Yang, B., Tang, J.L., Ma, X. , Chang, Y.W. & She, H.Y. (2019). Research and prediction on the sharing of WeChat official accounts' articles. In *Proceedings of The 19th International Conference on Electronic Business* (pp. 323-339). ICEB, Newcastle upon Tyne, UK, December 8-12.

Research and Prediction on the Sharing of WeChat Official Accounts' Articles

(Full Paper)

Bo Yang*, Renmin University of China, Beijing, China, yangbo_ruc@126.com
Junlin Tang, Renmin University of China, Beijing, China, junlinforever@ruc.edu.cn
Xi Ma, Renmin University of China, Beijing, China, hello_xi@outlook.com
Yawen Chang, Renmin University of China, Beijing, China, yawenchang@ruc.edu.cn
Huayang She, Renmin University of China, Beijing, China, shehuayang@ruc.edu.cn

ABSTRACT

With the development of mobile Internet, We Media was born. WeChat Official Account Platform is the largest we media platform in China. In WeChat social network, information can only be rapidly spread through the sharing operation of users. This paper takes WeChat official accounts as the object and uses logistic regression model to explore the influencing factors on sharing. After that, a prediction model is constructed based on logistic regression and support vector machine. The significance of this study is to propose the factors that influence WeChat official accounts' articles sharing, and to construct a sharing prediction model.

Keywords: WeChat Official Account, Sharing Behavior, Sharing Prediction.

*Corresponding author

INTRODUCTION

WeChat launched by Tencent in January 2011 is an instant messenger software based on mobile Internet. According to the 2018 WeChat Data Report, the number of active users reached 1.082 billion in 2018.

As an important component of WeChat system, WeChat official accounts have attracted more than 80% of users to subscribe. In the WeChat system, the relationship between users and official accounts is "follow and followed", in which the official accounts are information publishers and the users are information recipients. Following an official account, one can receive the messages from it and make some interactive actions, such as praise, comment, sharing and collection.

WeChat provides an appropriate place for dissemination of high-quality content. According to the WeChat Official Account Population Insight Report, more than half of WeChat users have shared articles published in Subscriptions. Actually, the subscribers of WeChat official accounts are impressionable. Most people will pay attention to a WeChat official account if it is recommended by others. Based on the "point-to-point" mode of transmission, the WeChat official account is relatively closed and private compared with forums, micro-blog and so on. Every WeChat official account can deliver its content to those who subscribed, which means disadvantages in mass dissemination and secondary spread. In WeChat social media network, information dissemination depends on users' sharing. Therefore, users' sharing behavior is critical to the transfer of information.

The relationship among WeChat users, WeChat official accounts and information is shown in the Figure 1.

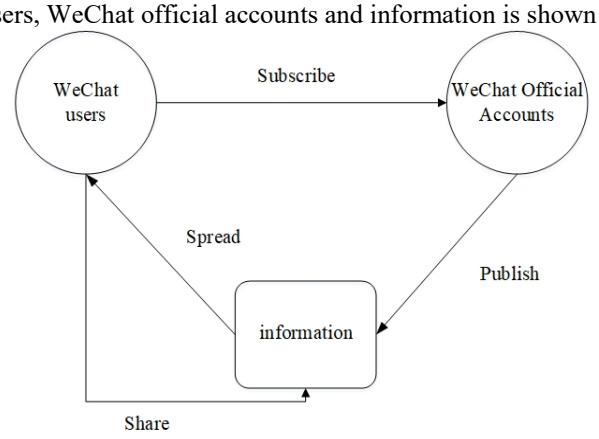

Figure 1: Relationship between WeChat users, official accounts and information

Due to the characteristics of this platform of WeChat, the official accounts can only deliver information to the users who follow them. Information dissemination can be chopper style through users' sharing. Therefore, users' sharing behavior is particularly important for the transmission of WeChat content. The mechanism of sharing can be deeply understood by this research on influencing factors, which can help account operations to grasp users' mind and real needs. Besides, the sharing

prediction model constructed in this article can accurately predict information dissemination rate and dissemination range, which has important practical guiding significance for study on WeChat users' behavior and interest, marketing, network public sentiment monitoring, and forecasting hot spots.

LITERATURE REVIEW

In recent years, many scholars have studied the sharing and reposting behavior of social media platforms such as Sina Weibo and Twitter from different technical perspectives and characteristics. Many excellent research achievements have came out. However, researches on the sharing behavior of WeChat official account are relatively less and lightly shallow.

Researches on WeChat official account

At present, the researches on WeChat official account mainly focus on the development status, user behavior, information dissemination. In terms of development status, many official accounts are studied as specific cases. Researchers summarize development characteristics, operation strategies and put forward problems and solving measures. In the research of user behavior, questionnaire survey is mostly used. Researchers pay attention to users' motivation, satisfaction, loss and willingness to continue subscribing. Many studies have introduced the "Uses and Gratifications" (Huang & Peng, 2014). Guan (2015) concluded that cognitive motivation and habitual motivation are the main motivation for users to use the official accounts and that information needs, social needs and entertainment needs are important factors in promoting users' willingness to use it. In the field of information dissemination, it is widely believed that the dissemination of WeChat official accounts has the characteristics of privacy, precision, deep reading and weak control (Jing & Zhou & Ma, 2014). It is not dominant in mass and rapid communication. Most of the researches focuses on the characteristics, process and mechanism of communication.

Research on the influencing factors of sharing

In recent years, users' sharing behavior in social media, such as Sina Weibo and Twitter, has attracted many scholars in the world. Current researches about WeChat official accounts concentrate on users' sharing motivation, the characteristics of information publisher and recipient and the characteristics of content. These research results have important reference significance for feature extraction of sharing behavior in this paper.

Apart from the above factors, some researches also include the characteristics of title, whether the official account is authenticated or not, and the frequency of delivery, based on the special mode of WeChat official accounts. Zhang and Li (2016) found that subject, article title, information source, presentation form and article length all affect the frequency of users' WeChat sharing. Wang (2015) put forward that whether the title could catch the user's eyes and hit the user directly has a decisive effect on the sharing rate of articles published in WeChat official accounts.

Research on the prediction of users' sharing

The researches on prediction of users' sharing behavior originated as Twitter became increasingly popular. Early studies focused on theoretical analysis about the causes, modes and contents of users' sharing behavior. Then, researchers use some machine learning algorithms such as logistic regression, support vector machine and neural network to construct the model, and predict the users' sharing behavior as a classification problem.

In order to improve the accuracy of the prediction on whether a user shares an article, researchers have explored more types of algorithms in recent years. Zhang, Lu and Yang (2012) proposed a method for predicting the reposting behavior of Weibo, with an overall hit rate of 85.9%. Wang, Feng, Jia, Zhu and Qin (2017) introduced neural network prediction algorithm in machine learning to build multi-feature prediction model. Zhou, Huang and Deng (2017) put forward a method by fusing anomaly detection with random forests. Others integrate the results of multiple machine learning algorithms according to some certain rules to obtain better prediction results than single learner (Zhang & Wang & Qin, 2018).

