MSBA7012 Group Project
Identifying Influential Bloggers for Makeup Brands

Subclass A, Group 3

Group Members

Lin Liwei	3035676030
Lin Hongying	3035674680
Wang Yanyuan	3035675684
Wang Yang	3035675282
Zhang Xinyi	3035674446
Zhou Zezhong	3035674812
Zhong Qitong	3035675529

1 Introduction

1.1 Project Idea

Product Promotion in social media platforms is very popular in recent years. According to TopKlout (2019), in 2020, the share of the social media market in the e-market will increase by 11.5%, and the market value of blogger-related e-market will reach 3,000,000,000 RMB.

The beauty industry, in particular, is embracing this influencer marketing strategy. More and more cosmetic brands are partnering with beauty influencers to share and promote their products on social media platforms. Based on general knowledge, the number of one blogger's active followers are positively correlated with his/her promotion effect. However, there might be some other factors, such as product information and post content, that would influence bloggers' promotion effect. Thus, it might be not accurate enough when brands choose bloggers only based on their social media follower numbers.

The purpose of this project is to explore a scientific model to select the best-performance beauty influencers for beauty brands to cooperate with regarding their product information.

1.2 Data and Methodology

Our analysis involves the following 3 aspects of data:

1. Products/Brands' features: brand, product name, the product's popularity, price in T-mall, score in T-mall, Hot-topic in Weibo, etc.

2. Bloggers' features: bloggers' user-id, name, gender, region, tag, follows, number of fans, number of posts in total, etc.

3. Posts' features: post id, contents, posts' picture/video, posted time, number of likes, comments, shares, tags in a post, whether it mentions others, etc.

The data mentioned above will be scraped on social media websites such as Weibo.com, T-mall.com and Cosme-De.com. Then we used the features we extracted from products /brands, bloggers and posts as the independent variables to estimate the promotion effectiveness of each post, where the effectiveness is measured by the weighted average of the number of likes, comments, and shares. The detailed determination of weights will be introduced later. The model will be expressed in this form:

$$Y_i \sim \underbrace{X_{i,1}, X_{i,2}, \dots, X_{i,m}}_{Bloggers' Features}, \underbrace{X_{i,m+1}, X_{i,m+2}, \dots, X_{i,p}}_{Brands' Features}, \underbrace{X_{i,p+1}, X_{i,p+2}, \dots, X_{i,q}}_{Posts' Features}$$

where, in the Classification model, Y_i is a factor variable with three levels, measured by $log(L+S+C)$ for the i_{th} post using K-means function and in Regression Model, Y_i is $log(L+S+C)$ in Random Forest model.

1.3 Key Findings and Results

We use linear model, random forest and XGBoost to do classification. The result of XGBoost is the best and the accuracy is 72.32%. We also get the feature importance for each variable, which provides insights for companies to make the most efficient plan in product promotion. As we can see from the result, blogger fan is the most important feature. However, there are other features that can be considered, such as tags on a post, the length of the post, or whether to include videos or pictures in the post, etc.

By applying our model, once we get the brand information and related product information, we can output the three-level recommended blogger list.

2 Data

Data Aspect	Data	Resource	Type	Example	Data Size(rows)
Promotion effect	No. of likes, comments, shares	Weibo Post	num	30,42	
Brand Information	Brand Name	cosme -de.com	char	Innisfree	624
	Hot Topic	Weibo topic	boolean	1 (has hot topic) 0 (has no hot topic)	
	Account Fans	Weibo Account	num	100,000	
	Account Post	Weibo Account	num	100,000	
Products Information	Brand Name	T-mall	char	Innisfree/悦诗风吟	20500
	Product Name	T-mall	char	悦诗风吟悦享鲜萃面膜	
	Popularity	T-mall	char	2790	
	Price	T-mall	num	216	
	Score	T-mall	num	4.7	
Bloggers Information	User-ID	Weibo	num	1968758563	3788
	Name	Weibo	char	李佳琦 Austin	
	Gender	Weibo	char	male	
	Region	Weibo	char	上海	
	Tag	Weibo	char	知名美妆博主	
	Follows	Weibo	num	282	
	Fans	Weibo	num	8910000	
	Posts	Weibo	num	907	
Posts	Post-ID	Weibo	num	4449134897482840	1000845
	Content	Weibo	char	“Dior限量气垫竟是国内先上!”	
	Picture/Video	Weibo	boolean	1 (has picture/video) 0 (has no picture/video)	
	posted time	Weibo	char	12/13/2019	
	No. of likes*	Weibo	num	42	
	No. of comments*	Weibo	num	30	
	No. of shares*	Weibo	num	53	

