THE TECHNOLOGISATION OF THEMATIC ANALYSIS: A CASE STUDY INTO AUTOMATISING QUALITATIVE RESEARCH

M. Constantinou[1], R. Polvara[1], E. Makridis[2]

[1]University of Lincoln (UNITED KINGDOM)
[2]University of Cyprus (CYPRUS)

Abstract

Thematic analysis is the most commonly used form of qualitative analysis used extensively in educational sciences. While the process is straightforward in the sense that a hermeneutic analysis is conducted so as to detect patterns and assign themes emerging from the data acquired, replicability can be challenging. As a result, there is significant debate about what constitutes reliability and rigour in relation to qualitative coding. Traditional thematic analysis in educational sciences requires the development of a codebook and the recruitment of a research team for intercoder reviewing and code testing. Such a process is often lengthy and infeasible when the number of texts to be analysed increases exponentially. To overcome these limitations, in this work, we use an unsupervised text analysis technique called the Latent Dirichlet Allocation (LDA) to identify distinct abstract topics which are then clustered into potential themes. Our results show that thematic analysis in the field of educational sciences using the LDA text analysis technique has prospects of demonstrating rigour and higher thematic coding reliability and validity while offering a valid intra-coder complementary support to the researcher.

Keywords: educational sciences, thematic analysis, topic modelling, Latent Dirichlet Allocation (LDA).

1 INTRODUCTION

Thematic analysis is the most commonly used induction-based hermeneutic content methodological approach in qualitative research [1]. Collected empirical data guide researchers and exploit the topic analysis process, as data are usually studied with limited reference to or without a specific theoretical framework. Thematic analysis is an exploratory, challenging, and time-consuming method that is useful when knowledge of a research topic is limited. It is emphasized that thematic analysis is used for data analysis of transcripts deriving from the most commonly used qualitative data collection approach, that is conducting interviews. Researchers need to examine and comprehend the data collected in order to decipher and interpret the transcripts in light of the relevant research question posed. The main purpose of thematic analysis is to systematically record codes and themes that emerge from the data transcribed in an attempt to verify and expand them [2] until thematic saturation is reached.

Thematic analysis of qualitative data acquired from interviews can be challenging due to the difficulty of determining the trustworthiness of the data assigned for each theme due to the subjective nature of their categorisation. For this reason, demonstrating rigour can help decrease bias, and increase the validity of results and therefore their potential replicability [3].

As previously discussed in the literature [3], [4] the manual codebook construction begins with identifying the initial themes and codes applied to the qualitative data until theme saturation is reached. Once codes are clearly defined, additional analysts are invited to review the themes and their interpretation. The intercoder reviewing process usually involves the distribution of interview scripts along with a well-defined codebook to the appointed data analysts who attempt to assign the interviewees' responses to the codes. The results from the intercoder testing are then reviewed and compared to the original coding, using the reliability metric which can be calculated by dividing the number of agreements by the total number of agreements plus disagreements [3]. The suggested minimum percentage indicating reasonable consent is 75%. In case of an inadequate level of agreement, discussion with data analysts allows negotiation of consensus between the coders that invites a second round of intercoder reviewing to reach adequate levels of agreement and finalise the codebook. The limitations of the complex and time-consuming thematic analysis as part of a qualitative methodological process pose numerous difficulties to the researcher in an attempt to demonstrate rigour and reliability to counterbalance subjectiveness.

This paper attempts to tackle the aforementioned challenges of qualitative thematic analysis by exploiting features of Natural Language Processing tools. In this work, we employ a Latent Dirichlet Allocation (LDA) model, to automatically detect topics from a textual input extracted from interviews. Similar approaches were performed using corpuses transcribed from online videos [5] and online magazines [6]. LDA topic modelling will be used indicatively as a quick quantitative tool as part of an initial content analysis phase, complementary to the manual thematic analysis, to guide the researcher by automatically organising long corpora in patterns of co-occurring words that can be interpreted by the researcher into meaningful themes. The utilisation of the algorithm, apart from automatically discovering potential topics in a transcript, can enhance interpretational reliability through probabilistic detection of corpora patterns that form a potential topic. Through discovering specific word patterns, interpretations of the transcript as a whole can be made, something that can assist the researcher in manually developing a codebook. However, word patterns cannot be interpreted alone or in isolation of the transcript, therefore this study does not intend to introduce LDA topic modelling analysis as an alternative method of thematic analysis but merely as a complementary method of support to the researcher.