Apart from prediction on users' sharing behavior, some scholars study sharing volume prediction. Li, Yu and Liu (2013) proposed a prediction model based on SVM algorithm, but the complexity of feature computation is not suitable for large scale data. Deng, Ma, Liu, Zhang (2015) applied BP neural network to predict user sharing volume in emergencies, but whether it can be widely used has not been discussed. Some scholars also proposed a model based on sharing intention and sharing influence by analyzing the simulated sharing process (Zhao & Liu & Shi & Wu, 2016).

RESEARCH METHODS

Based on the existing literature and theoretical research, there are many factors affecting information sharing. Most of the researches on Sina Weibo's sharing were studied from several dimensions, which include information publishers, recipients and content's characteristics. These dimensions are detailed as follows: influence power of an account, whether an account is authenticated, the number of fans, age of the accounts, credibility of the Weibo content, length of the microblog, whether it contains a URL, whether it contains Image, whether it contains video, frequency of publication, delivery time, delivery date, and so on. Considering the similarity of information dissemination between Weibo and WeChat Official Account Platform, this paper divide the influencing factors on WeChat official account users' sharing into three categories: information publishers (WeChat official accounts), recipients (users), and information as well. Meanwhile, we integrate WeChat official accounts' own characteristics in the three aspects. The influencing factors model that affects the users of WeChat official accounts sharing, is as shown in Figure 2.

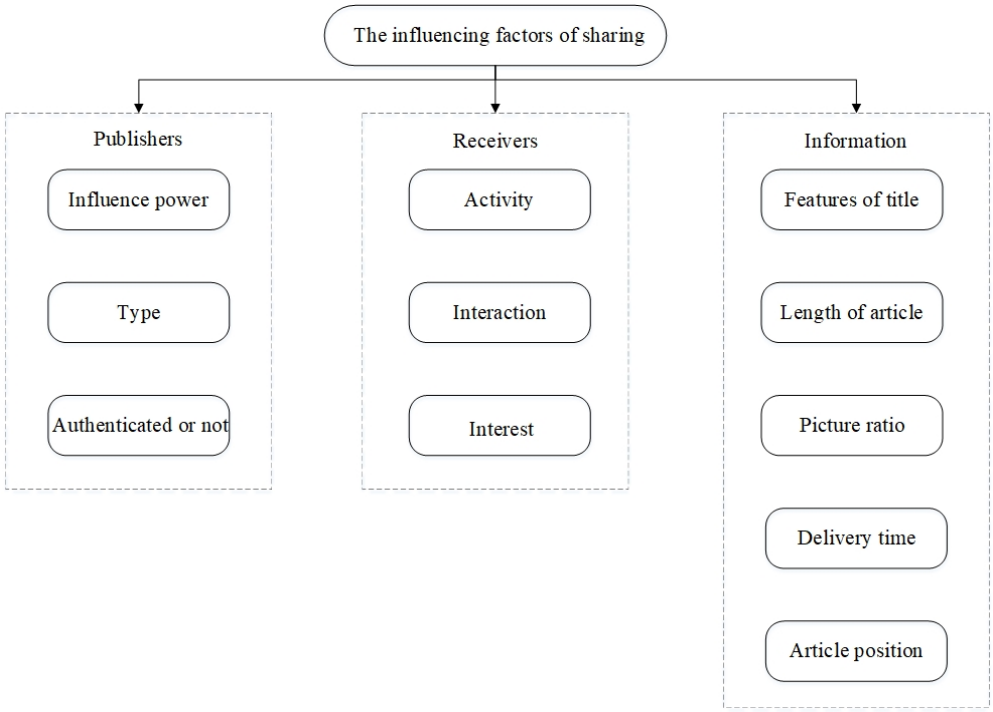

Figure 2: Three categories of the influence factors of sharing

Relevant technical methods

In this paper, the basic methods of logistic regression and support vector machine are used to quantitatively study the influencing factors of WeChat users' sharing. Logistic regression model is used in studying the relationship between independent variables (official accounts' characteristics, users' characteristics and characteristics of information) and dependent variables (sharing) because the dependent variables selected in this paper are categorized variables, and independent variables have both continuous variables and categorized variables. In the process of predictive sharing, this paper classifies the research problem into a typical binary classification problem. Support vector machine algorithm is selected to construct sharing prediction model.

Relevant data analysis methods

In the process of data analysis in this paper, most of the work is on data preparation because the adequacy of data processing greatly affects the effect of the model. The data analysis in this paper involves data cleaning, correlation analysis, regression analysis, model training and model checking. The whole data analysis method is shown in the Figure 3.

