2021 2nd International Conference on Artificial Intelligence and Education (ICAIE) | 978-1-6654-2492-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIE53562.2021.00146

Health Education Based on Natural Language Processing(NLP) for Infectious Disease Outbreak

Tao Jiang
School of Humanities and Management
Guilin Medical University
Guilin, China
tj290@uowmail.edu.au

Chaozhi Xu
School of Humanities and Management
Guilin Medical University
Guilin, China
zhi_qwerty@hotmail.com

Dan Liang
School of Public Health
Guilin Medical University
Guilin, China
744235127@qq.com

Yingjue Wei*
School of Humanities and Management
Guilin Medical University
Guilin, China
* Corresponding author: weiyinjgue@qq.com

***Abstract*—The purpose of this study is to test and use Natural Language Processing (NLP) to analyze epidemic case reports to establish an effective health education system. A total of 100 cases were randomly selected from the epidemiological case report of Feb 1, 2021 to May 15, 2021 published on the Chinese public media website. The NLP techniques are used to help the assessment team identify and summarize relevant issues. Infectious disease prediction system based on a small number of epidemic reports, in the shortest possible time to help the assessment team to summarize the relevant problems, for experts to make a judgment to provide a basis. We found that NLP technology can play a certain role in the analysis of epidemiological reports, which is based on mature languages of existing language libraries, and can effectively improve the analysis efficiency of experts. This preliminary study confirmed that NLP technology can be used to analyze the text of epidemic case reports and help experts quickly establish a health education system.**

Keywords-Heath Educaiton; Infectious Disease Outbreak; Natural Language Processing(NLP).

I. INTRODUCTION

The COVID-19 has brought severe challenges to the world's public health system. Facing the outbreak of various infectious diseases in the future, it is necessary to develop a rapid infectious disease prediction system based on natural language process for epidemiologists to help them quickly grasp the situation of the outbreak of infectious diseases[1-3]. Current research on the use of NLP technology in the field of COVID-19 shows that as long as the appropriate language library is selected, NLP technology can help everyone predict the outbreak of COVID-19 [1, 4-6]. Recent studies have shown that a large number of epidemiological investigation reports will be produced in the early stages of infectious diseases, which are useful for predicting the situation and health education [3, 7]. The lack of an intelligent analysis system in the field of health education makes it difficult for us to correctly evaluate based on existing epidemic case reports, and to make correct health education based on the current situation, and to quickly analyze the situation shown in infectious disease reports based on limited information.

Artificial intelligence has been used in health education, but 'understanding' related Natural Language Processing (NLP) is still a challenge [8]. With the outbreak of COVID-19, more and more NLP related techniques have been used in the analysis of epidemiological case reports [7]. The existence of these techniques makes it easy for experts to conduct reports: topic classification, topic mining, trend forecasting, crisis assessment, and text mining [5, 8-10]. The existence of NLP and other text mining technologies makes it convenient for researchers to quickly discover hot topics from reports, make a pre-assessment of the current situation, and possibly obtain pre-knowledge of infectious disease prevention and control from relevant reports[1, 2, 9]. With the lowering of the threshold for the use of these technologies, there are still a large number of epidemiological reports available on the public network, which makes it convenient for us to select reports within a certain period and conduct in-depth analysis through the use of NLP.

The health education is mainly reflected in the conscious adoption of healthy behaviors and lifestyles, elimination or reduction of risk factors affecting health, and prevention of diseases [6, 9]. Healthy behaviors and lifestyles include adequate sleep, balanced nutrition, exercise, etc. The groups of people who have undergone health education who adopt healthy behavior patterns have advantages in terms of system, physical and mental health, psychological factors, and healthy habits. [6, 7, 9-11]. The current multiple diseases are mainly related to health behaviors and psychological factors. Correct health education can effectively reduce exposure to disease-related factors. [1, 9]. During the COVID-19 outbreak, we can prevent the occurrence of diseases through situational awareness and pre-knowledge awareness.

During an infectious disease outbreak, how to quickly obtain pre-knowledge and use it for pre-evaluation is a key factor in epidemic prevention and control [7, 9]. The epidemiological case report is the key to correctly obtaining pre-knowledge and using this pre-knowledge to complete pre-assessment. As the pre-experimental part of future large-scale

978-1-6654-2492-9/21/$31.00 ©2021 IEEE
DOI 10.1109/ICAIE53562.2021.00146

research, this research aims to obtain the correct way of pre-evaluation and pre-knowledge by testing NLP technology for epidemiological case report analysis, and to establish a correct health education system based on the above understanding. The test aims to establish pre-assessment and pre-knowledge acquisition standards and research to establish consistent standards with the aid of NLP technology.

