Are we There yet? Thematic Analysis, NLP, and Machine Learning for Research

Elena Fitkov-Norris, Nataliya Kocheva
Kingston University, Kingston-Upon-Thames, UK

e.fitkov-norris@kingston.ac.uk
n.e.kocheva@kingston.ac.uk

Abstract. Thematic analysis is a well-established technique for qualitative analysis which is covered in traditional research methods training. The objective of thematic analysis is to elicit themes and significant topics from discursive data such as free style discussions and semi structured or unstructured interviews or comments. The approach is laborious and time consuming and requires a significant input from researchers for identifying and coding the themes although software tools such as NVivo, T-Lab and IRaMuTeQ can aid with results presentation. Recent developments in Machine Learning (ML) and Natural Language Processing (NLP) have boosted interest in text analytics and its applications to social science research. For example, automatic topic identification using ML NLP offers valuable insights in social media analytics. However, machine learning techniques conventionally rely on large data sets to enable the algorithm to elicit themes. More recent research efforts have turned to the performance of machine learning approaches with smaller data sets.

This study aims to compare and contrast the effectiveness of Machine Learning NLP vs human generated themes using the text analytics tools NVivo, T-Lab, IRaMuTeQ, as well as the low-code ML tool KNIME for automatically eliciting themes from academic literature review in the contexts of service operations management research and semi-structured customer interviews. Results indicate that the ML NLP approach has the potential to automatically detect research themes even with small data sets, although the results vary across the different tools and are dependent on the capabilities of the built-in text analytic algorithms. In particular, T-Lab offered the best mapping of machine learning derived topics to researcher themes, and KNIME proved the most robust software, able to derive meaningful topics even with very small sample sizes. The implications for training research students are also significant as they suggest that the inclusion of ML NLP tools and algorithms in the training curriculum of social scientists may be beneficial.

Key words: Thematic Analysis, NLP, Machine Learning, Qualitative Data Analysis, Comparative Review

1. Introduction

Qualitative data analysis (QDA) methods have played a significant role in research over the past few decades (Creswell, 2014), and the tools available for such analysis have been expanding rapidly (Meyer and Avery, 2009). Nowadays, text-analytics software has emerged as a valuable resource that enhances researchers' capabilities and enables more efficient quantification of qualitative data. Among the various techniques employed, Thematic Analysis (TA) stands out as a widely utilized approach, known for its ability to identify, analyse, and report patterns or themes within data (Braun and Clark, 2006). The objective of thematic analysis is to elicit themes and significant topics from discursive data such as free style discussions and semi-structured or unstructured interviews or comments. This approach allows researchers to delve deeper into the underlying meanings and concepts present in the data, moving beyond surface-level observations and exploring the interconnectedness of ideas. By employing coding and thematic analysis techniques, researchers can uncover rich insights and patterns that contribute to a more comprehensive understanding of the research topic. However, the approach is laborious and time consuming and requires a significant input from researchers for identifying and coding the themes although software tools such as NVivo, T-Lab and IRaMuTeQ can aid with results presentation.

Recent developments in Machine Learning (ML) and Natural Language Processing (NLP) have boosted interest in text analytics and its applications to social science research. For example, evidence suggest that the addition of automatic topic identification using ML NLP from social media analytics improves the predictive power of healthcare surveillance systems (Gupta and Katarya, 2020). However, machine learning techniques conventionally rely on large data sets to enable the algorithm to elicit meaningful themes. More recent research efforts have turned to the performance of machine learning approaches with smaller data sets and have shown promising results with as few as 300 examples in supervised machine learning (Riekert, Riekert and Klein, 2021). However, the majority of adopt an exploratory approach which is handled via unsupervised ML algorithms. The identification of themes in a text using NLP is an emerging exploratory approach which works well with large data sets (Davidson *et al.,* 2021; Kim *et al.,* 2022) but limited research has explored how unsupervised text-analytics approaches fit into the traditional landscape of qualitative methods of which traditionally deal with smaller data sets (Firmin *et al.,* 2017). As a result, the application of machine learning text analytics tools in the field of education remains relatively unexplored (Spector and Ma, 2019). Given the potential for AI to

revolutionise education, the case for incorporating ML and NLP tools in the curriculum should also be considered.

To this end, this study aims to facilitate further understanding of the effectiveness of unsupervised machine learning NLP tools for theme identification with smaller data sets and the implications for inclusion of ML and NLP in the curriculum. The research examines and compares the effectiveness of the bult in automatic topic identification engines of several social science software tools, namely NVivo, IRaMuTeQ, T-Lab and KNIME to human generated themes across two data sets. The data sets represent two common use cases in social science research: an academic literature review in the contexts of service operations management and semi-structured customer interviews discussing the attitudes of shoppers in a department store.

