Annotation Assistance for Thematic Analysis using Transfer Learning

Removed for Review

Abstract

Manual annotation of qualitative research data is costly and time-consuming. Recently, machine learning approaches have been introduced to assist such tasks. However, it remains challenging for a machine learning model to incorporate context, data scarcity, data imbalance and other aspects in thematic analysis. We employed transfer learning, combining the pre-trained models BERT and ResNet to propose an annotation assistance model and evaluated its accuracy and efficiency for semi-automatic annotation. We experimented on a dataset of focus group discussions between researchers and participants on perception towards robots in public spaces. We tested various training methods, including few-shot learning, data augmentation, and the use of different data modalities, to evaluate the impact of dataset size, data balance, and data modality on the proposed annotation assistance model's performance. The best-performing model achieved an average balanced accuracy of 59.89% for predicting thematic labels in researcher sentences and 48.67% for participant sentences.

1 Introduction

Data annotation is an essential task in machine learning and data-driven analysis. Except for unsupervised learning, developing a machine learning model often relies on high-quality datasets, which ideally include large, diverse and representative data along with accurate annotation. However, datasets collected in naturalistic and uncontrolled conditions often suffer from data imbalance issues and require intensive and costly manual data labelling labour. This is especially prevalent in the analysis of qualitative research data, which focuses on distilling subjective and high-level themes emerging in less structured data and currently relies on time-consuming manual annotation approaches.

A majority of data annotation tasks focus on objective labels developed for quantitative research, such as object bounding boxes for computer vision, or lower-level information, such as part-of-speech tagging for natural language processing. In contrast, there have been relatively limited labelled datasets for text comprehension at the sentence, paragraph, or higher semantic levels (Jain et al., 2020). However, such high-level, contextualised information is precisely what is required in qualitative research.

Qualitative research is critical in generating comprehensive understanding. In addition to obtaining users' thoughts and expectations, qualitative research allows researchers to use the data to improve the user experience (Søraa et al., 2023). For example, it can help improve aspects of a product that users perceive as failing (Horstmann and Krämer, 2019). A common qualitative research method is thematic analysis, in which thematic labels of data are developed by researchers based on the research themes and contextual analysis of qualitative data such as interviews. Existing thematic analysis approaches remain largely dependent on manual annotation, which is time-consuming and labour-intensive. Additionally, the diversity of topics in each qualitative research study imposes challenges for researchers to transfer thematic labels developed from one domain to another.

Prior work has shown that manually annotating the subjective labels of 369,436 reviews required a total of 14 weeks by 3 annotators (Qureshi et al., 2022). Likewise, qualitative research data labelling can also take several months, depending on the dataset's size (Stuckey, 2015). Motivated by reducing the labour costs associated with the annotation of qualitative data, we propose to use transfer learning methods to address the challenges of qualitative research data annotation. We demonstrated the feasibility of this approach using a representative dataset with ground-truth thematic labels and investigated the performance of transfer learning models on this dataset. This facilitates the future

development of semi-automatic annotation support to accelerate thematic analysis for qualitative research.

As qualitative research data is particularly affected by privacy and ethical issues, publicly available data is difficult to acquire. The scarcity of relevant data leads to a challenge for developing models for automated labelling. Additionally, qualitative research data is often not primarily aimed at training new machine-learning models. Therefore, such datasets are significantly smaller in size (Chen et al., 2018). Transfer learning is a common approach for tasks with insufficient data (Weiss et al., 2016). Pre-trained models that have previously learned patterns and features from large datasets can be fine-tuned to be applied to a new task and achieve better performance than learning from scratch. Considering the complexity of qualitative data and the benefits of having expert human annotators in the loop to generate high-level discoveries, we have decided to take the semi-automatic annotation method instead of fully automatic annotation for assisting thematic analysis. This allows researchers to correct labels (Mosqueira-Rey et al., 2023), thus ensuring annotation accuracy as well as efficiency (Desmond et al., 2021).

