Available online www.jsaer.com

Journal of Scientific and Engineering Research, 2022, 9(11):24-30

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Identification of the Cyber Violence Behavior Based on a Binary Logistic Regression Model

Yongwei Yang, Bin Song, Qiqi Li, Pengfei Yu

School of Mathematics and Statistics, Anyang Normal University, Anyang 455000, China

Abstract This paper uses the binary logistic regression model to identify news events and judge whether the events will develop into network violence events according to the selected indicators. We first select indicators that may develop into cyber violence incidents, and use Python crawlers to find relevant data for the selected indicators. Secondly, the SPSS software is used to analyze the correlation analysis between these indicators, and the indicators with strong correlation are selected as the variables in the binary Logistic regression equation. Finally, a binary logistic regression equation for identifying cyber violence is established, and the model is verified to make the regression equation more reliable and predictable.

Keywords binary logistic regression model, cyber violence, crawler technology

1. Introduction

Under the background of free speech liberalization and anonymity, the network platform is the main place of human entertainment in the new information age, and its hidden problems are becoming increasingly prominent. Among them, cyber violence is one of the most typical problems of today. Cyber violence refers to a kind of speech, text, pictures and video published on the Internet by netizens, which are defamatory, slander, violate reputation infringement, damage rights and interests. Peterson and Densley explored the individual-level correlates and risk factors associated with cyber violence, the group processes involved in cyber violence, and the macro-level context of online aggression [1]. With the influence of reinforcement sensitivity on cyber violence as the starting point, Zhang and Li analyzed punishment sensitivity and cyber violence and the relationship between the two by building a structural equation model [2]. Through a review of the current literature on social media use in teacher education, and a multi-disciplinary perspective on issues of cyber-violence, Nagle [3] discussed the ethical implications for teacher educators who want to use Twitter as a pedagogical tool and offer strategies to develop critical social media literacy practices. In the paper [4], the Theory of Planned Behavior (TPB) was used to identify the factors that affect cyber violence behaviors. The TPB antecedents of attitudes, subjective norms, and perceived behavioral control had a statistically significant and positive relationship with cyber violence behavior. However, there are relatively few literature studies on the early identification of cyber violence. Based on this, we establish a binary Logistic regression model for cyber violence identification to effectively reduce the adverse effects of cyber violence on society.

2. Using Python crawler technology to crawl Micro-blog data

Cyber violence is not only a negative topic, but also a violent topic. The comments in the topic obviously involve words of violence, insults, slander, and ridicule. The participants in cyber violence are not only individual netizens, but also the participation of some members of the public and the media. At present, Sina Micro-blog is a large user communication platform in China. It is a public platform to establish contact and conduct information interaction among users. Micro-blog users share text, video and other information, which contains a huge amount of data. At the same time, microblog is also one of the more commonly used software for college students.Therefore, we used the Python crawler technology to select some controversial topics on the microblog platform. By crawling three typical cyber violence incidents in microblog, including social news, emotional entertainment, and public events, we obtained the number of public comments, the number of public microblog reposts, the number of public likes, and the number of online news media in the incident. Through the analysis of these indicator data, we canidentify the occurrence of cyber violence incidents.

With the help of keywords, we crawl all the search results within a certain period of time on the microblog platform, and count the number of microblog posts, user types, microblog comments and other information in the search results, and output the results to an Excel table.

Table 1: Data on crawler results

	Number of public comments	**The number of public microblog reposts**	**The number of public likes**	**The number of online news media**	**The number of public blogs**	**The number of microblogs forwarded by the media**
Violation of Ali female employees	392631	87279	4249345	56	199	41375
The counterattack of the postgraduate re-examination of Peking Union	54409	10128	792214	15	32	4375
Lost wedding ring in taxi and asks for payment	2962	646	96733	16	248	252
Deng Lun's tax evasion incident	8340	1184	59602	10	569	72
China Eastern Airlines mu5735	2342	1487	14751	30	451	67
Women's figure skating Zhu Yi was questioned	526	100	1562	4	20	6
A Hangzhou girl was rumored to pick up an express package	564	826	5617	2	45	88
Taking pictures with naked backs in the community	42421	13621	1585184	21	491	12753
Jiang Ge's mother is questioned	45	21	242	0	2	0
Instant noodles violate food safety regulations	291	80	708	9	21	80
Luo Xiaomaomaozi drinks pesticide	12170	2212	189811	58	60	1370
Wei Ya's tax evasion	6661	2204	234506	30	488	982
Weilong's latiao packaging case	16103	10355	265222	56	557	1419
The cat in The Housewife	470	445	2283	0	30	0

3. Binary logistic regression models

The Logistic regression model is a model that establishes a regression formula for the classification boundary function and makes predictions according to the relevant factor data when the results need to be classified [5].

