Online Review Analysis Perception of Source Reliability on E-Cigarettes from the Perspective of Health Education among College Students in Chengdu, Sichuan

Li Wang¹,* Bingbing Qu², Lixin Huang³

¹Ya’an Polytechnic College, Ya’an, Sichuan, 625000, China
²Chengdu New Century Foreign Language School, Chengdu, Sichuan, 610000, China
³Chengdu Sport University, Chengdu, Sichuan, 610000, China
*Corresponding author

Keywords: Data mining; College student; Source Reliability; Health Education

Abstract: In order to understand the knowledge of Undergraduates in Chengdu about electronic cigarette, strengthen the health propaganda and education of electronic cigarette, prevent and control electronic cigarette, and build a smokeless campus environment, this paper aims to analyze the recognition of Undergraduates in Chengdu on the reliability of electronic cigarette sources based on data mining technology. Firstly, using random sampling method, a university in Chengdu was selected as the research object, and the attempted smoking rate, electronic smoke awareness rate and electronic smoke usage rate of college students were analyzed. The main factors that affect college students' awareness of electronic smoke were analyzed, and the reliability of electronic smoke sources was analyzed. Then, based on online review text mining method, the confidence of electronic smoke sources was analyzed. This paper builds a theme model to analyze college students' perception of the reliability of e-cigarette sources by using classification topics. Through experimental validation, the method proposed in this paper has higher data mining accuracy and efficiency, can effectively improve college students' awareness of electronic cigarette, strengthen the health promotion of electronic cigarette, and expect to provide support for subsequent related research.

1. Introduction

In recent years, the situation of tobacco control has become increasingly serious. Many countries take measures to prevent or control smoking in public places. As a new way to quit smoking, electronic smoking has been brought into the public's view and attracted the attention of adolescents. E-cigarettes have gradually become a way of addiction to nicotine among adolescents, and the combination of e-cigarettes with traditional cigarettes has frequently occurred among young people. Studies have shown that nicotine is the main component of electronic cigarette smoke, and nicotine can have a serious impact on infants and adolescents with incomplete brain development and even young adults, resulting in some adverse outcomes such as learning disability. In addition, electronic cigarettes can also produce many toxic substances during heating, which can irritate human eyes,
upper respiratory tract, lung and cardiovascular diseases, and even cause some effects such as carcinogenesis and teratogenesis. According to the related adult tobacco survey report, the public awareness rate of e-tobacco in our country shows a significant upward trend. Compared with previous years, the awareness rate of e-tobacco has significantly increased. The proportion of people who have used or are using e-tobacco has also increased. Young people are the main population currently using e-tobacco. Currently, there are limited studies on the reliability of e-cigarette sources among Chinese College students, but the recognition of the source reliability of e-cigarette by college students, a young group, has certain impact on the future smoking behavior of College students. Therefore, this paper aims to analyze the recognition of college students in Chengdu on the reliability of electronic cigarette sources.

The innovations of this paper are as follows: (1) Using random sampling method, a university in Chengdu was selected as the research object, and the attempted smoking rate, electronic smoke awareness rate and electronic smoke usage rate of college students were analyzed, and the main factors affecting college students' awareness of electronic smoke were analyzed, and the reliability of electronic smoke sources was analyzed. Then, using text mining method based on online comments, this paper analyzes the college students' perception of the reliability of e-cigarette sources in Chengdu, and builds a theme model to analyze the college students' perception of the reliability of e-cigarette sources by using classification topics. (2) Compared with other Undergraduates' cognitive analysis methods on the reliability of e-cigarette sources, the methods mentioned in this paper have higher data mining accuracy and efficiency, can effectively improve Undergraduates' awareness of e-cigarette, strengthen e-cigarette health promotion, and build a smoke-free campus environment.