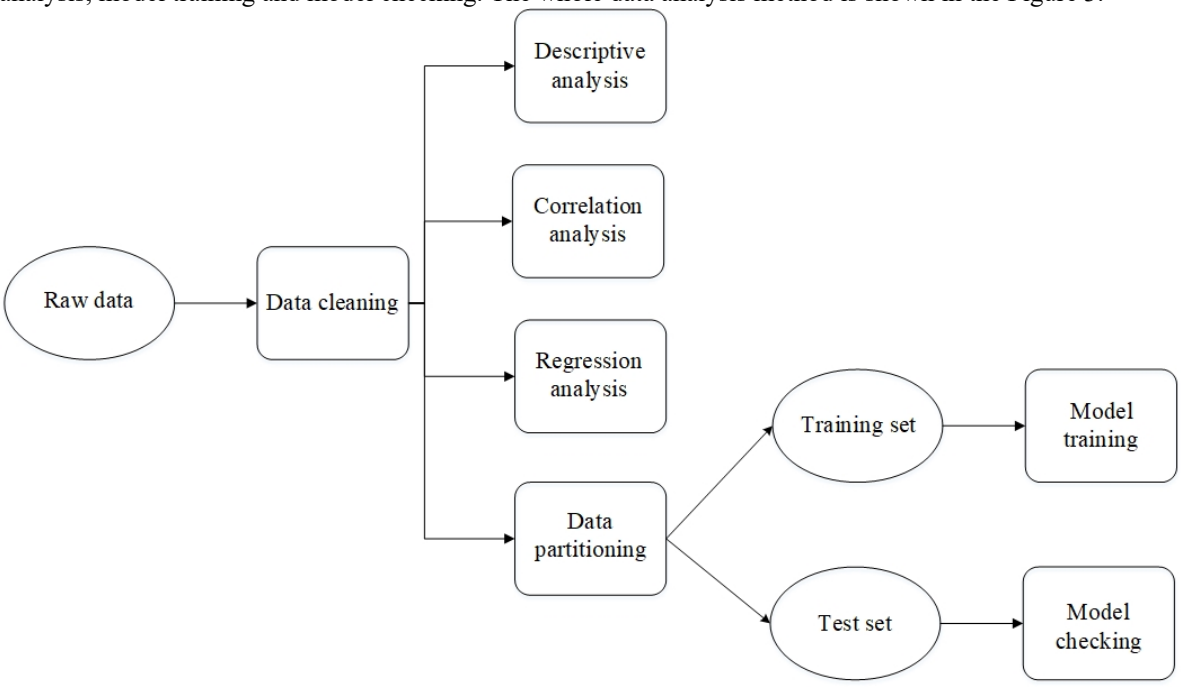

Figure 3: Data analysis method

DATA COLLECTION AND DESCRIPTION

Data Collection

With the help of Beijing Sootoo Internet Technology Company Limited who operates more than 100 WeChat official accounts, this paper selected 21 WeChat official accounts, the number of whose fans ranges from 100 thousand to 1 million and whose types include entertainment, emotion, adoration, fashion, health, finance, etc. These platforms' real operation data have been collected, including the identity number of accounts, the number of fans, article titles, article content, delivery time, article position, reading amount, sharing amount and other basic information. Partial data structure is shown in Table 1. This paper collected data from July 1, 2018 to December 31, 2018. During this period, the 21 official accounts published 13,000 articles, with a total reading quantity of 114 million and a total sharing quantity of 3.07 million. The number of fans, reading quantity, sharing quantity of the 21 official accounts are shown in Table 2.

Table 1: Partial data structure of published articles

Account ID	Delivery time	Title	Fans amount	Reading amount	Liking amount	…
gh_f0692ac72534	2018/07/01 17:16	The cat keeper is not a good keeper who doesn't allow cats enter air-conditioned rooms.	535,652	41,502	759	…
gh_f0692ac72534	2018/07/01 17:16	Speak out. What else is it that you cats take as a bed?	432,613	13,936	565	…
gh_87e04a3d2aea	2018/07/01 18:09	It is my greatest freedom to choose not to marry.	1,236,047	34,133	353	…

Table 2: Basic information Statistics of 21 WeChat official accounts

Account name	Follower amount	Article amount	Reading amount	Sharing amount
Movie Heaven	1,266,259	550	20,285	187
Creative Agency	1,235,968	829	19,008	448
Sales Tactics	1,113,732	656	10,167	504
Human Resources Management	1,095,755	853	9,202	459
Immortal Guo	1,074,336	807	11,338	133
Mood Signature	1,010,434	823	14,316	263
Encyclopedia of Health Fashion	652,339	617	10,621	683
Daily Yoga	604,659	929	8,209	175
Cat is coming	551,101	900	16,082	274
Short Novel	485,049	392	7,989	162
New Ideas	406,711	743	5,351	266
Constellations and Love Situation	402,065	606	7,821	69
Business personnel	384,785	652	2,851	126
Sales & Market	321,449	430	1,204	49
Selected Moments	317,430	754	1,857	76
An English Song Every Day	278,463	731	4,308	212
Financial Reference	230,886	275	3,860	67
Loving Fashion	229,717	328	3,521	35
Head News	204,020	600	1,451	69
Beautiful Trip	160,231	358	1,566	60
Music Valley	147,939	365	2,113	46

Data preprocessing

Data cleaning

The first step of data preprocessing is to clean up the original data, including data verification, outlier processing, identification processing of invalid samples and missing value processing. Data preprocessing provides high-quality and convincing data for subsequent data analysis and is the basis of all data activities. Because of the long time, some articles have been deleted by operators after being published, which may lead to large deviations in reading, sharing and other data. Meanwhile, some missing data were found by data checking. After preprocess, the effective data, 10,857 WeChat articles, were finally obtained.

Table 3: Statistics of data set

The number of accounts	The number of followers	The number of published articles	The number of sharing
21	12,173,328	10,857	2,543,396

Descriptive analysis

In the dimension of official account, the data obtained are analyzed and the information table of official accounts is formed. The table contains information collected by 21 official accounts from July 1, 2018 to December 31, 2018. According to the theme of the account, it can be divided into four categories: entertainment, emotion, life and sports and work information. Relevant descriptions are shown as follows.

Table 4: Four categories of collected official accounts

Category	Account name	The number of accounts	Average number of followers
Entertainment	Movie Heaven Creative Agency Short Novel Selected Moments An English Song Every Day	5	716,634
Emotion	Immortal Guo New Ideas Mood Signature Constellations and Love Situation Music Valley	5	608,297
Life and Sports	Encyclopedia of Health Fashion Daily Yoga Cat is coming Loving Fashion Beautiful Trip	5	439,609
Work Information	Sales Tactics Business personnel Human Resources Management Sales & Market Financial Reference Head News	6	558,438

Table 5: Descriptive analysis of partial data items belonging to official accounts' articles

(a) Entertainment

Entertainment	Minimum	Maximum	Mean	Standard deviation
read	119	181022	11478.12	15379.368
like	0	3403	117.25	216.903

comment	0	936	22.30	41.598
add_to_favor	0	1790	56.74	101.040
sharing	0	10698	245.55	540.003
N				2847

(b) Emotion

Emotion	Minimum	Maximum	Mean	Standard deviation
read	474	100626	9151.07	9288.599
like	0	1150	47.39	76.182
comment	0	1066	16.70	33.461
add_to_favor	0	1223	35.94	68.698
sharing	0	3685	159.29	297.452
N				2948

(c) Life and Sports

Life and Sports	Minimum	Maximum	Mean	Standard deviation
read	112	122249	10234.50	9649.130
like	0	2906	229.58	316.696
comment	0	1001	30.86	60.216
add_to_favor	0	2317	79.75	158.043
sharing	0	8040	260.25	508.324
N				2590

(d) Work Information

Work Information	Minimum	Maximum	Mean	Standard deviation
read	65	218070	5942.90	9142.689
like	0	1054	23.92	51.623
comment	0	127	5.00	9.432
add_to_favor	0	3675	85.38	191.146
sharing	0	13181	283.44	581.124
N				2472

Annotation: “read”, “like”, “comment”, “add_to_favor”, “sharing” represent reading amount, praising amount, comment amount, the number of adding article to favorites, and sharing amount respectively.

Sample recognition

The follower amount of different official accounts varies greatly, and the reading amount of published articles varies greatly. The sharing amount is limited by the reading amount. Therefore, articles with a large reading amount are more likely to have a large sharing amount.

This paper studies the sharing amount based on the reading amount and defines it as sharing rate. The calculation formula is shown in formula (1).

$$sharing\_rate_j = \frac{sharing_j}{reading_j} \tag{1}$$

In this formula, j denotes the No. j article. $sharing\_rate_j$ is the sharing amount of the No. j article. $reading_j$ is the reading amount of the the No. j article. The scatter plot of sharing rate in the whole data range is shown in Figure 4. It can be seen that there are abnormal values, which may lead to great contingency and externality. Therefore, it is necessary to eliminate abnormal data in this study.