II. METHODS

A. Data source

On May 15, 2021, 100 epidemic case reports (released on Feb 1, 2021 and May 15, 2021) published on official public media websites are selected for data analysis. These data are then used to verify the reliability and practical effects of using NLP technology to obtain the necessary situational awareness and pre-knowledge of the health education system.

B. Piloting Natural Language Processing(NLP) method

The qualitative analysis method is used to identify the important situation and knowledge of the epidemic case report in text format. For better batch processing of pre-knowledge and situational awareness, we have developed a program based on Stanford NLP. 'Keyword search' and 'Manually load content' were also used in the analysis process, with the purpose of mapping the pre-knowledge and situational awareness in the 100 reports to the identified criteria. Figure 1 outlines the entire analysis process.

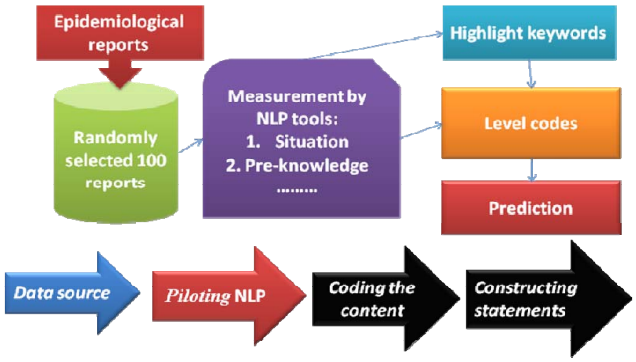

Figure 1. The four steps to analyse data from the case reports.

The principle of establishing measurement statements is to come from different regions, and we take a random sample of reports from different regions. In the end, we selected 5 reports as the establishment measurement statements.

The extraction of pre-knowledge is based on the analysis of each sentence, and NLP technology assists the entire analysis process. The purpose of this is to ensure that all pre-knowledge is fully and effectively extracted.

In order to confirm the validity of pre-knowledge, the four authors used a qualitative analysis method. They first determine the meaning of the sentence with the help of NLP. The sentence has multiple meanings will be divided into clauses. The meaning is fully extracted in advance, and it is also necessary to ensure that the sub-sentence is similar to the original structure.

For example, the infected person stated that the direct cause of the infection was that they did not take effective protection. We used NLP-assisted analysis to divide this sentence into three clauses: no situational awareness, inappropriate situational awareness, and no pre-knowledge of pollutant preferences awareness. At the same time retain the structure and wording of the clauses.

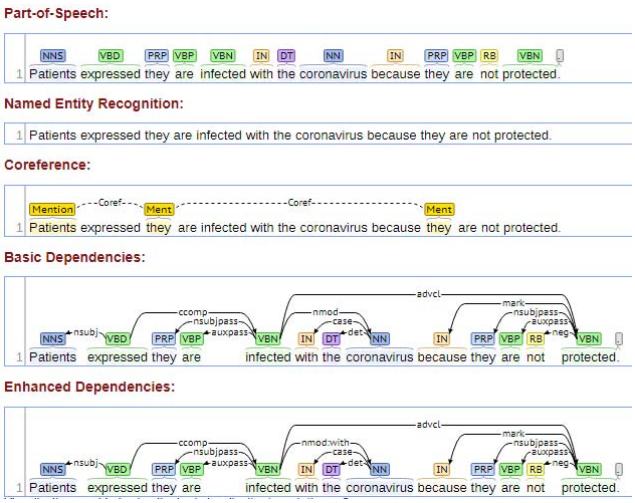

Figure 2. A program based on Stanford NLP are used to extract the pre-knowledge.

After that, all authors need to conduct a centralized review of the extracted content. The authors determine the measurement statement by examining the consistency of the pre-knowledge and the original sentence expression. For inconsistencies, it is necessary to vote to reach a consensus.

In order to ensure the reliability of the measurement statement, all authors also performed the same task on 5 identical case reports. We finally obtained 16 sub-sentence and recorded them in a spreadsheet, marking them as individual IDs.

In the end, the authors achieved a consensus rate of 93%. The main difference lies in how to mark pre-knowledge in accordance with clauses. Finally, the authors finally obtained a 100% consensus rate, and finally divided the 12 clauses into 36 sub-sentences covering pre-knowledge.

C. Coding the content

The sentence simplification was carried out in order to retain its main meaning through simplification. On the basis of the first-level code, some sentences are simplified to 'Patients indicate that their situational awareness is appropriate.'

