Topic Modelling for Focus Group Transcript Analysis: Comparing Model Families, Aggregation Levels, and Expert Judgment

Chao Yong Ng
Singapore University of Technology and Design
Singapore
chaoyong_ng@alumni.sutd.edu.sg
0009-0003-3434-8187

Andrew Z H Yee
Wee Kim Wee School of Communication and Information
Nanyang Technological University, Singapore
andrew.yee@ntu.edu.sg
0000-0002-7854-8792

Kwan Hui Lim
School of Business
Singapore University of Social Sciences
Singapore
khlim@suss.edu.sg
0000-0002-4569-0901

***Abstract*—The amount of text that researchers can code by hand limits the scale of focus group studies. We propose a semi-automated framework that identifies candidate themes in the transcripts for researchers to review and refine. At three aggregation levels, we compare four topic models: Latent Dirichlet Allocation (LDA), Word2Vec, Latent Feature LDA (LFLDA), and Sentence-BERT. Using 13 focus group transcripts from a study on mobile health technologies for older adults in Singapore, we evaluate the quality of the generated topics using automated coherence metrics and qualitative assessments by the principal investigators of the original study. Our results show that LFLDA achieves the best balance between coherence and interpretability on the 20 Newsgroups benchmark and three of the four focus group configurations. We trained each configuration once, so these comparisons need to be repeated with different random seeds. For LFLDA, the experts' average topic quality rating more than doubled when we grouped the transcripts by discussion theme. The rankings based on coherence often differed from those of the experts, suggesting that expert judgment should form part of model selection. From these results, we derive five recommendations for designing focus group analysis tools.**

***Index Terms*—focus groups, topic modelling, natural language processing, text mining, transcript analysis, decision support**

I. Introduction

Researchers in public health, user experience research, policy evaluation, and market research use focus groups to examine how participants understand social and behavioural issues [1]. Coding the transcripts accounts for much of the work after a focus group session. For example, grounded theory requires researchers to manually code the discussions a sentence at a time, then develop more general themes as they examine and re-examine the coded sentences, as a theme that emerges later may be applicable to earlier sentences, which is time-consuming [2].

NVivo, ATLAS.ti, and MAXQDA are examples of computer-assisted qualitative data analysis software (CAQDAS) used to record, annotate, and retrieve codes, while theme discovery remains a manual task. As such, transcript analysis for candidate topics is time-consuming and may not be replicable across different coders. Hence, an automated tool that proposes themes for researchers to examine and approve would speed up this manual task.

For unlabelled documents, words that frequently co-occur serve as an indication for topics, which serve as the basis for statistical models like Latent Dirichlet Allocation (LDA) [3] that are used to study news articles, academic abstracts, and survey responses. Other approaches use word embeddings [4, 5] or represent words along with their context using Transformer-based models [6, 7]. Many approaches use structured text like how focus group transcripts record the multi-turn interactions among multiple speakers. However, there are fewer studies on the results of different models and algorithms in more informal settings.

Another approach uses large language models (LLMs), such as GPT-4, for zero-shot thematic coding of qualitative data [8]. While LLMs are powerful, it may be challenging to explain how a particular theme appears or to replicate the results with another run of the LLM, and cloud-based LLMs also pose privacy concerns. On the other hand, traditional topic models provide intermediate results that researchers can examine and reproduce, such as word distribution for a topic and topic distribution for documents.

In this paper, we empirically compare four topic modelling families across three levels of transcript aggregation, using a semi-automated framework to support qualitative researchers in identifying candidate themes from focus group transcripts. The researchers review each proposed topic against their manual coding and use their domain knowledge to decide whether to retain, revise, or discard it. Topic model benchmarks typically compare coherence scores on curated corpora in which each document has a single author. In contrast, our study focuses on multi-party conversations, examining how different transcript aggregation levels affect topic modelling and how the generated topics are assessed by experts from the original study. We seek to identify themes in unlabelled focus group transcripts, where the discussion is unstructured and no reference labels are available for evaluation, based on the judgment of domain experts. Our main contributions

Accepted version. To appear in the Proceedings of the 2026 IEEE International Conference on Big Data (BigData), Phoenix, AZ, USA, December 2026.
© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

comprise the following:

1) An **empirical comparison of four topic modelling families** within a decision support framework in which researchers review and interpret the topics identified in focus group transcripts.
2) **Study-derived recommendations** for transcript aggregation and model selection, based on our evaluation of four corpus configurations from a single application domain.
3) **An analysis of agreement between coherence scores and expert judgments** that identifies where the two evaluations differ and what each contributes to the choice of a topic model.

We address three research questions:

RQ1: How do different topic modelling approaches (statistical, embedding-based, hybrid, and Transformer-based) compare in their ability to extract interpretable themes from focus group transcripts?

RQ2: How does the level of corpus aggregation (top-level, topic-level, stakeholder-level) affect the quality and utility of automatically extracted topics?

RQ3: How closely do automated coherence scores agree with experts' qualitative assessments, and what does this imply for the use of topic models to support research decisions?

The rest of this paper is organised as follows. Section II reviews related work, while Section III and Section IV describe the framework and experimental design, respectively. Section V reports the benchmark and focus group results, and Section VI discusses their implications and the study limitations. Section VII concludes the paper and outlines future work.

II. Related Work

A. Decision support for qualitative research

Decision support systems use text analytics to obtain information from collections of unstructured text. Coussement and Van den Poel [9] classified incoming customer emails by combining textual content with linguistic style features to support complaint management, while Pournarakis et al. [10] clustered large volumes of tweets into semantically coherent themes using an LDA-based model in order to inform brand awareness and brand meaning decisions. Both studies use text analysis to inform human decisions, which is also the purpose of our framework. Similarly, qualitative researchers may have interview transcripts, open-ended survey responses, and focus group discussions in quantities that make manual reading and coding difficult.