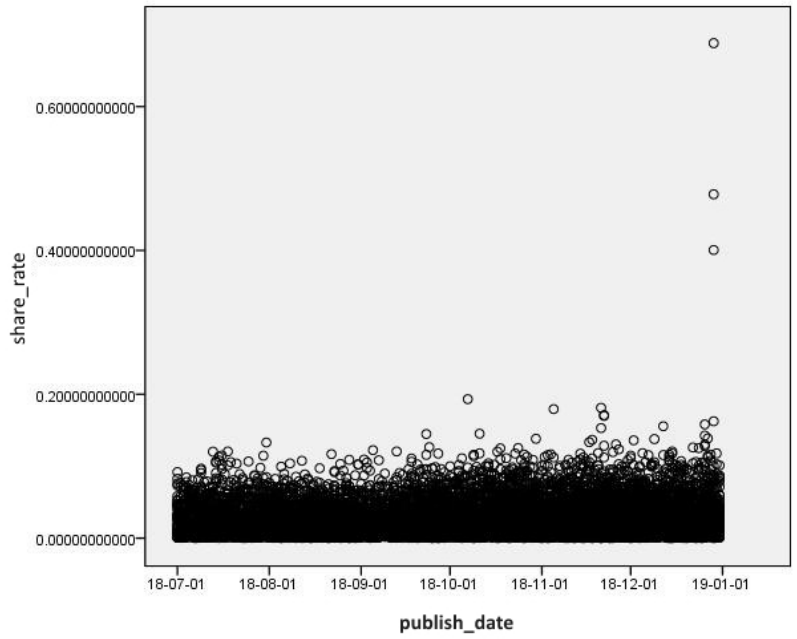

Figure 4: The scatter plot of sharing rate

The basic statistics of sharing rate after excluding outliers are shown in Table 6.

Table 6: Basic statistics of sharing rate

Sharing rate	Mean		.0261025232262
	Standard deviation		.02476798269887
	Minimum		0E-11
	Maximum		.19324090121
	Percentile	25	.0066921282412
		50	.0173247480017
		75	.0395813515349
	N		10854

It can be seen from Table 6 that the maximum sharing rate is 0.193, and the average sharing rate is 0.026. The upper quartile of the sharing rate is 0.0396. The above quartile can indicate the users' attitude to share to a certain degree. In order to calculate conveniently, the sharing rate of 4% or above is defined as high sharing, and the sharing rate below 4% is defined as low sharing. According to this definition, of the total 10,854 articles, 2,676 (25%) articles were high sharing, and 8,178 (75%) were low sharing.

ANALYSIS ON FACTORS AFFECTING SHARING ARTICLES

Characteristics of publisher

Official accounts' influence

The WeChat official accounts with larger reading amount have larger influence. Therefore, this paper considers five indicators, including follower amount, average reading amount, the largest reading amount, average sharing amount and the largest sharing amount. The WeChat official accounts' influence is calculated by different weights.

According to the entropy method, the weights are shown in Table 7.

Table 7: Weight of impact indicators

Indicator	Weight
Follower amount	0.1709
Average reading amount	0.1867
The largest reading amount	0.2016
Average sharing amount	0.2126
The largest sharing amount	0.2282

It can be seen from Table 7 that average sharing amount and the largest sharing amount have the largest weighting value of the influence score.

Type of official accounts

This paper considers that the type of official account (wx_type) has an impact on user's sharing behavior. The accounts of entertainment, emotion, life, sports and work information are marked as 1, 2, 3 and 4 respectively. The relationship between the four types and sharing rate is shown in Figure 5.

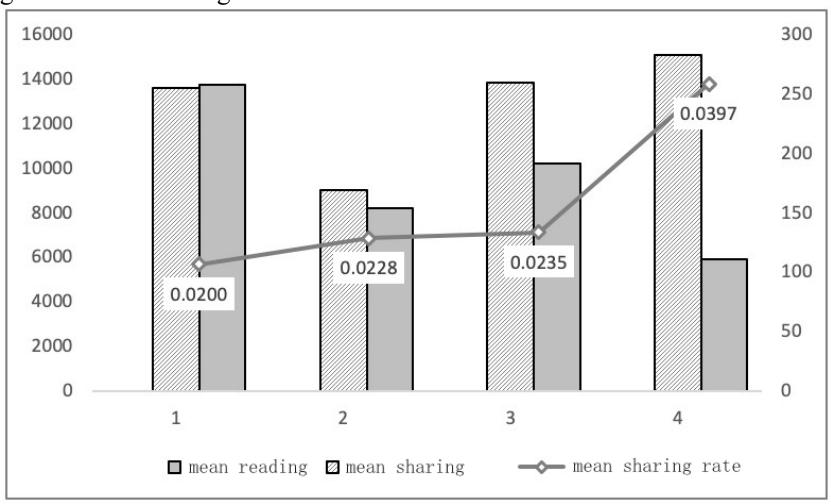

Figure 5: Relationship between account type and sharing rate

Authenticated or not?

This paper considers that whether it is authenticated (wx_certified) is another factor affecting sharing. If authenticated, wx_certified = 1, whereas wx_certified = 0. Table 8 is a basic description of unauthorized and authenticated account articles' sharing rates.

Table 8: A basic description of unauthorized and authenticated articles

	wx_certified	N	percentage	average	standard deviation
share_rate	0	2522	23%	0.02704	0.02451
	1	8332	77%	0.02582	0.02484

Characteristics of recipient

We define the user of the WeChat official account as the recipient. Researches have shown that the higher the recipient's activity and the stronger interaction lead to the stronger willingness of sharing. From the perspective of recipient, the degree of interaction between recipient and publisher is expressed as recipient's commenting, sharing, liking, and adding to favorites.

Characteristics of Information

Title expression

This article divides the title into three types according to different ways of expression: direct (title_type=1); rhetorical (title_type=2); exaggerated (title_type=3) and incomplete type of words (title_type=4). Different categories are independent of each other. Table 9 compares the average sharing rates of titles expressed in different ways. It can be seen that direct and exaggerated titles account for more, and average sharing rates of these two types are relatively high, but the overall difference is not significant.

Table 9: Descriptive Statistics of Title Expressions

	title_type	N	percentage	average	standard deviation
share_rate	1	4315	39.8%	0.0281	0.0258
	2	2214	20.4%	0.0217	0.0216
	3	3794	35.0%	0.0271	0.0255
	4	531	4.9%	0.0208	0.0199

Picture proportion

The article is divided into three types: text-based (img_ratio=1), half-image-based (img_ratio=2), and image-based (img_ratio=3) according to the proportion of pictures from small to large. Table 10 shows the proportion of three types of articles and their sharing rates.

Table 10: Descriptive Statistics of Picture Proportion

	img_ratio	N	percentage	average	standard deviation
share_rate	1	1750	16.1%	0.0356	0.0277
	2	6345	58.5%	0.0264	0.0249
	3	2759	25.4%	0.0195	0.0201

From Table 10, we can see that most of the articles are in the form of half picture and half text, which is also the current popular arrangement mode. However, the higher proportion of pictures may lead to the lower average sharing rate.