In addition to transfer learning, other practical solutions for addressing the challenge of limited data samples include zero-shot, few-shot learning and data augmentation (Parnami and Lee, 2022; Longpre et al., 2020; Kirk et al., 2023). Data augmentation helps to overcome the problem of class imbalance and increases a model's robustness (Feng et al., 2021). Similarly, feature extraction from multiple data modalities is shown to enhance model performance (Gandhi et al., 2023). Thus, we are motivated to explore whether these approaches can further contribute to semi-automatic annotation assistance for qualitative research.

In summary, this work contributes to qualitative data annotation and analysis by proposing a novel semi-automatic annotation assistance model using transfer learning, specifically:

- We assess the impact of data modalities, namely text, audio, and the multimodal combination of text and audio on using pre-trained models for semi-automatic thematic analysis.
- We evaluate the utility of textual data augmentation on qualitative research data to improve the performance of semi-automatic annotation.
- We investigate the feasibility of tuning a pre-trained model with a small number of data points, i.e., zero-shot and few-shot learning, for semi-automatic annotation of qualitative research data.
- We propose a machine learning model based on pre-trained BERT and ResNet for semi-automatic annotation and provide recommendations for future work.

2 Background

In qualitative research, coding is the process of analysing data, representing large volumes of text or speech data with concise single words or short phrases, and identifying the main topics covered in the data. As such, coding is crucial in qualitative data analysis as it allows researchers to quickly retrieve and categorize the data (Stuckey, 2015; Gillies et al., 2022). However, creating codes and labelling data can be very time-consuming. Several annotation tools exist to assist this task and improve annotation efficiency (John and Johnson, 2000; Banner and Albarran, 2009), such as Atlas.ti, NVivo, and MAXQDA. Nevertheless, these systems provide only an interface for manual data labelling, and researchers still need to develop the coding scheme and perform the annotation.

To increase labelling efficiency and quality, machine assistance has played an important role in data annotation tasks in recent years. For instance, INCEpTION (Klie et al., 2018) combines a recommendation system and can customise algorithms to choose the next data point to be annotated through active learning. AILA (Choi et al., 2019) incorporates machine learning models to predict important words in documents and highlights them. In addition to the NLP field, there are tools for annotating multimodal data, such as PEANUT (Zhang et al., 2023) and NOVA (Baur et al., 2020). The latter also proposes a Cooperative Machine Learning approach to track faces and skeletons in videos and identify people's emotions.

Machine-assisted methods have been applied recently to develop annotation platforms specifically designed for qualitative research. Cody (Rietz and Maedche, 2021) embeds supervised machine learning to assist users, allowing them to accept or reject

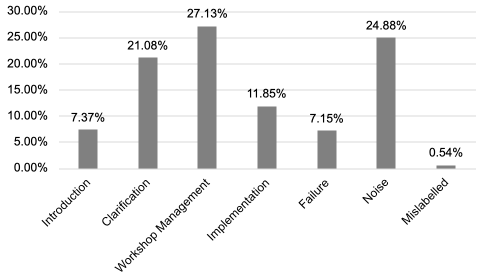
label suggestions generated by the system and iteratively update the model based on the users' choices. PaTAT (Gebreegziabher et al., 2023) uses a model to find semantically similar words to group data and provides users with similarity scores via a graphical user interface for reference. Both Cody and PaTAT use label recommendation systems and update models based on users' feedback to have a better predicting suggestion. However, both systems are only based on the similarity of sentences or frequently occurring words in sentences and cannot handle context understanding in natural language to aid higher-level thematic analysis.