2 METHODOLOGY

In this work, we propose an unsupervised classification method to identify topics based on the interviews collected and reported manually in text format. In particular, we use the Latent Dirichlet Allocation (LDA) algorithm [7] for extracting major topics from the given corpus of text data. LDA assumes that a given text is a mixture of topics, where each topic consists of a set of words. Hence, multiple parts of the given text could potentially share the same topic. In contrast to manual identification methodologies that are currently performed in educational literature, thematic analysis, LDA only identifies topics by the list of words associated with it, without assigning a label (*i.e.,* “education”, “research”, “students”) to the topic itself. For these purposes, LDA computes the probability that a given term is generated by a topic. Because mixtures are not discrete and separated groups, the same word can be generated by multiple topics with different probabilities. LDA acts in an iterative manner to calculate and refine such probabilities. This process continues until a steady state is achieved.

In what follows, we present the three major steps of LDA topic modelling followed in this work:

a) Data collection and preparation: consisting of the process of collecting the interviews and importing the corresponding text document that is to be analysed into the final document hereinafter called corpus;

b) Data cleaning and pre-processing: needed for purging the text from punctuation and stop-words, and lemmatizing (words in the third person are changed to first person and verbs in past and future tenses are changed into the present), as well as stemming the corpus (where words are reduced to their root form);

c) Model training: making use of the LDA algorithm to obtain the distribution of words for each topic, and the distribution of topics for each particular document (interview participant).

Once the document-term matrix (final corpus) has been computed, LDA can be used for training a topic model on it. In this work, we implemented LDA topic modelling using Python programming language and the Gensim open-source library for unsupervised topic modelling and natural language processing [9].

2.1 Data Collection and Preparation

Data analysed in this paper were collected from interviews conducted with members of staff at a university’s life science department whose laboratory lessons were observed. Interested participants were interviewed at their convenience after the practical work lessons were observed. The interviews were based on a semi-structured approach, where the researcher asked prompting questions that allowed interviewees to lead the conversation by responding and reflecting. Interviews were audio recorded and were subsequently manually transcribed to form the textual data that were used for the purposes of this study. The text was cleaned and pre-processed in the following step. It is worth mentioning that the textual data from the interviews were merged into a single document since distinguishing the participants (members of staff) is not desired in this context. Manual thematic analysis was already conducted prior to attempting to run the script and analyse it using Topic Modelling. The already analysed text using manual thematic analysis, the most frequently used

method for analysing qualitative interview data, will later be compared against the results generated by the topic modelling algorithm. The codebook developed from the manual thematic analysis has been validated through intercoder reviewing by three data analysts appointed for the purposes of the initial study; a time-consuming process presented as a limitation of manual thematic analysis and is considered to be potentially overcome by Topic Modelling. It is important to note that what is defined as themes in the manual thematic analysis is what recurs multiple times in the corpus as 'pragmatic tools' [10] with an intent to address the research questions posed for the study.

2.2 Data Cleaning and Pre-processing

During this step, textual data were transformed into a suitable format that could be used as input for training the LDA model. In particular, data cleaning processes were followed using the NLTK library [8], to remove numerical values, symbols, punctuations, and stop words based on the English vocabulary. Then, the pre-processed text was tokenized into a list of single words to convert our initial text into our final corpus which will serve as the input for training the LDA model.

2.3 Model Training

The training process of the LDA model is performed over the final corpus prepared using the data cleaning and pre-processing step, while the number of topics to be extracted is chosen to K=3. It is worth mentioning that, to interpret the meaning of each extracted topic, we keep the number of topics low such that there are no overlaps between topics. However, the selection of the number of topics for the LDA model is often made using grid search approaches to find the optimal number of topics K, which depends on the input dataset. In this work, the selection of training the LDA model using three topics results in the best performance as will be shown and discussed in the subsequent section.

3 EXPERIMENTAL RESULTS

Before applying LDA, interviews from 14 members of staff were all assembled into a single corpus. This is the text on which the researchers run the LDA to fit a topic model and discover how many topics are present. A word map was generated in order to have a visual insight into which words appear in the corpus along with their weight, before removing the stop words from the initial textual data. An illustration is presented in Figure 1. As seen on the word map, the bigger a word appears in the cloud, the higher its relevance and appearance frequency in the corpus. The pre-processing step is applied to the corpus in order to 'clean' text from punctuation and stop words, but also lemmatise and normalise the terms.