After completing the summary analysis of the primary code, we began to compile the secondary code. The secondary code is divided into three main bodies: situation information, coping strategies, and useful knowledge. For example, the patient's discomfort is classified as a 'predictive' theme.

D. Constructing measurement statements

The sub-sentences were constructed into four pre-knowledge measurement statements under the three themes: seven items in situation information, eight items in coping strategies and 13 items in useful knowledge. The 28 items in process were related to 21 pre-knowledge in infectious disease outbreak.

100 reports are mapped to the measurement statements. We randomly selected 100 reports epidemiological reports for use in the mapping process. Figure 1 described the whole mapping process. We developed a keyword search list for the three major measurement statements. We also developed an NLP tool to search pre-knowledge, extract information and export information to a text document or excel spreadsheet.

For example, we identified all the sentences in Section Pre-knowledge in these 100 case reports that contained the pre-knowledge about 'diagnose'. Then we read the selected content manually to examine if the sentence could be semantically mapped to the statement we intended to map. If a statement is mapped, we recorded its presence as "1" for this audit report; otherwise, "0" is recorded.

After the mapping process, we analyzed the percentage of matching for each statement to achieve the aim of examining the consistency of use of measurements by the assessment team.

III. RESULT

A. Extracted pre-knowledge statements

With the assistance NLP technology, the 100 reports found that the current pre-knowledge is mainly distributed in four areas. We divided the three pre-divided themes (situation information, coping strategies, useful knowledge) into four areas: overall correct situational awareness, no or incorrect situational awareness, correct pre-knowledge, no knowledge or no correct pre-knowledge (Table1).

TABLE I. PRE-KNOWLEDGE DURING INFECTIOUS DISEASE.

No	Pre-knowledge during infectious disease.	
	Pre-knowledge awareness	*Percentage*
1	Correct situational awareness	7%
2	None or incorrect situational awareness	32%
3	Correct pre-knowledge awareness	3%
4	None or incorrect pre-knowledge awareness	24%

In all 100 reports, the first two knowledges that appeared were 'no or incorrect prior knowledge' (32 reports) and 'correct situational awareness' (7 reports). Pre-understanding of situational awareness is also mentioned in those reports. They are 'pre-knowledge awareness in employee practice', used to manage pre-knowledge (3 reports) and 'unmonitored situation awareness' (24 reports). Pre-knowledge and situational awareness seem to be key factors in health education.

B. Discovering pre-knowledge

Figure 2 shows the results of knowledge mining and data mapping. Through the analysis of pre-knowledge, we find that NLP technology-assisted analysis can quickly let us master the pre-knowledge and situational awareness necessary for health education. Almost all epidemiological reports show that patients do not have a clear understanding of the situation and lack the necessary pre-knowledge. These problems are the reasons why our health education needs to be strengthened. Of course, effective coping strategies, a clear understanding of the situation, and a lot of useful knowledge have become the core of health education.

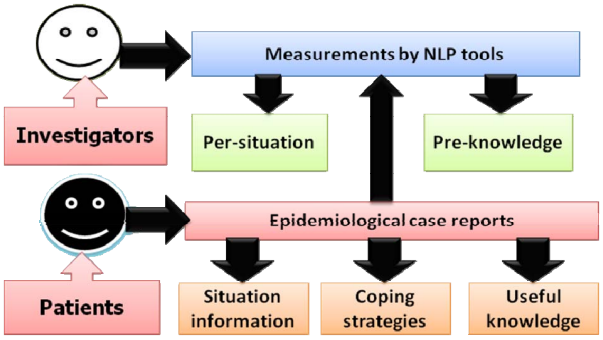

Figure 3. Situational awareness and pre-knowledge awareness for infectious disease.

IV. DISCUSSION

With the assistance of NLP, we have effectively improved our ability to analyze pre-knowledge and situational awareness in epidemiological reports. In the future, the research method optimization is needed to improve the ability to measure the mapping semantics of statements. The shortcomings of our pilot research method are the difficult to achieve automated rapid analysis.

Although NLP technology realizes the function of partial sentence, it still needs manual to realize perfect mapping for some complicated semantics. Our preliminary research found that the COVID-19 case report has a unique structure and laws, and it has become an important channel for us to obtain pre-knowledge. These pre-knowledge and eventually change into effective knowledge are also hot topics in the field of health education.

Of course, it is not enough to use NLP technology to assist mining based on the knowledge in 100 epidemic case reports. It can only verify the effectiveness of the research method. Fortunately, the number of epidemic reports is also limited in the early stage of an infectious disease.