Due to the scarcity of qualitative research data, coupled with the limitations faced by Cody and PaTAT, the use of large-scale language models and transfer learning can help master semantics and achieve good results on a small number of training samples (Ruder et al., 2019). Transformer-based BERT (Devlin et al., 2018) is a large language model for bidirectional and unsupervised language representations. It can autonomously learn from text without specific labels and has demonstrated outstanding performance in downstream NLP tasks, including high accuracy in detecting hate speech on social media (Mozafari et al., 2020) or applications in identifying fake news (Qasim et al., 2022), making it a popular pre-trained model option (Koroteev, 2021). Previous research has shown that combining BERT with zero-shot and few-shot learning methods can achieve good performance, even without further fine-tuning (Gupta et al., 2020). Thus, we are motivated to investigate applying BERT to support the thematic analysis of qualitative data.

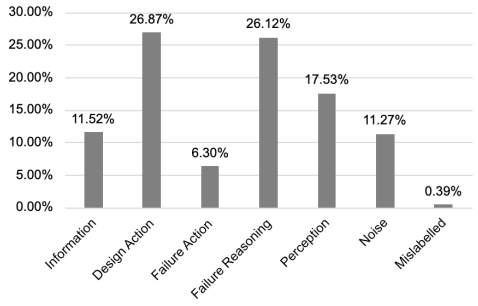
3 Methodology

3.1 Dataset

We employed a dataset collected in our previous work for exploring how transfer learning may benefit qualitative research, which contains audio-video recordings and transcripts of participatory design workshops on human-robot interaction (anon, 2020). These workshops encompass eight focus groups where researchers and participants engaged in discussion regarding the roles that robots could take in public spaces and active robot prototyping via Zoom video conferencing. The raw data comprises approximately 16 hours of video and audio recordings. ASR (Automatic Speech Recognition) transcription service by Amazon was used to generate textual transcripts with sentence timestamps for each workshop. These transcripts comprise 7,244 sentences, with 2,749 sentences contributed by participants and 4,495 sentences by researchers.



(a) Labels in Researcher sentences



(b) Labels in Participant sentences

Figure 1: Imbalanced sentence labels in the original dataset

This dataset has been manually annotated by researchers for thematic analysis, with a different set of sentence-level labels given to sentences spoken by participants ('P') or researchers ('R'). These sentences are categorised into 64 thematic labels, with 29 for 'R' sentences and 35 for 'P' sentences. These 64 labels are further aggregated into 11 thematic categories, 5 for 'R' and 5 for 'P', along with a 'Noise' label for disfluencies and other sentences irrelevant for thematic analysis. Each label and its corresponding thematic category were developed by researchers specifically for qualitative analysis of this data, defining their meanings and corresponding usage contexts.

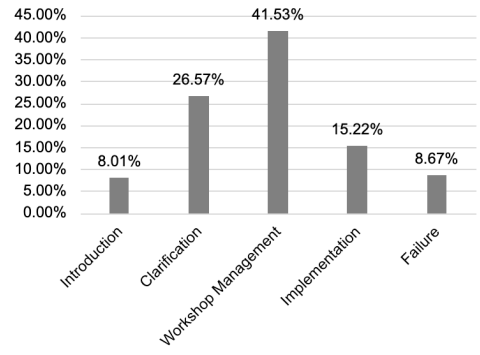
The proportion of sentences with each label is summarised in Figure 1. As shown here, the dataset contains imbalanced labels: for 'R' labels, 'Workshop Management' is the majority class with 27.13% of the sentences given this label; for 'P' labels, both 'Design Action' and 'Failure Reasoning' have a relatively large number of sentences, at 26.87% and 26.12% respectively.

3.2 Data Preprocessing

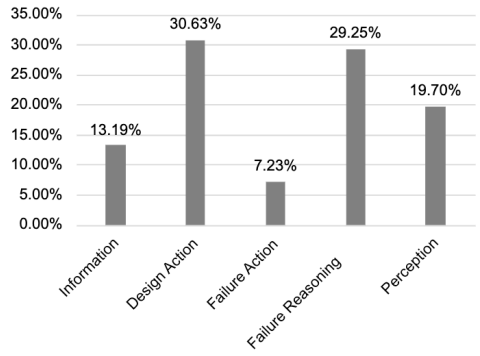
Textual Data There were a small number of labelling errors in the manual annotation, such as using a researchers' label for a participants' sentence. These mislabelled data points were excluded in our experiments. Additionally, data labelled as 'Noise' has been excluded. We prioritised the accurate identification of main theme categories as this is the most relevant for qualitative analysis.

