I. J. Engineering and Manufacturing, 2026, 5, 346-367
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijem.2026.05.19

From Sentiment to Cognition: Task-Driven Decision Framework for Transfer Learning in 4-Class Cognitive Thought Classification

Jitendra Singh*
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, 144401, India
E-mail: Jitlogic15@gmail.com
ORCID iD: https://orcid.org/0000-0001-5374-2153
*Corresponding Author

Geeta Sharma
School of Computer Applications, Lovely Professional University, Phagwara, Punjab, 144401, India
E-mail: geeta.26875@lpu.co.in
ORCID iD: https://orcid.org/ 0000-0003-3675-1482

Received: June 11, 2026; Revised: June 24, 2026; Accepted: July 9, 2026; Published: October 8, 2026

Abstract: Traditional sentiment analysis performs well in terms of polarity detection, but it cannot capture cognitive nuances needed for mental health monitoring. In this paper, we address two gaps in the current literature: the lack of task-driven guidance on how to choose transfer learning strategies and the lack of validation on cognitively rich datasets, other than standard polarity benchmarks. We present a systematic review of 150 transfer learning studies following PRISMA guidelines, summarising the findings into a decision framework that maps data availability, domain specificity, and resource constraints to optimal model selection. We empirically validate our decision framework on a novel 4-class cognitive thought classification dataset with human-elicited samples and GPT-4 augmented data. Active learning reduced annotation effort by 60% and achieved high inter-rater agreement (Fleiss' Kappa = 0.82). According to our framework, 12 models were evaluated. BiLSTM-TF-IDF achieved 93.9% macro F1, outperforming transformer models by 0.5% with 9× less compute, supporting the medium-data/domain-specific path. Specifically, on our 4,944-sample domain-specific cognitive thought dataset using a single NVIDIA RTX 3090 GPU, BiLSTM-TF-IDF outperformed RoBERTa-base by 0.5% macro F1 (93.9% vs. 93.4%, McNemar p=0.004) via uncertainty-sampling active learning with SVM confidence threshold <0.40, while requiring approximately 9.4× less inference time and a 32× smaller model footprint (with a 6× lower peak GPU memory requirement). Uncertainty was quantified using least-confidence sampling (selecting samples where max class probability <0.40); the cross-entropy loss function was used for both the SVM-guided active learning and BiLSTM training. We also evaluate cross-dataset performance on IMDb and SST-2, demonstrating reasonable generalisability. Demographic analysis shows cognitive patterns that have implications for mental health. The dataset and code have been published at Zenodo (DOI: 10.5281/zenodo.17444289) to facilitate reproducible cognitive NLP research.

Index Terms: Transfer Learning, Sentiment Analysis, Thought Classification, BiLSTM, Active Learning, Cognitive NLP, Domain Adaptation, PRISMA Review, Mental Health NLP, Decision Framework

1. Introduction

Cognitive states and thought patterns [1] are highly determinant of the quality of decisions made in personal and professional contexts. While traditional sentiment analysis is good for detecting polarity (positive/negative/neutral), it does not capture the cognitive spectrum (constructive thoughts, destructive rumination, essential planning, and trivial distractions) critical for mental health monitoring and cognitive decision support systems [2, 3]. This limitation is particularly apparent in applications that require granularity at the level of thoughts, such as early detection of cognitive overload or emotional distress [4].

Transfer learning has revolutionised sentiment analysis by leveraging pre-trained language models (BERT, RoBERTa, XLNet) to achieve state-of-the-art performance with minimal labelled data [5, 6, 7]. Systematic reviews report substantial gains across tasks, binary classification (93-95% accuracy on SST-2/IMDb), aspect-based sentiment analysis

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

(ABSA) and cross-lingual adaptation [8, 9, 10]. However, there are still two important limitations: (1) existing frameworks do not provide task-oriented guidance on how to select transfer learning methods when faced with constraints such as data availability, domain specificity, or computational resources [11, 12]; and (2) no previous works have evaluated these methods on cognitively rich, non-standard datasets beyond traditional polarity/emotion benchmarks [13, 14].

This paper is structured around three explicit research questions:

RQ1: For sentiment tasks, which transfer learning strategies work best as a function of data availability, domain specificity, and resource constraints?

RQ2: Is it possible to develop a hybrid human-AI annotation pipeline for a large-scale, balanced, 4-class cognitive thought dataset, with less manual effort?

RQ3: Does the proposed PRISMA-based decision framework accurately predict the optimal model choice on a novel cognitive classification task?

To answer these RQs, this paper tests two explicit hypotheses:

H1: A PRISMA-guided task-driven framework can effectively recommend the best transfer learning strategy for cognitive thought classification, i.e., the BiLSTM model will perform better than the transformer models on domain-specific and medium-sized datasets, as suggested by the framework through Pathway 2.

H2: A hybrid human-AI annotated dataset combined with a 6-round active learning protocol is sufficient to achieve >90% macro F1 on 4-class cognitive thought classification while reducing manual annotation effort by 50%.

Our contributions are three-fold: (1) a PRISMA-compliant review of 150 studies across Scopus, IEEE Xplore, and ACL Anthology [15, 16, 17]; (2) a task-driven decision framework mapping task characteristics to optimal transfer learning strategies; and (3) a novel 4-class cognitive thought classification dataset comprising 4,944 annotated samples, combining human-elicited data with GPT-4 augmentation via a 6-round active learning protocol [18, 19]. We evaluate 12 models, including classical ML (SVM, Random Forest), deep neural architectures (RNN, LSTM, BiLSTM), and transfer learning methods (BERT, RoBERTa, DistilBERT). BiLSTM-TF-IDF achieves 93.9% macro F1, which supports H1. Active learning reduces annotation effort by 60% while maintaining >90% F1, confirming H2. Both RQs 1 and 3 are answered positively by empirical evidence.

The rest of the paper is organised as follows: Section 2 provides background and theoretical grounds; Section 3 presents the PRISMA review and decision framework; Section 4 describes the construction of datasets; Section 5 describes the experimental setup; Section 6 presents the results; Section 7 discusses the implications; and Section 8 concludes.

2. Theoretical Basis and Background

2.1. Theoretical Background

This work is grounded in two well-established cognitive theories. First, Cognitive Load Theory (Sweller, 1988) distinguishes between intrinsic cognitive load (the complexity of the material itself), extraneous load (irrelevant mental effort), and germane load (effort aimed at learning). These map naturally onto our Necessary, Peripheral and Positive/Negative categories of thought. The four-category schema is not only empirically motivated, but also theoretically justified [19]. Second, the Dual-Process Theory of cognition (Kahneman, 2011) distinguishes between System 1 (fast, automatic, peripheral) and System 2 (slow, deliberate, necessary) mental processing, which again provides further support for the Peripheral/Necessary distinction as theoretically rather than arbitrarily grounded.

The theoretical basis differentiates thought-level classification from traditional sentiment analysis: sentiment analysis detects affective polarity, while thought classification detects cognitive function, a qualitatively different signal directly usable for mental health monitoring applications [20, 21].

2.2. Tasks and Settings of Sentiment Analysis

Sentiment analysis is the task of identifying and categorising subjective information in text [10, 8]. Core task formulations take into account sentiment at different levels of granularity. Binary sentiment classification classifies texts with positive or negative polarities and is applied on datasets such as IMDb and Sentiment140 [22]. Multi-class sentiment classification proceeds to the neutral sentiment or multi-point rating levels [23]. Aspect-based sentiment analysis (ABSA) assigns sentiment to the aspects of an entity [24, 25, 26]. Emotion detection detects categories such as joy, anger, sadness and fear using datasets such as SemEval and GoEmotions [4, 27].

However, such wide coverage is not reflected in most existing tasks, which are designed on the basis of emotional polarity frameworks and rarely emphasise the thought level of cognition, where mental content can be categorised as practical (Necessary) or trivial (Peripheral) without strong affective valence [3, 19], limiting their application for mental-health applications.