2. Related Work

In recent years, with the implementation of measures such as banning smoking in public places in China, relevant scholars have made extensive research on College Students' perception of the source reliability of e-smoking and obtained corresponding results. Ma M L explores the factors influencing college students' perception of the source reliability of electronic cigarettes. Select university students from 2015 to 2018 as the research object, and select one class for each grade as the research object. A questionnaire was used to investigate the reliability of e-cigarette sources. Comparing the recognition of source reliability of different characteristics of e-cigarette population, making a comprehensive analysis of the influencing factors, the number of college students trying to use e-cigarette is 100, and the number of college students interested in e-cigarette is 219. Among them, the number of male students is 177, the number of female students is 42, and the number of male students interested in e-cigarette is more than that of female students. The difference is statistically significant. Gender, family integrity and smoking status of family members are the factors that affect college students' recognition of the reliability of e-cigarette sources. Smoking recognition intervention can be carried out according to these indicators for college students' confidence of e-cigarette sources, but this method does not prevent and control e-cigarette [1]. Zhan F et al. To understand the situation and source reliability of e-cigarette use by college students in a city, formulate targeted e-cigarette intervention strategies to provide scientific basis. Using a probability sampling method in proportion to the population size, this study selected students from a university in a city as the research object, and conducted a survey using the Chinese side of the Global College Student Tobacco Survey. The contents of the questionnaire include e-tobacco use, e-tobacco advertising and e-tobacco source confidence perception. Statistical analysis was carried out using SAS9.3 and SPSS19.0 software, the weighting rate was calculated, and the test and multivariate logistic regression analysis were carried out in the modules of complex sampling. The
situation of e-tobacco use among college students in a city is quite serious, which should be paid attention to by the society, schools and relevant government departments, and related health education on the harm of e-tobacco should be strengthened. However, the process of this method is more complex [2]. Qin R et al. has understood and mastered the current situation of e-cigarette use and source reliability of college students in a city, in order to provide an important basis for promoting the construction of smoke-free environment in Colleges and universities. Using scale-to-size proportional probability sampling method, we randomly sampled the samples in various colleges and universities in a city until the sample size of the study data was met. The tool included 18,800 college students who knew electronic tobacco. The main content of the monitoring was the reliability of the source of electronic tobacco and the status of electronic tobacco use. Among college students who are aware of electronic cigarettes, female students are less likely to use electronic cigarettes than male students. After entering university, the number of students who use electronic cigarettes also shows an increasing trend. Nonsmoking by parents was negatively related to college students' distance from e-cigarettes, and college students who had not tried cigarettes were less likely to use e-cigarettes than those who had already tried cigarettes. The trend of electronic cigarette prevalence among college students should be checked, blanks and errors in college students' recognition of the reliability of electronic cigarette sources should be corrected, and innovative ways should be adopted to scientifically guide college students away from electronic cigarette, but this method has not put forward specific prevention and control measures [3]. Shi H W et al. It has understood the recognition and influencing factors of the reliability of electronic tobacco sources among college students in a city, and provided an effective reference for universities to carry out electronic tobacco control work and related medical education work. A random sample and questionnaire point difference method were used to investigate 1608 students in a university in a city. A total of 1608 questionnaires were issued and 1538 were effectively collected, of which 838 were boys and 700 were girls. Research and analysis show that the professional nature of college students in a city and a university is an important factor affecting college students' confidence in the source of electronic smoke. E-tobacco has a relatively high recognition rate among university students in a city, and there are a lot of hidden dangers and risks. Universities, families and society should work together to do a good job in publicity and education of public health knowledge, correctly guide students to resist the temptation of the outside world and create a smoke-free environment, but this method does not put forward specific measures to prevent and control [4].

3. Cognitive Analysis on the Credibility of E-Cigarette Source among College Students

3.1 Research Objects and Methods

(1)Research Objects
There are 1391 college students in an ordinary university in Chengdu.

(2)Sampling method
A University was randomly selected from all general colleges and universities in Chengdu, and divided into medical and non-medical majors according to their specialty hierarchy. According to the required sample size, college students were randomly selected in each grade as the object of this study. A questionnaire was used to conduct a questionnaire on College Students' confidence in the source of electronic tobacco.

(3)Investigation methods
The specific content of the questionnaire includes demographic characteristics, current smoking status and confidence perception of e-cigarette sources [5-6].

(4)Decision Method
The awareness rate of electronic smoke refers specifically to the percentage of people who have
heard of electronic smoke in the survey population, and the usage rate of electronic smoke refers specifically to the percentage of users of electronic smoke in the survey population.

(5) Statistical analysis

Import the questionnaire into Excel software, check and process the data in the questionnaire, and use SPSS22.0 software package to carry out data analysis [7-8]. The measurement data are expressed by standard deviation \((\bar{x} \pm s)\). The counting data are described by frequency and percentage. The sample rate is tested by \(\chi^2\). The grade data is tested by trend \(\chi^2\). The influence factors of college students using electronic cigarette are analyzed by classified multivariate logistic regression. The test level \(\alpha = 0.05\).

