Research on Public Opinion Analysis Methods Based on Knowledge Mapping and Deep Learning

Jinqiang Ma[1,a], Zheng Li[1,b], Lujie Li[2,c]

[a]wjxymjq@126.com, [b]lz_nefelibata@163.com, [c]411947275@qq.com

[1]China People's Police University, Langfang, Hebei, China
[2]Xiangning County Public Security Bureau of Linfen, Shanxi, China

Abstract. In order to effectively respond to public opinion events, it is necessary to integrate various information sources and construct a comprehensive and accurate knowledge base. Knowledge graph, as an emerging data structure, can be a good solution to this problem. This paper proposes a method that combines the advantages of knowledge graph and deep learning technology to improve the accuracy and professionalism of public opinion analysis. In the study, police-related public opinion data from different datasets are fused and de-duplicated through data preprocessing, and core words are extracted for theme mining and sentiment analysis. The optimisation algorithm model is incorporated into knowledge graph entities, and machine translation and entity recognition techniques are used to improve the reliability and pertinence of the algorithm. Experiments show that using the knowledge graph as the input and constraint of deep learning can improve the efficiency and accuracy of the algorithm. The research results have certain guiding significance for improving the professionalism and accuracy of police-related public opinion analysis.

Keywords: police-related public opinion; LDA model; stochastic gradient descent; relational mapping; sentiment analysis

1 Introduction

With the popularity of the Internet, people's access to information has become more and more extensive, and social media has become an important channel for people to learn about social hotspots and pay attention to events[1]. However, the development of the Internet has also caused some negative impacts[2] ,such as fraud, rumours and cyber violence. The emergence of these public opinions poses a threat to social stability and social stability[3][4]. Therefore, it is important to accurately identify and analyse public opinion, so as to detect and dispose of undesirable situations in a timely manner, thereby maintaining social stability and harmony.

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345358

2 Main Research Framework

This chapter proposes a framework for researching online public opinion on social public events by pooling deep learning text sentiment analysis and LDA text topic mining (as shown in Fig. 1).

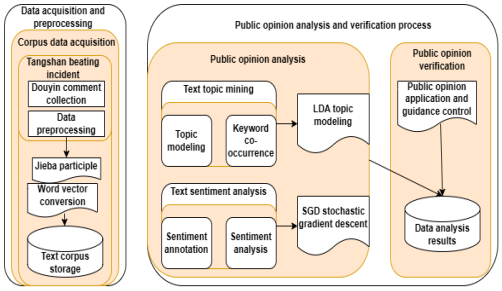

Fig. 1. Research framework

Taking police-related online public opinion as an example, the study takes the video comments on the topic of "Tangshan beating incident" on Douyin platform as the data base. First, the collected source data are preprocessed, then the LDA topic model is used for topic mining and intensity evolution analysis based on different periods of public opinion development, and then the SGD model is used to classify the unlabelled sentiment data into sentiment tendencies, so as to obtain all the data sets of both negative and positive sentiments. Finally, the specific focus of netizens at each stage is revealed.

3 Acquisition And Processing Of Public Opinion Data

3.1 Data Acquisition

In this paper, we crawl the data related to Douyin, firstly, we get the user's homepage or video chain that needs to be crawled, then we use python crawler code to send a request to the link and get the response data; we parse the response and extract the needed data, such as the user, video, and comment information, and so on.

Based on the collection requirements, a data acquisition field form was designed as shown in Table 1.

Table 1. Data field design

Field name	Meaning of field
Name	User ID
Time	Times
Like	Number of likes(on a website)
Content	Comments
title	News headline

When collecting data, two keywords are used for searching and the collection efficiency and accuracy are improved by changing IP addresses. After collecting one piece of information each time, the programme would hibernate for 3 seconds before the next collection to avoid being blocked by the website due to frequent visits. Eventually, the obtained police-related public opinion data are stored in CSV tables. The overall public opinion trend obtained by the crawler is shown in Fig 2:

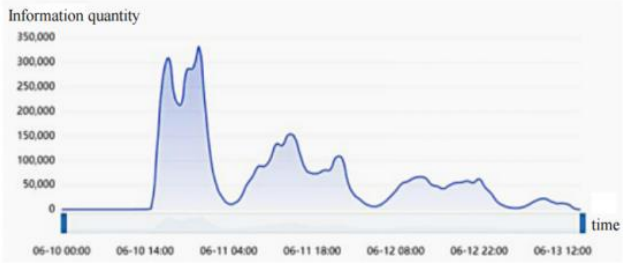

Fig. 2. Overall Public Opinion Trend Chart

According to the above chart, during the monitoring period, the incident reached its peak in the evening of 10 June, the public opinion fluctuated with the official response of Tangshan City, and began to go down after the arrest of the nine people involved in the case. By 15:00 on the 13th, there were a total of 5404,173 pieces of related information on the whole network (excluding follow-up comments and repetitions), with a total cumulative readership of 30 billion+ and over 10 million comments.

3.2 Date Cleansing

Then the crawled text to clean the relevant data. First of all, use the drop_duplicates() function in the pandas library to perform the de-duplication operation on the dataset, you can specify the removal of duplicate columns, or you can choose to retain the first or last duplicate record; by default, the drop_duplicates() function will determine whether all the columns are the same value, and if they are the same, then it is considered to be a duplicate record; after the de-duplication you can use the shape attribute to get the number of rows and columns in the dataset to confirm the effect of de-duplication. Delete one of the rows through the drop_duplicates() method (the first row is kept by default) to achieve the effect of de-duplication.

3.3 Jieba Chinese Segmentatin

The next jieba Chinese participle, the participle mode has full mode, precise mode and search engine mode three, of which the search engine mode for the keyword analysis of the highest degree of accuracy.

segement_list = jieba.cut_for_search(sentence)print(" Search Engine Segmentation Results: ", "/".join(segement_list))

Search Engine Subtitle Results: involvement/participants/ should/be/legally/responsible//otherwise/un/fair/

In order to reduce the dimensionality of the text and make the processing more efficient and accurate, deactivation operations are usually required. The list of stop words usually contains some common meaningless words, such as "the", "has", "is", etc., which can be selected and expanded according to the actual needs and domain knowledge. Removing stop words can save storage space, improve text processing efficiency and accuracy, and at the same time reduce the interference and noise of stop words for understanding the text. Therefore, de-duplication is a common operation step in text processing and analysis.

The general process of de-stopping words is to filter and screen the result of word splitting according to the pre-generated stop word list. After creating an empty string, we read the stopwords list line by line and store it in the list stopwords.txt, filter the elements in the original text according to each "word" unit, if the elements in the original text are not in the stopwords list, we do not stop using them and save the non-stopping words, and finally save the results of the corpus preprocessing as a file.

3.4 Keyword Extraction

In this paper, based on jieba participle, a statistical method is used to refine the keywords of police-related public opinion in order to collect data more efficiently. In the middle of this take the public opinion type as a topic with difference, each key pair of words should be one or more topics, and under each topic there is a keyword that matches it. As shown in Table 2.

Table 2. Examples Of Keywords

Category of public opinion	Examples of related topics	Examples of keywords
Public opinion related to the police	Political factors	Alleys, barbecue restaurants, fair,…
	Social Impact Class	Remote, indifferent, severely punished,…

After keyword collection of news articles related to police-related public opinion. Then the top 500 keywords were extracted from these articles. The data is tagged manually to filter out compliant police-related public opinion keywords and add them to the preliminary keyword database. This process can be repeated.

WordClouds library in Python is a word cloud generator function that can be used to generate word clouds for various topics to quickly understand the main keywords of the text. The key dictionaries in this paper are imported to generate specific word cloud images and saved in the "word cloud.png" file.

The size of the font is displayed according to the frequency of the words, i.e., the greater the frequency of the words, the larger the font, and the focus of people's attention is not only on "severely punished", "really", and "protected", but also on In addition to "severely punish", "really" and "protect", people also pay more attention to "fear", "call the police", "clothes", "indifference" and "voice". voice" and other topics. It can be seen that the theme of Internet citizens for this public incident focuses on positive attitudes and the process of the incident, for the public security organs of the fairness of law enforcement to give a very high hope, and in the

process of the incident, the vast majority of people formed the women are vulnerable thinking, mainly divided into the negative feelings of girls should not go out at night and the positive feelings of the women should be protected, and for the abuser's physical characteristics of the critics.