The paper starts with a brief introduction to thematic analysis, followed by brief outline of the analytical steps in a machine learning natural language processing (NPL) algorithm, before discussing the methods and data sets used in our analysis. A presentation of the analytical output from each tool is followed by a discussion of the results and conclusions are drawn as to the state of art of NLP and machine learning. Finally, recommendations for curriculum design and future research are highlighted.

2. Thematic analysis (TA)

Thematic Analysis (TA) is a QDA method used for identifying patterns or themes of meaning within qualitative data such as interview transcripts, media, focus groups based on the main themes within their research questions. The approach accommodates both small and large datasets and is versatile as it can be utilised for both theory-driven or data-driven analyses (Braun and Clarke, 2006). As pointed out by Guest, MacQueen, and Namey (2012), TA extends beyond simply counting the most frequent words or phrases in a qualitative data set but aims to identify both implicit and explicit ideas within the data and uses codes or themes to analyse and compare their co-occurrence, as well as to display relationships among them.

According to Braun and Clarke (2006) TA includes six phases: 1) Data Familiarization: requires the researcher to read the texts and generate a codebook, based on the research questions; 2) Manual Code Generation: based on the previously generated codebook and research questions; 3) Themes identification: grouping the manually generated codes; 4) Themes review: refine codes; 5) Themes naming and 6) Report. The steps are normally employed by researchers as a cycle of inductive development of themes and their subsequent inferential application to the existing data, while assessing their significance and validity (Strauss and Corbin, 1998). According to Braun and Clark (2006), TA can adopt inductive or deductive-theoretical approach in the search of the why's and how's. TA approaches are predominantly influenced by the typology proposed by Clarke et al. (2019), which highlights three key aspects: coding reliability, reflective practice, and the use of a codebook to facilitate transparency and replicability. However, despite the guidance of codebook rules, thematic analysis, can be challenging, especially for novice researchers, as they grapple with understanding the effectiveness of their coding techniques and whether they align with their research focus (Fonteyn *et al.,* 2008). The process of theme identification in traditional qualitative research methods is by default subjective and related to the experience, understanding and knowledge of the researcher, and may lead to issue with the replicability and the reliability of the analysis (Anderson *et al.,* 2019). Variations to this methodology have been suggested to attempt to minimise the subjective researcher bias (Milles and Huberman, 1994)

There is growing awareness that text analysis methods can benefit from the adoption of computer software as a support tool. For example, Hwang (2008) argues that the adoption of software during the analysis can enhance transparency and replicability by providing a structured framework for analysis as well as facilitating the organization and management of data. Consequently, it is more inductive, and data driven in comparison to TA. Furthermore, research tools can save time, particularly when working with a large dataset or collaborating in team setting. However, it is important to note that while text analysis software offers valuable support, it does not replace the need of critical thinking and researcher involvement in the analytical process (Fielding, 2002). The debate is even more pertinent with the emergence of Generative Pre-trained Transformer tools such as OpenAI's ChatGPT which has taken the research community by storm. So far, ChatGPT has demonstrated promising capabilities in generation texts in various domains, including computer code generation, text translation and interpretation (Sobania *et al.,* 2023; Tate *et al.,*2023), Although to the best of our knowledge, ChatGPT's capabilities as a thematic identification tool have not been explored, its potential to disrupt the existing research ecosystem is significant.

3. Machine Learning (ML) and Natural Language Processing (NLP)

Machine learning approaches attempt to identify topics in a set of texts automatically, following a predetermined set of text processing algorithms. Thus topic derivation is purely data driven and independent of the specific research questions. The approaches require significant text pre-processing to reduce the dimensionality of the task (Step 1) and the deployment of ML approaches such as association and clustering, to identify common word patterns, which are then grouped into topics (Step 2).

Step 1: Pre-processing and data cleansing. The pre-processing step involves removal of punctuation, numbers, capitalisation, short and stop words such as "the", "a" "and" as well as tokenisation, which assigns part of speech (POS) label for each word, and lemmatization, which identifies the lemma of a term by removing inflections.

The objective of the pre-processing step is to identify the set of words, referred to as a Bag of Words (BOW) that are used in the documents. The BOW may be processed further for topic identification (see Step 2), used for word frequency analysis, or presented in a word cloud.

Step 2: Word embedding (coding) and modelling using topic detection algorithms or language models, such as LDA, skip-gram or continuous bag of words (CBOW) to elicit topics or predict the next word in a sentence. The objective of this step is to encode the text (also referred to as embedding) into word vectors to enable the language models to learn the text. Language models predict and learn by calculating conditional probabilities of words occurring together, and word vectors facilitate the probability calculations. Embedding involves the representation of the text at word, sentence, paragraph or even document level into a set of vectors that can be learned by the NLP models (Mikolov *et al.,* 2013). Table 1 shows a simple representation of the sentence "I love shopping at KDS".