3.2 Data analysis

(1) Basic situation. A total of 1391 college students were investigated in this study. The valid questionnaires were 1389, and the valid rate was 99%. Among them, 542 were boys, 847 were girls, 705 were medical professions, 684 were other professions, 1125 were living below 1500 yuan per month, 264 were living above 1500 yuan per month. Among the college students surveyed, 444 attempted to smoke and 147 smoked.

(2) Analysis on the Credibility of E-Smoke Source and the Use Factors of College Students in Chengdu

Will you know if electronic cigarettes are used as a strain, where electronic cigarettes have been heard as 1, not as 0, used as 1, and not used as 0\([9-10]\]. Based on the results of single-factor analysis and independent variables of professional knowledge selection, factors such as gender, grade, professional and monthly living expenses were selected as the independent variables to be assigned. Multivariate logistic regression analysis was carried out according to Table 1.

<table>
<thead>
<tr>
<th>variable</th>
<th>Variable assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Whether friends or relatives have used electronic cigarettes female = 0; Male = 1</td>
</tr>
<tr>
<td>grade</td>
<td>Freshman = 1; Sophomore and above = 2</td>
</tr>
<tr>
<td>major</td>
<td>Other = 0; Medical = 1</td>
</tr>
<tr>
<td>Monthly living expenses</td>
<td>(\leq 1500) yuan = 1; &gt; 1500 yuan = 2</td>
</tr>
<tr>
<td>Do you try to smoke</td>
<td>No = 0; Yes = 1</td>
</tr>
<tr>
<td>Smoking</td>
<td>Never smoking = 0; Smoking = 1</td>
</tr>
<tr>
<td>Have friends or relatives ever used electronic cigarettes</td>
<td>No = 0; Yes = 1</td>
</tr>
</tbody>
</table>

Table 1 shows that whether or not to try to smoke and grade are the main factors for college students' perception of the reliability of e-cigarette sources. That is, the probability of trying to smoke is greater than that of non-smokers and first-grade College students, which is indicated by Table 2. Smokers and friends or relatives who have used electronic cigarettes are the main risk factors for college students to use electronic cigarettes \((P < 0.05)\), that is, college students who use electronic cigarettes for smokers and friends or relatives have a higher risk of using electronic cigarettes, which is indicated by Table 3\([11-12]\).
The results of multivariate analysis show that whether or not students attempt to smoke and grade are important factors that affect college students' perception of the reliability of e-cigarette sources. College students are in an important period of development from adolescence to adulthood, facing a variety of psychological changes, which increase the risk of college students trying tobacco and alcohol. College students in this period have more opportunities to contact electronic cigarettes.

### 3.3 Source Confidence Analysis

Conformance measure function and sequence randomness test $P$-value function were used to analyze the reliability of electronic cigarette source. When a certain kind of label $y_{w_e}$ is attached to the test sample $x_{n+1}$ of the electronic smoke source, the sequence $Z^{(n+1)} = \{z_1, z_2, \ldots, z_n, z_{n+1}\}$ of the test sample can be formed together with the training sample. The random $P$-value of $z_{n+1}$ in the test sequence of $Z^{(n+1)}$ is calculated, and the $P$-value can be calculated as:

$$p = \left[ i = 1, \ldots, n + 1 : a_i \geq a_{n+1} \right] / n + 1$$

(1)

Formula (1) in which $i$ represents the number of measures greater than $a_{n+1}$ in the non-uniform measure sequences $a_1, a_2, \cdots, a_n, a_{n+1}$, and the non-uniform measure $a_i$ of each e-smoke source sample can be obtained by using the measure function to construct the non-uniform measure of the source sample, as follows:
In formula (2), \( \sum_{j=1}^{K} d_{ij}^+ \) represents the sum of \( K \) nearest neighbor distances of the same category as \( x_i \) samples, and \( \sum_{j=1}^{K} d_{ij}^- \) represents the sum of \( K \) nearest neighbor distances of different categories with \( x_i \) samples \^[13-14\]. This p-value describes the degree to which the sequence of the current test sample conforms to the independent distribution similarity. The higher the similarity, the greater the possibility that the test sample belongs to the \( y_{w_i} \) category. The definition of source reliability is:

\[
C = p \times 100\% \tag{3}
\]