Table 2 Definition of data

3 Methodology

3.1 Data Collecting

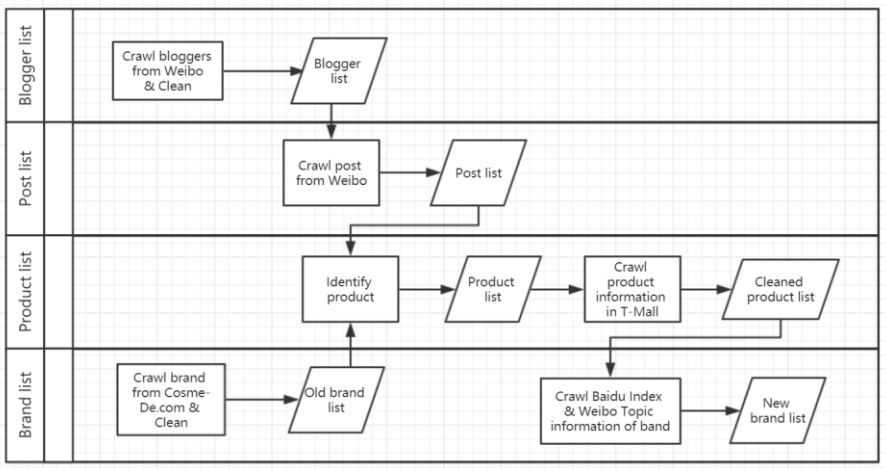


Figure3.1.1 Data Collecting Flowchart

1. Old Brand List

We firstly crawled most brand names on Cosme-De.com and then manually added some domestic cosmetic brands and medical cosmetology brands. During data processing process, we deleted 67 duplicated brands and manually deleted strange brand names such as Adidas and Mercedes-Benz that are not the brand that focuses on the make-up products. Containing the famous brand that not focusing on make-up products will have the potential to get bias result. Finally, we got our Old Brand List.

2. Blogger List

As there are no integrated blogger lists for our reference, we decided to crawl the corresponding bloggers by searching keywords in Weibo. In our project, we used 8 different keywords, including "美妆博主，美妆自媒体，时尚博主，时尚自媒体，美妆达人，时尚达人，微博美妆博主，微博时尚博主". For every keyword, we crawl the first 50 pages' blogger information (bloggers' user-id, name, gender, region, tag, follows, number of fans, number of posts in total). At the data cleansing process, we first delete the duplicated bloggers (someone might be in several keywords), and then delete the bloggers whose name contains "海淘，代购，折扣". By our experience, bloggers with names containing these words are mainly the online seller (sell products online), which will also give the bias information for our study. After the above-mentioned process, we got the blogger list.

3. Post List

After we built the blogger list, we used bloggers' user-id to crawl their original posts on Weibo from 2019/01/01 to 2020/01/01 (YYYY/MM/DD). During the crawling process, if it took too much time on one blogger, and the number of comments/likes/shares are very low, we would manually check the account and delete it from the blogger list if possible.