Human-in-the-loop machine learning involves an expert who reviews a system's suggestions and revises them where needed [11]. For a qualitative researcher, a list of suggested themes narrows the initial search. The researcher still has to decide what each theme means in the context of the transcripts and whether it needs revision.

B. Focus group methodology and analysis challenges

A focus group typically involves 6–12 participants in a structured group discussion guided by a moderator and a set of predefined topics [1]. Researchers often use grounded theory to code the transcripts [2], supported in part by participant comments that show how their views develop over the conversation. This also requires the researcher to be familiar with the discussion topic. Although involving additional coders can improve reliability, it can introduce new issues on agreement in terms of coding a passage.

Two further properties complicate analysis beyond the use of informal language and code-switching. First, interpreting one speaker's remark can require reading back several turns to see what the other participants were discussing. Second, as participants respond to one another, the conversation may move quickly between topics and depart from the order in the discussion guide.

C. Topic modelling: from statistical to neural approaches

Topic modelling is an unsupervised approach that identifies latent thematic structures in a collection of documents. The most widely used method is Latent Dirichlet Allocation (LDA), which models each document as a mixture of topics and each topic as a distribution over words [3]. LDA and related methods, including Probabilistic Latent Semantic Analysis [12], and Non-negative Matrix Factorization [13], use word co-occurrence statistics derived from a bag-of-words representation. This representation discards the order of words and the context in which each word occurs.

Some approaches combine word counts with distributed representations of the vocabulary. Latent Feature LDA (LFLDA), for example, uses Word2Vec embeddings in its topic-word distribution to account for semantic similarity as well as co-occurrence [4].

Transformer-based models use the surrounding words to represent a word in context, which can help distinguish its meanings in informal speech. Sentence-BERT produces an embedding for the whole sentence from these contextual representations [6, 7]. BERTopic [14] combines Sentence-BERT embeddings with clustering and class-based TF-IDF to produce topics. For short, informal social media text, ClusTop [15] derives topics by clustering word co-occurrence networks, with a subsequent extension incorporating word embeddings [16]. Deployed systems also use these models to identify events and topics in social data streams [17]. Neural models have performed well on standard benchmarks but remain less studied on informal conversations involving code-switching.

Castellanos et al. [8] investigated the use of LLMs for qualitative coding in health care research and found that LLMs can generate initial thematic summaries, although human verification is still required. Similarly, topic-focused summarisation methods that do not require labelled data [18] aim to generate concise summaries of text collections, which is closely related to the objective of topic modelling in our study.

D. Evaluation of topic models

Word co-occurrence measures allow automated comparison of topic models at scale. Among the commonly used coherence measures are C_V and C_{UCI} [19], defined in Section IV. Syed and Spruit [20] examined the performance of these measures across different document granularities. Other studies have found that coherence scores correlate with readers' judgments of topic quality in the particular settings they examined [19, 21].

The ranking obtained from an automated measure can differ from the ranking obtained from readers. Chang et al. [22] used a word-intrusion task and found that a model with higher held-out likelihood could produce topics that readers found harder to interpret. Coherence has a similar limitation, where frequent co-occurrence can give a topic a high score even when its top words repeat the same idea or express a theme of little substance. In addition, the analyst also needs to judge whether that theme bears on the research question.

Experts can rate a topic's interpretability, give it a label, or check how it relates to themes they have identified in the material. Although these judgments take longer to obtain than a coherence score, they could show how readers interpret the topic. We use both expert judgment and coherence scores to evaluate the topics.

E. Positioning of this study

Our study differs from previous work in three main aspects. First, we focus on multi-party focus group transcripts that contain informal language, code-switching, and conversational exchanges where the meaning of individual statements may depend on interactions between participants. Second, we investigate the effect of different transcript aggregation levels, namely top-level, topic-level, and stakeholder-level, on topic modelling performance using the same collection of transcripts. Third, we complement automated coherence metrics with qualitative assessments by the principal investigators of the original study to evaluate the usefulness of the generated topics for qualitative research.

III. Proposed Framework

Our modular pipeline processes conversational text to generate candidate topics, with researchers involved in choosing its configuration and evaluating the resulting topics.

A. Framework overview

Fig. 1 shows the overall architecture. The pipeline comprises three stages, namely (i) preprocessing, (ii) topic modelling, and (iii) clustering with keyword extraction. We use the same preprocessing and evaluation procedures for all four algorithms. Researchers choose the aggregation level (top, topic, or stakeholder) and the range of topic counts to examine. After modelling, researchers evaluate the quality and relevance of the generated topics. Researchers therefore set the scope of the model comparison and judge what the resulting topics mean in the context of the study.

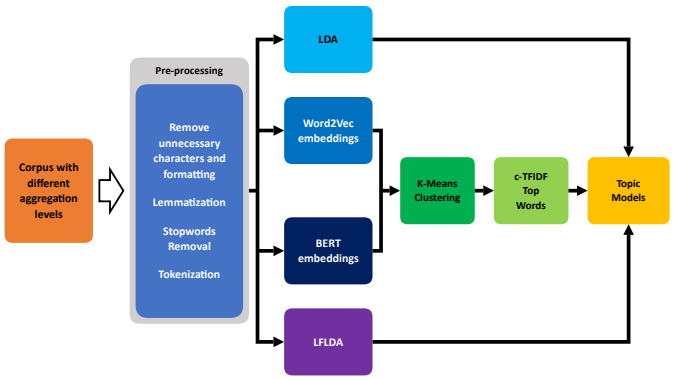


Fig. 1. Proposed framework for topic modelling of focus group transcripts. The three stages comprise transcript preprocessing, topic modelling with four algorithms, and clustering with keyword extraction.