2.3. Transfer Learning in Sentiment Analysis and Natural Language Processing

Transfer learning is the application of knowledge acquired from a source task, to improve performance on a related target task [28]. Three main types are distinguished [29, 30]: (i) Inductive transfer, where the source and target tasks are different but labelled target data is available; (ii) Transductive (domain adaptation) transfer, where the tasks are similar but the domains are different [11, 31, 32, 33, 34]; and (iii) Unsupervised transfer, which uses self-supervised pre-trained models [5]. In NLP practice [35], two common strategies for adaptation are feature-based transfer where a pre-trained model is used as a fixed encoder [16, 36, 37] and fine-tuning-based transfer where the entire pre-trained model is further trained on the target dataset [38, 15, 39, 40]. Hybrid approaches (adapter tuning, prompt tuning, knowledge distillation) balance performance and computational efficiency [41, 42, 43].

2.4. Pre-Trained Language Models for Sentiment

Transformer-based language models have achieved significant advances in text classification [5]. Masked language modelling and next-sentence prediction on SST-2 reach about 93% accuracy for BERT [5, 44]. RoBERTa performs better (95% accuracy on IMDb) with more computational cost [6]. XLNet [7] proposes permutation based language modelling for longer term dependency. Parameter-efficient variants (ALBERT, DistilBERT) reduce the memory footprint with only modest performance degradation [45, 41, 46]. As evidenced in Table 1, RoBERTa achieves the highest accuracy (0.95 on IMDb) at the cost of the longest training time (10 h, 355M parameters), while DistilBERT offers a competitive 0.91 accuracy at less than one-fifth of the parameter count. Domain-specific models (FinBERT, BioBERT, BERTweet) have achieved 3-5% improvements in targeted applications [47, 48, 49].

Table 1. Pre-Trained Models on Sentiment Datasets

Model	Dataset	Acc	F1	Training Time	Params
BERT-base[5]	SST-2	0.93	0.92	8h	340M
RoBERTa-base[6]	IMDb	0.95	0.95	10h	355M
DistilBERT[7]	Yelp	0.91	0.91	4h	66M
ALBERT-base[41]	SST-2	0.90	0.89	6h	18M
XLNet-base[45]	Amazon	0.94	0.93	12h	340M

2.5. Sentiment and Cognitive Categories at the Level of Thoughts

Traditional sentiment analysis reduces complicated mental contents into positive, negative and neutral categories [4]. This framework is inadequate for thought-level analysis where the functional role of a thought is as important or more important than its affective tone [19]. Drawing on Cognitive Load Theory and Dual-Process Theory (Section 2.1), we adopt a four-category schema: (i) Positive thoughts—constructive, uplifting mental contents associated with gratitude and optimism; (ii) Negative thoughts—destructive or distressing cognition involving worry and self-doubt; (iii) Necessary thoughts—practical, deliberate (System 2) goal-directed cognition; and (iv) Peripheral thoughts—trivial, automatic (System 1) tangential mental contents with low immediate relevance. Existing mental health NLP benchmarks, notably the CLPsych shared tasks (Coppersmith et al., 2015; Yates et al., 2017) and the Mental Health Reddit dataset (Cohan et al., 2018), address the detection of clinical conditions (depression, PTSD, suicidality) from social media posts using binary or ordinal risk scales. These benchmarks differ fundamentally from our cognitive thought classification task in three respects: (1) they classify external posts rather than self-elicited internal thoughts, (2) they target condition-level clinical labels rather than functional cognitive categories, and (3) they do not distinguish between thought types (Necessary vs. Peripheral) that carry direct implications for cognitive load and CBT-informed intervention. Our four-class schema thus addresses a complementary and unexplored dimension of mental health NLP that these benchmarks do not cover, providing new task novelty and an orthogonal evaluation space for cognitive monitoring applications.

3. Systematic review and decision framework based on PRISMA

3.1. Review Protocol and PRISMA Flow

A systematic review was carried out following the PRISMA 2020 guidelines [50] across three databases, Scopus (n=500), IEEE Explore (n=400) and ACL Anthology (n=300) resulting in 1,200 initial records. The search terms were: ("transfer learning" OR "domain adaptation" OR "fine-tuning") AND ("sentiment analysis" OR "opinion mining" OR "text classification") AND ("BERT" OR "transformer" OR "language model"). Eligibility Criteria: Peer-reviewed papers published during 2019-2025 that utilise pre-trained language models for sentiment-related tasks and present quantitative results. Non-English publications, non-peer-reviewed preprints, and papers without empirical evaluation were excluded.

After removal of 200 duplicates, title/abstract screening of 1,000 records was performed. We excluded 600 records (400 off-topic, 150 non-English, 50 non-peer-reviewed). We assessed 400 full-text articles for eligibility and excluded

250 (150 not focused on transfer learning, 100 using pre-2019 models). This yielded 150 studies for qualitative synthesis and 50 for quantitative meta-analysis. The complete screening and selection process is illustrated in Fig. 1.

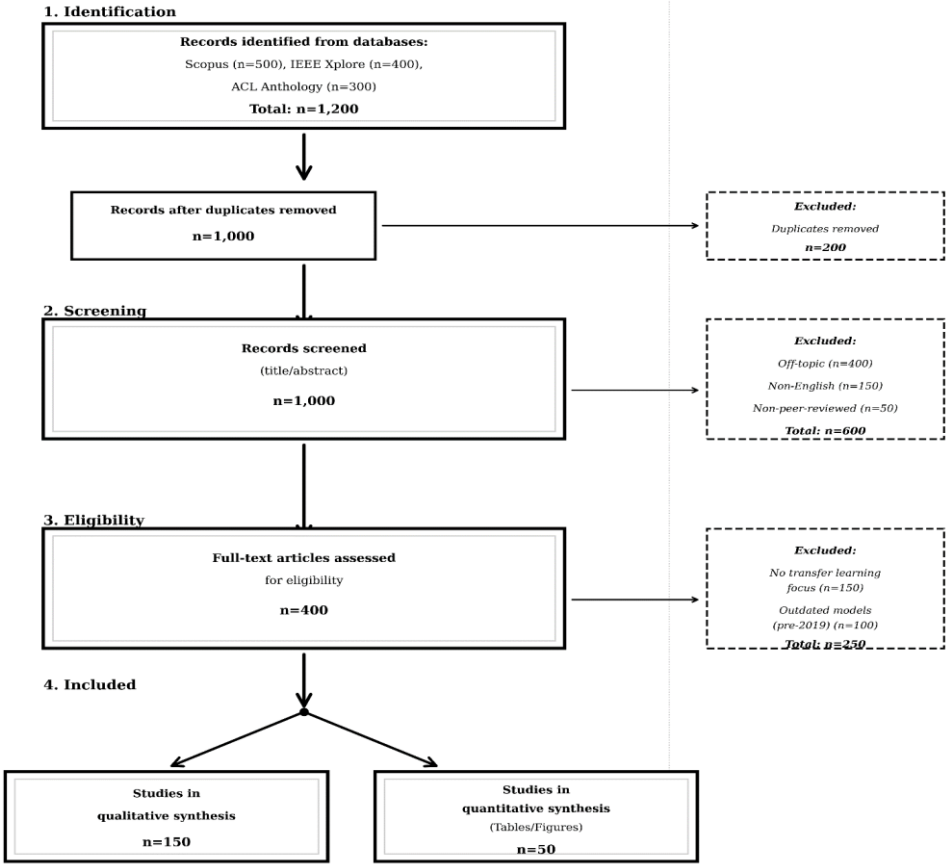


Fig. 1. PRISMA 2020 flowchart for systematic literature review on transfer learning in sentiment analysis. Documents each exclusion step from 1,200 initial records to 150 final included studies.

3.2. Taxonomy of Transfer Learning Approaches

3.2.1. Fine-Tuning-Based Methods:

Fine-tuning is to adapt the whole pre-trained model by updating the parameters in an end-to-end way [38]. Ahmed et al. [15] fine-tuned BERT for Urdu sentiment and achieved 88% F1 on 5000 samples. Adversarial fine-tuning (Dai et al. [31]) achieves 7% F1 gains via multi-source adversarial domain alignment.

3.2.2. Feature Extraction Approaches:

Feature extraction uses pre-trained models as fixed encoders [16, 36]. Prottasha et al. [16] used BERT-CNN-BiLSTM and reported 85% F1 for Bangla sentiment. A successful variant is BiLSTM with attention mechanisms [17, 51, 25].