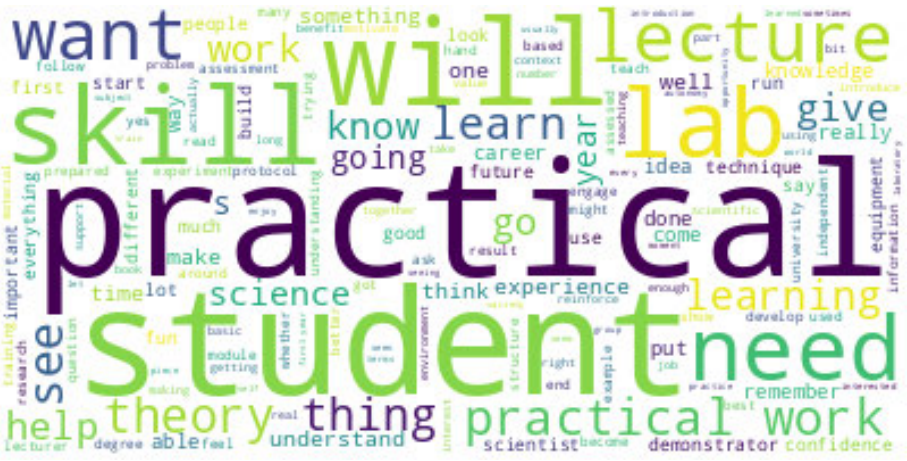

Figure 1. Word cloud generated out of the corpus.

Similarly, a histogram with the ten most frequent words was generated and presented in Figure 2.

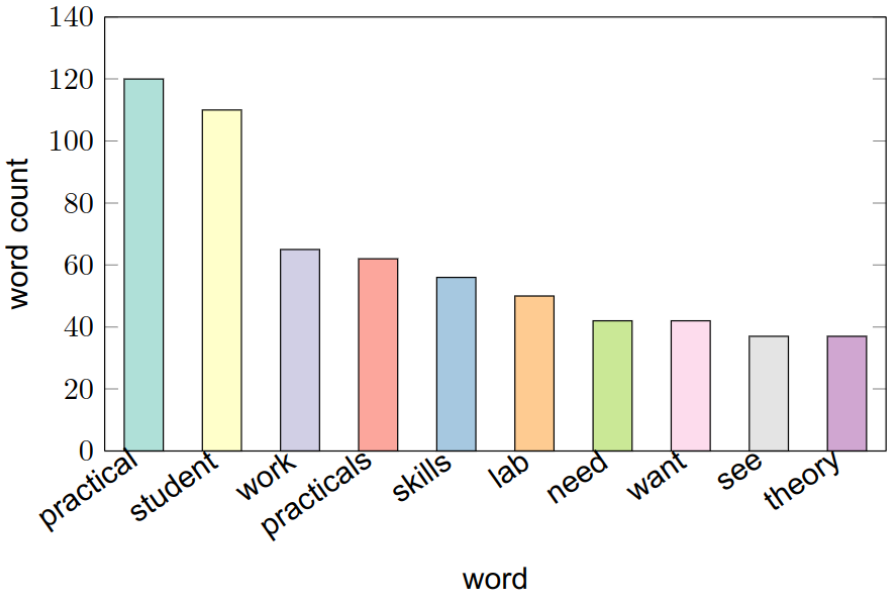

Figure 2. The ten most frequent words and their occurrence in the corpus document.

To evaluate and interpret the topics obtained by the trained LDA model, we use the visualisation package pyLDAvis in Python which is based on the LDAvis method presented in [11]. This method exploits the interpretation of the topics by extracting information from the trained LDA topic model in an interactive visualisation. The topics that were extracted by our trained LDA model are denoted by the bubbles shown in Figure 3. Clearly, all three topics are of equivalent significance, while they are scattered in different quadrants implying that the combination of words in the extracted topics does not overlap. The model maximising the inter-topic distance among the topics, without allowing overlapping was chosen as the best model for the purposes of this study. Based on empirical evidence from the results analysed, the optimal number of identified topics for the best result outputs was limited to three. The left part of Figure 3 illustrates the inter-topic distance map for the case of Topic 1. In this example, the visualisation generated for Topic 1 as well as Table 1(A) that shows the contribution of each word in its corresponding topic generates the words “students”, “lab”, “want, “lecture”, “repetition”, “time”, “remember”.

It is possible to infer that this topic is associated with claims that students in the laboratory like that with time, they can remember what they learned from the lecture through repetition in the laboratory. Indeed, in the manual thematic analysis, some key themes emerging were related to ‘repetition’ as members of staff commented on how important that is for undergraduates in order to master a skill in the laboratory and that through practical work students ‘remembered’ better of what they had learned. Additionally, for the effectiveness of practical work, members of staff also considered as important that lectures complimented practical work.

