Analysing Open-Ended Questions in ESPAD-MedSPAD bridge project: a manually labelled dataset

Giada Anastasi[1,2,†], Corrado Fizzarotti[1,3,†], Lorenzo Nelli[1,4,†], Rodolfo Cotichini[1,†], Elisa Benedetti[1,*,†], Claudia Luppi[1,†] and Sabrina Molinaro[1,†]

[1] *Institute of Clinical Physiology (IFC), Italian National Centre of Research (CNR), Pisa, Italy*

[2] *Institute of Computer Science, University of Pisa, Pisa, Italy*

[3] *Department of Linguistic and Cultural Studies, University of Modena and Reggio Emilia, Italy*

[4] *Department of Communication science, Humanities and International Studies, University of Urbino Carlo Bo, Italy*

Abstract

This contribution aims to explore the use of mixed method analysis in examining data obtained from some open-ended questions posed to 260 stakeholders in addiction policies, prevention and researchers from 47 countries. The reference study from which the data are taken is the ESPAD-MedSPAD bridge project, a pioneering initiative with the mission to assess the role of school surveys in policy making, prevention planning and evaluation in a broad spectrum of countries, spanning the European and Mediterranean regions. The study employs a two-pronged methodological design, combining quantitative and qualitative components. This mixed-methods approach is particularly advantageous when studying complex social contexts such as addiction. During the study, the open-ended questions are exposed to T2K tool analysis, which identifies the most frequent topic areas among the participants' answers. Following that, the experts identify the most relevant theme areas and subareas and manually label the comments. The end result is a dataset with labels for all open-ended questions provided by respondents. This will then serve as the foundation for subsequent analyses.

Keywords

Information Retrieval, Natural Language Processing (NLP), Open-ended Questions; School Survey, Policy, Prevention.

1. Introduction

The retrieval of information from open-ended survey questions is a fundamental component of the survey research process. Respondents can submit detailed and unstructured comments to open-ended survey questions, yielding significant insights into their ideas, feelings, and opinions. Open-ended questions let respondents express themselves in their own terms, lowering the possibility of bias imposed by predefined response possibilities. This can result in more honest

AIxIA'23: 22nd International Conference of the Italian Association for Artificial Intelligence, November 06–09, 2023, Rome, Italy

[*] Corresponding author.

[†] These authors contributed equally.

giadaanastasi@cnr.it (G. Anastasi); corradofizzarotti@cnr.it (C. Fizzarotti); lorenzonelli@cnr.it (L. Nelli); rodolfo.cotichini@cnr.it (R. Cotichini); elisa.benedetti@cnr.it (E. Benedetti); claudia.luppi@cnr.it (C. Luppi); sabrina.molinaro@cnr.it (S. Molinaro)

0000-0002-5854-3716 (G. Anastasi); 0000-0001-8616-314X (C. Fizzarotti); 0000-0002-3456-6411 (L. Nelli); 0000-0002-8162-1938 (R. Cotichini); 0000-0002-7345-3962 (E. Benedetti); 0000-0001-7221-0873 (S. Molinaro)

and genuine responses. The use of open-ended questions allows for a more personalised and tailored survey experience. Respondents can share their unique perspectives and experiences, which improves the overall quality of the data, as also evidenced by literature from other fields [1, 2, 3].

1.1. The ESPAD-MedSPAD project

The ESPAD-MedSPAD bridge project is a pioneering initiative with the mission to assess the role of school surveys in policy making, prevention planning and evaluation in a broad spectrum of more than 40 countries spanning the European and Mediterranean regions. Funded by the Council of Europe - Pompidou Group (PG-CoE) within the framework of the 2022 work programme of the Mediterranean Cooperation Network on Drugs and Addiction (MedNET), and reinforced by the support of the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), this study was conducted under the scientific supervision of the National Research Council (CNR) in collaboration with the EMCDDA.