4. Cognitive Analysis on the Credibility of E-Cigarette Sources among College Students Based on Online Comment Text Mining

4.1 Feature Extraction of Electronic Cigarette Text Data

Combining the above analysis results, this paper uses text mining method based on online comments to analyze college students' perception of the reliability of e-cigarette sources. Online reviews are a collection of document data made up of e-cigarette sources. Each document is made up of a number of terms and logical combinations. Based on the rating of e-cigarette terms in source reliability and the expression of specific topics, a text feature model is built for a large number of e-cigarette sources using vector space model, which transforms the text into data that quantifies the structure to expand feature mining \^[15-16\].

The basic principle of VSM is to comment on the fact that e-cigarette document \( D_i \) can represent a set of \( D_i = D(t_1, w_{i1}; t_2, w_{i2}; t_m, w_{im}) \), in which \( (t_1, t_2, \cdots, t_m) \) is a set of m-dimensional terms that are different from each other, \( (w_{i1}, w_{i2}, \cdots, w_{im}) \) the weight of the corresponding m-dimensional terms in the set, i.e. the degree of importance of the corresponding m-dimensional terms in e-cigarette source confidence document \( D_i \). \( w_{ij} \) is generally defined as the frequency function \( tf(d_{ij}) \) of the term \( t_j \) \( (j=1,2,3,\cdots,m) \) in the document \( D_i (i=1,2,3,\cdots,n) \).

TF-IDF is a weighting statistical method. Among them, TF is word frequency, refers to the number of times a given item appears in the electronic cigarette source file, IDF is the frequency of inverse documents, refers to the logarithm value of documents containing electronic cigarette word items after taking the inverse proportion in the total document, indicating the ability of electronic cigarette word items to distinguish documents \^[17-18\]. The higher the frequency of E-Cigarette Source Confidence term in the document, the fewer occurrences in other documents, indicating that the stronger the term distinguishes from the document, the greater the weight value. TF-IDF can be calculated as:

\[
a_i = \frac{\sum_{j=1}^{K} d_{ij}^+}{\sum_{j=1}^{K} d_{ij}^-}, i = 1,2,\cdots,n+1
\]
\[ TF - IDF = TF \times IDF = tf_i \times \log\left(\frac{N}{df_i} + 1\right) \]  

In formula (4), \( tf_i \) denotes the frequency of occurrence of term \( i \) in electronic cigarette from source, \( df_i \) denotes the number of documents with term \( i \), and \( N \) is the total number of documents.

4.2 Document Body Model

Document body model is an unsupervised machine learning statistical model for mining potential subjects in a set of documents. The probability theme model is mainly a word bag model, which considers that there is an intermediate topic \( Z \) between e-cigarette Related words \( W \) in the document. Documents are subject probability distribution, subject probability distribution or word probability distribution. Therefore, the vector space model of high-dimension documents and word items is mapped to low-dimension document and word space to explore the theme \([19-20]\) contained in the document of College Students' confidence in e-cigarette source. Document hierarchy relationships are shown in Figure 1.

![Figure 1: Document structure of subject model of e-cigarette source reliability](image)

Matrix conversion relationship of Undergraduates' perception documents on the reliability of electronic cigarette sources is shown in Figure 2.

In Figure 2, \( C \) represents the probability distribution of electronic cigarettes in the electronic cigarette source confidence recognition document, matrix \( \Phi \) Probability distribution of e-cigarette words under representative topics, matrix \( \theta \) represents the subject probability distribution in the electronic cigarette source confidence cognitive document, and the main purpose of analyzing the subject model is to derive the matrix by parsing document \( C \) \( \Phi \) and Matrix \( \theta \).
Combined with the above contents, this paper is based on the cognitive analysis of college students on the source credibility of e-cigarettes based on online comment text mining, which is specifically divided into text feature extraction and feature mining modeling. The process framework of the specific analysis is shown in Fig. 3. Thus, the cognitive analysis of college students on the source credibility of e-cigarettes based on online comment text mining is completed.