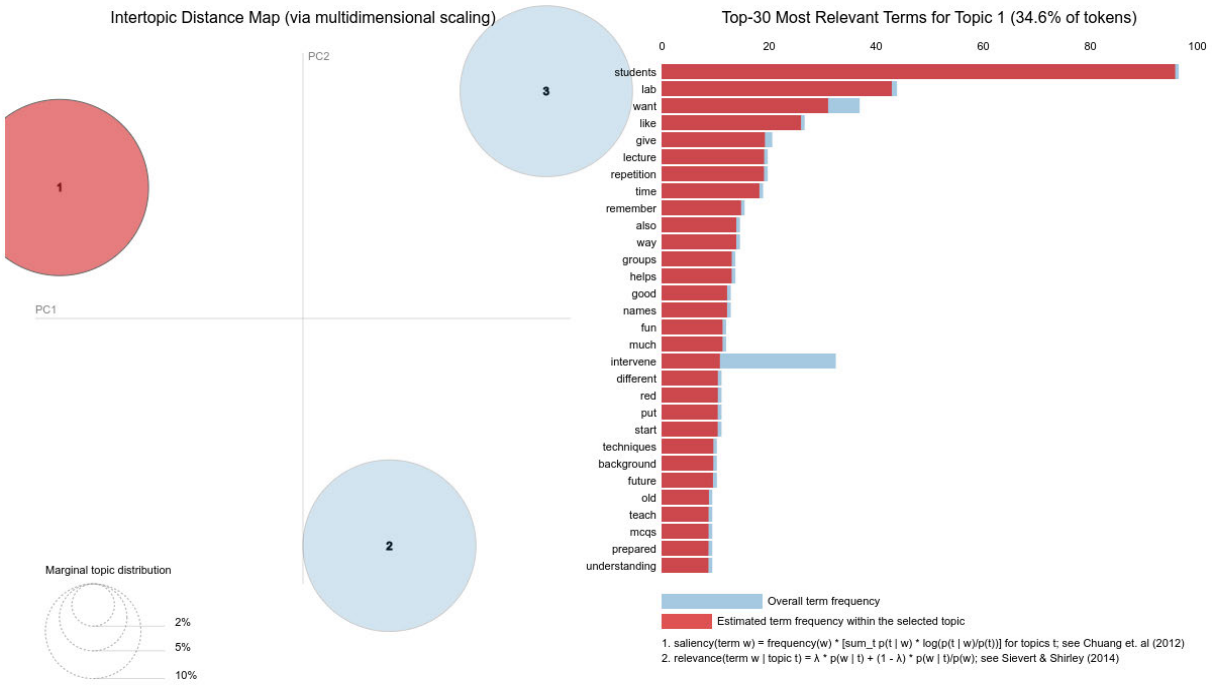

Figure 3. The ten most frequent words and their occurrence in the corpus document.

Therefore, in the way demonstrated above, the top ten words (together with their respective weight that contributes to the corresponding topic) were collected for each topic in Table 1 and were assigned to an interpreted meaning in a similar way as this is generally done in the manual thematic analysis.

Further to the interpretations related to the topics discovered in Table 1 (A), as previously discussed for Topic 1, it is inferred that Topic 2, shown in Table 1 (B), is associated with claims that in practical work lessons, students need to 'see' something and with the 'intervention' of someone, in this case members of staff, they can 'think' about their 'experiences'. Indeed, in manual thematic analysis the theme 'visualise', matching the word 'see' that was found in a topic after LDA training, emerged as members of staff highlighted that one of the main advantages of practical work for students was that they could combine lectures with experiments in order to see in real life through observables what they saw in books. Through 'Intervention' the importance of members of staff support in practical work lessons was also evident as they played an important role in conceptual understanding in the laboratory. The verb 'make' found in Topic 2, Table 1 (B), indicates the general goal of practical work emerging in manual thematic analysis referred to as 'Hands-On' that through which students had their learning experience 'reinforced' as they 'thought' about their 'experience'.

For Topic 3 shown in Table 1 (C), even though the interpretation is generalised, the combination of words 'skills', 'learn', 'know', 'understand', 'first', infer that with practical work students develop their skills and learn and that in order to know something they need to understand it first. This matches one of the key themes that emerged in manual thematic analysis, that of the development of skills which was highlighted repetitively in combination with the effective practical work had on the students learning experience through 'reinforcement'.