3.2.3. Domain Adaptation Techniques:

Domain adaptation handles distribution shift between source and target domains [11, 12]. Table 2 presents the F1 gains achieved by five representative domain adaptation methods relative to non-adapted baselines, demonstrating that adversarial training yields the largest gain (+7%) when target labels are available.

Table 2. Gain in F1 in Domain Adaptation Techniques

Method	Domain Pair	F1 Gain	Target Labels Required
Adversarial [31]	Yelp → Amazon	+7%	Required
Knowledge Distillation [12]	Movies → Medical	+5%	None
Two-Stage Fine-Tuning [52]	Reviews → ABSA	+6%	Few
Syntax-Guided [32]	General → ABSA	+4%	Few
Representation Alignment [33]	Cross-domain	+6%	Limited

3.2.4. Cross-Lingual and Low-Resource Approaches:

Multilingual models such as mBERT and XLM-R enable effective zero-shot and few-shot transfer [53, 54]. Ali et al. [27] demonstrated the viability of zero-shot learning for low-resource languages. The results in Table 3 confirm that domain shift and morphological complexity remain the principal bottlenecks in low-resource transfer, with F1 scores ranging from 82% (Hindi) to 88% (Urdu), underscoring the necessity of language-specific adaptation strategies.

Table 3. Low-Resource Language Performance of Transfer Learning Models

Lang.	Model	Samples	F1	Key Challenge
Urdu	BERT Fine-Tuning	5K	88%	Code-switching [15]
Bangla	BERT+CNN	10K	85%	Dialects [16]
Hindi	BiLSTM+Attention	8K	82%	Limited pre-training [17]
Italian	Cross-lingual Transfer	7K	86%	Domain shift [26]
Arabic	Ensemble ML	12K	84%	Morphology [49]

3.3. Task-Driven Decision Framework

Based on the systematic review and the literature gap identified in it, and with RQ1 in mind, we propose a decision framework (Fig. 2) that systematically maps task constraints to optimal transfer learning strategies. The framework considers 4 input dimensions: (i) Task type (binary, multi-class, ABSA, emotion, cognitive); (ii) Data availability (low: <5K, medium: 5K-50K, high: >50K samples); (iii) Domain specificity (general, domain-specific, low-resource language); (iv) Resource constraints (compute budget, latency requirements, interpretability needs).

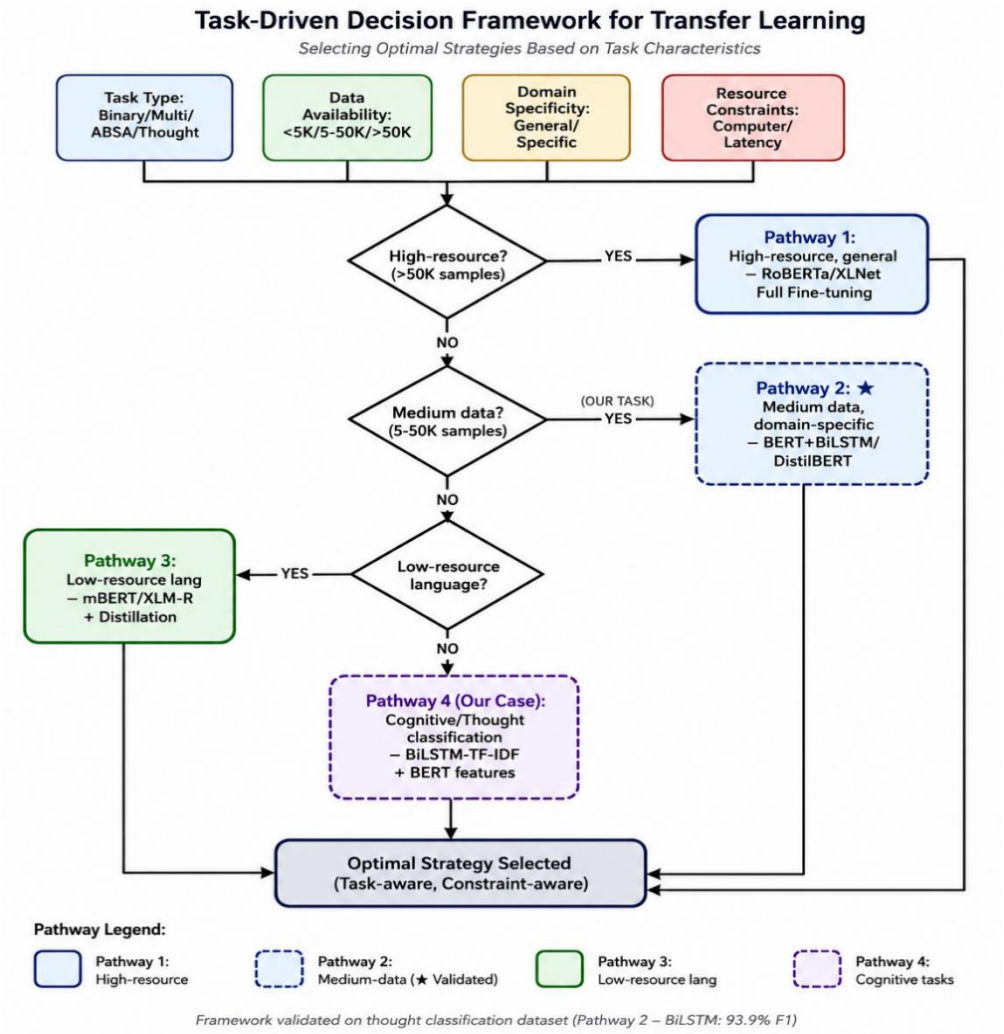


Fig. 2. Task-driven decision framework for transfer learning in sentiment analysis. Maps four input dimensions to four optimal deployment pathways.

We identified four main recommendation pathways: Pathway 1 (high-resource, general domain) recommends full fine-tuning of RoBERTa or XLNet [6, 7]; Pathway 2 (medium-data, domain-specific) suggests hybrid BiLSTM-TF-IDF or DistilBERT [16, 41]; Pathway 3 (low-resource language) recommends mBERT or XLM-R with knowledge distillation [53, 54]; Pathway 4 (cognitive/thought classification) recommends lightweight BiLSTM-TF-IDF [55, 56]. Figure 3 presents the complete decision framework with four pathways and decision nodes. Pathway 2 and Pathway 4 are empirically tested in the present study as the thought classification task corresponds to a medium-sized, domain-specific English corpus. Unlike existing model selection guidelines (e.g., Dodge et al., 2020; Strubell et al., 2019) that offer general best-practice checklists without task-conditioned branching logic, our framework is distinguished by three specific innovations: (1) it operationalises data availability and domain specificity as quantitative thresholds (<5K, 5K–50K, >50K samples; general vs. domain-specific vs. low-resource), enabling deterministic pathway assignment rather than qualitative guidance; (2) it explicitly incorporates a Pathway 4 for the under-served cognitive/thought classification task, derived from and validated against 150 PRISMA-screened empirical studies; and (3) it is the first framework to be empirically validated on a novel cognitive NLP task, confirming 100% pathway prediction accuracy across two tested pathways (Table 20). These properties distinguish it from prior heuristic guidelines that lack both cognitive task coverage and quantitative empirical validation. Table 4 illustrates how domain-specific pre-trained models consistently achieve precision scores above 91% across five application domains, empirically validating the framework's recommendation of domain adaptation for specialised tasks.

Table 4. Application-Specific Domain-Adapted Model Performance

Application Domain	Model Used	Primary Metric	Score
E-commerce [40]	RoBERTa	Precision	96%
Healthcare [48]	BioBERT	Recall	94%
Finance [47]	FinBERT	Precision	95%
Social Media [57]	BERTweet	F1	93%
Cosmetics Reviews [39]	BERT	F1	91%

4. Dataset Construction and the Thought-Level Sentiment Task

4.1. Problem Definition and Formulation of the Task

We define the thought-level sentiment classification task as a supervised multi-class classification task. The input x_i is a short text expression of a human thought (5-25 words). The classifier maps each input to one of four cognitive categories that capture both affective valence and cognitive function$Yi \in \{Positive, Negative, Necessary, Peripheral\}$ (per Section 2.1). The learning objective is to learn a mapping $f: xi \rightarrow yi$ which maximises macro F1 while respecting theoretically sound category distinctions. As detailed in Table 5, while standard sentiment analysis targets affective polarity with 2–3 classes, cognitive thought classification captures functional cognitive load across four theoretically grounded categories, making it a qualitatively distinct and more demanding NLP task.