Article length

In this paper, 0-500 words are defined as short text (text_words=1), users can read the text in one minute; 500-4000 words are medium text (text_words=2), users can read in less than 10 minutes; more than 4000 words are long text (text_words=3), users need to spend a long time for reading. The sharing rate data should be shown in Table 11.

Table 11: Descriptive Statistics of Article Length

	text_words	N	percentage	average	standard deviation
share_rate	1	2111	19.4%	0.0176	0.0196
	2	7293	67.2%	0.0272	0.0247
	3	1450	13.4%	0.0330	0.0282

Table 11 shows that 500-4000-word medium-long text articles take up the majority, while the short and long text forms are relatively few, which is consistent with the reading habits of users.

Delivery time

Unlike other online social platforms, Wechat official account is designed to cultivate readers' reading habits. Generally, the delivery time is fixed in a day. Because of the different schedules of users from Monday to Sunday, there may be great differences in reading behavior, which may lead to fewer reading opportunities on Monday to Friday weekdays, and more on weekends. That is to say, the sharing rate in different days of a week is likely to be different. Therefore, it is worth exploring the effect of publishing date. Monday to Sunday are represented with 1-7, and the number of articles and sharing rate in a week are shown in Figure 6.

Figure 6: Number and sharing rate distribution of articles on different date

The picture shows that official accounts are mainly delivered on Monday and Sunday. Average sharing rate on Friday is the lowest in a week.

Article's position

Each official account can only deliver a message once a day. Each delivery is limited to six items. The first one is called "headline". Headline is not only at the top of delivery, but its card display enlarges the cover image, making it more attractive. Therefore, the reading and sharing amount of headline is generally higher than the others. Therefore, this paper regards article's location as one of the factors affecting user's sharing behavior. The integer 1-6 is used to represent the position order of the article in the current delivery. Fig. 7 is a scatter plot of the delivery position and sharing rate.

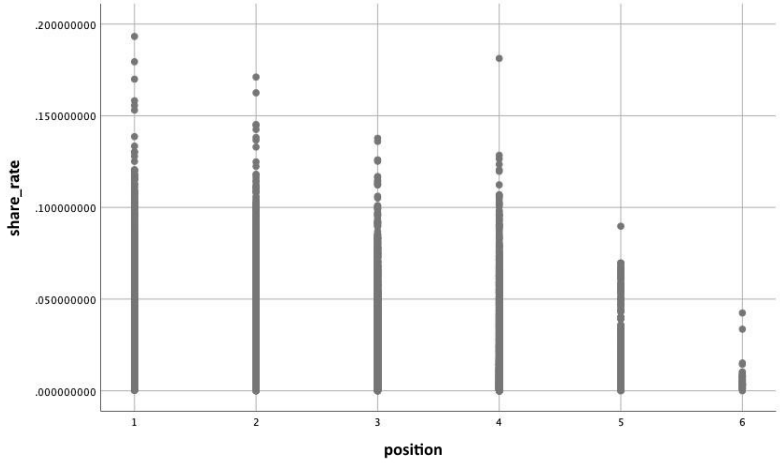

Figure 7: The scatter plot of delivery position and sharing rate

From Figure 7, it can be seen that the sharing rate decreases step by step as the delivery position moves backward, that is, the lower delivery position is, the lower sharing rate is.

Logistic regression analysis of influencing factors

This paper studies the influencing factors of WeChat articles' sharing, so the article sharing is set as the interpreted variable. According to what mentioned before, the interpreted variable is defined as the second categorical variable: high sharing (Share=1, sharing_rate >= 4%) and low sharing (Share=0, sharing_rate < 4%), so the article sharing problem is transformed into a typical two-class problem, and this paper considers to use logistic regression model to establish the relationship between sharing and influencing factors. According to the above analysis, the influencing factors are summarized as shown in Table 12.

Table 12: Summary of influencing factors on sharing

	Variable	Description	Source/Range of Value
Features of publishers	X_1	Influence power of account	Weighted Number of Fans, Articles, Readers and Sharers
	X_2	Account type	Entertainment=1, Emotion=2, Life and Sports=3, Work Information=4
	X_3	Whether is authenticated or not	Yes=1, No=0
Features of recipients	X_4	Interaction status	Mean amount of Comments, Sharing, Liking and Collecting
Features of information	X_5	Style of title	Direct=1, Rhetorical=2, Exaggerated=3, Unfinished=4
	X_6	Picture ratio	Text-based=1, Half picture and half text=2, picture-base=3
	X_7	The length of article	0-500 words, short article=1; 500-4000 words, medium-length article=2; more than 4000, long article=3
	X_8	Delivery time	The time when article is published
	X_9	Delivery data	Mon.=1, Tue.=2, Wed.=3, Thu.=4, Fir.=5, Sat.=6, Sun.=7
	X_{10}	Article position	The first=1, the second=2, the third=3, the fourth=4, the fifth=5, the sixth=6, the seventh=7

Collinear Diagnosis

Before performing logistic regression analysis on the data, a multi-collinearity test is first performed on the selected influencing factors. It can be directly judged according to the correlation coefficient between independent variables. When the correlation coefficient approaches 1, it means there is a strong correlation. Pearson correlation coefficient of sharing factors are shown in Table 13 by using SPSS software.

Table 13: Correlation coefficient of explanatory variable

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
X_1	1									
X_2	-0.01	1								
X_3	0.42	-0.01	1							
X_4	0.50	-0.12	0.31	1						
X_5	0.01	0.04	0.00	0.04	1					
X_6	0.09	-0.29	0.17	0.11	0.20	1				
X_7	-0.16	0.19	-0.16	-0.09	-0.14	-0.62	1			
X_8	-0.26	-0.07	-0.02	-0.18	-0.12	-0.12	0.08	1		
X_9	0.00	0.00	0.00	0.01	-0.01	0.00	-0.01	0.05	1	
X_{10}	0.23	0.00	0.21	-0.36	-0.01	0.19	-0.27	-0.04	-0.02	1

The absolute value of correlation coefficient between most variables is small, but the correlation coefficient between the image ratio and the length of the article reaches -0.62. Therefore, the multicollinearity test is continued by using the tolerance (TOL) and variance expansion factor (VIF) indicators. The value is defined as:

$$TOL_{(j)} = 1 - R_j^2 \ (j = 1, 2, \ldots, m) \tag{2}$$

Among them, R_j is the correlation coefficient between the independent variable and other m-1 independent variables. If there is serious collinearity, that is $R_j \approx 1$, there is $TOL_{(j)} \approx 0$. Generally speaking, a tolerance of less than 0.2 can be considered as a sign of the existence of multiple collinearity. The variance expansion factor is the reciprocal of the allowable value. Normally, when VIF ≥ 5, it can be considered that there is collinearity between the independent variables. The collinearity statistics of independent variables are shown in Table 14.

