Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/106/20241242

A Review of Research on Coding Methods for Open-ended Text Responses in Survey Questionnaires

Song Wang

Bur Oak Secondary School, 933 Bur Oak Avenue, Markham , Ontario, L6E 1G4, Canada

wangsong20061019@gmail.com

Abstract. In fields such as social sciences and market research, open-ended questions can collect richer data information, but how to effectively count and analyse these text answers becomes a key issue. The study mainly explores the three coding methods of open-ended questions in questionnaires, including the definition, process, and application of manual coding, semi-automatic coding, and automatic coding. According to existing literature and data, manual coding has high flexibility and accuracy, but it is inefficient when processing large-scale data; semi-automatic coding combines manual coding and machine learning technology, which can improve efficiency while maintaining a certain degree of accuracy; automatic coding relies on natural language processing technology and deep learning models, which greatly improve coding efficiency, but there is a problem of insufficient accuracy when facing complex semantics. Future research can focus on improving the accuracy of automatic coding through deep learning, developing intelligent semi-automatic systems that reduce manual intervention, and incorporating real-time feedback mechanisms for continuous misappropriation.

Keywords: Open-ended Questions, Manual Coding, Semi-automatic Coding, Automatic Coding, natural language processing (NLP).

1. Introduction

Open-ended questions in questionnaires play an important role in fields such as social sciences and market research. According to Xiao Fuqun, the significance of open-ended questions in questionnaires stems from their ability to free respondents from predetermined answers, thereby facilitating the collection of richer data and enabling researchers to uncover unexpected new information[1]. For example, in consumer feedback, open-ended questions allow respondents to freely express their true feelings about products or services, rather than being limited to preset options. However, how to count and analyze the answers to numerous open-ended questions has also become a key problem. For example, in a market survey, if hundreds or thousands of consumers each describe their experience in different languages and expressions, how can researchers effectively summarize this information and draw meaningful conclusions? While data collection holds enormous potential, there are also huge challenges in how to integrate these answers. This study explores the limitations of manual coding, semi-automatic coding, and automatic coding in dealing with these challenges by describing their definitions, processes, and applications, and proposes suggestions for improvement. By analyzing

DOI: 10.54254/2755-2721/106/20241242

these coding methods, people can not only improve coding efficiency, but also increase the accuracy of results, thereby better serving the needs of social sciences and market research. Additionally, this study will explore the future development direction of coding methods, with the goal of providing guidance for future research and enhancing researchers' comfort level with complex open data.

2. Manual coding

Manual coding, one of the earliest text processing methods, is the basis for coding open-ended text answers and originated in the early stages of qualitative research. When researchers began to collect data using questionnaires with open questions, how to deal with this messy text data became a key issue. Early survey research primarily relied on manual coding, which required researchers to manually record each answer, classify it, and quantify it. Manual coding is a systematic process that aims to convert open-ended text answers in questionnaires into structured data, which usually includes nine steps. Firstly, researchers gather all questionnaire data to fully record all respondents' answers. Next, they determine the size of the data sample. . If the data set is small, such as a few hundred questionnaires, all of them are coded; if the data set is large, with thousands or tens of thousands of questionnaires, a portion is extracted for processing. In the third step, researchers carefully read each questionnaire answer to understand its surface and deep meaning. Open-ended questions in survey questionnaires vary in complexity. Simple, objective open-ended questions usually lead to clear and direct responses, while complex questions may produce answers with multiple layers of meaning and diverse expressions. For more objective open-ended questions, researchers can directly record the categories and their frequencies in the answers. However, for more complex text answers, researchers need to further distinguish their deeper meanings and record them in detail. Subsequently, researchers categorize all text answers to transform unstructured data into structured data that is easy to analyze. Next, according to the specific purpose of the study, these classified text answers are sorted and coded. For those answers that cannot be classified or appear less frequently, the researchers will classify them into the "other" category. Finally, the researchers output the coding results to provide a basis for subsequent statistical analysis[2]. Although this process is accurate, it is less efficient when processing large-scale data, so it has gradually been combined with semi-automatic and automatic coding methods in recent years.