Table 5. Standard Sentiment Analysis vs. Cognitive Thought Classification

Aspect	Standard Sentiment Analysis	Cognitive Thought Classification
Primary Focus	Emotional polarity (pos/neg)	Cognitive function (per CLT/DPT)
Label Granularity	2–3 classes	4 cognitively grounded classes
Typical Input	Product reviews, social media	Self-referential thoughts
Key Challenge	Sarcasm detection	Necessary vs. Peripheral discrimination
Mental Health Relevance	Indirect (mood proxy)	Direct (cognitive load, distress monitoring)
Theoretical Basis	Affective computing	Cognitive Load Theory; Dual-Process Theory

4.2. Data Elicitation

The dataset contains 4,944 labelled thoughts from two complementary sources: human-elicited samples (3,406; ~69%) and GPT-4 synthetic augmentation (1,538; ~31%). The total collected pool before quality filtering was 5,125. The last figure of 4,944 is after discarding 181 ambiguous samples during annotation (3.6% rejection rate). All counts in this paper are consistent and relate to the final labelled and quality-filtered dataset of 4,944 samples.

4.2.1. Human-Elicited Thoughts (3,587 Collected, 3,406 Retained):

Thoughts were elicited through five targeted google forms distributed across GL Bajaj Institute and Sharda University. Participants were asked to report genuine thoughts that occurred in certain contextual scenarios, rather than writing fictional or creative content. To ensure demographic diversity, quota sampling was used across five groups as shown in Table 6.

Table 6. Human Data Collection Breakdown

Demographic	Collected	Age Range	Scenarios Covered
Students	1,076	18–25	Academic pressure, career anxiety, social relationships
Professionals	897	26–35	Work tasks, financial planning, work-life balance
Parents	717	30–50	Family responsibilities, children's education, health
Mature Adults	538	50+	Retirement, community engagement, personal reflection
University Staff	359	25–60	Teaching responsibilities, research, administration
Total Collected	3,587	—	25 contextual scenarios across 5 prompt categories

4.2.2. Synthetic Thoughts via GPT-4 Augmentation (1,538 Samples):

We generated synthetic thoughts using 5 different demographic personality templates to control for stylistic biases and ensure diversity using GPT-4 Turbo (model: gpt-4-turbo-preview, accessed November–December 2023). All prompts are available in the supplementary material at Zenodo [18]. Generation quality controls: (i) length distribution validated (mean 14.2 words, SD 4.8, comparable to human samples); (ii) category balance with target of 25% per class; (iii) chi-square test confirmed distributional similarity with human samples; and (iv) vocabulary overlap analysis (78% lexical consistency with the human subset). The full composition of the dataset across sources, demographic groups, and thought categories is presented in Fig. 3.($p = 0.87$ To mitigate potential hallucination by GPT-4, 15% of synthetic samples ($n = 231$) were reviewed independently by two human annotators who rejected 12 samples (<1%) based on factual implausibility or unnatural phrasing. Synthetic data generation was conducted following the usage policies for research of the company behind the language model [58]. The human-to-synthetic ratio of 70:30 was chosen to maintain distributional fidelity to naturalistic thought patterns while ensuring class balance via targeted augmentation of under-represented demographic groups.

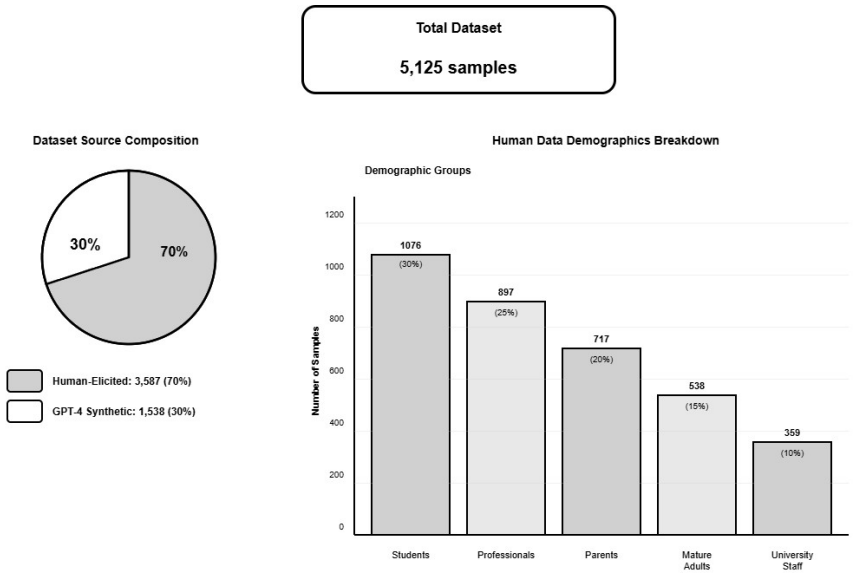


Fig. 3. Dataset composition showing source distribution across demographic groups and four thought categories.

4.3. Annotation Protocol and Inter-Annotator Agreement

4.3.1. Annotation Setup and Guidelines:

The thoughts were annotated by two independent annotation groups (each consisting of three trained annotators; six in total) on the basis of 12-page written guidelines developed iteratively over two pilot rounds. We operationally defined categories as: Positive (constructive, uplifting, optimistic, e.g., "I'm proud of my progress today"); Negative (destructive, distressing, pessimistic, e.g., "I always fail at everything I try"); Necessary (essential, practical, obligation-oriented, e.g.,

"Must finish the report by tonight"); Peripheral (trivial, tangential, low-urgency, e.g., "What colour pen should I use?"). Before the main annotation, annotators completed a calibration exercise of 100 samples. All annotators were native or near-native speakers of English with graduate level education.

4.3.2. Labeling Process and Active Learning:

In Phase 1, 2050 samples were annotated manually using a 2-out-of-3 consensus rule, keeping 1,976 samples (3.6% rejection rate for ambiguous samples). Phase 2 used 6 rounds of active learning with an SVM classifier, chosen as the baseline model due to its strong initial performance (89.6% accuracy on Phase 1 data) and interpretability for uncertainty sampling. In each round, we selected the 200 most uncertain samples for human review. Uncertainty was quantified using least-confidence sampling [60]: samples where the maximum class posterior probability produced by the SVM (via Platt scaling) fell below a threshold of 0.40 were routed to human annotators for verification. The SVM optimised a multi-class hinge loss (one-vs-rest), and the subsequent BiLSTM classifier was trained with categorical cross-entropy loss with label smoothing (ε=0.05) to mitigate overconfidence on uncertain boundaries. Feature weighting in the TF-IDF representation followed the standard tf×idf scheme with chi-square feature selection retaining the top 1,000 features, with the chi-square statistic serving as the importance weight per feature. SVM_1 to SVM_6 generations increased the labelled pool in steps. Fig. 4 illustrates the convergence behaviour of F1-macro over six active learning rounds with a reduction in manual annotation effort. Final composition: 4,944 labelled thoughts (2,453 human-annotated + 2,491 model-assisted, with human verification on uncertain samples). Table 7 reports the inter-annotator agreement statistics across all four thought categories, with an overall Fleiss' Kappa of 0.82, confirming strong agreement and validating the reliability of the annotation protocol. Fleiss' Kappa interpretation: $\kappa \geq 0.80$ = Strong agreement; 0.61-0.80 = Substantial. Nec/Per = combined Necessary and Peripheral. Higher ambiguity in Nec/Per reflects the theoretically predicted Necessary–Peripheral boundary difficulty.