The binary Logistic regression model studies the relationship between categorical variables and multiple factors directly, and is characterized by the occurrence ratio $\frac{p}{1-p}$ of events. If $\lim_{p\to 0^+}(\frac{p}{1-p})=0$, then $\lim_{p\to 1^-}(\frac{p}{1-p})=+\infty$.

Use the probability p of event occurrence to establish the logistic regression, the formula is as follows:

$$p=\frac{e^z}{1+e^z}=\frac{e^{b_0+b_1x_1+\cdots+b_nx_n}}{1+e^{b_0+b_1x_1+\cdots+b_nx_n}}. \tag{1}$$

If the event occurs, the probability is:

$$p(Y=1\,|\,x)=\pi(x)=\frac{1}{1+e^{-g(x)}}.$$

If the event does not occur, the probability is:

$$p(Y=0\,|\,x)=1-p(Y=1\,|\,x)=1-\frac{1}{1+e^{-g(x)}}=\frac{1}{1+e^{g(x)}}.$$

Where $g(x)=b_0+b_1x_1+\cdots+b_nx_n$.

The ratio of the probability of a story event occurring and not occurring is:

$$\frac{p(Y=1\,|\,x)}{p(Y=0\,|\,x)}=\frac{p}{1-p}=e^{g(x)}. \tag{2}$$

Taking the logarithm of the above formula, we get:

$$\ln\left(\frac{p}{1-p}\right)=g(x)=b_0+b_1x_1+\cdots+b_nx_n,$$

where, $b_0,b_1,\cdots,b_n$ are the regression coefficients, $x_1,\cdots,x_n$ are the main factors that affect cyber violence.

4. Applications of the binary Logistic regression model

4.1. Correlation analysis between indicators

In order to analyze whether there is a correlation between the factors that affect cyber violence incidents, we take the six factors that affect cyber violence as variables, and establish a two-variable correlation analysis model. This model not only gives the correlation between two variables and can be extended to the comparison between multiple factors, but also has a certain measure of the strength of the correlation. The specific steps are as follows.

Step 1 Draw scatter plots.

In order to facilitate the observation of whether there are regular changes between the two variables, we select the data between the two variables and make scatter plots using SPSS software. The number of public posts and the number of network media are taken as an example below.

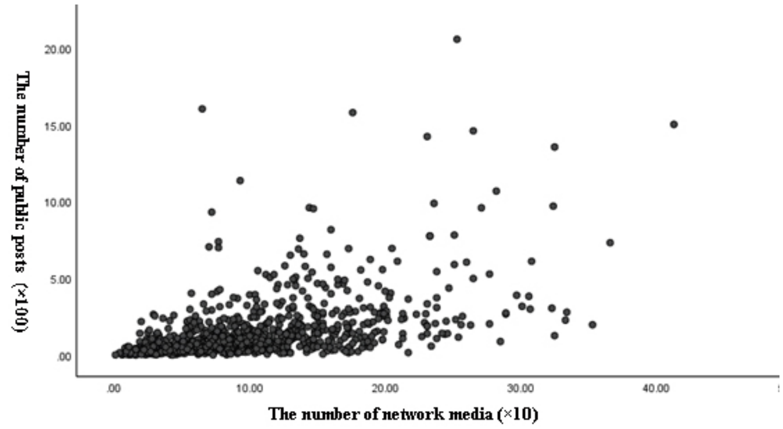

Figure 1: Scatter plot between the number of public posts and the number of network media

From Figure 1, it cannot be seen whether there is a correlation between the number of public posts and the number of network media.We need to further calculate the correlation coefficients in order to determine whether there is a correlation between them and the magnitude of the correlation.