Table 1. One-hot encoding for the sentence "I love shopping at KDS"

	1	2	3	4	5
I	1	0	0	0	0
love	0	1	0	0	0
shopping	0	0	1	0	0
at	0	0	0	1	0
KDS	0	0	0	0	1

Each of the words in the table are represented by their respective vectors. For example, the word "shopping" is represented by the vector {0,0,1,0,0}. One advantage of word vector presentations is that one can identify relationships between words such as "London" is to the "UK" as "Tokyo" is to "Japan" and "Paris" is to "France" to represent the relationship of capital city. See Mikolov et al. (2013) for a comprehensive overview.

The most popular Topic Modelling algorithm is LDA, which stands for Latent Dirichlet Allocation (Blei, Ng, and Jordan, 2003).

- *Latent* reflects that the algorithm can detect "hidden" topics. The texts are processes and the LDA algorithms identifies topics from within the texts that are not known in advance.
- *Dirichlet* stands for the Dirichlet distribution the model uses as a prior to generate document-topic and word-topic distributions.
- *Allocation* allocates the latent topics identified in the analysis to each text.

4. Method

The study adopts a comparative approach utilising two small data sets to test the automated machine learning capabilities of four text analysis software tools: NVivo, T-Lab and IRaMuTeQ and KNIME. The results compared the capacity of each tool to generate a word cloud (showing the BOW from each corpus) and thematic/topic extraction, which reflect an output from step 1 and step 2 from the machine learning process, respectively. In the context of thematic analysis, the word cloud generation corresponds to phase 1 and 2 of Braun and Clarke's (2006) method, while the thematic/topic extraction covers phases 2 and 3.

The two data sets used for testing are two common data types that arise in social science research. The first data set consists of bibliographic entries gathered as part of a literature review using keywords in Scopus. The dataset consists of 203 highly relevant and influential publications on service operation management, between 2007

and 2020. The analysis focused on the abstract of each publication to extract the BOW and themes withing the corpus. The second data set is very small and consists of nine semi-structured interview transcripts from a convenience sample of female customers of Kingston Department store (KDS) conducted in 2005. Due to the small sample size, the interviews were initially analysed as a whole, and then structured per question as suggested by Troung (2022) to fully challenge the capabilities of each tool to handle small data. Only the themes extracted by Q2: "Why do you shop at KDS" were shown for illustrative purposes but the results were consistent across all interview questions.

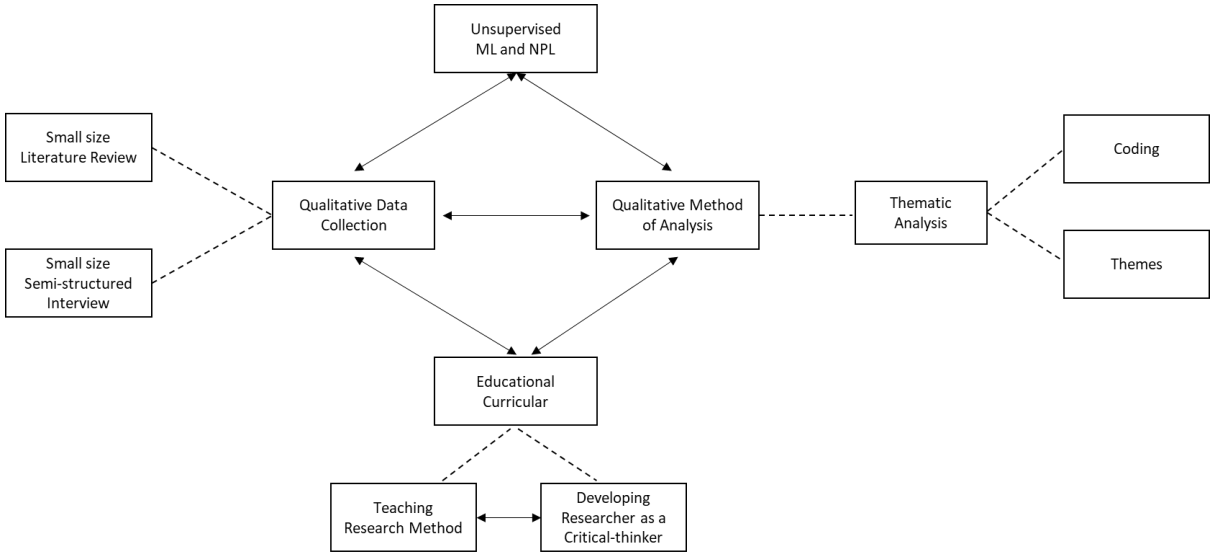

Figure 1: Mapping ML and NLP for empowering thematic analysis in educational curriculum

To ensure parity of comparison with machine learning, the capacity of NVivo, T-Lab, IRaMuTeQ and KNIME to automatically identify and group common themes, without any additional human input was tested. The word clouds were created with the top 100 most frequent words for the literature review and the 50 most common words for the interviews. The thematic analysis results were run with standard settings for each tool and were compared to the human generated themes from each data set as shown in Table 2 and Table 3 below. The coverage of human generated themes for each tool was reported and mean coverage as well as the coverage standard deviation and per theme coverage to allow for direct comparison between tools.