B. Transcript aggregation

We examine the effect of different transcript groupings on the thematic information available to the models. We group the transcripts at the three levels shown in Fig. 2. At the **top level**, we pool all transcripts in one corpus. This gives the broadest coverage of themes, although mixing discussion topics can introduce noise. At the **topic level**, transcripts are grouped according to the discussion themes in the focus group guide, such as "Technologies for Health" and "Challenges with Technologies", to maintain thematic coherence within each corpus. At the **stakeholder level**, each topic-level corpus is further divided according to participant groups, namely seniors, volunteers, and staff of Senior Activity Centres (SACs). This results in smaller corpora that are more focused on specific themes and stakeholder groups.

With two discussion topics and three stakeholder groups, we obtain nine possible corpus configurations: one at the top level, two at the topic level, and six at the stakeholder level. Researchers can use these groupings to examine a theme across participants, focus on a particular stakeholder group, or consider the collection as a whole.

C. Preprocessing

We first preprocess the text before applying the topic modelling approaches. We tokenize the transcripts and use lemmatization to map grammatical variants such as "running" and "ran" to the same base form. We then discard common stopwords and tokens shorter than three characters. The final filter uses document frequency. We exclude words found in fewer than 10% of documents, where occurrences are sparse, and those found in more than 50%, where the words have little value for distinguishing documents.

D. Topic modelling approaches

For each corpus, we compare four model families through LDA, Word2Vec, LFLDA, and Sentence-BERT. These represent statistical modelling, word embeddings, a combination of statistical modelling and word embeddings, and contextual embeddings, respectively.

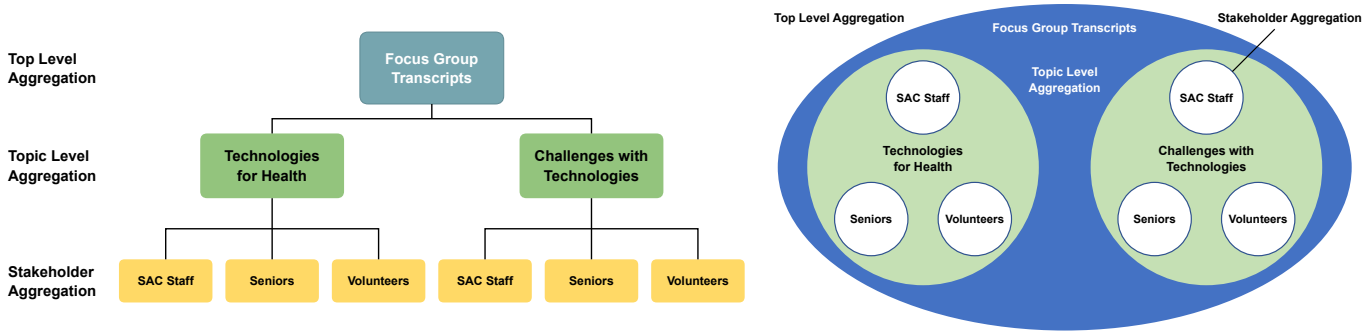


Fig. 2. Hierarchy of dataset aggregation levels (left) and set diagram showing the relationship between corpus configurations (right).

Latent Dirichlet Allocation (LDA) models a document as a mixture of latent topics and a topic as a probability distribution over words, using a generative probabilistic formulation [3]. LDA estimates document-topic and topic-word distributions from the observed words. The Dirichlet prior parameters are α for document-topic distributions and β for topic-word distributions. Smaller prior values result in sparser distributions, whereas larger values produce smoother distributions.

Word2Vec represents words using dense vector embeddings that capture semantic similarities based on word co-occurrence patterns. For both this model and LFLDA's latent feature component, we use the publicly available 300-dimensional vectors trained on Google News. Averaging the word vectors in each document gives us a document representation without having to learn a separate embedding for it [5]. We then group the documents into candidate topics by clustering these representations (Section III-E).

Latent Feature LDA (LFLDA) extends LDA by replacing the standard Dirichlet multinomial topic-word distribution with a two-component mixture, in which one component is the standard LDA multinomial and the other is a latent feature component based on Word2Vec embeddings [4]. A hyperparameter λ controls the balance between these two components. The resulting model combines LDA's representation of word co-occurrence with semantic similarity from Word2Vec. We use the same Google News vectors for its latent features.

Sentence-BERT uses the Transformer architecture to produce contextual sentence embeddings that capture the meaning of a sentence as a function of all its words and their positions [6, 7]. Unlike Word2Vec, which generates a fixed embedding for each word, Sentence-BERT produces contextual embeddings that capture the meaning of sentences based on their constituent words and surrounding context. We use a pre-trained Sentence-BERT model from the Sentence-Transformers library [7] to encode each document into a fixed-length embedding, which is then clustered to identify topics.

E. Clustering and keyword extraction

Word2Vec and Sentence-BERT produce document embeddings, which we convert into discrete topics through clustering and keyword extraction.

We use **K-Means clustering** to group semantically similar document embeddings into k clusters, each representing a candidate topic [23]. Varying the cluster count k during the experiments therefore varies the number of topics.

To find the terms that best describe each cluster, we use **class-based TF-IDF (c-TF-IDF)** [14]. We concatenate the documents in each cluster to form one pseudo-document per cluster and compute TF-IDF across the resulting collection. The terms with the highest scores for a cluster serve as the top words for that topic.

LDA and LFLDA estimate a distribution over words for each topic, so we can select the highest-probability words directly without computing c-TF-IDF. For Word2Vec and Sentence-BERT, we use identical values of k and the same procedures for clustering and extracting keywords. The choice of embedding is the only difference between these two pipelines. A comparison with LDA or LFLDA also involves a difference in how the topics are formed, since these models estimate topic-word distributions and the embedding methods group documents into clusters.