Several purposes drive this endeavour: firstly, to collect data on prevalence of psychoactive substance use and risk behaviour. Second, it is aimed at using the data it collects to identify needs and priorities, making its findings available to anyone who wants to implement evidence-based policies. Thirdly, the resulting data it acquires can serve as a monitoring mechanism to evaluate the effectiveness of existing prevention strategies and programmes. Finally, this project also aims to formulate prevention actions and strategies in the field of education and to contribute to the public discourse on substance use and risk behaviour, in particular through the involvement of the media.

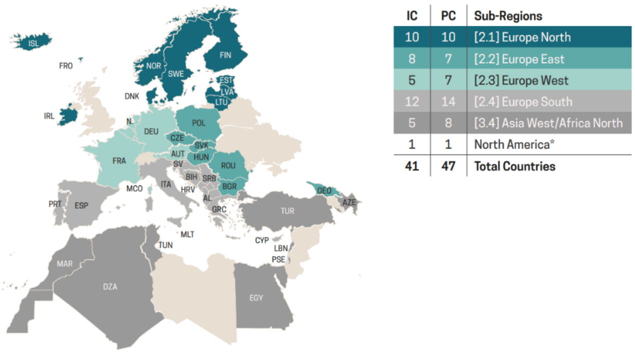
ESPAD-MedSPAD bridge benefited from the valuable contribution of more than 250 experts who lent their expertise to assess the current challenges associated with the use of school survey results and to develop strategies and actions to address them. Therefore, its contents faithfully represent the current state of play, based on the information and insights provided by policy makers, policy experts, prevention and harm reduction specialists as well as scientists actively involved in conducting or using school surveys. Ultimately, the results of this study will serve as a catalyst for improved evidence-based decision-making in drug policies, prevention strategies, education initiatives and public discourse on these critical issues.

2. Materials and Methods

2.1. The online ESPAD-MedSPAD survey

A total sample of 260 stakeholders from 47 countries participated in the online survey. In particular, as shown in Figure 1, the sample was composed by respondents from the following subregions: 10 countries from North Europe, 7 countries from Eastern and Western Europe respectively, 14 countries from Southern Europe, 8 countries from West Asia and North Africa, and the United States.

The questionnaire was created using a mixed method approach that combines qualitative and quantitative questions, leveraging the strengths of each data type while mitigating their weaknesses. This approach allows for a thorough exploration of stakeholders' perceptions



IC	PC	Sub-Regions
10	10	[2.1] Europe North
8	7	[2.2] Europe East
5	7	[2.3] Europe West
12	14	[2.4] Europe South
5	8	[3.4] Asia West/Africa North
1	1	North America*
41	**47**	**Total Countries**

Figure 1: Geographical coverage of the project by sub-region based on the United National classification from [4]

regarding the specific domains under investigation. The questionnaire is structured into four distinct sections, each linked to the primary areas of interest in the examination of school survey data: policy, prevention, training, and public opinion and media. These sections incorporate both multiple-choice and open-ended questions. In order to analyse the quantitative and qualitative data collected through the questionnaire, a hybrid approach combining automated algorithms and human judgement was developed to maximise data extrapolation and understanding.

2.2. T2K

T2K, which stands for Text-to-Knowledge, is a sophisticated system at the intersection of natural language processing, statistical analysis and machine learning. Its main function is to uncover the domain-specific information contained in textual data, transform this information into a structured graphical format and meticulously index document collections, making them easily accessible to users. This multifaceted tool is invaluable in the field of information extraction and comprehension, as it enables a range of tasks such as term extraction, taxonomic chain generation, extracted named entities and knowledge graph visualisations [5]. By leveraging these capabilities, T2K can provide a comprehensive and accurate representation of a complex and multi-dimensional field such as addiction. From an operational point of view, we handed this tool the dataset containing the open-ended responses from the stakeholders involved. Having set the parameters for the natural language analysis, we let the tool identify which topics were most frequently mentioned. This was the starting point for the ontological work that allowed us to define the categories within which to group the comments. Networks such as T2K are definitely useful for analysing complex datasets such as ours as they use Bayesian phylogenetic systems that allow reconstructing language family trees. Considering the different countries involved in our study, it is important to use a comparative-typological analysis and to consider data from multiple languages in order to understand the interactions between different levels of grammar in sentence structure. This approach allows the analysis of large amounts of data and the identification of patterns and trends that may not be easily discernible through traditional methods. The implications that interest us most are in the field of natural language processing (NLP). In this case, raw text data must be prepared by tokenization, normalisation and extraction of their salient linguistic features. This ensures that the text is transformed into a format suitable for subsequent analysis, reducing noise and variations in expression. The next stage involves the use of deep learning models, such as recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, which are often augmented with attention mechanisms and are designed to handle the complexities and idiosyncrasies of multiple languages.