We concatenated partial sentences produced by ASR in the generated transcript. For example, the sentence: "I'll just quickly share my screen" was split into three separate sentences in the transcript due to speech pauses: "I'll just quickly", "share", and "my screen". Therefore, we manually concatenated such sentences during preprocessing. Specifically, we concatenated two adjacent sentences with the same role (participant or researcher) and the same thematic label, provided their duration did not exceed 5 seconds.

Audio Data The original audio comprises complete recordings of the eight online meetings. We aligned the timestamps based on the processed text transcript to segment the recordings into sentence-level audio files. For each sentence, we extracted Mel-Frequency Cepstral Coefficients (MFCC) (Wang et al., 2016) using librosa (McFee et al., 2015) without re-sampling the original audio and saved the Mel spectrogram plots as the audio data.



(a) Researcher's sentences



(b) Participant's sentences

Figure 2: Label class distribution after preprocessing

Our preprocessing resulted in 2,273 'R' sentences and Mel spectrogram plots, 1,812 'P' sentences and Mel spectrogram plots. The distribution of all thematic labels after preprocessing is shown in Figure 2. Finally, we randomly split 'R' and 'P' data into training and test sets, with 80% for training and 20% for testing. We manually adjusted the test set to include a balanced number of samples for each thematic label and this test set was used in all the experiments.

3.3 Machine Learning Models

We developed machine learning models for three different modalities: text, audio, and multimodal models combining text and audio data.[1]

Text We utilised the pre-trained BERT-base tokenizer and text classification model from Hugging Face (Wolf et al., 2020), which has a 12-layer structure with 110 million parameters.

Audio As described in Section 3.2, audio features were represented as Mel spectrograms. For the audio model, we chose CNN (Convolutional Neural Network) architectures that demonstrate exceptional performance in visual classification tasks. More specifically, we employed ResNet models with residual layers, including the pre-trained ResNet50 and ResNet152 models (He et al., 2016), with weights set to 'IMAGENET1K_V2' sourced from TorchVision (maintainers and contributors, 2016). ResNet exclusively utilised spectrograms generated from audio files as its input.

Multimodal For the multimodal approach, we combined BERT and ResNet to extract text and Mel spectrogram embeddings separately, then concatenated them. We then add fully connected layers to predict the labels from the concatenated audio and textual features. The multimodal model can be further divided into two approaches: one using the pre-trained BERT from Transformers and pre-trained ResNet from TorchVision, and the other trained on our dataset with custom BERT and ResNet models.

[1] See https://github.com/LINK-REMOVED for source codes and a detailed description of the thematic label definitions and coding scheme.

3.4 Experiments

We conducted experiments to evaluate the performance of the thematic label classification model's performance across three aspects: Data Modality, Unbalanced Classes, and Available Number of Samples. Depending on the specific experiments, adjustments were made to the training set, while the test set remained the same for all models.

Baseline model: As shown in Figure 2, there were substantial differences in the distribution of each thematic label. We established a baseline classification model by predicting all test sentences as the majority label.

Text models: We compared different text-based models using different sizes of available training data, namely Zero-shot, 5-shot, 10-shot, 80% original text, and Augmented text models. All Text models were configured using a pre-trained BERT model and a pre-trained BERT tokenizer. Except for the Zero-shot model, the remaining models used the same parameters during training. The tokenized text input had a length of 256, the batch size was set to 16, and the learning rate was 0.00002 (2e-5). The training process was carried out over 20 epochs. The set of labels to predict was known to all models.

The *Zero-shot* model was not fine-tuned with samples from the training set and solely relied on the pre-trained BERT for classification on the test set.