Table 2: Literature Review Researcher Themes and Subthemes

Themes	Sub-themes	Number	%
1.The Operation Managers and Service Strategy (TOMSS)		**61**	**35%**
	Managing Operations (MO)	30	17%
	Service Strategy (SS)	31	18%
2.Service Design and Delivery Service (SDDS)		**27**	**17%**
	Service Design (SDgn)	10	6%
	Service Delivery (SD)	17	10%
	Location and Layout (LL)	2	1%
3.Managing Service Operations (MSO)		**67**	**39%**
	Managing Capacity (MC)	7	4%
	Planning-Scheduling-Control (PSC)	8	5%
	Managing Service Inventory (MSI)	4	2%
	Managing Service Quality (MSQ)	15	9%
	Managing Service Supply Chain (SSC)	30	17%
4.Improving Service Operations (ISO)		**4**	**2%**
Reviews		**14**	**7%**

Table 3: KDS Interviews Researcher Themes and Coding Frame

	Coding Frame	
N:	Code Name	Description
Theme 1	Motivation for visiting	What drives this target segment into the store? This include various touchpoints/encounters, for example: facilities, product, service performance and the level of satisfaction/dissatisfaction of it.Sub-Themes Convenience,Product, Experience, Facilities (such as car park, coffee shop etc.), Service performance
Theme 2	Motivation for purchasing	The ins and outs of what makes this target group buying from KDS store? What is preventing them checking out more frequently? Sub Themes: Brands,Offers,Price,Quality of the Products,Range of the Products,Staff
Theme 3	Value for money	Not only economic value such as price, promotion but also emotional value that influence their buying decision, i.e. positive past experience, parent value in term of their input in the know/heritage; time as value in term of easy services such as having automated check out, returns etc. and the fact that is one-stop-shop. In general value here is treated as everything that affect the benefits and cost of offerings in the personas head.
Theme 4	Sophistication	Putting their money where the brand personality's is. Is there ongoing relationship with the store or with offered best-loved brands? Is there sense of loyalty, trust and satisfaction? How these target segment identify its social status? Sub Themes: Store Image, Designer collection, Aesthetics, Feeling important, Time (automated checkout)
Theme 5	Competitors	Other similar shops, marketplaces and online option used for purchasing any items Sub-Themes: Boutiques, Local market, Online, Similar shops, Specialist shops
Theme 6	Frequencies	Unveil important characteristics of the population's behaviour, alongside the duration of visits. Can help determin store loyalty Sub-Themes: Regular (twice or more per month), Monthly; Occasional (seasonal, special occasion-few times a year), Rare (one or two times per year)

5. Automatic Text Analytics Results

The world clouds generated for the Literature Review data by each tool are shown in Figure 2 through to Figure 5 below and the themes extracted, when possible, are shown in Figures 6 thought to Figure 9. Similarly, the world clouds generated for the whole Interview data and Q2 only from the Interview data by each tool, are shown in Figure 10 through to Figure 13 and the themes extracted, when possible, are shown in Figures 14 thought to Figure 17, respectively.

5.1 Text Analytics of Literature Reviews: Word Clouds

Figure 2: NVivo: Word Cloud Literature Review

Figure 3: IRaMuTeQ: Word Cloud Literature Review

Figure 4: KNIME: Word Cloud Literature Review

Figure 5: T-Lab: Word Cloud Literature Review

5.2 Text Analytics of Literature Reviews: Themes

Cannot be done automatically without user input.