5. Experimental result

In order to verify the effectiveness of the cognitive analysis of Chengdu University Students on the source reliability of e-cigarettes based on data mining technology proposed in this paper, a simulation experiment was conducted. Table 4 shows the detailed parameters of the experiment.

<table>
<thead>
<tr>
<th>Experimental parameters</th>
<th>Specific configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>operating system</td>
<td>Win10</td>
</tr>
<tr>
<td>CPU</td>
<td>Intel Pentium g4600</td>
</tr>
<tr>
<td>Memory</td>
<td>8GB</td>
</tr>
<tr>
<td>Display card</td>
<td>Yingtong rx560d 4GB</td>
</tr>
<tr>
<td>simulation environment</td>
<td>Matlab 2019</td>
</tr>
</tbody>
</table>

Fig. 4 and Fig. 5 show the accuracy comparison of data mining for analyzing college students'
cognition of e-cigarette source reliability by using the method proposed in this paper and traditional methods for analyzing college students' cognition of e-cigarette source reliability.

Figure 4: Data mining accuracy of College Students' cognition of e-cigarette source reliability of the method proposed in this paper

Through the analysis of Fig. 4 and Fig. 5, it can be seen that the accuracy of data mining for analyzing college students' cognition of e-cigarette source reliability by using traditional data mining methods fluctuates greatly. With the gradual increase of data samples, the accuracy of data mining for e-cigarette source reliability cognition varies from high to low, but the overall accuracy of data mining does not exceed 80%. Using the method proposed in this paper to analyze college students' cognition of e-cigarette source reliability, the accuracy of data mining has been relatively stable, and has not fluctuated greatly with the increase of data samples, and has been stable at more than 80%. This shows that the method proposed in this paper can effectively analyze the cognition of college students in Chengdu on the source credibility of e-cigarettes, and improve the cognition.
of college students on e-cigarettes. Fig. 6 shows the time comparison between the methods proposed in this paper and the methods proposed in literature [1] and [2] to analyze the cognitive data mining of college students on the source reliability of e-cigarettes.

![Figure 6: Time comparison of data mining for analyzing college students' cognition of e-cigarette source reliability by different methods](image)

As can be seen from Figure 6, the highest time to mine the confidence recognition data of e-cigarette sources for undergraduates in Chengdu is 37s using the method mentioned in document [1], the highest time to mine the confidence recognition data of e-cigarette sources for undergraduates in Chengdu is 29s using the method mentioned in document [2], and the highest time to mine data using the method mentioned in this paper is 15s. Compared with the other two methods, the time of data mining is twice smaller. This shows that the data mining capabilities of the methods mentioned in this paper are relatively strong, which can effectively improve the efficiency of mining the confidence of college students in Chengdu on the source of electronic cigarette. Figure 7 shows the comparison of the extraction accuracy of undergraduate's cognitive data on the reliability of e-cigarette sources using the methods proposed in this paper with those in literature [3] and literature [4].

Analysis of Figure 7 shows that, through five experiments, when extracting the characteristics of undergraduate students' confidence in the source of electronic cigarette, the accuracy of extracting different features of data volume has corresponding differences. It can be seen that the methods in document [3] and document [4] still have deficiencies in extracting data features, resulting in lower accuracy of feature extraction results. However, the method proposed in this paper is not affected by the amount of data when extracting the characteristics of Undergraduates' cognitive data on the reliability of e-cigarette sources, and the accuracy of extracting data characteristics is relatively stable. Therefore, the data extraction capability of the methods presented in this paper is relatively strong.
6. Conclusions

In recent years, with tobacco control being incorporated into the health China action by the state, the smoke-free city policies of many cities have been gradually implemented, and the results of tobacco control have been slightly effective. In the group of College students, attention should be paid not only to the popularity of some new tobacco products such as traditional tobacco and electronic tobacco, but also to the college students with different characteristics, to carry out targeted tobacco control work, to explore tobacco control measures that meet the characteristics of College students, and to improve the social tobacco control laws and regulations as well as the construction of a multi-channel social tobacco control environment. In order to reduce the tobacco use rate of Undergraduates in Chengdu and even the whole country. Therefore, this paper aims to analyze the college students' recognition of the reliability of e-cigarette sources in Chengdu, and proves through experiments that the methods proposed in this paper can effectively analyze the college students' recognition of the reliability of e-cigarette sources, so as to strengthen the related safety education of e-cigarette.

References