The *Few-shot* models included 5-shot and 10-shot models, where five and ten samples, respectively, were selected from the training set for each of the thematic categories in each of the 'P' and 'R' data to fine-tune the model.

The *80% original text* model used all of the available training data (80% of the whole dataset) to fine-tune the pre-trained BERT model.

The *Augmented text* model used augmented training data with artificial training samples generated for the smaller thematic categories to create a balanced training set. Specifically, we used TextAttack (Morris et al., 2020) for text data augmentation, which generate paraphrased sentences of the original samples. Using the number of samples in the majority class as the reference, we generated sentences for the remaining classes for them to contain a relatively equal number of samples as the majority class. For example, the 'Workshop Management' label had the highest number of training samples in 'R', with 763 samples, while the 'Failure' label had only 157 samples. We performed data augmentation on these 157 samples, generating an additional 4 Augmented text samples for each original sample. This resulted in a total of 785 training samples for the 'Failure' label. We then randomly selected 606 samples from the Augmented text samples and combined them with the original 157 samples so that the size of the augmented set of data with the 'Failure' label matches the size of the 'Workshop Management' label.

Audio models: We employed pre-trained ResNet50 and ResNet152 for the audio models which predicts thematic labels based on audio information represented visually as the Mel spectrogram plots for each sentence time-aligned with the text model inputs. Similar to the Text models, we used a batch size of 16 and trained the models for 20 epochs, while the learning rate was set to 0.001.

Text+Audio multimodal model: The multimodal model combined both text and audio data as its inputs. The text and audio components had different learning rates same as the text-only and audio-only models. There outputs are then concatenated and passed through fully connected layers to generate the predictions. We developed two types of multimodal models:

1. The *multi-base* model used the pre-trained BERT and ResNet152 models without fine-tuning them on the training set.

2. The *multi-trained* model combined the best-performing fine-tuned models trained in the text-only and audio-only experiments.

Evaluation metrics: Due to the unbalanced classes, we employed the weighted F1-score and balanced accuracy metrics implemented with the scikit-learn library (Pedregosa et al., 2011) for performance evaluation. Balanced accuracy is computed as the average recall across all labels.

4 Results

4.1 Classification Performance Overview

Table 1 and Figure 3 provide an overview of all models' performance on the test set reported as balanced accuracy. As shown here, thematic analysis is a challenging task for text, audio, or multimodal models. For the specific dataset of focus group

Model	'R' labels	'P' labels
Baseline	20.00	20.00
Text		
Zero-shot	20.00	20.80
5-shot	30.91	27.12
10-shot	34.14	39.92
80% Original	58.06	**48.67**
Augmented	57.25	43.46
Audio		
ResNet152	42.00	26.59
ResNet50	41.53	28.12
Text + Audio		
Multi-base	**59.89**	48.33
Multi-trained	58.32	41.44

Table 1: Balanced accuracy (%) on the test set for all models.

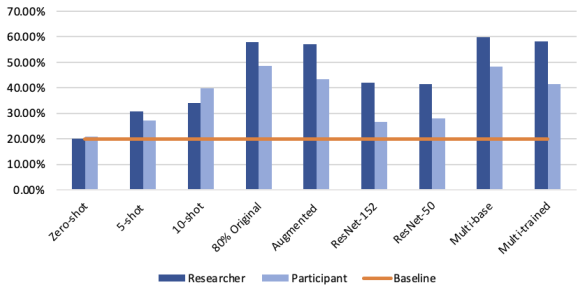


Figure 3: Balanced accuracy on the test set for all models.

discussion between researchers and participants that we experimented on, it is more difficult for the automatic models to predict thematic labels in participants' sentences than in the researchers' sentences, which may be due to the diversity in the topics discussed and the individual differences in speech styles and phrasing by participants.

Pre-trained models with the appropriate fine-tuning and modality fusion can achieve improved performance, thus having the potential to assist human annotators by suggesting auto-predicted labels when integrated into an annotation tool. However, human insights are still required and a semi-auto, collaborative annotation approach should be taken for thematic analysis. In Sections 4.2, 4.3, and 4.4 we will further discuss how data modality, class balance, and training sample availability influence the performance of thematic label prediction.