4. Before-Cleaned-Product List

For every blogger, we got a list of their posts within one year, and we tried to identify products mentioned in posts and those posts contain products will be regarded as promotion posts. To define the products with its brand, we created an identification method:

1. Created an end words dictionary, containing the possible end-words of a products, such as 眼膜，高光，隔离.
2. Created a brand dictionary containing all the brand name (Chinese & English).
3. Loaded three dictionaries mentioned above into the user-defined dictionaries into the tokenizer Jieba **with very high frequency** (ensuring the name will be tokenized with the highest priority every time).

4. Created a stop-words list containing possible useless wordings between brand name and end-words. For example, 限定.
5. After tokenization, we found the index of the brand names and product names we have found. Upon drop the stop-words, if the distance between brand name index and end-words index is not very long, 6 tokens for example, we combined tokens from the brand name to end-words and got our final products. For example, from the post "跟 YSL 的浪漫约会还原了摩洛哥一座花园，护肤代言人谢霆锋和昆凌皮肤状态真的太好了！ YSL 最的夜皇后精华来了推荐熬夜一族都用起来,给你的脸'一夜回春'的良药"we finally extracted the product name "YSL 夜皇后精华".

5. Product list

With lists of products identified and its corresponding promotion post, we searched the product information in T-mall. There are two reasons for this step, 1) Get more detailed information about the product for our further analysis (the product information) 2) Delete the wrongly identified product through our identification method. It is a convenient way to detect our identification mistake because if the product doesn't show in Taobao, it's very likely to be a wrong name or non-exist items. In short, we ask Taobao to do mistake detection for us.

6. Brand List

When we get the right product information in T-Mall, we also have the product's brand name. Collecting all the brand names from the identified products, we obtain a new brand list, which is a subset of the **Old Brand List.** As some of the brands shown in T-mall only have the English name or Chinese name, we can use **Old Brand List** to make them complete. Then we search the brand names in Weibo, to get the brands' account information, such as the number of fans and the number of posts, and check whether the brand has a hot topic (微博超话), which is a standard to evaluate the frequency of people talked in Weibo.

After getting all the data mentioned above, we store them in MySQL Server. **Blogger List** and **Post List** are connected by blogger id. **Post List** and **Not-Cleaned-Product List** are connected by post id. **Old Brand List** and **Not-Cleaned-Product List** are connected by brand id. **Not-Cleaned-Product List** and **Product List** are connected by post id. The ER Diagram of the database is shown as follows:

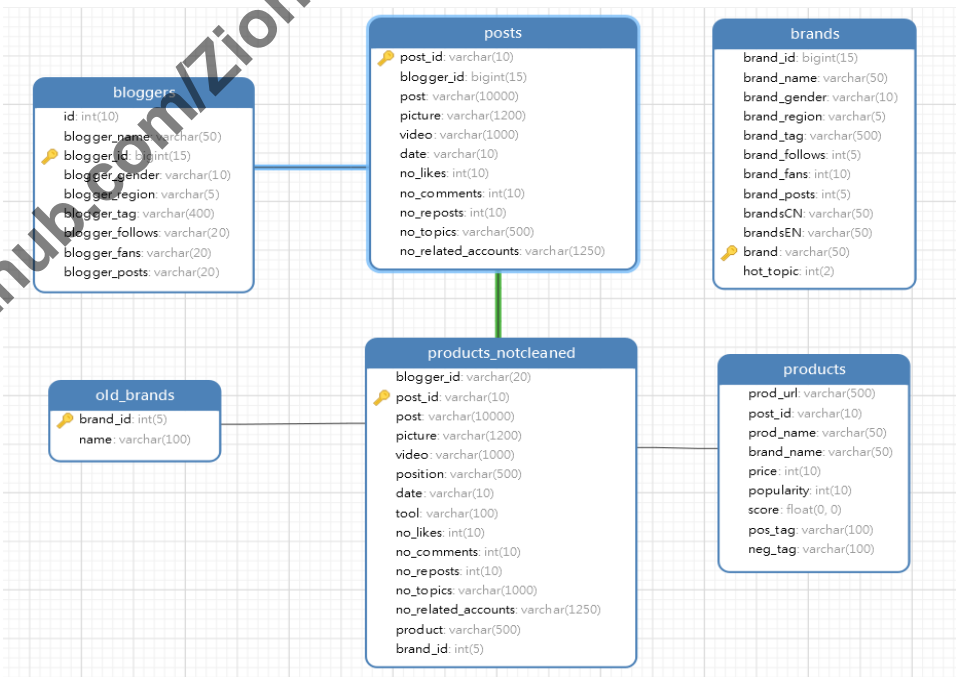


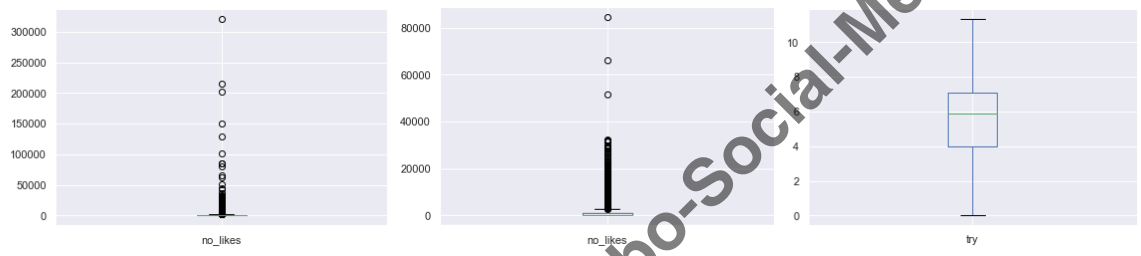
Figure3.1.2 Database ER Diagram

3.2 Data Preprocessing and Exploration

We extracted data from Blogger list, Post List, Product List and Brand list as training data and there are in total of 15908 rows of data. Then we do the data preprocessing: We transformed some variables into dummy variables, such as transform gender (0 for male and 1 for female) and also the brand's hot topic(0 for no and 1 for exist). For the blogger_tag, picture and video information, we transfer them as counts. For example, the picture variable equal to 6 means the post contains 6 pictures. For the post content, we first tokenized the post, and then count the number of tokens, which is a measurement of the length of posts.

After plotting the box plot of no_like, no_comment and no_repost, we found out that there are several outliers and most of these outliers are lottery posts (providing a lucky draw for the interaction users). Because posts with and without lottery differs a lot, we need to separate them. Now our focus is on the post without a lottery, so we need to exclude the lottery posts from our data if we detect the lottery posts. Our detecting method is to search the keywords such as '抽', '揪' and use Regular Expression '摸.个', etc.

Then we noticed the difference between top bloggers and average bloggers are huge. When we take logarithm on the number of likes (no_like for short), no_comment and no_repost, the final box plot looks good. So later we will always take logarithm on the dependent variable.



include lottery Weibo *exclude lottery Weibo* *use log*

Figure3.2.1 Box-Plots of data of different transformation

Then we drew the heatmap (*Figure3.2.2*) of the correlation between variables. As we can see from the graph, no_likes, no_reposts and no_comments have a high correlation as expected. Brand_posts, brand_fans and brand_hot_topic also show correlation. Other features are not so correlated. Notice that the correlation we mentioned above is just the linear correlation, not meaning that the variable can't be the predictors for the target. We can also choose some non-linear model to explore the relationship.