4 LDA-based Model For Topic Analysis

Thematic analysis and modelling is carried out to make it easier to understand and make use of textual data[5]. In modern society, the amount of text data we are dealing with is getting bigger and bigger, from news reports, social media comments to scientific papers, business reports, etc., all of them contain a lot of information and knowledge, which are scattered in the huge amount of text data[6]. Text data can be better understood and utilised by processing and analysing it, and extracting from it the topics, features and relationships hidden within the data.The LDA model is a powerful text analysis tool with many advantages, including the discovery of potential topics, improved performance in text categorisation, scalability, applicability to multi-language processing, and provision of interpretability. It can automatically discover potential topics in text data without manual labelling and can improve the performance of text classification. In addition, the topics and word distributions extracted by LDA can provide interpretability to help understand the meaning and inner structure of text data.

4.1 Confusion Analysis

LDA topic models can be judged by their perplexity, with smaller values proving that the model is better. The perplexity reflects the accuracy of the model's prediction of textual data that does not appear in the training data, i.e., the degree of uncertainty of the model in identifying which topics are contained in the document. Thus the lower the value, the higher the certainty and the better the final clustering result. By looking at how many different iterations correspond to the DEGREE OF CONFUSION, THE NUMBER OF ITERATIONS CORRESPONDING TO the smallest degree of confusion in the curve is the optimal number. The optimal number of topics is determined by plotting the Perplexity-Coherence-Topic line graph as shown in Fig 3.

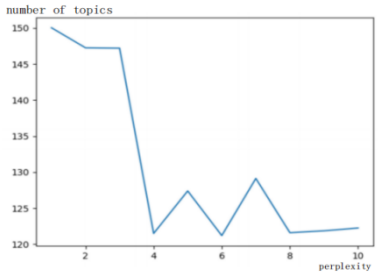

Fig. 3. Degree of confusion in LDA topic extraction

In this figure, the number of topics is taken as the horizontal coordinate, and the perplexity is taken as the vertical coordinate for the analysis, from the previous concepts, we can know that

the lower the perplexity is, the better, but through the line graph, it shows that it is not that the more topics are the better, and when there are too many topics, this model has been overfitted. And it is found that when the number of topics exceeds 3, the perplexity value of the model decreases rapidly and reaches the first minimum at 4. So, we consider the final number of themes in 4-5.

4.2 LDA Topic Moduelling

The model assumes that each word in the text is generated by combining potential topics, and that each topic in turn consists of a set of words. Using LDA, it is possible to represent the text as a distribution of event perspectives, while representing the individual topics as distributions of individual words. The specific modelling process is as follows:

Standardization -Load the file, split the content of each text, remove punctuation, and filter words shorter than 2 and in the deactivation list.

Word Form Reduction --Change words from third person to first person and verbs in past and future tenses to present tense. Reducing words to root of a word in the past and future tense forms, making dictionaries and generating Dictionaries to report words and the number of times these words occur, some of the hot words are shown in Table 3.

Table 3.. Hot word dictionary

Doc	Quantity	Doc	Quantity
Punish	311	Alarm	106
Schoolgirl	215	Help	103
Girl	184	Roadside	93
Clothes	152	Harass	87
Man	127	Speak	73

TF-IDF - Using the models.TfidfModel model, create a tf-idf model object of 'bow_corpus' and save it to tfidf, perform tfidf on the whole corpus transformer and name it 'tfidftransformer-2'.

Training the model - The LDA model training process is written as a function, using a multicore-based ldamulticore model trained and saved in "mymodel2", which calculates the occurrence of words under each topic and their relative weights.

Document Classification - Sample documents are classified using the LDA TF-IDF model, and the test document is accurately classified by the model as the one with the highest likelihood of the topic, indicating that the classification is accurate. If the classification error exists and the error is large, through continuous optimisation of the model, the appropriate tuning parameters, until the topic requirements are met, the topic classification is shown in Table 4.