Table 14: Collinearity statistics of explanatory variables

Variable	Tolerance	Variance Expansion Factor
X_1	0.514	1.946
X_2	0.877	1.14
X_3	0.742	1.348
X_4	0.454	2.204
X_5	0.939	1.064
X_6	0.547	1.83
X_7	0.579	1.728
X_8	0.996	1.004
X_9	0.887	1.127
X_{10}	0.552	1.81

Since the tolerance of all variables is greater than 0.2, the variance expansion factor is less than 5, so there is basically no multicollinearity problem between the explanatory variables affecting the sharing, and all variables can be included in the regression model.

Logistic regression model construction

In this paper, 10 factors are extracted from the three dimensions of publisher's feature, recipient's and information's as explanatory variables, and the sharing is used as the dependent variable to construct the regression model. According to the previous analysis, the sharing is a binary categorization variable. When the value is 1, it is high sharing, and when it is 0, it is low. Therefore, this analysis of the factors affecting sharing is a discrete selection problem. Given independent variable $X_1, X_2, \cdots X_n$, if $P(P \in [0,1])$ is the incidence of high sharing, then $1-P$ is the incidence of low sharing. So the binary logistic regression model is defined as:

$$Logit(P) = \ln\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n \tag{3}$$

In the formula, $P/(1-P)$ is a ratio, that is, the probability of an event occurring to the probability of non-occurrence. This paper refers it to the ratio of incidence of high sharing and low sharing. β_0 is a constant term, β_i represents the regression coefficient of each variable. When β_i is a positive value, it means each factor has a positive influence on sharing; when β_i is a negative value, it has a negative influence.

Through the operation it can be obtained:

$$P = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)]} \tag{4}$$

This is consistent with the sigmoid function, it also reflects the nonlinear relationship between the probability P and the dependent variable. Limited to 0.5, when P>0.5, it is classified as high sharing; otherwise classified as low sharing.

Logistic regression model results analysis

According to the established model, this paper uses SPSS 25.0 to realize the calculation of this model. Table 15 shows the overall fitting effect of this model, in which the chi-square ratio statistic is 2992.546 and P=0.000<0.05 meaning it reaches a significant level and indicating that at least one of these explanatory variables is a model pair. The correct percentage of the overall classification of this model is 74.8%, indicating that this regression model has higher accuracy.

Table 15: Test of Logistic regression fitness

Chi square	Log likelihood	Cox & Snell R^2	Nagelkerke R^2	Percentage of correctness
2992.546(0.000)	10881.046	0.241	0.334	74.8%

In this paper, the Backward LR method of logistic regression is used to estimate the model parameters, that is, all variables are entered into the regression equation, and then the independent is eliminated whose value of the p is maximum according to the probability value of the maximum likelihood estimation statistic. This process is repeated until remaining independent variables are significant. In the first step, all variables enter the regression test and the results are shown in Table 16.

In Table 16, B and Sig respectively represent the regression coefficient and the significance level, and Wald is used to test whether the independent variable has an influence on the dependent variable. The larger the Wald value is, the more significant the effect is. Exp(B) is an odds ratio (OR value), which indicates the change ratio of the event occurrence ratio caused by each unit of the independent variable when the other conditions are unchanged.

Table 16: Logistic regression results of variables

Variable	B	S.E	Wald	df	Sig	Exp(B)
X_1	0.285	0.011	719.272	1	0.000	0.752
X_2			1152.752	3	0.000	
X_3	0.906	0.09	101.547	1	0.000	2.475
X_4	2.697	0.078	1201.092	1	0.000	14.839
X_5			84.17	3	0.000	
X_6			2.567	2	0.277	
X_7			52.682	2	0.000	
X_8	0.040	0.016	6.495	1	0.011	1.041
X_9	0.002	0.012	0.024	1	0.876	1.002
X_{10}	-0.326	0.026	156.885	1	0.000	1.386
Constant	-3.443	0.365	88.746	1	0.000	0.032

Except the independent variable picture ratio and delivery date, the other 8 variables have significant effects on the dependent variable (P<0.05). After 2 rounds of optimization test and screening, the final variable regression results are obtained, as shown in Table 17.

Table 17: Logistic regression results of the optimized variables

Variable	B	S.E	Wald	df	Sig	Exp(B)
X_1	0.285	0.011	720.49	1	0.000	0.752
X_2			1299.789	3	0.000	
X_3	0.905	0.09	101.631	1	0.000	2.473
X_4	2.698	0.078	1202.788	1	0.000	14.854
X_5			85.904	3	0.000	
X_7			99.977	2	0.000	
X_8	0.040	0.016	6.4	1	0.011	1.041
X_{10}	-0.325	0.026	156.143	1	0.000	1.384
Constant	-3.298	0.353	87.137	1	0.000	0.037

After optimization test, we can know that type, authentication, user interaction degree, title expression, article length, delivery time and position are significant for sharing (P<0.05).

The influencing factors can be summarized as follows from three aspects: characteristics of publisher, characteristics of recipient and characteristics of information.

(1) Characteristics of publisher
From the regression results, influencing power, account type and whether is authenticated have a significant impact on sharing behavior. Among them, influencing power has a positive impact on sharing rate (B=0.285). From the perspective of the OR value, sharing rate will increase by 0.752 times for each unit increase of official accounts' influence. It indicates that the greater influence of the official account have, the more easily shared.

An authenticated account's articles is easier to be shared than articles published by the unauthenticated. From EXP(B)=2.473, the sharing probability of an article belong to an authenticated account is 2.473 times than an unauthenticated account's.

Compared with work information one official account type, entertainment, emotion, life and sports have lower sharing influence. The possible reason is that WeChat has strong working attributes. Therefore, they express their work sympathy or acquire professional skills through sharing behavior.

(2) Characteristics of recipient
Table 5-12 shows that the regression coefficient of the "user interaction degree" variable is 2.698, indicating that user interaction has a significant positive impact on sharing. For each unit of user interaction, the sharing probability increases by 14.854 times. Therefore, cultivating user interaction is one of the channels that lead to high sharing.

(3) Characteristics of information
Parametric regression results show that only the title expression, article length, delivery time, and delivery position have a significant impact on sharing. Among them, the "direct" title is more likely to produce high sharing than other expressions.

For the length of the article, 500-4000 words are most likely to produce high sharing. It is the best for touching users that the article is moderately long, the text is between 500 and 4000 words, and the reading time is about 10 minutes.

The article's position has a negative influence on sharing. The lower delivery position, the worse sharing effect. The probability of high sharing decreases by 1.384 times with the article ratchets down another notch. This is related to the way operator arranges the content. The headline delivery display is more conspicuous, so the operator should put higher quality content.

The effect between image scale and sharing is not significant, which may be related to the way users read article. Unlike the short and fast reading mode of Weibo, WeChat articles are relatively long, and users pay more attention to the content.