4.2 Influence of Data Modality

The performance of models using different data modalities for classifying each thematic label is shown in Table 2. Here we report the results of the multi-base multimodal model in the 'T+A' column. In the 'Audio' column, we reported ResNet152 results for 'R' and ResNet50 results for 'P'.

Label	Text	Audio	T+A
'R'esearcher labels			
Introduction	**43.90**	25.00	40.00
Clarification	70.23	55.17	**75.56**
Workshop Management	80.66	81.16	**86.32**
Implementation	40.00	29.51	**46.62**
Failure	48.78	21.74	**55.56**
'P'articipant labels			
Information	55.32	33.33	**60.38**
Design Action	**55.93**	33.33	51.79
Failure Action	**37.04**	12.50	34.78
Failure Reasoning	**51.85**	33.63	50.94
Perception	38.71	23.88	**41.18**

Table 2: Weighted F1-scores (%) of models using different data modalities with 80% of original data as the training set.

The Text+Audio model yields the best performance in predicting the 'R' labels except for the 'Introduction' class. The Audio model shows a significant performance gap compared to the Text model in predicting the 'R' labels, but combining the text and audio information yields benefits. For 'P' labels, the Text model achieved better performance in 3 out of 5 labels, while the multimodal model yielded better performance in predicting the other two thematic labels. Similar to the 'R' label results, the Audio model had worse performance when used alone, but contributed to predicting some labels when combined with the text data.

As shown here, similar to previous studies on data annotation, our work also indicates that utilising multimodal data is beneficial in transfer-learning assisted thematic analysis.

4.3 Influence of Unbalanced Classes

Table 3 reports the performance of text models trained with original data containing unbalanced classes and augmented data containing paraphrased sentences of original data to increase the number of samples in non-majority classes.

For 'R' labels, the model trained with balanced training data including paraphrases of original sentences achieved better results for all labels except for 'Introduction'. This improvement is particularly noticeable for the minority classes, such as the 'Failure' label. For the majority class 'Workshop Management,' the improvement is less sub-

Label	Origin.	Aug.
'R'esearcher labels		
Introduction	**43.90**	37.04
Clarification	70.23	**74.13**
Workshop Management	80.66	**82.11**
Implementation	40.00	**44.07**
Failure	48.78	**57.14**
'P'articipant labels		
Information	55.32	**62.22**
Design Action	**55.93**	50.39
Failure Action	**37.04**	35.29
Failure Reasoning	**51.85**	46.30
Perception	**38.71**	33.85

Table 3: Weighted F1-scores (%) of text models trained on original ("Origin.") and augmented training data ("Aug.")

stantial. For 'P' labels, using augmented training data did not yield improved performance except for the 'Information' label. The interesting performance difference in models using original vs. augmented training data for analysing themes in the researchers' or participants' sentences suggests that paraphrasing may be a more suitable data augmentation approach for qualitative data with more structure and fewer individual variances.

4.4 Influence of Available Training Samples

Table 4 reports the performance of text models with different amounts of training samples available for fine-tuning the pre-trained BERT model. 'Baseline' refers to always predicting the majority class.

For both 'R' and 'P' labels, the Zero-shot learning models (i.e., no fine-tuning) only predicted the majority class. This indicates that due to the significant difference in how thematic analysis is conducted for qualitative data compared to objective text classification and summarisation tasks, existing pre-trained language models are not directly applicable even when the thematic label sets are known. Therefore, a semi-automatic approach with close human guidance as opposed to a fully automatic approach is required in thematic analysis.

However, as soon as a small number of training samples become available (5-shot and 1-shot learning models), the pre-trained models can be fine-tuned to achieve better prediction performance. With more training samples available, including augmentation with paraphrasing, the models can achieve more accurate predictions and, thus potentially reduce the human annotator's workload.