IV. Experimental Design

A. Datasets

Our evaluation uses two datasets for the four approaches. We begin with 20 Newsgroups, obtained through `fetch_20newsgroups` in `scikit-learn`, which contains approximately 18,000 posts in 20 categories, including `sci.space` and `rec.sport.hockey`. The training and test subsets are combined for the experiments. We use the benchmark to establish how the methods perform on structured text by individual authors before applying them to discussions with several speakers.

The second dataset consists of 13 focus group transcripts from a prior social science study on the role of mobile health technologies in the lives of older adults in Singapore. The study involved three participant groups (seniors, SAC volunteers, and SAC staff) and examined how seniors and community partners view the role of health technologies in their lives, together with the barriers and opportunities involved in developing these technologies. The transcripts were collected as part of an earlier independent study. In this work, we use

these existing transcripts for computational secondary analysis without collecting additional data from human participants. The discussion guide covered two broad themes ("Technologies for Health" and "Challenges with Technologies"), with different prompt questions for each stakeholder group, while the actual flow of conversation depended on the dynamics of each session.

B. Corpus configurations

Following the segmentation strategy in Section III-B, we selected four corpus configurations that cover all three aggregation levels:

1) **Focus Group Transcripts** (top-level): the complete set of all 13 transcripts.
2) **Challenges with Technologies** (topic-level): transcripts grouped by the "Challenges with Technologies" discussion theme.
3) **Challenges – SAC Staff** (stakeholder-level): the subset of "Challenges with Technologies" transcripts from SAC staff participants.
4) **Challenges – Volunteers** (stakeholder-level): the subset of "Challenges with Technologies" transcripts from volunteer participants.

These four configurations provide at least one representative corpus from each aggregation level. Due to time and resource constraints, the remaining five of the nine possible configurations were not tested. Table I presents the corpus statistics after preprocessing.

TABLE I
DATASET STATISTICS AFTER PREPROCESSING.

Corpus	# Documents	# Tokens	Vocabulary
Focus Group Transcripts	2,085	23,155	2,897
Challenges with Technologies	403	5,661	1,356
Challenges – SAC Staff	79	1,494	558
Challenges – Volunteers	184	3,285	986

The smallest corpus has 79 documents and the largest has over 2,000, so the aggregation-level comparison also involves different amounts of text for modelling.

C. Model configuration and training

LDA and Word2Vec were implemented using Gensim, while LFLDA used the LFTM package. For Sentence-BERT, we used HuggingFace Transformers together with the Sentence-Transformers library. All four models were trained across a topic range of $k = 2$ to $k = 50$ with a step size of 2, with hyperparameters varied over the following grids:

- **LDA:** $\alpha, \beta \in \{0.01, 0.31, 0.61, 0.91\}$, together with Gensim's named asymmetric and symmetric priors.
- **LFLDA:** α, β, and λ each drawn from $\{0.01, 0.31, 0.61, 0.91\}$, producing 1,600 model variants.
- **Word2Vec:** We varied only k, with 50 models produced.
- **Sentence-BERT:** We used the Word2Vec configuration, also producing 50 models.

For each corpus configuration, the experiments involved training at least 2,100 variants. Training was performed once for each variant, giving one run for every parameter setting.

D. Evaluation framework

We assessed model performance using automated coherence metrics and expert ratings.

1) Quantitative evaluation: Two coherence metrics were used to measure the statistical consistency of generated topics:

- C_V: Computed using normalised pointwise mutual information (NPMI) within a sliding window over the reference corpus [19, 20]. Higher scores indicate greater semantic coherence among the top words of a topic.
- C_{UCI}: Based on pairwise pointwise mutual information (PMI) of top word pairs computed across the full corpus [19]. Like C_V, higher scores indicate more coherent topics.

Earlier studies have reported an association between each coherence measure and human ratings of topic quality [19, 21]. For the benchmark evaluation, we ranked the models for each algorithm separately on each measure and took the top five from each ranking. This gave the evaluators ten candidates per algorithm. For the focus group corpora, the representative-document protocol described below required more evaluation time, so we selected the top two models under each metric and removed duplicates, which yielded three or four candidates per algorithm.

2) Qualitative evaluation: Two principal investigators of the original focus group study evaluated the candidate models qualitatively. We revised the evaluation protocol for the focus group experiments in response to their feedback on the benchmark evaluation.

Benchmark evaluation (20 Newsgroups). Each topic model was presented to the evaluators as a list of top-10 keywords per topic. Evaluators assessed two dimensions:

- **Topic Quality:** The clarity and interpretability of topics, rated on a scale of 1 (not interpretable) to 10 (clearly interpretable), based on how easily the evaluator could assign a meaningful label to the set of top words.
- **Topic Coverage:** The fraction of the 20 known newsgroup categories that evaluators could match to one or more of the generated topics.

We averaged the two evaluators' scores to obtain Topic Quality. For Topic Coverage, we counted the unique category labels assigned to a model's topics and divided this count by the 20 known categories.

Focus group evaluation. After the benchmark experiment, the principal investigators reported that 10 keywords provided insufficient context for labelling focus group topics, as individual words were less discriminative than in newsgroup text. They also considered Topic Coverage inappropriate for inductive analysis, where codes emerge from the data. We removed Topic Coverage from this evaluation and provided three representative documents for each topic to give the investigators context for their assessment. These were the

documents with the highest similarity to the topic centroid. We retained the 1–10 Topic Quality scale, with each investigator rating every topic using that scale. The model-level Topic Quality scores reported for the focus group experiments combine the ratings from both investigators. We did not compute an inter-rater reliability statistic for these ratings.

V. Results

We report the 20 Newsgroups coherence scores and expert ratings first, then examine the focus group results at each of the three aggregation levels.