Manual coding is a fully manual operation with unique characteristics of high flexibility and precision, which enables it to play an important role in fields that require in-depth understanding and detailed analysis. In literary works and text analysis, researchers often need to conduct in-depth analysis of literary works such as novels and poems. Through manual coding, the themes, symbolism, emotional tone and style characteristics in the text can be captured, and these subtleties are often difficult to accurately identify through automated tools. However, manual coding is not without its limitations. Wu Jie et al. pointed out that manual coding has a high omission rate in comparison with computer coding. The study's coding protocols were strictly followed and experienced coders were used, although the percentage of missed codes was still as high as 20% to 30%. This suggests that even with the accuracy that comes with manual coding, major omissions can still occur due to human mistake, weariness, or oversight — especially when coding big volumes of data[3]. In addition to the risk of omissions, manual coding is also very time-consuming and labor-intensive. When faced with large-scale data, such as processing tens of thousands of open-ended questionnaires in market research, manual coding is obviously inefficient. This coding method relies entirely on manual operations, making it difficult to expand quickly, which is exactly what modern research urgently needs.

3. Semi-automatic coding

Semi-automatic coding is a coding method that combines manual operation with automation technology. It uses a small amount of manually annotated data to train the machine learning model, and then applies the trained model to a large-scale data set for coding. This method aims to improve coding efficiency through automation while maintaining manual coding's accuracy and flexibility. The process of semi-automatic coding includes multiple steps. The first step is to extract a sample from all

DOI: 10.54254/2755-2721/106/20241242

the data of the open questions to generate a data set of open text answers; the second step is to divide this dataset into a training set and a test set; the third step is to manually encode the training set. According to the preset coding standards, the researchers manually encode each text answer in the training set to ensure the accuracy and consistency of the initial coding; the fourth step is to use the already encoded training set data to train a machine learning model so that the model can encode according to the text content; the fifth step is to apply the trained model to the test set and use the uncoded data of the test set to verify the model to evaluate the coding ability of the model; the sixth step is to check the performance of the model on the test set to determine whether the model has reached the expected accuracy and consistency standards; the seventh step is to use the trained model to automatically encode a large-scale open text answer dataset after the model performance reaches the standard; the eighth step is that for some data that the model cannot accurately encode, it still needs to be supplemented by manual coding to ensure the accuracy of the coding; finally, the ninth step is to output all the coding results to provide a reliable basis for subsequent data analysis and research[2].

In He Zhongshanyue's research, semi-automatic encoding has improved efficiency to a certain extent. The reason is that manual encoding can be used to process text answers that are difficult to classify, while automatic encoding is used to process text that is easy to classify. The polynomial gradient enhancement algorithm is the main manifestation of this progress. The algorithm first automatically encodes the text answer, but still relies on manually adjusting parts that are difficult to process automatically[4].

A semi-automatic encoding method based on text mining and polynomial enhancement has been proposed. Compared with manual coding, this method not only improves coding efficiency, but also maintains high accuracy. The reason is that it can automatically encode text that is easy to classify, while manually processing text that is difficult to classify. In the author's study, with an accuracy rate of 80% as the goal, approximately 47% to 58% of the data can be automatically classified[5].

Semi automatic coding is highly suitable for educational research. It has the advantages of balancing efficiency and accuracy, and strong scalability. With the help of semi-automatic coding technology, educational researchers can quickly organize and analyze student evaluations and course feedback in the classroom, determine effective teaching methods and areas for improvement, and ultimately improve teaching quality.

Li Yanyan et al. used the semi-automatic coding function of the Vinca tool in "Collaborative Learning Interaction Analysis Tool and Its Case Study". The semi-automatic coding Vinca tool automatically recommends coding by displaying prompt words in the text, which greatly reduces the manual workload while still retaining space for manual review and adjustment[6]. Similarly, Hu Shengli and Zhang Songlin pointed out that the semi-automatic encoder can effectively capture the potential characteristics of users and projects by embedding auxiliary information in the input layer and coding and decoding the data[7]. This method illustrates the efficiency improvement of semi-automatic coding in information extraction and processing.

The semi-automatic encoding itself also has some limitations. For example, using manual encoding to train a model in a semi-automatic process may consume a significant amount of time. If the initial manual encoding contains errors, it is likely to reduce the credibility and efficiency of the entire encoding process. Because such errors may affect the accuracy of the model.