Table 4. Theme naming

Social impact	Characteristics of the abuser	Sequence of events	Public reaction
law enforcement	Short sleeves	Partner	reconciliation
criminal	fatso	Steps	Abominable

Public security management	Black clothes	Drag it out	Underworld forces
Picking quarrels and provoking trouble	Coat	Spank	Security
Public order Detention	remoteness Drunk	harassment Bystanders	zero tolerance justice
Fixed-term imprisonment	villain	counterattack	Punish the murderers severely
Crime	shorts	corners	Bottom lines
injuries	Barbecue restaurant	Great gentlemen	Shield
Sentenced	Greem clothes	Bench	Heavy punishment

Topic strength indicates the degree to which a text topic receives attention, using entropy to measure the degree to which the distribution of topics is concentrated, and then the entropy to calculate the document weights, a defined topic will have a high probability distribution across one or a few topics, and a low probability distribution or no distribution across other topics. Conversely, if 1 document has a more even distribution across topics, the document does not have a clear bias towards the topic. In this section, they are given a weight of 1. Depending on the distribution of the document's topics, the document will have a higher score for the exact topic tendency; and a document with overly generalised content will have a lower score. The topic strength is calculated as shown in Fig 4.

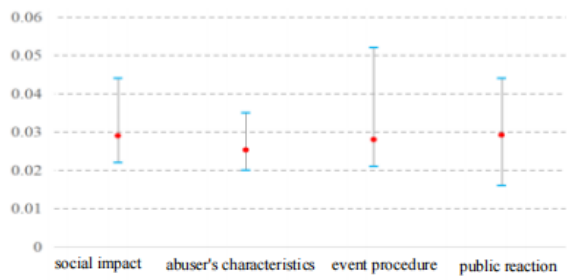

Fig. 4. Theme intensity map

From the above figure, it can be concluded that the lowest value of social impact is in the middle of 0.02 to 0.03, the highest value is near 0.045, and the average value is less than 0.03; similarly, the weight of the abuser's characteristics is located at around 0.025, and the course of the incident is close to 0.03, but its highest level reaches above 0.05. Overall, it seems that the highest average intensity of public reaction is located at 0.03, indicating that the public reaction has the highest degree of relevant opinion tendency, reflecting the public's zero tolerance for such evil forces as the Tangshan barbecue shop beating incident and their high desire for a sense of security.

4.3 Relationship Mapping

Topic word co-occurrence analysis can assess the correlation between topic words by counting the frequency of their occurrence in the same text, thus highlighting potential relationships and themes in the text. And the analysis can improve text comprehension, discover potential relationships, assist data mining and facilitate text classification by identifying the correlation between words or phrases that appear more frequently in the text.

Gephi Support for Interactive Analysis is a common graph drawing tool that allows interactive analysis and exploration through nodes and edges on the graph. Relationship graphs, like knowledge graphs, require the collection of data on nodes and edges before they can be mapped, and these data will form the basis of the relationship graph. Each node will have weights, and every two nodes will be connected by an edge, and this connecting line will also have weights. Some edges will have low weight values and can be treated as more minor relationships. These secondary relationships can lead to a lot of redundant information in the network, so measures need to be taken to reduce this redundancy.

Below are the general theme co-occurrence Gephi analysis steps:

Step 1: Prepare the data

First of all, you need to prepare the text data to be analysed, and carry out text preprocessing and topic model analysis to extract the keywords, topics and the correlation relationship between them. Usually, we can use the text analysis tool in Python to carry out preprocessing and topic model analysis, use Python's re module to remove punctuation marks in strings, and then count the co-occurring relationships and frequency of topics in each sentence and save them as text.

Step 2: Import data into Gephi to create a network diagram

Import the processed data into Gephi. Create a mesh graph in Gephi, where nodes represent keywords and edges express co-occurrence relationships among them. Nodes can be sized and coloured according to their degree (i.e., the number of edges adjacent to them) to reflect the importance of the node.

Step 3: Perform analysis and visualisation

Analysis and visualisation is performed using the analytical tools and plugins available in Gephi, such as degree centrality analysis, community discovery, node labelling, etc. Parameters such as node size, colour, labels, etc., and edge colour, width, etc., can be adjusted as needed to better present the data and analysis results.

The figure shows that the four theme words with the largest circle prove the highest frequency, and there are corresponding links between the nodes scattered with the theme words as the centre. The thickness of the connecting lines represents the frequency of the connection, and the probability of "drunk" and "remote" appearing together is much higher than the others in the characteristics of abusers, while the sub-themes in the characteristics of abusers are more closely connected to the sub-themes of the event process, which can be seen when the public is aware of the fact that It can be seen that when the public is concerned about a character with relevant characteristics, it will be associated with violence and harassment of women, which will increase the fear of and concern for such men. Secondly, it is hoped that the public security authorities will investigate the incident as soon as possible and announce the details of how the incident was

handled. It is believed that the victims' failure to speak out is a concealment of the society and a harbouring of the evil forces, and that only when the persons involved in the case are brought to justice as soon as possible and sanctioned by the law will the public be able to give them an explanation.