Figure 6: NVivo: Themes Literature Review

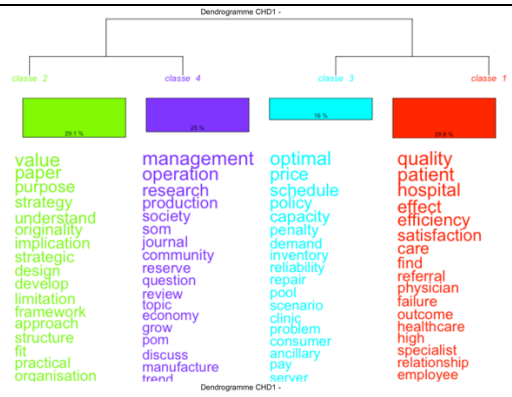

Figure 7: IRaMuTeQ: Themes Literature Review

	Topics Extracted
Topic 1	patient, time, system, model, cost, capacity, service, referral, wait, scheduling
Topic 2	service, strategy, system, paper, finding, study, process, performance, purpose, research
Topic 3	service, product, firm, provider, pricing, policy, price, customer, consumer, model
Topic 4	management, research, operation, service, operations, production, manufacturing, supply, risk, chain
Topic 5	service, paper, operation, research, purpose, study, management, process, finding, literature
Topic 6	quality, service, performance, customer, research, datum, effect, hospital, satisfaction, examine

Figure 8: KNIME: Themes Literature Review

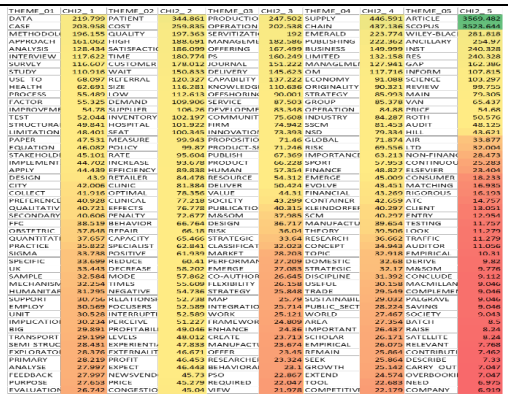

Figure 9: T-Lab: Themes Literature Review

5.3 Text Analytics of Whole Interview Data: Word Clouds

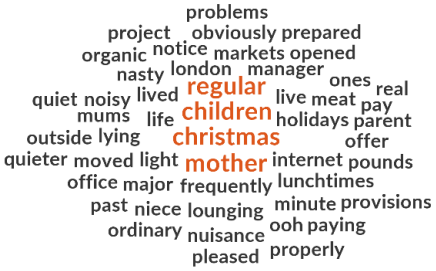

Figure 10: NVivo: Word Cloud Whole Interviews

Figure 11: IRaMuTeQ: Word Cloud Whole Interviews

Figure 12: KNIME: Word Cloud Whole Interviews

Figure 13: T-Lab: Word Cloud Whole Interviews

5.4 Text Analytics of Whole Interview Data: Themes

Cannot be done automatically without user input.

Figure 14: NVivo: Themes Whole Interviews

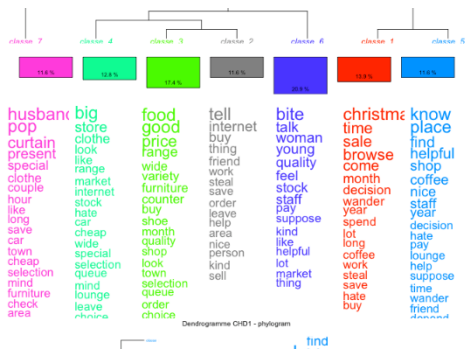

Figure 15: IRaMuTeQ: Themes Whole Interviews

	Topics Extracted
Topic 1	hat, west, summer, kitchen, polish, experienced, lady, courteous, polite, extremely
Topic 2	time, shop, stock, look, buy, clothes, feel, try, sometimes, sale
Topic 3	friend, we've, library, live, handbag, counter, store, odd, furnishings, occasion
Topic 4	buy, there's, shop, i've, quality, christmas, don't, nice, lot, staff
Topic 5	kds, hate, tuesday, specific, pop, park, feed, bit, nail-bar, somehow
Topic 6	curtain, talk, lounge, unless, car, able, store, plenty, finish, prescription

Figure 16: KNIME: Themes Whole Interviews

THEME_01	CHI2_1	THEME_02	CHI2_2	THEME_03	CHI2_3	THEME_04	CHI2_4
STOCK	62.097	SPEND	59.732	QUALITY	16.545	KDS	29.546
PRODUCT	36.255	LONG	32.782	GOOD	8.434	REASON	28.836
INTRODUCE	21.846	STORE	25.803	YEAR	8.114	INTERVIEW	25.702
CONSIDER	21.846	LOOK	20.229	PARK	7.2	CHOOSE	25.702
CHRISTMAS	17.529	MARKET	15.552	TILL	6.29	MIDDLE_AGED	21.383
NICE	15.965	INTERNET	11.526	QUEUE	6.29	JAN	21.383
RANGE	11.899		0	POLITE	6.29	RIVAL	21.383
KDS	8.969		0	VARIETY	6.29	LADY	16.205
	0		0	BROWSE	6.29	SHOP	13.777
	0		0	ASPECT	6.29	BUY	9.01

Figure 17: T-Lab: Themes Whole Interviews

6. Analysis and Discussion

6.1 Literature Review Results Analysis

The results show that all four tools can generate word clouds and reflect the most common words in the literature review data set (see Figure 2 through Figure 5), although the interpretability of word cloud results for IRaMuTeQ, KNIME and T-Lab is limited as the softwares do not colour code or remove most common terms automatically. NVivo, on the other hand, utilises colours and common words removal effectively to reflect the diversity of the literature on the topic (see Figure 2).