The topics found after LDA training match the general meaning inferred from the most frequent words occurring in the corpus (Figure 2). Reading through the interview transcripts, what was highlighted was the role of practical work in helping students develop their skills as well as matching their need in wanting to see theory live through observables.

Table 1 (A) (B) (C): Top-ten words and their weights have been reported for all the topics found by LDA. An interpretation is provided.

Table 1 (A) - Topic 1		
Probability	***Words***	***Topic Meaning***
0.062	students	Students in the lab like that with time, they can remember what they learned from the lecture through repetition in the laboratory. **Manual Analysis themes**: Repetition, Remembering, Combining Lectures-Laboratory
0.028	lab	
0.020	want	
0.017	like	
0.012	give	
0.012	lecture	
0.012	repetition	
0.012	time	
0.010	remember	
0.009	also	

Table 1 (B) - Topic 2		
Probability	***Words***	***Topic Meaning***
0.070	*practical*	In practical work lessons, students need to see something, and with intervention [of members of staff] they can think about their experiences. **Manual Analysis themes:** Visualisation [of theory through observables], Support by members of staff, Hands-on, Reinforcement
0.025	*need*	
0.022	*see*	
0.018	*things*	
0.015	*something*	
0.014	*intervene*	
0.013	*think*	
0.010	*make*	
0.010	*done*	
0.010	*experience*	

Table 1 (C) - Topic 3		
Probability	***Words***	***Topic Meaning***
0.037	*work*	*With practical work, students develop their skills and learn. In order to know something, they need to understand it first.* ***Manual Analysis themes**: Skills, Reinforcement.*
0.036	*practicals*	
0.032	*skills*	
0.020	*learn*	
0.019	*learning*	
0.017	*go*	
0.016	*know*	
0.013	*wear*	
0.012	*understand*	
0.011	*first*	

4 CONCLUSIONS AND FUTURE WORK

Topic Modelling allows the grouping of words into topics, without providing a natural language interpretation of what the topic entails. However, by analysing the words along with their order which is specified by their respective weight, it was attempted to assign a meaningful interpretation to each topic. Hence, the proposed method is not entirely automatised but is suggested to be used as an identifier, a first step prior to conducting manual thematic analysis. The use of LDA and Topic Modelling allows the researcher to save time in reading and annotating large volumes of text by suggesting potential words whose interpretation has to be associated with a theme. The LDA model provides a probabilistic output based on the corpus provided therefore even though the interpretation of the words co-occurring is still subjective, the words allocated in each topic by the LDA remain unbiased.

A promising extension of this work is to enrich the data input with additional interviews such that a better-trained LDA model is obtained, and hence more meaningful topics are extracted. Moreover, another interesting extension of this work is to extend the current method by other Natural Language Processing (NLP) approaches to generate potentially meaningful sentences to support the researcher identify topic meanings and evaluate the quality of the topics found by the LDA model.

ACKNOWLEDGEMENTS

On behalf of the team, we express our sincere gratitude for the grant awarded to this project by the University of Lincoln Doctoral School, UK. This allowed for a creative, productive and flourishing interdisciplinary collaboration with inputs from the fields of educational sciences, computer sciences and computer engineering, something that contributed towards reaching the desired team goals for the scope of completing this project.

REFERENCES

[1] Ritchie J, Lewis J, Elam G. Designing and selecting samples in: Qualitative Research Practice. A Guide for Social Science Students and Researchers.

[2] Galanis P. Data analysis in qualitative research: Thematic analysis. Archives of Hellenic medicine. 2018 May 1;35(3):416-21.

[3] Roberts K, Dowell A, Nie JB. Attempting rigour and replicability in thematic analysis of qualitative research data; a case study of codebook development. BMC medical research methodology. 2019 Dec;19(1):1-8.

[4] DeCuir-Gunby JT, Marshall PL, McCulloch AW. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. Field methods. 2011 May;23(2):136-55.

[5] Daniel C. Thematic exploration of YouTube data: A methodology for discovering latent topics. Muma Business Review. 2019 Feb 13;1:141-55.

[6] Blei DM. Probabilistic topic models. Communications of the ACM. 2012 Apr 1;55(4):77-84.

[7] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.

[8] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc." 2009.

[9] Rehurek R, Sojka P. Gensim - Python Framework for Vector Space Modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic. 2011;3(2).

[10] University of Huddersfield. Themes and codes [Internet] [cited 2023 January 13] Available from: https://research.hud.ac.uk/research-subjects/human-health/template-analysis/technique/themes-and-codes/\

[11] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces 2014 Jun (pp. 63-70).