There are reasonable and unreasonable demands, both for women and for law and order enforcement, for example, "Women should not be out alone at night for barbecues... Why don't the police go out immediately when they receive a call is it to harbour evil forces?" are negative sentiments; "Street patrols should be strengthened to improve our sense of security" is a positive sentiment. For this reason, by highlighting the sentiment of each theme through sentiment annotation and analysis, and further analysing and judging the sentiment tendency of thematic network public opinion, we can study the attitude of each party towards each theme in time, and find out the unfavourable signs that can enable the department to formulate corresponding plans in time and scientifically, and take the initiative in guiding the development of network public opinion.

5 Analysing the emotional tendency of the text

Public opinion emotional tendency analysis is essentially an analysis of emotional tendency of Web texts[7]. It assesses the public's attitude and emotional inclination towards an event by analysing the text content of Web comments, news reports, etc., and judging the emotional inclination therein. If the text tendency is divided into two types of problems, positive sentiment and negative sentiment, deep learning methods can be used for sentiment recognition and prediction[8][9]. This method can effectively improve the accuracy and efficiency of sentiment analysis and help people better understand the public's views and emotional tendencies towards an event.

5.1 Experimental Methodology

Sentiment subject-based tendency analysis of thematic online public opinion is to judge the positive and negative attitudes of public opinion on a certain topic by analysing the sentimental tendency of comments, news reports and other texts about the topic on the Internet. Therefore, this section adopts the accuracy rate P (Precision), recall rate R (Recall), and F-value (F-Measure) as the test indexes, which are calculated as follows:

$$P = \frac{The\ correct\ number\ of\ documents}{Number\ of\ documents\ returned}$$

$$R = \frac{The\ correct\ number\ of\ documents}{Total\ number\ of\ documents}$$

$$\text{F-value} = \frac{2 \times P \times R}{\text{P} + \text{R}}$$

In order to show the practical effect of police-related public opinion tendency recognition in sentiment ontology, this section uses a set of a total of 2257 already labelled comments as a dataset and divides it into two parts: the training set and the test set. In this section, these comments are manually labelled with positive and negative tendencies and two experiments are conducted to verify the practical effectiveness of the algorithm.

Experiment 1: BP neural network algorithm is used. The input data is fed into the neural network, and the output of the neural network is finally obtained by calculating the obtained data of each neuron step by step, comparing with the real label value and calculating the error value. According to the error value, the error signal of each neuron is calculated step by step, the error signal is propagated back to the previous layer, and the weights and bias values of the current layer are iterated according to the signal and the output value of the current layer.

Experiment 2: Using the SGD gradient descent algorithm, a model "for training machine learning". For each training sample, the gradient is calculated for that sample and used to replace the parameter values with new ones. This process can be repeated many times until a certain number of iterations are reached or a certain precision of convergence is achieved.

5.2 Experimental Results And Analyses

According to the above experimental steps, the results are shown in the Table 5.

Table 5. Sentiment analysis experimental results

SGD	Precision	Recall	F1-score	Support
0	0.82	0.74	0.78	373
1	0.77	0.84	0.80	389
Accuracy			0.79	762
Macro avg	0.79	0.79	0.79	762
Weighted avg	0.79	0.79	0.79	762
BP	Precision	Recall	F1-score	Support
0	0.77	0.77	0.77	366
1	0.78	0.78	0.78	396
Accuracy			0.77	762
Macro avg	0.77	0.77	0.77	762
Weighted avg	0.77	0.77	0.77	762

Experiment 1 accuracy and recall are only 77% and 77%, so the BP neural network based model fails to reach the experimental results in sentiment classification applications.

Experiment 2 accuracy and recall are substantially improved over Experiment 1, thus using the SGD algorithm for sentiment propensity analysis yields high accuracy and recall.

Emotions were labelled for each theme using the Origin tool, as shown below.