Three out of the four tools (IRaMuTeQ, KNIME and T-Lab) were able to auto identify topics within the literature review data set, without any human input. However, NVivo lacks this capability. Consulting the manual suggested that users need to manually code at least 10% of the inputs with themes before the automated coding could be enabled. This would require significant input from researchers, directly proportional to the sample size. The remaining three tools, T-Lab, IRaMuTeQ, and KNIME were able to highting themes consistent with the manually generated ones, such as service operation, service quality, strategy, and value (see Table 4).

Table 4: Literature Review Themes Results Analysis

Human Derived Themes	NVivo	IRaMuTeQ (4 topics)					T-Lab (5 topics)						KNIME ML (6 topics)						
for Literature Review		Topic 1	Topic 2	Topic 3	Topic 4	Total theme coverage	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Total theme coverage	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Total theme coverage
Theme 1: Operations Strategy	n/a	x	x			100%			x			100%		x		x		x	100%
Theme 2: Service Design & Delivery	n/a	x				30%		x	x			70%	x					x	70%
Theme 3: Operations Management	n/a			x	x	100%		x				100%	x		x			x	90%
Theme 4: Innovation	n/a					0%			x			75%							0%
Mean Coverage	n/a					**58%**						**86%**							**65%**
Coverage SD	n/a					**0.51**						**0.17**							**0.46**
Coverage (per topic used)	n/a					**15%**						**44%**							**13%**

T-Lab offered excellent coverage at 86% overall across just 2 topics. This was followed by KNIME at 65% and IRaMuTeQ, at 58%. The average efficiency of each topic for theme coverage is again highest in T-Lab and it is worth noting that both T-Lab and KNIME derived at least one topic which was not related to the human derived themes. The main issue, however, is that each software defaulted to different number of topics and thus although the three software tools can be used for thematic analysis, the results may be inconsistent, depending on the choice of software.

The results are encouraging as they suggest that ML tools may offer effective support to social science researchers conduction literature reviews in identifying themes from the research. However, the results are highly depended on the tool used and suggest that specialist tools (such as T-Lab) may offer an advantage in automated elicitation of meaningful themes over more generalist tools such as IRaMuTeQ and KNIME. However, interpreting the results poses challenges as the correspondence between themes and topics in not one-to-one across all software tested. This means that as a rule several topics contribute to each theme (for example, in IRaMuTeQ results, both topic 1 and topic 2 contribute to Theme 1), and/or each topic contributes to more than one theme (topic 1 in IRaMuTeQ contributes to both Theme 1 and Theme 2). This indicated that each software uses different criteria to the researcher to derive similar themes. This is perhaps unsurprising, but highlights the need to further research to enable researchers to find tune machine learning algorithms to group and identify topics based on the researcher's predefined criteria.

6.2 Interview Results Analysis

Working with the much smaller data set of semi-structured interviews, three out of the four software tools (IRaMuTeQ, KNIME and T-Lab) were able to produce word clouds that indicated aspects of the user experience such as range, product, quality, and staff (see Figure 10 through to Figure 13). However, the word cloud produced by NVivo offered little insight, suggesting that the software is less able to cope well with smaller data samples. The results from the topic extraction (see Table 5) show that the performance of the three tools able to complete the analysis, either stayed the same or deteriorated as the data sample size decreased. In fact, despite a marginal improvement in theme coverage (to 59%) IRaMuTeQ offered very little generalisation across the interviews by extracting 7 topics out of the 9 interviews which suggest that each individual interview is treated as a separate topic. This significantly limits the utility of the tool to meaningfully extract themes across interviews. As a rule, the granularity of topic coverage across each theme increased as the data size decreased, and each theme was described by higher number of topics and each topic also contributed to more themes. The results suggest that ML is not yet able to handle small data samples and elicit meaningful topics/themes automatically.