SHARING PREDICTION MODEL AND EXPERIMENTAL ANALYSIS

Prediction Model Construction

Logical Regression Model

In this logistic regression analysis, the dependent variable Sharing is a dichotomous variable whose values are Sharing=1 and Sharing=0, representing high sharing and low sharing respectively. The independent variables that affect the value of Sharing are $x_1, x_2, \ldots, x_m$. The logistic regression model can be expressed as:

$$f(x,\beta)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_m x_m)}} \tag{5}$$

In equation (5), $\beta=(\beta_1,\beta_2,\ldots,\beta_m)$ is the parameter of this model, namely the regression coefficient. x_i is the No. i eigenvalue of the record x, and m is the number of features. The logistic regression model is used to calculate the regression parameters: suppose the probability that an article is high sharing is $f(x,\beta)$; then get the probability that the text is low sharing is $1-f(x,\beta)$. The probability of y under conditions of x and β is shown in the equation (6):

$$P(y\mid x,\beta)=\begin{cases} f(x,\beta), y=1 \\ 1-f(x,\beta), y=0 \end{cases} \tag{6}$$

The possibility obtained from the entire training set can be expressed as follows:

$$P(X, y, \beta) = \prod_{i=1}^{N} f(x^i, \beta)^{yi} \left(1 - f(x^i, \beta)\right)^{1-yi} \tag{7}$$

Logarithm of equation (7) can be transformed to:

$$\ln P\left(X, y, \beta\right) = \sum_{i=1}^{N} \left(y_i \ln f\left(x^i, \beta\right) + \left(1 - y_i\right) \ln\left(1 - f\left(x^i, \beta\right)\right)\right) \tag{8}$$

x^i is the No. i record in data set X, y_i is the classification result of the No. i record, and N is the number of records in the training set. This logistic regression model parameters can be obtained by calculating the maximum value of $\ln P\left(X, y, \beta\right)$, based on the maximum likelihood estimation method and iterations until convergence.

Support Vector Machine Model

In general, linear separable data is relatively rare. For linear inseparable samples, kernel function is generally introduced to map them to higher dimensions to solve the problem. Therefore, this paper introduces the kernel function to construct the prediction model. Table 18 lists several commonly used kernels.

Table 18: Common Kernel Functions

kernel function	Expression	Explain
Linear Kernel Function	$\kappa\left(x_i, x_j\right) = x_i^T x_j$	Linear separable support vector machine
Polynomial Kernel Function	$\kappa\left(x_i, x_j\right) = (x_i^T x_j)^d$	One of the commonly used kernel functions of linear inseparable SVM
Gaussian Kernel Function	$\kappa\left(x_i, x_j\right) = \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2})$	The most mainstream kernel function of nonlinear classification SVM
Sigmoid Kernel Function	$\kappa\left(x_i, x_j\right) = \tanh(\beta x_i^T x_j + \theta)$	One of the commonly used kernel functions of linear inseparable SVM

Input: The training sample is $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, where x is the m-dimensional eigenvector and y_i is the binary output with a value of +1 or -1.

Output: Separate hyperplane parameters and decision functions.

The algorithm process is as follows:

(1) Construct the constrained optimization problem by selecting the appropriate kernel function and the penalty factor C.

$$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{m} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j)$$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{9}$$

$$0 \le \alpha_i \le C$$

(2)Solving with SMO algorithm, and obtaining the corresponding α^* vector with the minimum formula.

(3)Find out that all S support vectors satisfying $S = \{i \mid 0 < \alpha_i < C, i = 1, 2, \ldots, m\}$, solve equation (10) to get b corresponding to each support vector (x_i, y_i), and b^* is the average of all b, as expressed in equation (11).

$$y_s(\sum_{i \in S} \alpha_i y_i \kappa(x_i, x_j) + b) = 1 \tag{10}$$

$$b^* = \frac{1}{S}\sum_{s\in S}\left(y_s - \sum_{i\in S}\alpha_i y_i \kappa(x_i, x_j)\right) \tag{11}$$

The final classification model is

$$f(x) = \sum_{i=1}^{m}\alpha_i^* y_i \kappa(x_i, x_j) + b^* \tag{12}$$

In the process of training model, different kernel functions and parameters are used for comparative training in order to get the most suitable kernel functions and parameters. In actual operation process, the best is to use Gaussian kernel function, with the gamma value of the kernel function parameter is 0.1 and the penalty coefficient C is 4.

Analysis of Experimental Results

Evaluation Indicators

In order to evaluate the effectiveness of the sharing behavior prediction model, this paper uses the commonly used evaluation indicators of classification algorithm: Accuracy, Precision, Recall, F_1 (F_1 score):

$$Accuracy = (TP + TN)/(TP + FP + FN + TN)\times 100\%$$

$$precision = TP/(FP+TP)\times 100\%$$

$$recall = TP/(TP + FN)\times 100\%$$

$$F_1 = 2\cdot\frac{precision\cdot recall}{precision + recall}$$

Where TP represents the sample size predicted to be high sharing and actually high sharing; FP represents the sample size predicted to be high sharing but actually low sharing; TN is the sample size predicted to be low sharing and actually low sharing; FN is the sample size predicted to be low sharing but actually high sharing.

ROC curve is drawn in the form of curve with sensitivity (TPR) as ordinate and specificity (FPR) as abscissa. The larger the coverage area under ROC curve is, the better the prediction accuracy is and the better performance of the prediction model is.

Comparison of experimental results of classification algorithm

According to the characteristics shown in Table 12, the logistic regression model and the support vector machine model are respectively used for training and prediction, and then the prediction results are evaluated. The confusion matrix is shown in Table 19. The values of accuracy, precision, recall and F_1 in these two models are shown in Table 20. The ROC curve is shown in Figure 7.

Table 19 Prediction result confusion matrix

		Predictive high sharing	Predictive low sharing
Logistic Regression	Actual high sharing	550	2377
	Actual low sharing	230	589
Support Vector Machine	Actual high sharing	625	162
	Actual low sharing	132	687

Table 20 Comparison of experimental results

Prediction model	Accuracy	Precision	Recall	F_1 score
Logistic Regression	0.709	0.713	0.719	0.716
Support Vector Machine	0.817	0.809	0.839	0.824

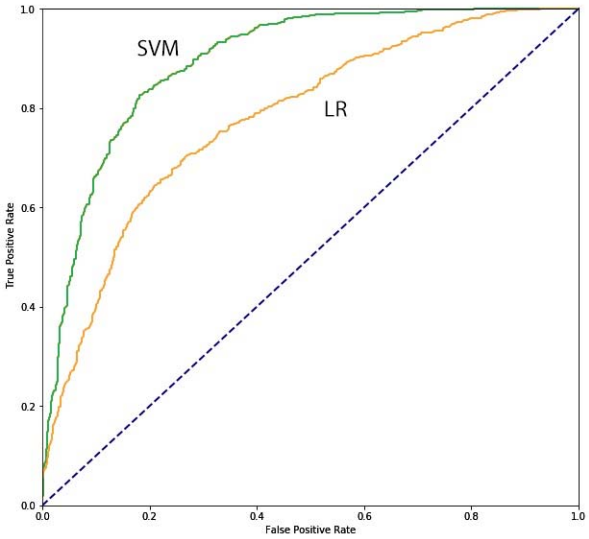

Figure 7: ROC Curves of Different Models

From Table 20, it can be seen that the prediction accuracy of support vector machine is 81.7%, while the prediction accuracy of logistic regression is only 70.9%. In terms of each evaluation index, support vector machine is superior to logistic regression model. What's more, it can be seen that the coverage area under the ROC curve of the support vector machine (SVM) model is larger, which indicates that the prediction accuracy of the SVM model is higher and the model performance is better.