2.3. A Hybrid Approach Combining Automated Algorithms and Human Judgement

Methodologically, it was decided to subsequently proceed with a manual labelling of the results of the linguistic analysis performed with T2K. This hybrid approach between linguistic analysis software and human agents in data labelling combines the strengths of automated algorithms and human judgement to improve the accuracy and efficiency of the labelling process. This is for several reasons. Firstly, such classification allows the construction of categories and the identification of specific linguistic patterns that might not be captured by automated methods. Secondly, ontological second-order work allowed the specific variations of the different stakeholder groups to be taken into account. Lastly, this kind of double-labelling minimised the semantic slippages due to the different cultural backgrounds of the sample respondents. In practice, starting from the macro areas that the T2K analysis provided us with, we assigned each comment to one of them. This allowed us to obtain the graphs that will then be commented on in the results section. We then worked on defining sub-categories in order to account for the diversity within each group. By combining the language patterns extracted from the text with additional contextual information, it is our opinion that the hybrid approach achieved better results. The output of this hybrid study provided thus far is a dataset including all of the open-ended answers supplied by the participants, each labelled by the experts based on the subject areas indicated by the automatic T2K method. The labelling is separated into two levels: a first, broader level that represents a fairly broad thematic area, and a second level that focuses on the particular of the comment. Depending on their substance, answers may belong to more than one topic category. The dataset was subjected to preliminary analysis, the findings of which are provided in the report. To highlight the relationships between the various subjects, semantic graphs were developed using Python. Each node in the graphs represents the first and second-level topic areas, the size of which relies on the number of occurrences of the relevant topic area. The arcs reflect the relationships and the thickness of the arcs depends on the number of times the two connected subject areas are addressed in the same comment by one of the experts. After obtaining the manually labelled dataset and doing an initial analysis of the most relevant themes and their correlations, some possibilities for future study include automating the entire analysis using AI techniques. According to the literature, some potential future work includes the use of social network analysis or machine learning approaches to understand opinions on a certain issue [6, 7, 8]. These studies would then allow us to connect the topics raised by hybrid analysis to the category of experts who raised the issue.