Table 5: Whole Interview Data Themes Results Analysis

Human Derived Themes	NVivo	IRaMuTeQ (7 topics)								T-Lab (4 topics)					KNIME ML (6 topics)						
for Interviews		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Total theme coverage	Topic 1	Topic 2	Topic 3	Topic 4	Total theme coverage	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Total theme coverage
Theme 1: Motivation for visiting	n/a	x	x			x		x	75%	x		x		100%	x	x	x		x	x	50%
Theme 2: Motivation for purchasing	n/a	x	x	x				x	50%	x		x		75%		x	x	x	x	x	50%
Theme 3: Value for Money	n/a			x		x	x	x	50%	x		x		50%	x	x	x	x			50%
Theme 4: Sophistication	n/a	x				x		x	25%					0%		x			x		25%
Theme 5: Competitors	n/a		x		x				80%		x		x	70%			x				25%
Theme 6: Visit frequency	n/a	x					x		75%	x		x		75%	x	x	x	x	x		25%
Mean Coverage	n/a								**59%**					**62%**							**38%**
Coverage SD	n/a								**0.22**					**0.35**							**0.14**
Coverage (per topic used)	n/a								**9%**					**16%**							**7%**

It is worth noting that the interviews were analysed a whole rather than per question which may have adversely affected the results. Empirical evidence has shown that breaking up semi-interviews and analysing per question rather than as whole may lead to better thematic analysis results (Troung, 2022). However, this reduces the data sample size even further and IRaMuTeQ and T-Lab were not able to complete the analysis and returned a matrix identification error.

KNIME proved to be the most robust software under small sample size conditions when data were presented per question (Q2 – "Why shop at KDS") and was able to derive four topics. The themes identified align meaningfully with theme 1, and 2 from the coding book (motivation for visiting and motivation for purchasing) (see Table 6). However, although the topics aligned fully with the human derived themes, the topics were fragmented across the themes making it difficult to interpret the results in the context defined by the researcher. This result confirms previous findings by Troung (2022), that analysis of semi structured interviews carried out on a question per question basis may offer advantages when utilising ML for automatic theme detection. This

may be in part because questions are framed in a particular context and ML algorithms are optimised to work well within specific context rather than across different contexts. The granularity of the results mapping of themes to topics and the resulting challenges in interpretation reiterate the finding that further research is required as to the best approach to facilitate the ability of researchers to pass topic definitions to machine learning algorithms, in order to improve their automated topic identification accuracy.

Table 6: Interview Data Themes Results Analysis of Q2 Why do you shop at KDS?

Human Derived Themes	NVivo	IRaMuTeQ	T-Lab	KNIME ML (4 topics)						Topics Extracted
Q2				Topic 1	Topic 2	Topic 3	Topic 4	Total theme covereage	Topic 1	shop, able, time, feel, staff, bit, stock, park, nice, sale
Theme 1: Motivation for visiting	n/a	n/a	n/a	x	x	x		100%	Topic 2	people, nice, tell, steal, warm, live, there's, off-hand, car-park, sell
Theme 2: Motivation for purchasing	n/a	n/a	n/a	x	x	x	x	100%	Topic 3	there's, food, lot, shopping, staff, suppose, coffee, they're, regular, sometimes
Coverage (per topic used)	n/a	n/a	n/a					**25%**	Topic 4	day, hat, you're, nice, suppose, rest, quiet, café, market, noisy

The results show that ML tools, can be effective in identifying topics that align reasonably closely with human derived themes. Thus, adoption of ML may allow researchers explore and understand qualitative data faster and with less ambiguity. This is particularly pertinent in the early stages of Thematic Analysis which require researchers to pre-examine the data, highlight potential points of interest, and assign labels. Adopting ML and NLP can help resolve coding inconsistencies and improve the transparency, reliability, and reproducibility of the analysis, particularly when working as part of team. Creating codebooks and thematic coding are complex and laborious processes, requiring extensive training and multiple coding revisions, even when following good practice guidelines and coding rules. ML and NPL tools can help automate these processes, minimising the need for manual input, and code creation and revisions.

However, overreliance on the ML tools and lack of understanding of their limitations may result in incomplete or misleading results. Moreover, de-contextualized features such keywords and word counts cannot replace the nuances captured as a result of immersion in the text and identification of latent or implicit representation of the concept under study. This is particularly true in the case of underrepresented constructs or emerging themes, which would be ignored by ML software as outliers. Therefore, there is need to offer social science researchers appropriate training in the capabilities and limitations of ML software and equip them with valuable skills that allow them to leverage ML tools to enhance their research practices.

7. Conclusions and Recommendations

In conclusion, ML NLP tools such as KNIME could provide effective support in identifying topics from small datasets, which align with human derived themes and thus can be utilized as support tools in thematic analysis. The results show that specialised software such as T-Lab can offer advantages in eliciting themes from academic abstracts with at least 200 observations, while generalist ML NLP software such as KNIME are robust even with much smaller data samples. The paper contributed to furthering our understanding of the effectiveness of ML and NLP tools in social science research with small data sets and highlighted the need for effective research skills training and development for qualitative researchers. Future research could focus on further examination of the limitations of software tools and the potential disruptive effect in the area of text analytics of Generative Pre-trained Transformer tools such as OpenAI's ChatGPT.

