class through twitter content. In *ACL*.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying user income through language, behaviour and affect in social media. *PloS one*, 10(9).

Delip Rao, Michael J Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM*, pages 598–601.

Steffen Rendle. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):57.

Alois Stutzer. 2004. The role of income aspirations in individual happiness. *Journal of Economic Behavior & Organization*, 54(1):89–109.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Michael Trusov, Randolph E Bucklin, and Koen Pauwels. 2009. Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing*, 73(5):90–102.

Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2015. Prism: Profession identification in social media with personal information and community structure. In *Chinese National Conference on Social Media Processing*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *ACL*.

Chi Wang, Rajat Raina, David Fong, Ding Zhou, Jiawei Han, and Greg J. Badros. 2011. Learning relevance from heterogeneous social network and its application in online targeting. In *SIGIR*.

Jinpeng Wang, Wayne Xin Zhao, Yulan He, and Xiaoming Li. 2015. Leveraging product adopter information from online reviews for product recommendation. In *ICWSM*.

Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *SIGIR*, pages 523–530. ACM.

Nancy Y Wong and Aaron C Ahuvia. 1998. Personal taste and family face: Luxury consumption in confucian and western societies. *Psychology and Marketing*, 15(5):423–441.

Yang Xiao, Wayne Xin Zhao, Kun Wang, and Zhen Xiao. 2014. Knowledge sharing via social login: Exploiting microblogging service for warming up social question answering websites. In *COLING*.

Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. 2011. Like like alike: joint friendship and interest propagation in social networks. In *WWW*, pages 537–546. ACM.

Reza Zafarani and Huan Liu. 2009. Connecting corresponding identities across communities. In *ICWSM*.

Yongzheng Zhang and Marco Pennacchiotti. 2013. Predicting purchase behaviors from social media. In *WWW*.