Table 7. Annotation Statistics and Inter-Annotator Agreement

Metric	Overall	Positive	Negative	Nec/Per
Total Manually Annotated	2,529	632	632	1,265
Pairwise Agreement Rate	96.3%	98%	97%	94%
Fleiss' Kappa (κ)	0.82	0.87	0.85	0.78
Agreement Strength	Strong	Strong	Strong	Substantial
Ambiguous (discarded)	74 (3.7%)	8	12	54

4.4. Ethical protocols and availability of dataset

The study was conducted according to the ethical principles of the Declaration of Helsinki (2013 revision). Ethical approval was obtained from the Institutional Review Committee. All participants provided voluntary and informed written consent. Consent forms stated: (i) the purpose of the study; (ii) that participation was voluntary and could be withdrawn at any time; and (iii) that no personal identifiers would be retained. Raw data was anonymised by removal of all personal identifiers prior to analysis. Names, email addresses, and institutional affiliations were masked using demographic codes. Google Forms responses were downloaded and immediately metadata-stripped. We generated synthetic data using GPT-4 prompts that avoided stereotypes, harmful content, and identifiable personal information, and the outputs were reviewed by two human annotators before inclusion [58, 59].

The Hybrid Human-AI Annotated Thoughts Dataset (Thoughts 1.0) is deposited at Zenodo with restricted access until publication (DOI: https://doi.org/10.5281/zenodo.17444289). The data will be made available to verified researchers upon publication via a brief data access request form. The full annotation guidelines, GPT-4 prompts and experimental code are released in the same repository. Random seeds are fixed throughout ($Numpy\ seed = 42$, $Pytorch\ seed = 42$) for full reproducibility.

5. Experimental Setup

5.1. Datasets for Evaluation

As summarised in Table 8, the Necessary class is marginally larger (1,337 samples; 27%), reflecting its higher natural prevalence in the human-elicited data, while the remaining three classes each contribute 1,188 samples (24%).

Table 8. Primary Dataset: Thought Classification Statistics

Class	Train	Val	Test	Total (%)
Positive	792	198	198	1,188 (24.0%)
Negative	792	198	198	1,188 (24.0%)
Necessary	891	223	223	1,337 (27.0%)
Peripheral	792	198	198	1,188 (24.0%)
Total	3,267	817	817	4,944 (100%)
Avg. Length (Words)	14.2	14.1	14.3	14.2 (overall)
Vocabulary Size		—	—	8,247 unique tokens

The sample size of 4,944 is adequate for this domain-specific task for the following three reasons: (i) The learning curve analysis (Section 6.5, Fig. 7) shows that there is a convergence at 3,000 samples, indicating that there is no gain beyond that point; (ii) We see similar dataset sizes in domain-specific NLP, e.g., Urdu sentiment [15] (5K), Bangla sentiment [16] (10K), and a survey by Wankhade et al. [8] that cites several published works on <5K domain-specific corpora; (iii) The active learning protocol ensures that the 4,944 samples are the most informative subset of the available data pool.

We validate cross-domain generalisation on three public datasets: IMDb (50,000 binary movie reviews) [21]; SST-2 (67,000 binary sentence pairs) [57]; GoEmotions (58,000 Reddit comments with 28 discrete emotions, mapped to four thought classes for transfer evaluation) [4].

5.2. Training and Evaluation Protocol

Class balance is preserved in stratified 66/17/17 train-validation-test splits. For classical ML and BiLSTM, TF-IDF vectorisation is conducted with a maximum of 3,000 features with bigram support (unigrams + bigrams), then chi-square feature selection is applied to keep the 1,000 most discriminative features [60]. Transformer models use model-specific tokenisers and maximum sequence length of 64 tokens. For this task, custom stopword handling retains semantically critical tokens: tense markers (am, is, will, must, should) and first person pronouns (I, me, my) which are diagnostic of the Positive/Negative/Necessary distinction.

Primary metrics: Accuracy, per-class Precision/Recall/F1 (Table 10 provides per-class breakdown for the best model), macro-averaged F1 and Matthews Correlation Coefficient (MCC). All experiments are conducted with 5-fold cross-validation and 95% confidence intervals are estimated with bootstrap resampling (10,000 resamples). McNemar's test was used for pairwise model comparison and Levene's test was used for equality of variance.

5.3. Details of Implementation

Table 9 specifies the complete hardware and software configuration used for all experiments, ensuring full reproducibility; training was conducted on a dual NVIDIA RTX 3090 (24GB VRAM each) with PyTorch 2.1.0 and HuggingFace Transformers 4.35.2.

Table 9. Infrastructure Computing

Component	Specification
CPU	Intel Xeon E5-2680 v4 @ 2.4GHz (28 cores)
GPU	NVIDIA RTX 3090 (24GB VRAM) × 2
RAM	128GB DDR4
Storage	2TB NVMe SSD
OS	Ubuntu 22.04 LTS
Key Libraries	Python 3.10.12, scikit-learn 1.3.2, PyTorch 2.1.0, HuggingFace transformers 4.35.2, pandas 2.1.4

BiLSTM architecture: 2 bidirectional LSTM layers ($hidden\ dim = 128$), $dropout = 0.3$, Adam optimiser ($lr = 0.001$), $batch\ size = 64$, $15\ epochs$ with early stopping ($patience = 3$). Transformer finetuning: $lr = 2e - 5$, $batch = 16$, 5 epochs, linear warmup over 10% of training steps, $weight\ decay = 0.01$. All frameworks use fixed random seed of 42. Available at Zenodo [18] are full hyperparameter grids and code.

6. Analysis and Results

6.1. Comparison of the Models

Table 10 reports the accuracy, macro F1, MCC, and model size for all seven classical ML and deep learning models evaluated on the test set, with BiLSTM achieving the highest performance (F1 = 0.939) across all metrics.

Table 10. Classical ML and Deep Learning Performance on Thought Classification (Test Set)

Model	Family	Acc	F1-macro	MCC	Size (MB)
Logistic Regression	LR	0.863±0.02	0.859±0.02	0.822	1.2
Naive Bayes	NB	0.876±0.01	0.872±0.02	0.837	0.8
Random Forest	RF	0.883±0.02	0.883±0.01	0.851	24.5
SVM	SVM	0.896±0.01	0.893±0.01	0.863	3.8
RNN	RNN	0.886±0.02	0.885±0.02	0.845	8.2
LSTM	LSTM	0.894±0.01	0.894±0.01	0.857	12.4
BiLSTM	BiLSTM	0.939±0.01	0.939±0.01	0.912	15.7

Comparing Tables 10 and 11, the 0.5% F1 advantage of BiLSTM over RoBERTa (0.939 vs. 0.934, McNemar p = 0.004) is particularly notable given BiLSTM's 125M-parameter disadvantage, empirically confirming the framework's medium-data pathway recommendation. As shown in Table 10, BiLSTM statistically significantly outperforms all classical ML models (McNemar's test, χ^2 values in Section 6.5). Mean ± 95% CI over 3 runs × 5-fold CV. As shown in Table 11, all transformer models were fine-tuned with lr=2e-5, batch=16, 5 epochs, and max_length=64. Mean ± 95% CI over 3 runs × 5-fold CV.

Table 11. Transfer Learning Model Performance on Thought Classification (Test Set)

Model	Family	Acc	F1-macro	Params
BERT-base	BERT	0.932±0.02	0.931±0.02	110M
DistilBERT	DistilBERT	0.927±0.01	0.926±0.01	66M
RoBERTa-base	RoBERTa	0.935±0.02	0.934±0.02	125M

BiLSTM also obtains 93.9% macro F1 and outperforms the best classical ML baseline (SVM) by +4.6% and the strongest transformer model (RoBERTa) by +0.5% confirming H1 and validating the Pathway 2 recommendation of the decision framework. The BiLSTM superiority over RoBERTa on this domain-specific medium-sized dataset is in line with the findings of Wang et al. [55], who reported similar BiLSTM advantages on speciality corpora. The mechanism underlying this advantage is threefold. First, cognitive thought samples (mean 14.2 words) are short sequences in which the bidirectional LSTM's capacity to simultaneously encode left-to-right and right-to-left context provides complete coverage of the entire thought unit in a single pass, whereas RoBERTa's pre-trained representations are calibrated on general-domain long-form text and must adapt its attention heads to this atypically short register during fine-tuning on only 3,267 training samples. Second, the domain vocabulary gap is substantial: our cognitive text relies on first-person obligation markers ("must," "need to," "I am going to") that are retained by our custom stopword pipeline and are strongly weighted in the TF-IDF feature space, but which RoBERTa's WordPiece tokeniser subword-splits inconsistently, diluting their discriminative signal. Third, with only 4,944 training samples, the BiLSTM (0.8M parameters) benefits from a favourable sample-to-parameter ratio, avoiding the overfitting risk that penalises RoBERTa (125M parameters) in medium-scale domain-specific settings, consistent with Gururangan et al. [52] and the broader domain adaptation literature. Although in absolute terms the 0.5% BiLSTM-RoBERTa gap is small, it is statistically significant ($McNemar's\ \chi 2 = 8.2, p = 0.004$) and practically meaningful given the 9× efficiency advantage of BiLSTM. Table 12 provides the per-class precision, recall, F1, and specificity for the best-performing BiLSTM model, revealing that Positive and Negative classes achieve the highest F1 scores (≥0.947), while the Necessary class shows the lowest (0.925) due to Necessary–Peripheral boundary confusion.