5 Discussion

5.1 Transfer learning for thematic analysis

Our experiments demonstrate that transfer learning has the potential to support thematic analysis of qualitative data, especially when data augmentation and multimodal fusion are adopted in fine-tuning the pre-trained models.

Sample Size: In interviews and focus group discussions that qualitative data is usually collected from, the time researchers and participants spend on discussing each theme is often uneven, resulting in unbalanced classes for automatic prediction. Our experiments showed that text data augmentation using paraphrasing is a promising approach to the unbalanced data issue and for increasing the performance of labelling themes with less available data. However, data augmentation for text data is less explored compared to image data augmentation in computer vision. Paraphrasing often yields sentences without substantial differences from the original sentences, which results in the model learning features that are already present in the original data, unable to effectively discover new patterns to distinguish various themes. Thus, better methods need to be developed to create artificial training samples that are diverse and believable. Furthermore, while few-shot and zero-shot learning have shown promising results for text classification and summarisation tasks, our experiments indicate that thematic analysis is a more challenging task and is not suited for a fully automatic annotation pipeline.

Multimodal fusion: Our experiments indicate that fusing information from text and audio modalities can lead to better performance for thematic label prediction. Interestingly, the pre-trained multimodal models without fine-tuning achieved better performance than the models trained on our dataset. Further research is required to investigate additional improvements to the multimodal model, such as including video data or adopting other modality fusion architectures than simple concatenation.

Dataset specific influences: The models performed differently when predicting thematic labels for researchers and for participants. Overall, the models achieved better results for 'R' labels than 'P' labels. Several factors specific to the dataset we conducted our experiments on may have contributed to this difference. Firstly, the data was collected from focus groups with mul-

Speaker	Label	Baseline	Zero-shot	5-shot	10-shot	Origin.	Aug.
'R'	Introduction	0.00	0.00	18.60	10.26	**43.90**	37.04
	Clarification	0.00	0.00	15.00	44.00	70.23	**74.13**
	Workshop Management	59.01	59.19	57.33	69.27	80.66	**82.11**
	Implementation	0.00	0.00	22.99	20.00	40.00	**44.07**
	Failure	0.00	0.00	19.15	21.05	48.78	**57.14**
'P'	Information	0.00	0.00	25.93	47.06	55.32	**62.22**
	Design Action	51.64	51.64	38.71	37.62	**55.93**	50.39
	Failure Action	0.00	0.00	14.29	25.00	**37.04**	35.29
	Failure Reasoning	0.00	0.00	20.51	38.55	**51.85**	46.30
	Perception	0.00	0.00	**40.00**	35.90	38.71	33.85

Table 4: Weighted F1-scores (%) of text models with different sizes of available training data.

tiple researchers and participants in each session. Participants had active discussions with occasional overlapping speech, which posed difficulties for ASR transcription. Secondly, the same group of researchers followed a semi-structured approach to organise the discussion and asked a similar set of questions across the 8 focus group sessions. Participants, on the other hand, were more spontaneous and reflective in phrasing their discussion, and a different set of participants engaged in discussion on a diverse range of topics specific to each focus group session and their personal backgrounds. Thus, it is a more challenging task for the model to identify themes shared across such a diverse population and phrasing in participants' sentences.

5.2 Limitation and Future Work

We used ASR-generated transcripts, which can have limited accuracy compared to manual transcription, especially as the speakers did not wear close-up microphones. This may have limited the text models' abilities to identify representative features from the auto-generated transcripts. Moreover, the current models are evaluated on one specific dataset, and it would be beneficial to expand our evaluation to other qualitative research datasets to understand the robustness and generalisability of the models with cross-corpora analysis. Lastly, user studies are required to evaluate the difference in annotation efficiency and quality between manual annotation and semi-automatic annotation using the proposed model for assisting thematic analysis. For this, we plan to embed the proposed model in annotation tools, such as Label Studio (Tkachenko et al., 2020-2022), and compare the annotation time and quality differences between the manual and semi-auto annotation approaches.