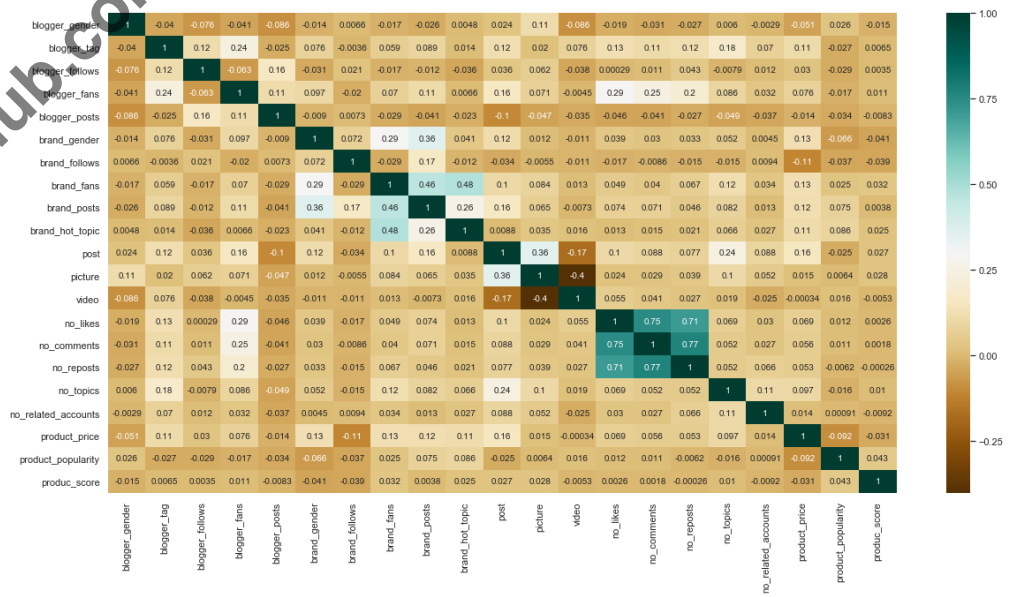


Figure3.2.2 Heatmap of correlation between variables

3.3 Model Training

As now there are three variables: number of Likes(L), Shares(S) and Comments(C) that can be used to measure the promotion effect, we came up with three ways to transform this three-dimensional data into a one-dimensional variable as our dependent variable.

Since each Weibo ID can make plenty of comments and shares but only one like on a single post, to prevent duplication, we first considered to use the number of likes as promotion effect measurement, after logarithm, that is: $log(L)$.

The second way is to sum the number of Likes, Shares and Comments up, and after logarithm, we have: $log(L + S + C)$.

Except for simply adding three variables together, we also used Principal Component Analysis to give different loadings to them. By using the PCA, we found that there is a combination of three variables that can contain around 75% of the total variance, so we think it's enough to be set as the performance measurement, the combination is 0.6073975 × L + 0.6172959 × C + 0.5000140 × S. After logarithm, we have: $log(PCA_1(L + S + C))$.

Now we have generated three datasets with different dependent variables, then we apply both Regression and Classification model on them.

For the Regression model, we tried Linear Regression and Random Forest Regression.

For the Classification model, we first applied K-means algorithm on dependent variable, to set the Y_i into different groups. With many tries, we found 3 groups is a fairly good choice. Figure below (Figure3.3) shows the result of the classification of $log(L + S + C)$.

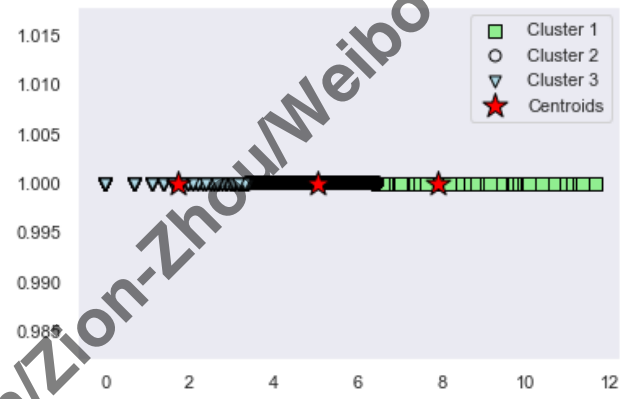


Figure3.3 Example of scatter plot of three classes

Then we used Logistics Regression, Random Forest, Extreme Gradient Boosting, Balanced Bagging, CatBoost, Naive Bayesian to build model. The performance of each the model are shown in *Table3.4.*