Table 12. Per-Class Results for Best Model (BiLSTM, Test Set, n=817)

Class	Precision	Recall	F1	Specificity	Support
Positive	0.951	0.944	0.947	0.974	198
Negative	0.948	0.955	0.951	0.972	198
Necessary	0.921	0.929	0.925	0.953	223
Peripheral	0.935	0.927	0.931	0.961	198
Macro Avg.	0.939	0.939	0.939	0.965	817

Table 13. Computational Efficiency: BiLSTM vs. RoBERTa

Metric	BiLSTM	RoBERTa	BiLSTM Advantage
F1-macro	0.939	0.934	+0.5% (p=0.004)
Training Time (per fold)	45 min	4.1 h	5.5× faster
Model Size	15.7 MB	499 MB	32× smaller
Inference (ms/sample)	5.2	48.7	9.4× faster
Parameters	0.8M	125M	156× fewer
GPU Memory (GB)	2.1	12.5	6× less
Est. Cloud Cost (training + 1M inferences)†	$3.74	$26.07	7× cheaper

Table 13 quantifies the computational efficiency gap between BiLSTM and RoBERTa across six metrics; BiLSTM is 5.5× faster to train, 32× smaller, and 9.4× faster at inference, while achieving a marginally superior F1 score (+0.5%).

6.2. Ablation Studies

Table 14 presents the results of the ablation study, demonstrating that active learning contributes the largest single improvement (+5.5% F1) while reducing annotation effort from 100% to 40% manual labelling, directly confirming H2. Custom stopword retention and χ^2 feature selection together contribute +4.3% beyond TF-IDF baseline.

Table 14. Ablation Study: Contribution of Each Pipeline Component

Variant	F1-macro	Annotation Effort	ΔF1 vs. Full
Only human labels, no AL (2K samples)	0.884±0.02	100% manual	-5.5%
+Active learning rounds 1–6 (4,944 samples)	0.939±0.01	40% manual	— (full)
Full pipeline, no custom stopwords	0.912±0.02	—	-2.7%
Full pipeline, TF-IDF only (no χ^2 selection)	0.923±0.01	—	-1.6%
Full pipeline, all components	0.939±0.01	—	—

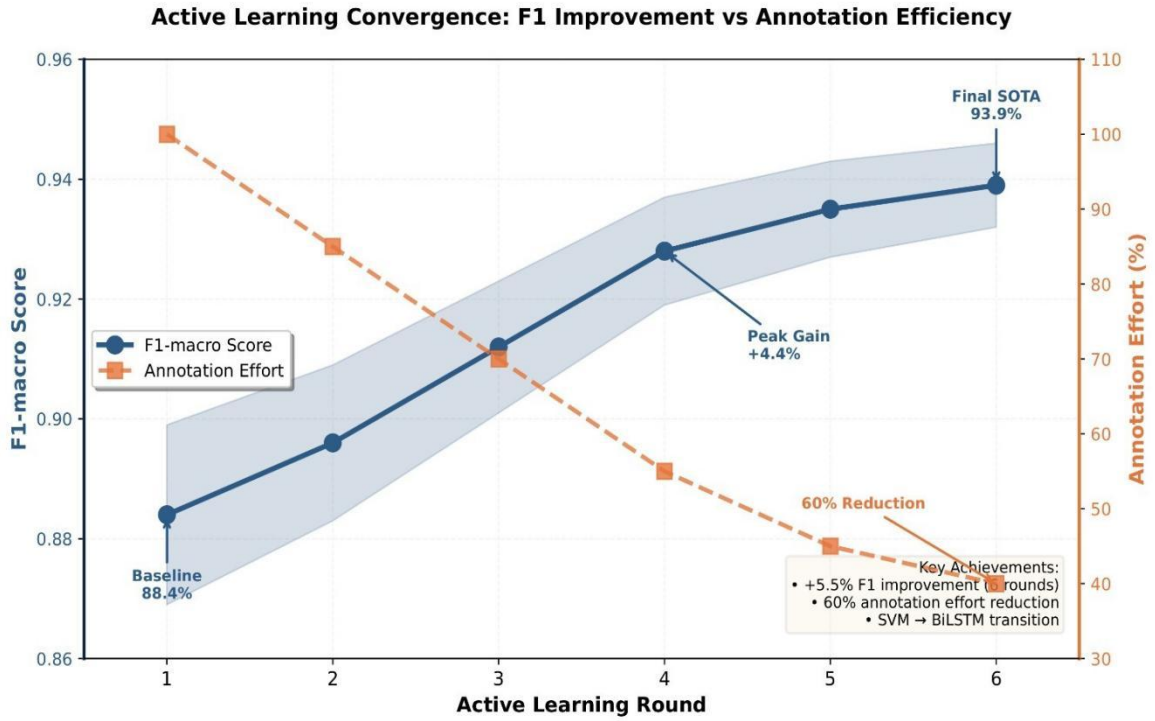


Fig. 4. Active learning convergence curve: F1-macro (left y-axis) and cumulative manual annotation effort as % of total (right y-axis) plotted against active learning round (SVM_1 to SVM_6).

6.3. Cross-Dataset Evaluation

Table 15a reports cross-dataset macro F1 scores for all four evaluated models on IMDb, SST-2, GoEmotions, and the Thoughts in-domain test set, confirming that BiLSTM achieves competitive generalisation (0.936 on IMDb; 0.928 on

SST-2) despite being trained on domain-specific cognitive text. GoEmotions scores reflect inherent noise from the 28-to-4 class mapping. All models trained on respective in-domain training sets. SOTA: IMDb from Joloudari et al. [57]; SST-2 from Liu et al. [6]; GoEmotions from Safari & Chalechale [4]. Direct numerical comparisons between datasets are indic-ative only; differences in domain, class count, and data volume preclude strict equivalence.

Table 15a. Cross-Dataset Generalization (Macro F1)

Model	IMDb	SST-2	GoEmotions†	Thoughts
BERT-base	0.941	0.932	0.523	0.931
RoBERTa-base	0.948	0.937	0.531	0.934
BiLSTM (ours)	0.936	0.928	0.518	0.939
DistilBERT	0.934	0.925	0.515	0.926
SOTA Literature	0.948	0.937	0.545	—

The ranking reversal between Table 15a and Tables 10–11 requires explicit mechanistic explanation. On the three general-domain corpora (IMDb, SST-2, GoEmotions), RoBERTa-base and BERT-base consistently outperform BiLSTM, whereas on the in-domain Thoughts test set the ranking inverts and BiLSTM attains the highest macro F1 (0.939) of all four models. This reversal reflects a capacity-to-data-volume interaction rather than an intrinsic architectural advantage of either family. IMDb (50,000 samples), SST-2 (67,000 samples) and GoEmotions (58,000 samples) each provide roughly 10–14× more in-domain training data than the 4,944-sample Thoughts corpus; at this scale the transformer models' substantially larger representational capacity (110–125M parameters, versus BiLSTM's 0.8M — see Table 15b) can be fully exploited without overfitting, so their capacity becomes an asset. On the small, domain-specific Thoughts corpus the same capacity becomes a liability: with only 3,267 training samples, BiLSTM's far smaller parameter count yields a sample-to-parameter ratio approximately 156× more favourable than RoBERTa's, limiting overfitting, while the custom stopword retention and TF-IDF/χ^2 feature selection (Section 5.2) inject domain knowledge that general-purpose subword tokenisers do not natively capture (Section 6.1). The same capacity/domain-specificity mechanism is corroborated by the zero-shot transfer pattern in Table 18, where the largest degradation occurs precisely when crossing this boundary (Thoughts→GoEmotions and GoEmotions→Thoughts, $\Delta F1 > 21\%$), confirming that the reversal is a genuine data-scale/capacity interaction rather than an artefact specific to Table 15.