Step 2 Select the correlation coefficient formula

The correlation coefficient is a numerical feature of the degree of correlation between the research variables, generally represented by r,

$$r = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(x)}\sqrt{Var(y)}}.$$

Before calculating the correlation coefficient, it is necessary to determine the choice of the correlation coefficient. Generally, there are three types of correlation coefficients:

(1) Pearson correlation coefficient. Calculate the correlation analysis between continuous variables or equally spaced variables, but it requires the variables to obey the normal distribution, so it is necessary to test whether the variables obey the normal distribution before calculation. The formula for calculating the sample correlation coefficient is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

(2) Spearman correlation coefficient. It is the nonparametric form of Pearson correlation coefficient. It is calculated according to the rank rather than the actual value of the data. It is suitable for ordered data and equidistant data that do not satisfy the assumption of normal distribution. The calculation formula is as follows:

$$r = 1 - \frac{6\sum d^2}{n(n^2-1)'},$$

where d is the difference between the ranks of each pair of observations after taking the ranks of x and y respectively, and n is the number of all observation pairs.

(3) Kendall rank correlation coefficient. It is a statistic to measure the degree of correlation between two ordinal variables or two rank variables, and it is a method for calculating the correlation coefficient belonging to a nonparametric test.

The normal distribution diagram is used to characterize whether each factor obeys the normal distribution. The normal distribution Q-Q diagram of the number of online media is given below.

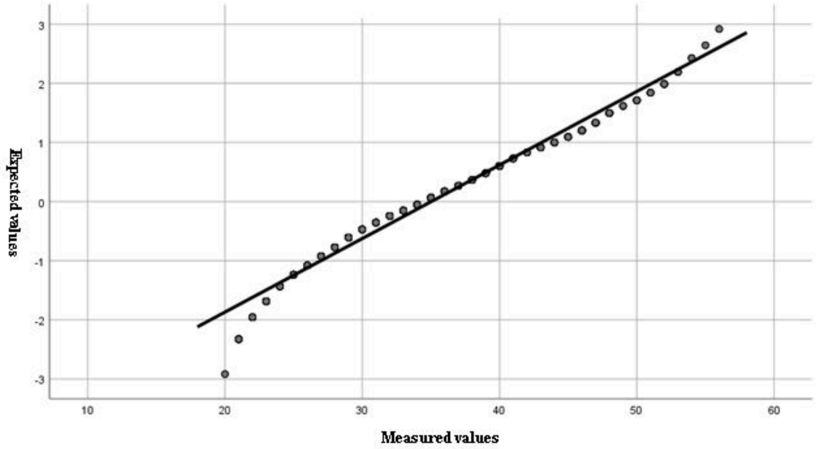

Figure 2: The normal distribution Q-Q diagram of the number of online media

According to Figure 2, not all of the factors in the influencing factors of network violence meet the normal distribution, so the Pearson correlation coefficient cannot be used to measure it, and we use the Spearman correlation coefficient to calculate the correlation coefficient.

Step 3 Determine the correlation coefficient between the various factors.

In this paper, SPSS software was used to calculate the correlation, and the correlation results between the factors were obtained as shown in Table 2.

Table 2: Correlation matrix between factors

	Number of public comments	The number of public microblog reposts	The number of public likes	The number of online news media	The number of public blogs	The number of microblogs forwarded by the media
Number of public comments	1.000	0.952**	0.969**	0.741**	0.618*	0.895**
The number of public microblog reposts	0.952**	1.000	0.960**	0.792**	0.653*	0.880**
The number of public likes	0.969**	0.960**	1.000	0.752**	0.640*	0.920**
The number of online news media	0.741**	0.792**	0.752**	1.000	0.622*	0.724**
The number of public blogs	0.618*	0.653*	0.640*	0.622*	1.000	0.475
The number of microblogs forwarded by the media	0.895**	0.880**	0.920**	0.724**	0.475	1.000

The correlation coefficient is used to evaluate the degree of correlation. If $0 < r < 0.3$, there is basically no correlation between the two variables. If $0.3 < r < 0.5$, there is a weak correlation between the two variables. If $0.5 < r < 0.7$, there is a strong correlation between the two variables. If $0.7 < r < 1$, there is a very strong correlation between the two variables [6].