3.4 Model Selection

Regression (R^2)			
Model	log(L+S+C)	log (PCA(L+S+C))	log (No. of Likes)
Linear Regression	0.4219	0.427	0.427
Random Forest	**0.8160**	0.8165	0.8167
Classification (*Kappa*)			
Model	log(L+S+C)	log (PCA(L+S+C))	log (No. of Likes)
Logistics Regression	0.4171	0.4147	0.4184
Random Forest	0.7143	0.6760	0.6869
Balanced Bagging	0.6959	0.6687	0.6777
eXtreme Gradient	**0.7276**	0.6926	0.7211

Boosting			
CatBoost	0.7144	0.6702	0.6888
Naive Bayesian	0.3574	0.3326	0.3487

Table3.4 Performance of different models

We used R-squares and Kappa respectively as the criteria for accuracy of regression and classification models. As shown in *Table3.4*, Random Forest under the dependent variable of log (No. of Likes) and Extreme Gradient Boosting under dependent variable of log(L+S+C) are the best performers. However, we further considered that using only variable of number of likes as promotion effect is not comprehensive, and the R-square are similar under Random Forest, we finally chose Random Forest and eXtreme Gradient Boosting under dependent variable log(L+S+C) as our final model.

Key Challenges & Solutions

1. Anti-web scraping in Weibo and T-mall.
   Solution: When we used url ending with '.com', Weibo didn't allow us to crawl such a large amount of data. Then, we tried url ending with '.cn' and successfully crawled a lot of data, as this kind of url is used by mobile phone and the information website transport back now is simply in JSON format.
2. Identify products. As there is no complete product list online, it became very hard to identify products in the post.
   Solution: We extract possible ending words of products from the posts we crawled and built an endwords dictionary to identify possible product. Then we try several posts to get possible words appearing between brand name and end word, and built stopword dictionary to clean those words. Besides, we check the accuracy of products by comparing the result searched from T-Mall and products we identified. If they are the same, the product is right, and vice versa.
3. It took a long time to crawl post from Weibo as the data size is more than one billion rows.
   Solution: We crawled post separately and make use of the iMacs in the library. Meanwhile we use shared documents to keep records of the work in case some mistakes happened.

4 Results

1.Feature importance

The feature importance graph is as below.

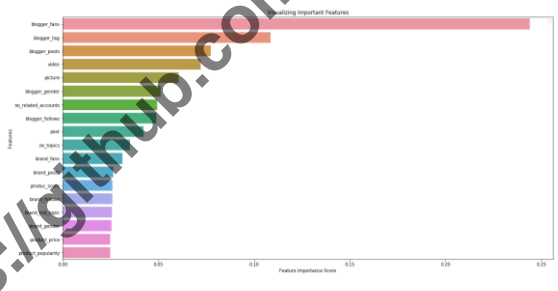


Figure4.1 Feature importance of variables for classification

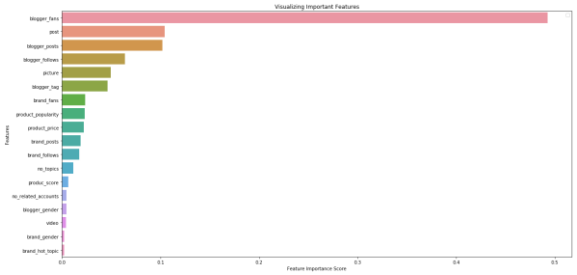


Figure4.2 Feature importance of variables for regression

As shown in both *Figure4.1 and Figure4.2,* number of bloggers' followers is the most important feature in both the Classification and Regression model, which means that the traditional way of selecting bloggers based on their number of followers is rational. However, there are more features such as bloggers' tag and post content should be considered. For example, if a blogger sends more posts in one year, the promotion effect will be better which probably because more frequent posts attract more active users. Besides, whether to

include video, picture and the length of the post are also important features. Based on that, we can recommend different post forms to different brands to improve the promotion effect.