4. Automatic coding

With the in-depth research of deep learning in the field of natural language processing, automatic coding has gradually become a popular research topic. This method is particularly suitable for preprocessing text answers to open-ended questions in large-scale questionnaires. The current automatic coding has achieved almost 100% automation. It uses pre built big data models to encode open text answers. The use of intelligent algorithms such as machine learning and artificial intelligence reduces coding costs while improving coding quality. Automatic encoding has significantly improved the efficiency and consistency of encoding compared to previous encoding methods. However, improving its accuracy and generalization ability across different fields and

DOI: 10.54254/2755-2721/106/20241242

problem types remains a major challenge it currently faces. The steps for automatic encoding are as follows: The first step is to identify unresolved issues. The purpose of recognition is to determine whether the processed text is the answer to an open-ended question. If not, then other text answers need to be classified. The second step is to determine whether it belongs to a specific topic. If yes, continue processing; if not, perform other processing. The third step is to extracttextual answers. Extract the text answer data of the open question and prepare for subsequent processing. The fourth step is to extract the main emotions and sentiment analysis. Extract the main themes and sentiment information from the extracted text answers. The fifth step is to automatically encode the answers to the open questions, and use the automatic coding system to encode the extracted text. Step six involves updating the vocabulary. According to the new vocabulary or unmatched text content found during the coding process, the relevant coding model is updated. The seventh step is matching verification to determine whether the coding result matches the expected coding standard or vocabulary. Finally, step eight output the result. If the matching verification is successful, the final coding result is output; if it does not match, the unmatched text is sent back to the vocabulary update module for reprocessing until the match is successful[2].

Natural language processing (NLP) technology is the cornerstone of automatic coding. This technology plays an important role in the encoding process. The importance of this role is particularly evident when dealing with complex textual data in open-ended questionnaires. Firstly, NLP technology analyzes text to identify low-level tasks. The purpose of recognition is to ensure accurate decomposition and processing of textual information. Then, the system uses NLP technology to perform advanced tasks. The purpose is to understand the core information and context in the text. Named entity recognition and word sense disambiguation are typical representatives of these tasks. These processes enable the system to convert text data into structured encoding based on predefined rules or through machine learning algorithms. The application of NLP technology has continuously improved the accuracy and consistency of automatic encoding. This progress has been made with the development of big data and machine learning technologies, especially when dealing with ambiguous or ambiguous medical texts[8].

Song Fan et al. pointed out that by combining deep learning convolutional neural networks (CNNs) with long short-term memory networks (LSTMs) and introducing a bidirectional memory conduction mechanism, the accuracy and efficiency of ICD automatic coding will be effectively improved. Studies have shown that this method performs well in dealing with long sequence dependencies and complex causal relationships in medical texts. Compared with other coding models, the average Macro-F1 value reached 63.2% and the Micro-F1 value was 69.9%. In addition, the speed of automatic coding has also been significantly improved, with an average coding time of 0.05 seconds for the test set, which is ahead of the efficiency of manual coding and semi-automatic coding[9].

Lai Jianzhi, in his paper "Research of Urban Shared Bikes Parking Area Automatic Coding Method," proposed an automatic coding method based on ArcGIS geoprocessing tools and ArcObject programming interface, aiming to solve the problem of precision management of shared bicycle parking areas. He used spatial analysis technology to automatically code shared bicycle parking areas on both sides of the road, and determined the order of parking areas through buffer analysis and overlay analysis. The study found that this method performed well in terms of coding continuity and accuracy, with a high degree of automation, and provided strong technical support for the subsequent management of shared bicycles[10].

Zhang Jing pointed out in the article "Automatic Coding Method and Application of Evaluation Questions in Questionnaire Surveys" that compared with manual coding, automatic coding achieves more efficient text processing by using natural language processing technologies such as topic modeling (LDA) and sentiment analysis, improves efficiency, and saves costs by reducing the need for manual intervention. However, Zhang Jing also pointed out that automatic coding has limited accuracy when facing complex texts with diverse meanings. In addition, automatic coding also depends on the degree of perfection of the topic model and sentiment dictionary. If the model and dictionary are not comprehensive enough, the accuracy of the results may be reduced[11].