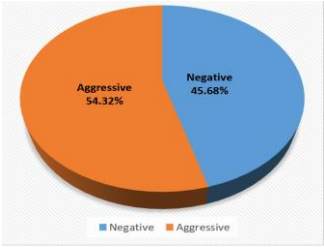

Fig 5. Sentiment labelling of public responses

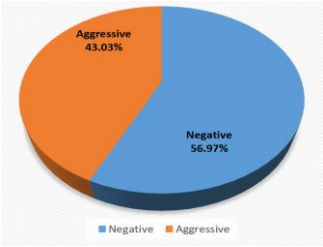

Fig 6. Sentiment labelling of social impacts

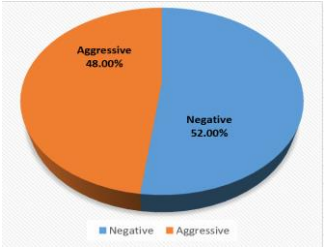

Fig 7. Affective labelling of abuser's characteristics

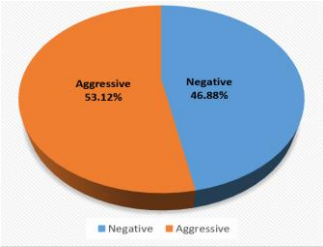

Fig 8. Affective labelling of the course of the incident

5.3 Thematic and Emotional Mapping

After further research, it can be found that the above core thematic terms coincide with those obtained from the sentiment analysis of the Tangshan beating incident. Among them, the covariance between the course of the incident and the abuser's characteristics is the strongest. And the proportion of negative emotions between public reaction and abuser characteristics is greater than 50%, indicating that there is a close connection between abuser characteristics and public reaction in terms of negative emotions. As women are less able to resist, and it is night time, women travelling alone do pose a certain danger, and being big and having tattoos can also give people a sense of fear. But calling for women to go out less at night and spurning men with tattoos, flowery clothes and big bellies, although it reduces the probability of this kind of thing happening to a certain extent, is not the root of the problem. Only by abandoning the "victim's guilt theory" and starting with the system, can we achieve the good development of social security.

6 Conclusions

With the development of society, public security organs are facing more and more complicated police-related public opinion events, and the dissemination of these events on online platforms has formed an extremely complicated public opinion environment. In order to cope with this situation, One Net Two Weibo, as a representative of the new media, has the feasibility to guide and control police-related public opinion by assessing the sentiment according to the connection between different topics in the knowledge graph of police-related public opinion. However, there is still room for improvement, because it does not take into account objective factors such as the

direction of attention and browsing structure of individuals or organisations that disseminate public opinion information.

Acknowledgments. Research on the security protection system of critical information infrastructure under the overall national security concept(ZDZX202004).

References

[1] ZHOU Qian, WANG Jie, YU Jianbin. A method for perceiving police-related public opinion information based on knowledge graph[J]. Computer Science, 2018, 45(4): 269-275.

[2] Camacho D,Panizo-Lledot N,Bello-Orgaz G,et al. The Four Dimensions of Social Network Analysis: An Overview of Research Methods, Applications, and Software Tools[J].Information Fusion, 2020, 63:88-120.

[3] CHEN Xiangyu, LI Tao, ZHANG Tianming, et al. Cross-media analysis of police-related public opinion based on knowledge graph[J]. Journal of China Public Security University: Social Science Edition, 2020, 38(5): 21-29.

[4] Barachi M E,Alkhatib M,Mathew S,et al. A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change[J].Journal of Cleaner Production, 2021(5):127820.

[5] Hosseini H, Bagheri E. Learning to rank implicit entities on Twitter [J]. Information Processing&Management, 2021.58(3)102503 .

[6] ZHONG Jinhong, LI Mengping, XUAN Zhanxiang. Analysis of public opinion on epidemic prevention and control in colleges and universities based on LDA model [J]. Information Technology and Informatisation, 2022(12):4.

[7] WANG Tianxiang. Research on the application of gradient descent algorithm in deep learning [D].Zhejiang Normal University,2019:10-30.

[8] MA Zhekun, TU Yan. Research on the content monitoring of emerging topics of network public opinion based on knowledge graph [J]. Intelligence Science, 2019(2):7.

[9] CHEN Yongfeng, KANG Jiahao. Research on the Public Security Organs' Strategies for Responding to Negative Online Public Opinion in Eliminating Organized Crime and Evil: A Case Study of the Tangshan Barbecue Restaurant Beating Incident [J]. Journal of Shanxi Police Academy, 2023, 31(2):7.