The mapping results also show that not all pre-knowledge has been mentioned, but the report will reflect the current situation. The pre-knowledge measurement is limited to two statements that measure awareness of infectious situations. The advantage of the current health education system is that NLP technology can be used to quickly distinguish pre-knowledge and make scientific judgments on the situation. The useful knowledge in the pre-knowledge may play a certain role in the

prevention and control of transmission of infectious diseases, and provide necessary protection in the absence of sufficient scientific basis.

The limitation of the research at this stage is the lack of a large amount of data support and the review of the research results by relevant experts. At the same time, although NLP technology automatically solves the phrasing problem, it cannot automatically summarize complex concepts independently. Future research will study other methods based on artificial intelligence.

V. CONCLUSION

This prospective study tested the use of NLP technology to assist in the establishment of the pre-knowledge base and situational awareness necessary for health education. By testing 100 epidemiological reports published in Chinese public media from January 1 to May 15, 2021, we have verified the reliability and effectiveness of the system. We found that the extraction of pre-knowledge and situation information covered in the COVID-19 summary report can effectively provide useful knowledge, response methods, and situation judgments for the establishment of a health education system.

We recommend that epidemiological case reports provide objective pre-knowledge as much as possible, and through scientific verification, it becomes important available knowledge in health education. The advantage of NLP technology is that it does not miss any pre-knowledge that may be ignored by humans. Of course, the epidemiological case report is structured and stated in accordance with the existing measurement statements.

The intelligence level of analysis methods needs to be improved, and other artificial intelligence-based analysis technologies will be considered in the future to improve the analysis capabilities of NLP technology. Further research will consider extraction of personal heath measurement results.

ACKNOWLEDGMENT

Fund Project: Guangxi Bagui Scholars; The Risk Management System for Aged care Services in Guilin(2021KY0501); Project for Middle-aged and Young Teachers in the Universities in Guangxi Province (2021KY0519); Scientific research ability improvement project of young and middle aged staff (2018glmcy014).

REFERENCES

[1]M. Roche, "COVID-19 and Media datasets: Period- and location-specific textual data mining," (in eng), Data Brief, vol. 33, p. 106356, Dec 2020.

[2]P. López-Úbeda, M. C. Díaz-Galiano, T. Martín-Noguerol, A. Luna, L. A. Ureña-López, and M. T. Martín-Valdivia, "COVID-19 detection in radiological text reports integrating entity recognition," (in eng), Comput Biol Med, vol. 127, p. 104066, Dec 2020.

[3]S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and R. N. K. S, "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study," (in eng), JMIR Public Health Surveill, vol. 6, no. 2, p. e18828, Apr 14 2020.

[4]T. S. Shen, A. Z. Chen, P. Bovonratwet, C. L. Shen, and E. P. Su, "COVID-19-Related Internet Search Patterns Among People in the United States: Exploratory Analysis," (in eng), J Med Internet Res, vol. 22, no. 11, p. e22407, Nov 23 2020.

[5]H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," (in eng), IEEE J Biomed Health Inform, vol. 24, no. 10, pp. 2733-2742, Oct 2020.

[6]J. L. Izquierdo, J. Ancochea, and J. B. Soriano, "Clinical Characteristics and Prognostic Factors for Intensive Care Unit Admission of Patients With COVID-19: Retrospective Study Using Machine Learning and Natural Language Processing," (in eng), J Med Internet Res, vol. 22, no. 10, p. e21801, Oct 28 2020.

[7]M. Odlum et al., "Application of Topic Modeling to Tweets as the Foundation for Health Disparity Research for COVID-19," (in eng), Stud Health Technol Inform, vol. 272, pp. 24-27, Jun 26 2020.

[8]D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study," (in eng), J Med Internet Res, vol. 22, no. 10, p. e22635, Oct 12 2020.

[9]E. M. Soltan, S. M. El-Zoghby, and H. M. Salama, "Knowledge, Risk Perception, and Preventive Behaviors Related to COVID-19 Pandemic Among Undergraduate Medical Students in Egypt," (in eng), SN Compr Clin Med, pp. 1-8, Nov 9 2020.

[10]E. Massaad and P. Cherfan, "Social Media Data Analytics on Telehealth During the COVID-19 Pandemic," (in eng), Cureus, vol. 12, no. 4, p. e7838, Apr 26 2020.

[11]N. Zheng et al., "Predicting COVID-19 in China Using Hybrid AI Model," (in eng), IEEE Trans Cybern, vol. 50, no. 7, pp. 2891-2904, Jul 2020.