DOI: 10.54254/2755-2721/106/20241242

5. Discussion

Despite the continuous progress of manual, semi-automatic, and automatic coding methods, each method still faces significant challenges in specific situations, which limits its effectiveness. Although automatic coding is extremely efficient, it has difficulties maintaining the accuracy and flexibility of manual coding. Automatic coding relies on pre-built models, which means that its effectiveness is limited by the quality and comprehensiveness of the training data.

With the continuous increase in data volume and the continuous advancement of text analysis technology, there will be new opportunities for coding methods for answers to open-ended questions. Future research can be explored from the following aspects:

- Although automatic coding has greatly improved coding efficiency, there is still room for improvement in accuracy when facing complex and diverse texts. Future exploration can further integrate deep learning technology with natural language processing (NLP) technology to improve the accuracy of automatic coding.
- Future coding methods can rely more on human-computer interaction, that is, combining the accuracy of manual coding with the efficiency of automatic coding on the basis of the automatic coding model to build an intelligent semi-automatic coding system. While ensuring the accuracy of coding, the need for manual intervention is further reduced.
- Future coding systems can incorporate real-time feedback mechanisms, allowing the system to continuously self-learn and iterate based on new data or manual corrections by users, thereby improving the efficiency and accuracy of coding.

6. Conclusion

This study essentially explores the evolution of open-ended question answer coding methods from manual coding to automatic coding. Although manual coding has high accuracy and flexibility, it is time-consuming and difficult to adapt to large-scale data sets. Semi-automatic coding achieves a balance between efficiency and accuracy by combining manual coding and machine learning. Automatic coding is most efficient when processing large-scale data, but sometimes lacks accuracy and contextual understanding. However, there are also some shortcomings in this study. To improve this study, the future can focus on exploring how to better integrate semi-automatic and automatic coding methods and optimize the model to better handle complex text data. Simultaneously, real-time feedback mechanisms and more comprehensive training data sets can enhance the coding system's accuracy and adaptability across various fields.

References

[1] Xiao, F. Q. (2007). Coding open-ended questions in survey research. Statistics and Decision, 3(233), 73-74. https://doi.org/10.13546/j.cnki.tjyjc.2007.05.033

[2] Liu, P., & An, J. (2023). A review of text answer coding methods for open-ended questions in questionnaires. Statistics and Application, 12(5), 1464-1476. https://doi.org/10.12677/sa.2023.125150

[3] Wu, J., Kors, J. A., Herpen, G., & Zhang, R. J. (1998). Comparison of computerized Minnesota coding and manual coding: A performance evaluation.Chinese Journal of Cardiac Arrhythmias, 2(2), 92-95.

[4] He, Z. (2020). On the automatic coding of text answers to open-ended questions in surveys (Doctoral dissertation, University of Waterloo). http://uwspace.uwaterloo.ca/xmlui/handle/10012/15633

[5] Scholau, M., & Couper, M. P. (2016). Semi-automated categorization of open-ended questions. Survey Research Methods, 10(2), 143-152. https://doi.org/10.18148/srm/2016.v10i2.6213

[6] Li, Y., Liao, J., Wang, J., & Huang, R. (2007). A tool for analyzing interaction in CSCL and a case study. Open Education Research, 13(4), 94-99. https://doi.org/10.19697/j.cnki.1673-4432.202103008

DOI: 10.54254/2755-2721/106/20241242

[7] Hu, S., & Zhang, S. (2021). A hybrid deep learning algorithm for recommendation system. Journal of Xiamen University of Technology, 29(3), 49-55.
[8] Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. Journal of the American Medical Informatics Association, 18(5), 544-551. https://doi.org/10.1136/amiajnl-2011-000464
[9] Song, F., Yang, X., Wang, Y., & Yu, J. (2023). ICD automatic encoding method based on bidirectional memory conduction. Chinese Journal of Health Informatics and Management, 20(6), 977-984. https://doi.org/10.3969/j.issn.1672-5166.2023.06.019
[10] Lai, J. (2022). Research of urban shared bikes parking area automatic coding method. Journal of Geomatics, 47(2), 136-138. https://doi.org/10.14188/j.2095-6045.2019572
[11] Zhang, J. (2021). Automatic coding method of evaluation questions in questionnaire survey and its application. Tianjin University of Commerce, 1-45.