A. Benchmark on 20 Newsgroups

1) Quantitative results: Table II summarises the coherence scores across the ten candidate models (top five by C_V, top five by C_{UCI}) for each algorithm, together with their variance and standard deviation. LDA achieved the highest average C_{UCI} score (2.491), while LFLDA yielded the highest average C_V score (0.723). LDA's high mean C_{UCI} was accompanied by a variance of 0.26, equal to LFLDA, indicating considerable variation in topic formation across parameter settings. LFLDA's moderate C_V variance (7.43×10^{-4}) suggests more consistent scores across settings than LDA's.

TABLE II
Coherence scores of candidate models on 20 Newsgroups: average, variance and standard deviation for each metric.

	C_V			C_{UCI}		
Model	Avg.	Var.	SD	Avg.	Var.	SD
LDA	0.719	1.47×10^{-3}	0.04	**2.491**	0.26	0.51
Word2Vec	0.472	3.22×10^{-4}	0.02	−0.873	0.14	0.37
LFLDA	**0.723**	7.43×10^{-4}	0.03	1.782	0.26	0.51
Sentence-BERT	0.401	1.82×10^{-4}	0.01	−0.406	0.21	0.46

2) Qualitative results: Table III presents the qualitative evaluation results. LFLDA achieved the highest average Topic Quality score (8.0 out of 10), followed by Word2Vec (6.56). Word2Vec obtained the highest Topic Coverage (0.56), indicating that its topics covered the broadest range of the 20 known categories. The preferred model differed by criterion: Word2Vec covered more categories, while LFLDA had better individual topics.

TABLE III
Qualitative evaluation on 20 Newsgroups. Coverage is the proportion of the 20 categories matched; Quality is rated on a 1–10 scale.

Model	**Coverage**	**Quality**
LDA	0.36	5.4
Word2Vec	**0.56**	6.56
LFLDA	0.35	**8.0**
Sentence-BERT	0.28	5.2

The benchmark rankings differed between automated metrics and expert assessments, with Word2Vec ranking third in C_V, second in Topic Quality, and first in Coverage. LFLDA and Word2Vec were the two best models qualitatively, producing topics whose top words could be readily mapped to recognisable newsgroup categories. Table IV shows example topics from these two models.

B. Application to focus group transcripts

We applied each algorithm to the four focus group corpus configurations defined in Section IV. For qualitative assessment, the investigators rated only Topic Quality from each topic's three most representative documents, using the revised protocol in Section IV-D.

1) Top-level aggregation: Table V reports both the coherence-score variance (left) and the qualitative evaluation (right) for the top-level corpus (all 2,085 documents). For these sparse focus group corpora, the unnormalised PMI measure C_{UCI} often takes negative values. We therefore interpret C_{UCI} as a relative ranking within this evaluation, not as an absolute coherence level. LFLDA had the smallest C_V variance, while only LDA had a smaller C_{UCI} variance, indicating consistent quality across hyperparameter settings. Word2Vec had reasonable C_V scores, but its mean C_{UCI} of approximately -6.2 suggests weak co-occurrence among its topic words. The experts gave Sentence-BERT the highest average Topic Quality rating and LFLDA the next highest. LDA and Word2Vec both scored 1.0, the scale minimum for topics that the experts could not interpret. The experts thus favoured Sentence-BERT despite its weak coherence scores. We examine this and other differences in model rankings in Section V-D.

LFLDA provided the best balance at this level, with the second-highest Topic Quality and the lowest C_V variance, although Sentence-BERT achieved higher Topic Quality.

2) Topic-level aggregation: The "Challenges with Technologies" corpus (403 documents) was used for topic-level analysis. LFLDA achieved the highest average C_V score (0.80 over its five top models by C_V) with the lowest C_V variance (3.97×10^{-6}), confirming its consistency.

Table VI presents the qualitative results. Word2Vec received the highest average Topic Quality (6.00) but with zero variance. When we compared the three representative documents presented to the evaluators for each Word2Vec cluster, we found nearly the same set each time. Most of the 403 short utterances mapped to a small region of the embedding space, with little separation between the K-Means clusters. Based on our examination of the clusters, one possible explanation is that by averaging the Word2Vec vectors for utterances with few content words, it results in the positioning of their document embeddings near the centroid of the space. The same utterances appeared as representative documents for different clusters, so the Quality score reflects repeated judgments of one set of documents and provides little basis for distinguishing the topics. We exclude Word2Vec when selecting the best topic-level model for this reason. LFLDA achieved the second-highest average Topic Quality (5.10) with the highest variance (7.47), as its best model achieved a score of 9.0 whereas the other models scored between 3.11 and 5.0.

TABLE IV
EXAMPLE TOPICS FROM WORD2VEC AND LFLDA ON 20 NEWSGROUPS, WITH TOP WORDS AND INFERRED CATEGORY LABELS.

Model	Inferred Label	Top Words
Word2Vec	talk.politics.mideast	nazi, palestinian, turkish, armenian, jewish, arab, jew, israeli, muslim, israel
Word2Vec	soc.religion.christian	catholic, homosexuality, atheism, christianity, homosexual, church, religious, atheist, religion, christian
Word2Vec	sci.space	surface, satellite, planet, detector, shuttle, spacecraft, solar, space, earth, orbit
LFLDA	sci.space	space, writes, would, article, Space, NASA, system, orbit, launch, year
LFLDA	sci.med	patient, disease, doctor, article, study, drug, health, treatment, medical, writes
LFLDA	rec.sport.hockey	game, team, player, play, hockey, year, season, goal, writes, period

TABLE V
FOCUS GROUP TRANSCRIPTS AT TOP-LEVEL AGGREGATION: COHERENCE-SCORE VARIANCE (LOWER INDICATES MORE CONSISTENCY ACROSS HYPERPARAMETERS) AND QUALITATIVE EVALUATION.