In order to further verify the validity of the Gaussian kernel function selected in the support vector machine model, this paper selected three other different kernel functions for comparative experiments, and calculated the results of each evaluation index as shown in Table 21.

Table 21: Prediction results of different kernels

	Accuracy	Precision	Recall	F_1 score
Linear Kernel Function	0.721	0.717	0.737	0.727
Polynomial Kernel Function	0.797	0.796	0.803	0.800
Gauss Kernel Function	0.817	0.809	0.839	0.824
Sigmoid Kernel Function	0.573	0.579	0.559	0.569

Characteristic Importance Comparison

In order to further investigate the impact of various features on the prediction results of users' sharing behavior, this paper presents some features that have significant effects on the dependent variable after using full features respectively, and the prediction results after removing some types of features, are as shown in Table 22.

Table 22: Prediction effect of different features

	Accuracy	Precision	Recall	F1-score
Use all features	0.817	0.809	0.839	0.824
Using optimized features	0.832	0.818	0.862	0.839
Unused Publisher Characteristics	0.729	0.722	0.762	0.742
Unused recipient characteristics	0.737	0.734	0.761	0.747
Unused information features	0.752	0.758	0.755	0.756

SUMMARY AND PROSPECT

This paper takes the WeChat official accounts as research object, draws lessons from researches that are relatively mature and focus on microblogs' sharing, extracts the possible influencing factors from the three dimensions: information publisher, information recipient and information characteristics. With the support of some official accounts' real operation data, we explore the influence of various factors on users' sharing behavior.

What this study have found are as follows. First, account influence, account type, and whether account is authenticated in the characteristics of the information publisher have a significant effect on sharing. Compared with the official accounts of entertainment, emotion, life and sports, the official accounts of work information deliver articles with high sharing probability. Second, the interaction between users and WeChat official accounts has a significant positive impact on sharing behavior. That is, the higher the interaction is, the higher the probability of sharing is. Third, title expression, article length, delivery time and

delivery position in information features importantly determine sharing. Fourth, the characteristics of publisher, recipient and information have great influence on sharing, and the importance is not different.

Without doubt, there are still many shortcomings in this study. Based on the existing research results, we can further study and improve the following two aspects in the future. First, the influencing factors in this paper are mainly constructed from three aspects: official accounts' features, users' features and information's features. However, due to the technical limitations, the text content characteristics of the article are not taken into account. In addition, other factors including the originality of the article and the frequency of delivery should also be considered. The improvement of indicators will be the focus of future research. Second, this paper defines the threshold of high sharing and low sharing, and takes it as a classification variable. For the validity of the whole experiment, the definition between high sharing and low sharing is very important. In the future, the boundary value will be adjusted to obtain the optimal boundary value.

ACKNOWLEDGMENT

Financial support from the Science and Technology Plan Project of Beijing (No.Z171100000117009) and NSFC (No.91546125) is acknowledged.

REFERENCES

Deng Qin, Ma Yefeng, Liu Yi, Zhang Hui. (2015). Prediction of Weibo Reposting Volume Based on BP Neural Network. *Journal of Tsinghua University (Natural Science Edition).* 55(12), 1342-1347. (in Chinese)

Guan Bin. (2015). Research on Influencing Factors of User's Willingness on Using WeChat Official Account. (Doctoral dissertation). Central China Normal University. (in Chinese)

Han Xinming. (2018). A WeChat Moments Information Dissemination Model Based on Behavior Analysis. *Modern intelligence*. 38(07), 62-66. (in Chinese)

Huang Chujun, Peng Qilin. (2014). Research on Motivation and Communication Effect of University WeChat Official Accounts: An Empirical Analysis on Central South University's WeChat official account. *Southeast communication*. 2014(08), 122-124. (in Chinese)

Jing Ming, Zhou Yan, Ma Danchen. (2014). The Ways, Characteristics and Reflections of Wechat Communication. *News and Writing.* 2014(7), 41-45. (in Chinese)

Li Yinle, Yu Hongtao, Liu Lixiong. (2013). Prediction Method of Weibo Reposting Scale Based on SVM. *Computer Applied Research.* 30(09), 2594-2597. (in Chinese)

Nor Athiyah A, Dai N, Yuko T, Yuko M. Why I Retweet? Exploring User's Perspective on Decision-Making of Proceedings of Information Spreading during Disasters. The 50th Hawaii International Conference on System Sciences, 2017.

QuestMoblie. (2018). WeChat Official Account Population Insight Report. Retrieved from http://www.questmobile.com.cn/research/report-new/18/ (accessed June 2019) (in Chinese)

Tencent. (2018). 2018 WeChat Data Report. Retrieved from https://support.weixin.qq.com/cgi-bin/mmsupport-bin/getopendays/ (accessed June 2019). (in Chinese)

Wang Haiyan. (2015). Analysis on Editing and Operating Strategies of WeChat Official Account in Traditional Media. *Friends of Editors*. 12 (2), 1-2. (in Chinese)

Wang Zhifeng, Feng Xiwei, Jia Qiang, Zhu Rui, Qin Hang. (2017). Weibo Reposting Prediction Based on Multi-feature Neural Network. *Journal of Liaoning University of Petroleum and Chemical Technology*. 37(06), 47-50. (in Chinese)

Yuanfeng C, Dan Z. (2013). Understanding Factors Influencing Users' Retweeting Behavior---A Theoretical Perspective. Proceedings of the Nineteenth Americas Conference on Information Systems, Chicago, Illinois, August 15-17, 2013.

Zhou Xianting, Huang Wenming, Deng Zhenrong. (2017). Weibo Reposting Behavior Prediction Based on Anomaly Detection and Random Forest. *Computer Science*. 44(07), 191-196+220. (in Chinese)

Zhang Yinyu, Li Wu. (2016). Which WeChat Articles Are Easier to Share: An Exploratory Study Based on Content Analysis. *See and Listen*. 2016(02), 117-118. (in Chinese)

Zhang Yang, Lu Rong, Yang Qing. (2012). Research on Reposting Behavior Prediction in Weibo. *Journal of Chinese Information Science*. 26(4), 109-114. (in Chinese)

Zhang XiaoWei, Wang Wei, Qin Dongxia. (2018). Prediction of Weibo User Reposting Behavior Based on Integrated Learning. *Journal of Henan Normal University (Natural Science Edition).* 46(02), 111-116. (in Chinese)

Zhao Huidong, Liu Gang, Shi Chuan, Wu Bin. (2016). Reposting Quantity Prediction of Weibo Based on Reposting Propagation Process. Chinese Journal of electronics. 44(12), 2989-2996. (in Chinese)