References

Anderson, R., Taylor, S., Taylor, T. and Virues-Ortega, J. (2021) "Thematic and textual analysis methods for developing social validity questionnaires in Applied Behaviour Analysis", *Behavioural Interventions*, 37(3), pp. 732–753. doi:10.1002/bin.1832.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3, pp.993-1022.

Braun, V. and Clarke, V. (2006) "Using thematic analysis in psychology", *Qualitative Research in Psychology*, 3(2), pp.77-101

Clarke, V., Braun, V., Terry, G. and Hayfield N. (2019) *Thematic analysis*. In Liamputtong, P. (Edn.) Handbook of research methods in health and social sciences (pp.843-860). Springer. https://doi.org/10.1007/978-981-10-5251-4_103.

Davidson, J. E., Ye, G., Parra, M. C., Choflet, A., Lee, K., Barnes, A., Harkavy-Friedman, J. and Zisook, S. (2021) "Job-Related Problems Prior to Nurse Suicide, 2003-2017: A Mixed Methods Analysis Using Natural Language Processing and

Thematic Analysis", *Journal of Nursing Regulation*, *12*(1), pp.28-39. https://doi.org/10.1016/s2155-8256(21)00017-x

Fielding, N. G. (2002) *Interviewing*. London. Thousand Oaks. CA: Sage.

Firmin, R.L., Bonfils, K.A., Luther, L., Minor, K.S. and Salyers, M.P. (2017) "Using text-analysis computer software and thematic analysis on the same qualitative data: A case example.", *Qualitative Psychology*, 4(3), pp.201-210. https://doi:10.1037/qup0000050.

Fonteyn, M. E., Vettese, M., Lancaster, D. R. and Bauer-Wu, S. (2008) "Developing a codebook to guide content analysis of expressive writing transcripts", *Applied Nursing Research, 21*(3), pp.165–168. https://doi.org/10.1016/j.apnr.2006.08.005

Guest, G., MacQueen, K.M. and Namey, E.E. (2012) *Applied Thematic Analysis*. 1st edn. Los Angeles: Sage.

Gupta, A. and Katarya, R. (2020) "Social media based surveillance systems for healthcare using machine learning: A systematic review", *Journal of Biomedical Informatics*, *108*, pp.103500. https://doi.org/10.1016/j.jbi.2020.103500.

Hwang, S. (2008) "Utilizing qualitative data analysis software a review of atlas. ti.", *Social Science Computer Review*, 26, pp.519-527.

Kim, K., Ye, G. Y., Haddad, A. M., Kos, N., Zisook, S. and Davidson, J. E. (2022) "Thematic analysis and natural language processing of job-related problems prior to physician suicide in 2003–2018", *Suicide and Life-Threatening Behaviour*, *52*(5), pp.1002–1011. https://doi.org/10.1111/sltb.12896.

Meyer, D. Z. and Avery, L. M. (2009) "Excel as a Qualitative Data Analysis Tool", *Field Methods*, 21(1), pp.91-112. https://doi.org/10.1177/1525822X08323985.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013b) "Distributed Representations of Words and Phrases and their Compositionality", *arXiv*. https://doi.org/10.48550/arxiv.1310.4546.

Miles, M.B. and Huberman, A.M. (1994) *Qualitative data analysis: an expanded sourcebook*. Sage Publications.

Riekert, M., Riekert, M. and Klein, A. (2021) "Simple Baseline Machine Learning Text Classifiers for Small Datasets", *SN Computer Science*, *2*(3), pp.178. https://doi.org/10.1007/s42979-021-00480-4.

Sobania, D., Briesch, M., Hanna, C. and Petke, J. (2023) "An analysis of the automatic bug fixing performance of ChatGPT", arXiv. https://doi.org/10.48550/arXiv.2301.08653.

Spector, J.M. and Ma, S. (2019) "Inquiry and critical thinking skills for the next generation: From Artificial Intelligence back to human intelligence", *Smart Learning Environments*, 6(1). https://doi:10.1186/s40561-019-0088-z.

Tate, T. P., Doroudi, S., Ritchie, D. and Xu, Y. (2023) "Educational research and AI-generated writing: Confronting the coming tsunami", *EdArXiv*. https://doi.org/10.35542/osf.io/4mec3.

Troung, D. (2022) *NVivo R1 (MAC) - coding purpose, coding process, and how to perform interactive coding in NVivoDothang*, *YouTube*. Available at: https://www.youtube.com/watch?v=HoLjzktKh40&list=PLRDHFVPFwalDu6EKuq4MXCvMjPztD840D&index=10 (Accessed: 06 June 2023).