6 Conclusion

We addressed the challenge of annotation assistance in qualitative research by investigating the efficacy of using pre-trained models with various fine-tuning approaches for thematic analysis. Specifically, we evaluated few-shot learning, data augmentation, and multimodal fusion considering three main aspects that can influence the thematic label classification performance: size of available training samples, modality fusion, and class balance. On a dataset of focus group discussions, the transfer learning model achieved a balanced accuracy of up to 59.89% for predicting a set of thematic labels, with weighted F1-scores of up to 86.32% for predicting individual labels. Our work demonstrates the potential of adopting transfer learning to support qualitative research and reduce the human annotator's workload in the complex task of thematic analysis via annotation assistance.

References

anon. 2020. Reference removed for anonymity.

Davina J Banner and John W Albarran. 2009. Computer-assisted qualitative data analysis software: a review. *Canadian journal of cardiovascular nursing*, 19(3).

Tobias Baur, Alexander Heimerl, Florian Lingenfelser, Johannes Wagner, Michel F Valstar, Björn Schuller, and Elisabeth André. 2020. explainable cooperative machine learning with nova. *KI-Künstliche Intelligenz*, 34:143–164.

Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst.*, 8(2).

Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila:

Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA. Association for Computing Machinery.

Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. 2021. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. 2023. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Marco Gillies, Dhiraj Murthy, Harry Brenton, and Rapheal Olaniyan. 2022. Theme and topic: How qualitative research and topic modeling can be brought together. *arXiv preprint arXiv:2210.00707*.

Aakriti Gupta, Kapil Thadani, and Neil O'Hare. 2020. Effective few-shot classification with transfer learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1061–1066, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Aike C Horstmann and Nicole C Krämer. 2019. Great expectations? relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology*, 10:939.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. Scirex: A challenge dataset for document-level information extraction. *arXiv preprint arXiv:2005.00512*.

Winsome St John and Patricia Johnson. 2000. The pros and cons of data analysis software for qualitative research. *Journal of nursing scholarship*, 32(4):393–397.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. 2023. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

MV Koroteev. 2021. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*.

Shayne Longpre, Yu Wang, and Christopher DuBois. 2020. How effective is task-agnostic data augmentation for pretrained transformers? *arXiv preprint arXiv:2010.01764*.

TorchVision maintainers and contributors. 2016. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Archit Parnami and Minwoo Lee. 2022. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, Abdulwahab Ali Almazroi, et al. 2022. A fine-tuned bert-based transfer learning approach for text classification. *Journal of healthcare engineering*, 2022.

Muhammad Aasim Qureshi, Muhammad Asif, Mohd Fadzil Hassan, Ghulam Mustafa, Muhammad Khurram Ehsan, Aasim Ali, and Unaza Sajid. 2022. A novel auto-annotation technique for aspect level sentiment analysis. *Comput Mater Contin*, 70(3):4987–5004.

Tim Rietz and Alexander Maedche. 2021. Cody: An ai-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Roger Andre Søraa, Gunhild Tøndel, Mark W Kharas, and J Artur Serrano. 2023. What do older adults want from social robots? a qualitative research approach to human-robot interaction (hri) studies. *International Journal of Social Robotics*, 15(3):411–424.

Heather L Stuckey. 2015. The second step in data analysis: Coding qualitative research data. *Journal of Social Health and Diabetes*, 3(01):007–010.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Wenbo Wang, Sichun Li, Jianshe Yang, Zhao Liu, and Weicun Zhou. 2016. Feature extraction of underwater target in auditory sensation area based on mfcc. In *2016 IEEE/OES China Ocean Acoustics (COA)*, pages 1–6.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3(1):1–40.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zheng Zhang, Zheng Ning, Chenliang Xu, Yapeng Tian, and Toby Jia-Jun Li. 2023. Peanut: A human-ai collaborative tool for annotating audio-visual data. *arXiv preprint arXiv:2307.15167*.