	Coherence variance		Topic Quality		
Model	C_V	C_{UCI}	Avg.	Variance	SD
LDA	1.42×10^{-5}	$\mathbf{4.89}\times10^{-5}$	1.00	0	0
Word2Vec	2.00×10^{-3}	2.35	1.00	0	0
LFLDA	$\mathbf{4.99}\times10^{-6}$	6.47×10^{-4}	2.39	1.29	1.14
Sentence-BERT	1.00×10^{-3}	0.46	**5.10**	0.93	0.96

Sentence-BERT was the most consistent model qualitatively (variance 0.35), with an average of 3.25.

TABLE VI
AVERAGE TOPIC QUALITY AND ITS VARIANCE FOR "CHALLENGES WITH TECHNOLOGIES" AT TOPIC-LEVEL AGGREGATION.

Model	Avg. Quality	Variance	SD
LDA	2.21	0.89	0.94
Word2Vec	6.00	0	0
LFLDA	**5.10**	7.47	2.73
Sentence-BERT	3.25	0.35	0.59

After excluding the anomalous Word2Vec result, LFLDA provided the best balance of quantitative and qualitative performance at the topic level.

3) Stakeholder-level aggregation: Two stakeholder-level corpora were evaluated: Volunteers (184 documents) and SAC Staff (79 documents).

Volunteers. LFLDA again achieved the highest average Topic Quality but with the largest variance, while Sentence-BERT was second. LDA and Word2Vec both scored 1.0, indicating uninterpretable topics. Quantitatively, LFLDA had the highest C_V scores with the lowest variance (6.06×10^{-6}).

SAC Staff. LDA achieved the highest average Topic Quality at this level, although its very high variance indicates inconsistent results across model variants. LFLDA obtained a lower average score with lower variance, followed by Sentence-BERT. Quantitatively, LFLDA had the highest C_V scores with the lowest variance (1.98×10^{-6}), consistent with results at other aggregation levels.

Table VII summarises the qualitative results across both stakeholder corpora.

TABLE VII
AVERAGE TOPIC QUALITY AT STAKEHOLDER-LEVEL AGGREGATION.

	Volunteers		SAC Staff	
Model	Avg. Quality	SD	Avg. Quality	SD
LDA	1.00	0	5.43	2.98
Word2Vec	1.00	0	1.00	0
LFLDA	**5.40**	1.80	4.05	2.12
Sentence-BERT	2.71	0.25	2.24	0.46

C. Impact of aggregation level

We examine performance across the three aggregation levels using LFLDA, which provided the best balance of coherence and expert-rated quality on three of the four configurations. Table VIII shows the average coherence scores and Topic Quality for LFLDA at each level.

TABLE VIII
PERFORMANCE OF LFLDA ACROSS AGGREGATION LEVELS. EACH COHERENCE SCORE IS AVERAGED OVER THE FIVE LFLDA MODELS SHORTLISTED UNDER THAT METRIC, AND TOPIC QUALITY OVER THE QUALITATIVELY RATED LFLDA CANDIDATES, AT EACH LEVEL. STAKEHOLDER-LEVEL VALUES ARE AVERAGED ACROSS THE VOLUNTEERS AND SAC STAFF CORPORA.

Aggregation Level	C_V (avg)	C_{UCI} (avg)	Topic Quality
Top-level	0.63	0.41	2.39
Topic-level	**0.80**	-1.25	**5.10**
Stakeholder-level	0.76	-3.30	4.73

Topic-level aggregation produced the best results on C_V and on expert-rated Topic Quality, with C_V 27% higher than at the top level (0.80 vs. 0.63) and average Topic Quality more than doubled (5.10 vs. 2.39). In contrast, C_{UCI} decreased monotonically from the top level to the stakeholder level and did not follow the expert ratings. The stakeholder-level averages lay between the other two levels for C_V and Topic Quality. Grouping by theme appears to balance coherence with corpus size: combining all themes dilutes the thematic signal, while further division into stakeholder corpora of 79–184 documents leaves insufficient lexical diversity for reliable topic identification.

D. Agreement between coherence scores and expert ratings

The model rankings from coherence scores did not consistently match those from expert ratings. For 20 Newsgroups,

the evaluators placed Word2Vec second in Topic Quality and first in Coverage, compared with third on C_V and fourth on C_{UCI}. For the top-level focus group corpus, Sentence-BERT was first in Topic Quality but fourth on C_V. LFLDA ranked second in Topic Quality and first on C_V. Word2Vec was first in Topic Quality and third on C_V at the topic level, although the repeated-document issue described in Section V-B excludes it from selection. These comparisons support using expert assessment alongside automated metrics to select models for qualitative analysis.

VI. Discussion

A. Summary of findings

RQ1: Model comparison. LFLDA achieved the best balance between quantitative coherence and expert-rated topic quality on the 20 Newsgroups benchmark and three of the four focus group corpus configurations. On the benchmark dataset, LFLDA obtained the highest average C_V and Topic Quality scores, with moderate variance in coherence scores (Table II and Table III). For the focus group transcripts, LFLDA performed favourably across both evaluation criteria at the top level, topic level, and volunteer stakeholder level. However, for the SAC staff corpus, LDA achieved a higher average Topic Quality score than LFLDA, although its variance was substantially higher, making it difficult to identify a clear best-performing model (Table VII). The combination of statistical word co-occurrence information and semantic embeddings may have contributed to LFLDA's better performance compared to LDA, Word2Vec, and Sentence-BERT across most of the focus group configurations. Nevertheless, as each model configuration was evaluated using only a single run, further experiments with multiple controlled random seeds are required to validate these findings.

Sentence-BERT scored poorly on automated coherence metrics across all configurations but received the highest expert Topic Quality rating at the top level. The experts appear to have recognised semantic relationships in the topics that the word co-occurrence scores missed. We evaluated the embeddings and clustering together (Section III-E), so we cannot attribute this result to either stage alone.