According to Table 2, there is a strong correlation between the selected six indicators. Therefore, the following is a binary logistic regression model based on these indicators.

4.2. Solution of the binary logistic regression model

We use the binary logistic regression model to make predictions about whether the selected events will have a network violence phenomenon or not, and substitute the values of the main factors into the regression formula to get the predicted results. The analysis is obtained by the binary logistic regression using SPSS software. Table 3 shows the coefficients influencing the occurrence of violent events in the regression model.

Table 3: Variables in the equation

	B	Standard error	Wald	Degrees of freedom	Conspicuousness	Exp(B)
Number of public comments	-1.59	217.008	0	1	0.994	0.204
The number of public microblog reposts	-0.835	604.064	0	1	0.999	0.434
The number of public likes	0.003	0.378	0	1	0.994	1.003
The number of online news media	0.073	15.938	0	1	0.996	1.075
The number of public blogs	-0.349	49.598	0	1	0.994	0.705
The number of microblogs forwarded by the media	-0.01	4.428	0	1	0.998	0.99
Constant	48.444	6298.702	0	1	0.994	1.0934

The formula for the probability of cyber violence can be obtained from formula (2). From Table 3, it can be obtained that the constant in the above equation is 48.444. The main factors are the number of public comments, the number of public microblog retweets, the number of public likes, the number of online news media, the number of public blogs, and the number of microblogs forwarded by the media. The corresponding coefficient values are -0.010, 0.073, 0.003, -0.835, -1.590, -0.349, respectively. Putting these coefficients into formula (1), the binary logistic regression equation can be obtained:

$$p=\frac{e^{48.444-0.010x_1+0.073x_2+0.003x_3-0.835x_4-1.590x_5-0.349x_6}}{1+e^{48.444-0.010x_1+0.073x_2+0.003x_3-0.835x_4-1.590x_5-0.349x_6}}\ .$$

Import the data of event-related indicators into the above formula, and you can get whether they are cyber-violence. If the probability is greater than 0.5, it is recorded as cyber-violence. Otherwise, cyber-violence does not occur.

Table 4: The Omnibus's test of the model coefficients

	chi-square	Degrees of freedom	Conspicuousness
Steps	247.633	4	0
Blocks	247.633	4	0
The model	247.633	4	0

Omnibus's comprehensive test of the model coefficient is a global test of the model. According to Table 4, the Sig value of the test results is less than 0.05, indicating statistical significance, the obtained regression equation is reliable, and can be used to judge whether the event will develop into a network violence event.

5. Conclusion

This paper takes cyber violence as the research topic, and uses Python crawler technology to crawl microblog data. Then, the relationship between the influencing factors of cyber violence is analyzed. Finally, a binary logistic regression equation for identifying cyber violence is established. The research results of this paper will help to identify cyber violence incidents in time, stop cyber violence in time, and reduce the harm to the parties and the negative impact to the society.

Acknowledgment

The works described in this paper are partially supported by the **Undergraduate Innovation Foundation Project of Anyang Normal University** (No. 202210479040, 202210479046) and Henan Provincial National College Students' Innovation and Entrepreneurship Training Program Project (No. 202210479003, 202210479004).

References

[1]. Peterson J, Densley J. Cyber violence: What do we know and where do we go from here? *Aggression and violent behavior,* 2017, 34: 193-200.

[2]. Zhang J, Li Z. The Influence of Chinese University Students' Reinforcement Sensitivity on Cyber Violence. *Humanities and Social Sciences Letters*, 2021, 9(4): 341-350.

[3]. Nagle J. Twitter, cyber-violence, and the need for a critical social media literacy in teacher education: A review of the literature[J]. Teaching and Teacher Education, 2018, 76: 86-94.

[4]. Alotaibi N B, Mukred M. Factors affecting the cyber violence behavior among Saudi youth and its relation with the suiciding: A descriptive study on university students in Riyadh city of KSA. *Technology in Society,* 2022, 68: 101863.

[5]. Harrell F E. Binary logistic regression. Springer, Cham, 2015: 219-274.

[6]. Mehrolia S, Alagarsamy S, Solaikutty V M. Customers response to online food delivery services during COVID-19 outbreak using binary logistic regression. *International journal of consumer studies*, 2021, 45(3): 396-408.