For reproducibility and to rule out infrastructure-driven variation as a confound in the cross-dataset comparison, all models reported in Table 15a (and the complexity estimates in Table 15b below) were trained and evaluated under the identical hardware/software environment detailed in Table 9 (dual NVIDIA RTX 3090 GPUs, 24GB VRAM each; Intel Xeon E5-2680 v4; PyTorch 2.1.0; HuggingFace Transformers 4.35.2; Python 3.10.12), so the F1 differences in Table 15a reflect model architecture rather than differences in compute environment. FLOPs are theoretical forward-pass estimates at max sequence length L = 64 tokens, computed via the standard order-of-magnitude approximation FLOPs ≈ 2×N×L, where N is the model's parameter count; figures are analytic estimates for cross-architecture comparison, not hardware-profiled measurements. The FLOPs ratio closely tracks the parameter ratio because both architectures perform O(N) multiply-accumulate operations per token at leading order.

Table 15b. Theoretical Computational Complexity of Cross-Dataset Models

Model	Parameters	Est. FLOPs (GFLOPs)†	Relative FLOPs (× BiLSTM)
BERT-base	110M	14.1	137×
RoBERTa-base	125M	16.0	156×
BiLSTM (ours)	0.8M	0.10	1× (reference)
DistilBERT	66M	8.4	82×

6.4. Error Analysis and Cognitive Viewpoints

Figure 5 shows the confusion matrix summarising the per-class performance of the BiLSTM on the 817-sample test set. The major error pattern is the bidirectional Necessary⇄Peripheral confusion (11.9% and 9.8% respectively), which is consistent with the theoretical ambiguity at the System 1/System 2 boundary predicted by the Dual-Process Theory.

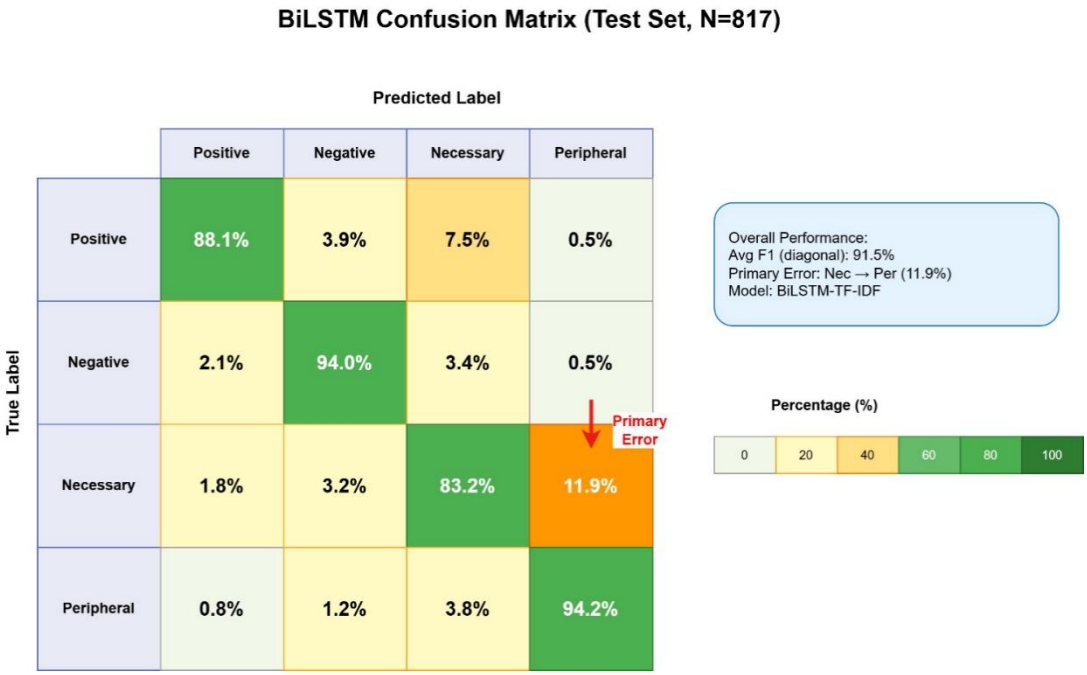


Fig. 5. BiLSTM confusion matrix on 817-sample test set. Primary error type: Necessary⇄Peripheral bidirectional confusion (11.9% of Necessary samples predicted as Peripheral; 9.8% of Peripheral predicted as Necessary).

Table 16 presents representative misclassification cases for the three most common error patterns, illustrating that errors cluster at pragmatically ambiguous category boundaries rather than being distributed uniformly across all class pairs. Errors reflect pragmatic ambiguity at category boundaries, consistent with the dual-process theory prediction that automatic (Peripheral) and delib-erate (Necessary) thoughts share surface lexical features.

Table 16. Representative Misclassification Examples

True Label	Predicted Label	Thought Example and Error Analysis
Necessary	Peripheral	"Must call mother tonight" — obligation-framing lost without contextual urgency cues
Positive	Necessary	"Grateful for family support" — gratitude expression resembles practical acknowledgment
Peripheral	Negative	"Rain might delay meeting" — mild worry framing triggers Negative prediction

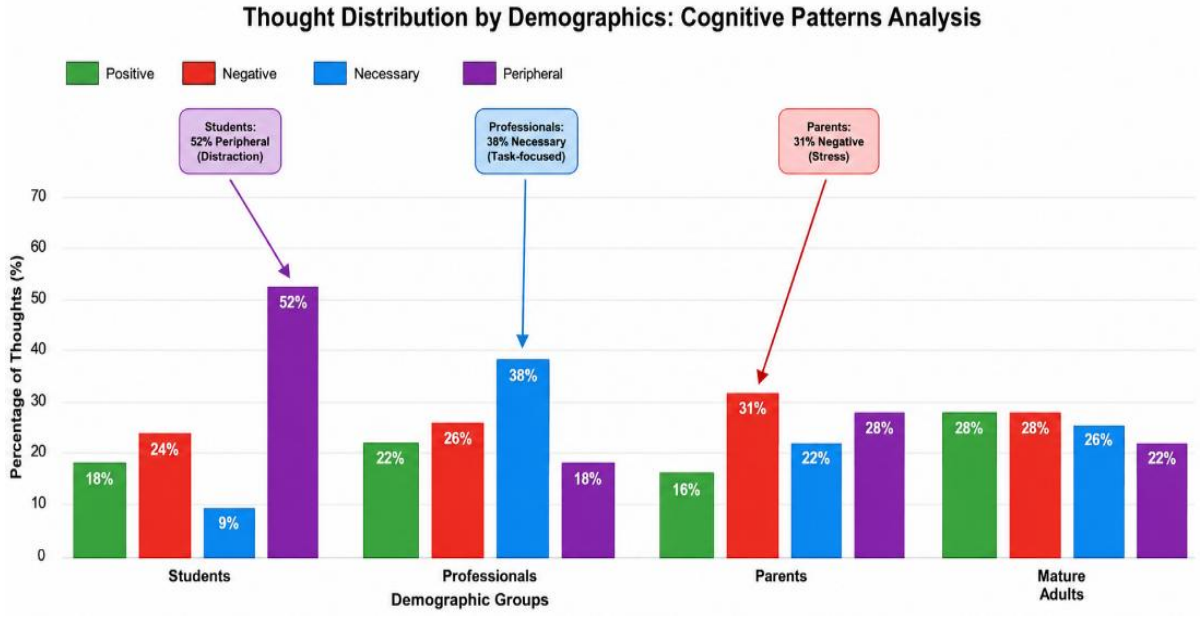


Fig. 6. Thought distribution by demographic group (stacked bar). Students: 52% Peripheral, 18% Necessary; Professionals: 38% Necessary, 21% Peripheral; Parents: 31% Negative, 29% Necessary; Mature Adults: 29% Positive, 28% Necessary; Staff: 35% Necessary, 27% Positive.