RQ2: Aggregation and topic quality. Topic quality varied substantially with the level at which we grouped the transcripts. For LFLDA, grouping by discussion theme gave both a higher C_V score and a higher expert rating than pooling all transcripts or subdividing them by stakeholder group (Table VIII). The corpus grouped by theme retains lexical diversity while restricting the discussion to a coherent theme. Combining all themes at the top level dilutes the thematic signal, while the focused stakeholder corpora contain only 79–184 documents and have limited lexical diversity for identifying topics reliably. As we move to smaller groups, the corpora also become more uniform in theme. The comparison therefore cannot establish how much of the observed difference is due to corpus size and how much is due to thematic grouping. A subsequent experiment could subsample the top-level corpus to the size of the topic-level corpus to examine this confound.

RQ3: Metric divergence. The comparisons in Section V-D show disagreement between coherence scores and expert judgment in several experiments. Both C_V and C_{UCI} are based on word co-occurrence, which is only one aspect of a topic's usefulness in qualitative research. Determining whether a topic is interpretable, whether its documents concern the same theme, and whether it fits what is known about the domain requires a separate assessment. We therefore recommend that researchers consider the experts' reading of the topics alongside the coherence scores when choosing a model.

B. Implications for decision support in qualitative research

The results support using topic modelling during the *exploratory* phase of qualitative analysis, particularly inductive coding. The principal investigators reached the same conclusion, and their feedback led us to drop Topic Coverage and assess Topic Quality from representative documents (Section IV-D). Two requirements follow for applying the framework. First, researchers should model transcripts grouped by the topics in the discussion guide. In our evaluation, this grouping more than doubled LFLDA's average Topic Quality and raised C_V by 27%. Second, domain experts should assess the shortlisted models before their topics are used for analysis.

Our evaluation supports these recommendations on topic quality alone. Decision quality, analysis time, and confidence in the themes were not measured and need evaluation before deployment with practitioners.

C. Study-derived recommendations for focus group analysis tools

Based on the evaluation of 13 transcripts from one study on mobile health for older adults in Singapore and the 20 Newsgroups benchmark, we provide five recommendations for focus group analysis tools. Further studies are needed to establish whether these recommendations generalise to other settings.

1) **Give hybrid models priority in the initial comparison.** Across most configurations, LFLDA offered a better balance between coherence and expert-rated quality than either the statistical model (LDA) or the embedding models (Word2Vec and Sentence-BERT). Combining these approaches may help a model handle the noise and informal language found in focus group discussions.
2) **Group transcripts by discussion theme before fitting a model.** Of the three levels, topic-level aggregation gave LFLDA the highest average Topic Quality, so researchers should structure their transcript data around the themes of the discussion guide rather than running models on the full, unsegmented corpus. The confounding effects of corpus size and thematic homogeneity discussed under RQ2 require further control before this recommendation can be generalised.
3) **Present representative documents alongside topic keywords.** The principal investigators found that key-

words alone provided insufficient context for reliable interpretation of focus group topics. We therefore presented representative documents, which provide the surrounding context needed to interpret conversational topics.

4) **Do not rely on automated coherence metrics as the sole selection criterion.** Researchers can use C_V and C_{UCI} to exclude clearly poor models, but the repeated differences from expert ratings mean that selection also requires human assessment. At least the shortlisted candidates should be reviewed before a model is selected.
5) **Retain qualitative coding when using the tool for initial exploration.** In the focus group experiments, LFLDA achieved its highest average Topic Quality score of 5.40 out of 10 on the Volunteers corpus. Although some individual models generated interpretable topics, the highest average score across model variants was only slightly above the midpoint of the rating scale. These results suggest that the framework is not sufficiently reliable to replace manual qualitative coding. Instead, the framework should support the initial exploration of transcripts, allowing researchers to subsequently examine the original text and validate, combine, separate, or discard the suggested topics.

D. Challenges of conversational and multilingual data

All models showed reduced average Topic Quality on the focus group transcripts compared to the 20 Newsgroups benchmark (LFLDA: 8.0 on benchmark versus 5.10 average at the topic level, with a per-model peak of 9.0). Language register is a likely contributor to the lower scores. Participants frequently spoke Singlish, an English creole with vocabulary from Malay, Mandarin, Hokkien, and Tamil, and switched languages within individual utterances. Word2Vec was trained on Google News and Sentence-BERT on general-domain English. Neither corpus covered the Singlish used in the discussions, which may limit how well the embeddings capture the meanings of words in this setting. The same difficulty may arise in communities that use other non-standard English varieties or switch between languages in conversation. One way to adapt the models is to continue pre-training on conversations from the population of interest. Another is to fine-tune the models on a representative sample of that population's transcripts.

E. Limitations

Single application domain. All focus group transcripts used in our evaluation were obtained from a single study on mobile health technologies for older adults in Singapore. Although we also evaluated the models on 20 Newsgroups, the findings from the focus group experiments may not generalise to other discussion topics, participant populations, or cultural contexts. Due to the scope of this study, we did not evaluate the framework on additional focus group datasets involving different research topics and participant demographics. Given the growing research interest in older adults' use of digital health technologies [24], future work could examine transcripts from other studies in this area, as well as different application domains, to further validate the findings on transcript aggregation.

Small expert panel and single-run protocol. The generated topics were evaluated by two investigators from the original study. Their familiarity with the transcripts informed the assessment, although their involvement in the original study and the lack of blinding limit what we can infer about the ratings an independent panel would give. We trained each configuration in a single run. For LDA and LFLDA, both stochastic models, the variance we report is across hyperparameter settings and additional follow-up could investigate the effects of different random seeds on the best configuration. For more robustness, future work can involve a larger, independent panel under blinded conditions to determine if LFLDA consistently provides the best balance between coherence and topic quality.

Subset of possible aggregation configurations. Four of the nine possible corpus configurations were tested. The selected configurations represent all three aggregation levels, although evaluation of the remaining five would provide further evidence of the relationship between aggregation level and model performance.