3. 3. Preliminary results and conclusions

The approach presented here is important for several reasons. The literature on the relationship between manual and automatic labelling of datasets is decidedly vast: it would be impossible to effectively account for it here. Certainly, the choice of one approach over the other is a function of several factors: from the amount of available resources to the theoretical colouring of the study to which the method is applied. Considered in general, manually labelled datasets are important for information retrieval as they not only form the basis of machine learning model training but also serve as a benchmark against which to compare the results of automated systems. At some point in the process, manual labels always serve as a reference, allowing for the calculation of metrics such as precision, recall, F1 score, and accuracy, which are critical for evaluating the quality of retrieval algorithms. Also, from a technical perspective, manual labelling allows the dataset to be customised to align with the unique requirements of the task. This is especially important when pre-existing datasets do not adequately cover the desired domain. This is what was done in this context where human intervention helped us define a second-level ontology and allowed us to account for the particular nuances of the domain we were dealing with. Manual labelling allows for quality control to ensure that the dataset accurately represents the desired concepts or relevance criteria. This is one of those cases where manual labelling provides higher quality labels as it relies on human beings who can understand the context and nuances of the data better than machines. In any case, it is not in our aspirations to propose a method that is entirely manual. Here we are presenting the advantages of a hybrid approach. It is true that manual labelling guarantees high quality labels, but it is time consuming and may not be feasible for large data sets. Automatic labelling, on the other hand, can process large amounts of data quickly. A hybrid approach can therefore exploit the speed of automatic labelling and the quality of manual labelling. In addition, a hybrid approach can, in our opinion, reduce the bias inherent in monist solutions. For instance, manual labelling can be subjective and influenced by the personal biases of the labeller, while automatic labelling can perpetuate existing biases in the data or algorithms. A hybrid approach can help mitigate these biases and adaptively correct various errors. A final point is that a hybrid approach can also facilitate continuous learning. An hypothetical system can start with manual labelling, use the labelled data to train a machine learning model for automatic labelling, and then continue to improve the model with further manual labelling of data that the model finds difficult to label. This iterative process can improve the accuracy and efficiency of the labelling process. The first phase of the hybrid approach, done with the T2K tool, is critical for various reasons, the most important of which is data preparation. This entails gathering and preparing raw data for processing, such as text or speech. This covers tasks such as data cleansing, special character removal, tokenization, stop word removal, normalisation, and more, to ensure that the data is organised for analysis. NLP data frequently contains a huge amount of information, and this stage aids with reducing the data's complexity by focusing on the most important parts and minimising noise. Exploratory analysis is also possible with preliminary data preparation, which can uncover crucial trends, patterns, and insights that influence further data processing and model building. This stage also detects and removes noise in the data, such as typing errors, incomplete data, or irrelevant information, improving the data's overall quality. Another significant feature of this initial phase is data standardisation, which ensures consistency and comparability between data from diverse sources, which is vital in applications such as named entity identification or machine translation. Very different models and different equilibria can instantiate from the just described scheme: it just depends on where the manual and automatic elements are placed. In our case, we had a language model do the semantic heavy lifting and used the results to build a more colourful ontology on the labels identified by T2K. The two systems influence each other and reinforce each other, coming to consolidate the basic idea that it is the iteration between systems that is the best way to provide a correct approximation of a multicultural, complex and emotionally connoted reality such as the one we analysed.

Acknowledgments

We are very thankful to our research group at IFC-CNR for introducing us to this challenging project.

References

[1] L. Marcinowicz, S. Chlabicz, R. Grębowski, Open-ended questions in surveys of patients' satisfaction with family doctors, Journal of Health Services Research & Policy 12 (2007) 86–89.

[2] E. Riiskjær, J. Ammentorp, P.-E. Kofoed, The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective, International Journal for Quality in Health Care 24 (2012) 509–516.

[3] A. O'Cathain, K. J. Thomas, " any other comments?" open questions on questionnaires–a bane or a bonus to research?, BMC medical research methodology 4 (2004) 1–7.

[4] E. Benedetti, R. Cotichini, S. Molinaro, C. Fizzarotti, E. Colozza, G. Anastasi, L. Nelli, The use of school surveys in policy and prevention planning and evaluation. results of the 2022 espad - medspad bridge project, 2023. URL: https://rm.coe.int/medspad-bridge-electronic-en/1680ab1f7e.

[5] F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni, T2k^ 2: a system for automatically extracting and organizing knowledge from texts, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014, pp. 2062–2070.

[6] A. Romascanu, H. Ker, R. Sieber, S. Greenidge, S. Lumley, D. Bush, S. Morgan, R. Zhao, M. Brunila, Using deep learning and social network analysis to understand and manage extreme flooding, Journal of Contingencies and Crisis Management 28 (2020) 251–261.

[7] S. Verma, S. Vieweg, W. Corvey, L. Palen, J. Martin, M. Palmer, A. Schram, K. Anderson, Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 5, 2011, pp. 385–392.

[8] S. Cresci, A. Cimino, F. Dell'Orletta, M. Tesconi, Crisis mapping during natural disasters via text analysis of social media messages, in: Web Information Systems Engineering–WISE 2015: 16th International Conference, Miami, FL, USA, November 1-3, 2015, Proceedings, Part II 16, Springer, 2015, pp. 250–258.