Since the brand's related data are also important features, our recommendation of bloggers will vary from one brand to another. Meanwhile, since brands' related data is not one of the top important features in our model, we can conclude that even small brands can reach a good promotion effect if they can successfully select the right bloggers and post form.

2. Model Application

We used the brand Origins and their product '炭瓷膜' as an example of our model application. After putting all the information of bloggers, brand, and post form in, the classification model has output three classes of blogger lists (*Table4.2.1*) and the regression model has output a ranking list of recommended bloggers (*Table4.2.2*).

Class 1	**Class 2**	**Class 3**
萌大雨 YUYU	小米苏酥	Fancy 范欣
missfaye	ZY 大暖	MisA7
kittywenny	壹十七少	喜欢包包的木石
羽晗 winnie	滚滚君卖萌不高冷	猪小角
潘朵拉 Pandore	JeccyCao	滚滚大妹阿徐

Table4.2.1 Output of classification model for Origins case

Blogger List
ai 媚儿
韩恩汐汐
刘佳妮丶
玉米酱罐
林小宅-

Table4.2.2 Output of regression model for Origins case

Table 4.2.1 shows five bloggers of each class, with class 1 stands for the best promotion effect and class 3 stands for the least. Based on the result, we will recommend the marketing team of Origins to negotiate with bloggers in Class 1 for their mask promotion.

Table 4.2.2 is the ranking list of the top five recommended bloggers with decreasing promotion effect for Origins case. The marketing team can choose a specific number of bloggers from top to bottom to cooperate with base on their budget and preference.

5 Limitations and Future Work

5.1 Limitation

Accuracy

1.Miss products while doing identification

Now we can't locate the product if it uses non-structural name such as its nickname. Because it didn't happen often, so we ignored it now.

2.Exclude lottery posts

We used keywords such as 抽/揪/摸.位/摸.个/抱.个 to identify lottery posts; however, some keywords might be missed and many lottery information might be posted in comments. Until now, we can't detect the lottery posts from comments due to the lack of comments data.

3.Online seller

Our goal is to get true make-up bloggers, but some online sellers or non-related bloggers might be included in our blogger list.

4.Fake information
Some bloggers would buy fake “reposts, likes and comments” to increase the post-heat so sometimes the measurement is not accurate.

5.Measurement of promotion effect
We used the number of likes, comments and reposts as a measurement of promotion effect. However, these numbers only show how many people have been reached and made reaction, but we cannot track whether people bought the product because of these posts. We hope to find a more efficient way to measure the promotion performance of a post.

Lack of Information

1.Network within bloggers/brands
Bloggers may follow each other and it will form a social network. We can further analyze related properties such as centrality.

2.Video/Picture contents transformation
We didn’t get detailed information on video/picture such as what did the blogger to say in the video that attached to the post.

3.Information from other platforms
We didn’t include information from other platforms such as Xinlang Index.

4.Social related events
We ignored the potential effects caused by social related events, such as the product quality problem of the make-up products which ruins the reputation of the brands.

Dynamic Factor Issue
Some of the data we collected is dynamic data, for example, fans in blogger data, and popularity, price, score data of products in T-Mall. If we can extract the dynamic data, we can make up-to-date recommendations.

5.2 Future Work

1.Fans Portrait
We can portrait the fans profile of both make-up brands and bloggers, and by doing this we can make a more accurate match between brands and bloggers based on their tags and characteristics.

2.Classifier Improvement
Classifier improvement can improve classification Levels. By now we classify all posts into three levels. If we classify them into more levels, the recommendation we made will be more accurate and applicable.

3.Sentiment Analysis
In the future, we can do sentiment analysis based on post comments, product comments, and Weibo Hot Topics, and analyze user attitudes of each brand and product.

6 Work Allocation

1.Data Scraping: Coding: Lin Liwei, Wang Yanyuan, Zhou Zezhong; Scraping: All members
2.Data Preprocessing: All members
3.Modeling: All members
4.Report and PowerPoint: All members
5.Presentation: Lin Liwei, Wang Yanyuan, Zhou Zezhong