No comparison with more recent models. We selected four topic modelling approaches to represent different model families while retaining the ability to inspect their intermediate outputs. The framework allows researchers to examine topic-word distributions, document-cluster assignments, and coherence scores, as well as reproduce the experimental results. Other approaches, such as BERTopic [14], also provide interpretable intermediate outputs and should be considered in future comparisons. LLMs offer a further means of summarising the themes in a transcript, although it can be difficult to trace how a summary was produced, and the themes may change across runs. An LLM may also follow the coding instructions inconsistently [25]. Using a cloud service may require researchers to share sensitive transcripts with an external provider. We did not evaluate zero-shot LLM coding in this study. Comparing our framework with BERTopic and zero-shot coding by a locally hosted LLM would allow us to examine the quality and interpretability of the topics, the reproducibility of the results, and how each method handles confidential transcripts.

Other limitations. We did not isolate the effect of clustering and keyword extraction from the effect of the embeddings in the embedding-based pipeline (Section III-E). The implementation also processes data in batches, with no support for interactive analysis. The observed disagreement with expert ratings suggests that the word co-occurrence measures C_V and C_{UCI} may miss higher-order semantic relationships. Topic diversity and embedding-based coherence are two complementary measures to examine in a further evaluation.

VII. Conclusion

In this paper, we presented a semi-automated framework that uses topic modelling to support the analysis of focus group transcripts. We compared four approaches at three levels of

transcript aggregation, using coherence scores and the experts' assessments to judge the topics at each level.

LFLDA provided the best balance of coherence and expert-assessed topic quality on 20 Newsgroups and on three of the four focus group configurations. For LFLDA, grouping the transcripts by discussion theme more than doubled average Topic Quality relative to modelling the full corpus. Further division by stakeholder group gave lower average quality than theme-level grouping. Coherence and expert ratings did not always agree: at the top level, Sentence-BERT had the lowest C_V rank but the highest Topic Quality rank.

For exploratory inductive analysis, the results favour grouping transcripts by discussion theme and considering LFLDA among the initial candidates. Researchers should be able to inspect representative documents when evaluating these candidates and read the transcripts closely to assess the suggested themes, with coherence scores forming only part of that assessment.

The model comparison needs confirmation through repeated runs with different random seeds and assessment by a larger independent panel. We plan to study how the framework affects researchers' work through a web interface where they can adjust parameters, inspect topics, and give feedback. We also plan comparisons with BERTopic and LLM-based thematic coding. Domain-adaptive pre-training or fine-tuning may help the models interpret Singlish and other non-standard varieties. Finally, the findings and recommendations need evaluation on transcripts from studies with different topics, participant populations, and cultural contexts before their generalisability can be established.

Acknowledgements

We thank the participants and research team of the original Singapore study on mobile health technologies for older adults. The focus group transcripts collected in that study formed the secondary data used for our analysis.

References

[1] T. O. Nyumba, K. Wilson, C. J. Derrick, and N. Mukherjee, "The use of focus group discussion methodology: Insights from two decades of application in conservation," *Methods in Ecology and Evolution*, vol. 9, no. 1, pp. 20–32, 2018.

[2] Y. Chun Tie, M. Birks, and K. Francis, "Grounded theory research: A design framework for novice researchers," *SAGE Open Medicine*, vol. 7, p. 2050312118822927, 2019.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015.

[5] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[7] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[8] A. Castellanos, H. Jiang, P. Gomes, D. Vander Meer, and A. Castillo, "Large language models for thematic summarization in qualitative health care research: Comparative analysis of model and human performance," *JMIR AI*, vol. 4, p. e64447, 2025.

[9] K. Coussement and D. Van den Poel, "Improving customer complaint management by automatic email classification using linguistic style features as predictors," *Decision Support Systems*, vol. 44, no. 4, pp. 870–882, 2008.

[10] D. E. Pournarakis, D. N. Sotiropoulos, and G. M. Giaglis, "A computational model for mining consumer perceptions in social media," *Decision Support Systems*, vol. 93, pp. 98–110, 2017.

[11] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: a state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2023.

[12] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, no. 1–2, pp. 177–196, 2001.

[13] D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2000.

[14] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[15] K. H. Lim, S. Karunasekera, and A. Harwood, "ClusTop: A clustering-based topic modelling algorithm for Twitter using word networks," in *Proceedings of the IEEE International Conference on Big Data*, 2017, pp. 2009–2018.

[16] W. Mu, K. H. Lim, J. Liu, S. Karunasekera, L. Falzon, and A. Harwood, "A clustering-based topic model using word networks and word embeddings," *Journal of Big Data*, vol. 9, no. 1, p. 38, 2022.

[17] W. Mu, M. Li, and K. H. Lim, "A social data-driven system for identifying estate-related events and topics," in *Advances in Social Networks Analysis and Mining (ASONAM)*, ser. Lecture Notes in Social Networks. Springer Nature Switzerland, 2025, pp. 443–449.

[18] W. Mu and K. H. Lim, "Label-free topic-focused summarization using query augmentation," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[19] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408.

[20] S. Syed and M. Spruit, "Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation," in *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 2017, pp. 165–174.

[21] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 215–224.

[22] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in Neural Information Processing Systems*, vol. 22, 2009, pp. 288–296.

[23] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[24] A. Z. H. Yee, B. Kwok, J. Ng, G. P. Loy, L. Y. Ng, J. R. H. Sng, S. Perrault, K. H. Lim, and K. Subburaj, "Why older adults integrate digital health technology into their everyday lives: The role of basic psychological needs," *Human Behavior and Emerging Technologies*, vol. 2025, p. 9711661, 2025.

[25] H. Lim and K. H. Lim, "The atomic instruction gap: Instruction-tuned LLMs struggle with simple, self-contained directives," in *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion '26)*, 2026, pp. 1037–1045.