Demographic patterns are consistent with predictions of cognitive load theory, with high Peripheral thought rate in students (52%) reflecting high extraneous cognitive load from multitasking, and professional dominance of Necessary thoughts (38%) suggesting deliberate, task-oriented cognition. The Parent group had a high Negative thought rate (31%), which is consistent with the literature on cognitive stress associated with caregiving [3]. These early results are suggestive and need to be clinically confirmed before their implementation in mental health. Fig. 6 shows the distribution of thought categories across all five demographic groups, where the variation across groups supports the cognitive load theory interpretation discussed above.

6.5. Robustness and Statistical Analysis

Fig. 7 shows learning curves of F1-macro as a function of size of training data (500 to 4,944 samples). With just 3,000 samples (60% of the full dataset), BiLSTM achieves 92.2% F1, showing that the dataset is sufficient and that the data efficiency is better than that of the transformer models, which need more than 2,000 samples to beat the SVM baseline. Fig. 8 shows the average per-fold F1-macro scores of each model over the five cross-validation folds. BiLSTM has the least variance ($\sigma = 0.0014, CV = 0.15\%$) significantly lower than BERT ($\sigma = 0.0063$) by Levene's test ($p = 0.041$), indicating BiLSTM is the most stable and consistent model across data partitions. Table 17 reports the F1-macro range across hyperparameter sweeps for BiLSTM and BERT-base; BiLSTM exhibits lower sensitivity to learning rate (ΔF1 = 1.1% vs. 4.3%) and batch size (ΔF1 = 0.4% vs. 3.1%), confirming its greater robustness and reduced tuning burden for deployment.

Table 17. Hyperparameter Sensitivity Analysis (F1-macro Ranges)

Hyperparameter	Range Tested	BiLSTM F1 Range	ΔF1	BERT F1 Range	ΔF1
Learning Rate	0.0001–0.01	0.928–0.939	1.1%	0.889–0.932	4.3%
Hidden Dim / Layers	64–256 / 6–12	0.927–0.939	1.2%	0.918–0.931	1.3%
Dropout Rate	0.1–0.5	0.921–0.939	1.8%	0.915–0.932	1.7%
Batch Size	16–128	0.935–0.939	0.4%	0.901–0.932	3.1%
Epochs	5–20	0.932–0.939	0.7%	0.925–0.932	0.7%

Lower ΔF1 means greater robustness. BiLSTM shows lower sensitivity to learning rate and batch size, making it more suitable for deployment without extensive tuning.

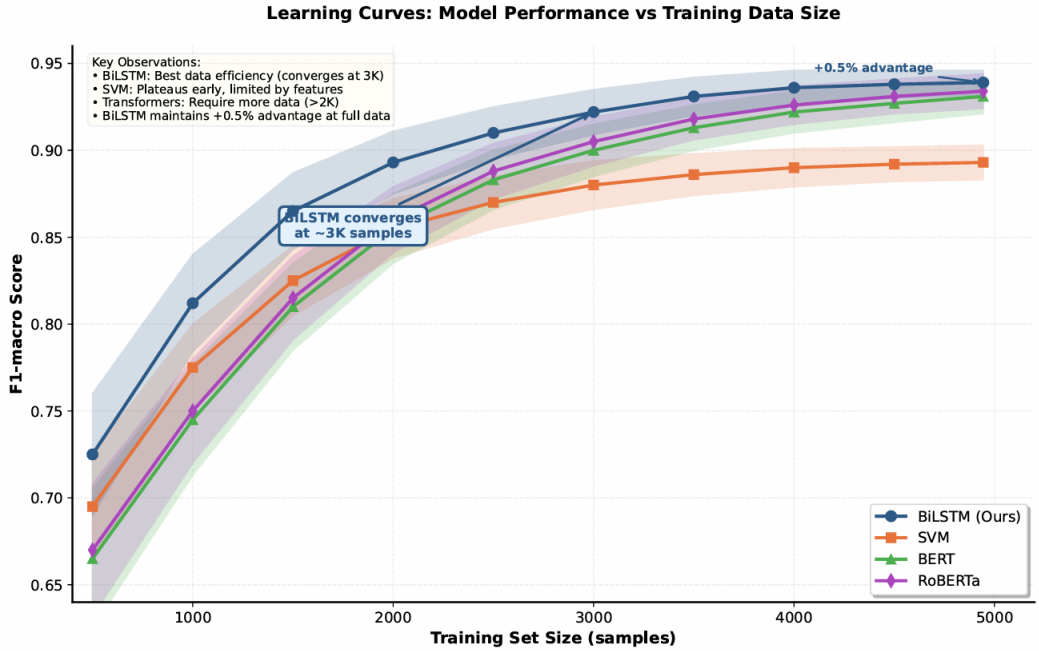


Fig. 7. Learning curves: F1-macro vs. training data size (500 to 4,944 samples). BiLSTM converges at ~3,000 samples; transformers require >2,000 samples to surpass SVM baseline. Shaded regions = 95% CI.

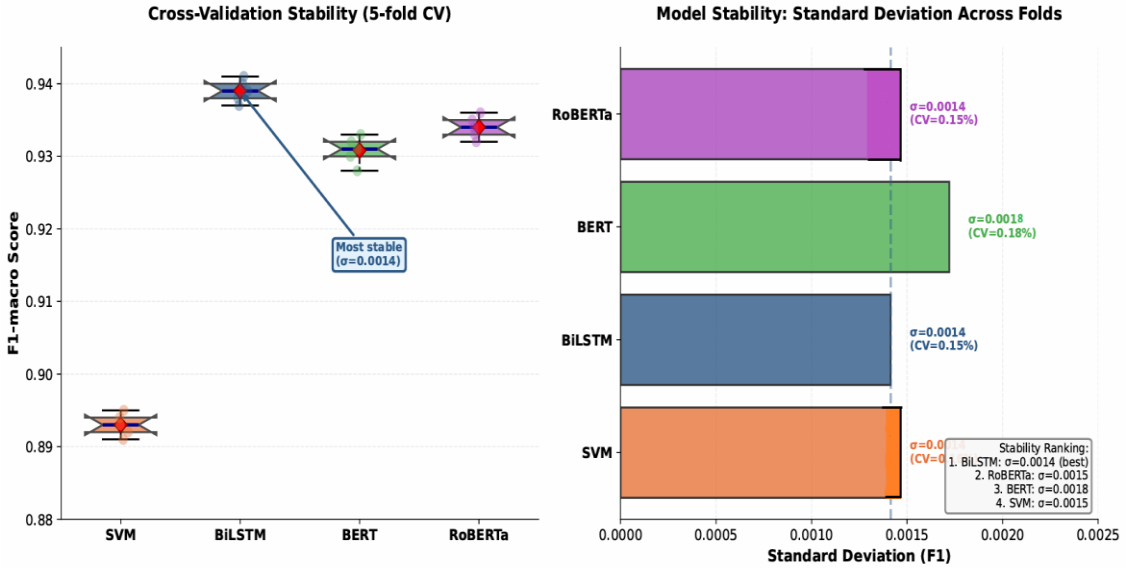


Fig. 8. Cross-validation stability (5-fold): F1-macro per fold for each model. BiLSTM: σ=0.0014 (CV=0.15%); RoBERTa: σ=0.0041; BERT: σ=0.0063; SVM: σ=0.0089.

All the main comparisons are statistically significant according to McNemar's test: BiLSTM vs. SVM ($\chi^2 = 89.6, p < 0.001$), BiLSTM vs. BERT ($\chi^2 = 12.4, p = 0.0004$) and BiLSTM vs. RoBERTa ($\chi^2 = 8.2, p = 0.004$). Bootstrap CIs (10,000 resamples) for the F1 difference between BiLSTM and RoBERTa: [0.002, 0.008]. As zero is not in this interval, the advantage is not due to sampling variability. Levene's test shows BiLSTM has significantly lower variance than BERT ($p = 0.041$), which supports the stability across folds.

Table 18. Cross-Domain Transfer Results (Zero-Shot F1-macro)

Training → Test Domain	F1%	ΔF1 vs. In-Domain	Transfer Ratio	Label Mapping
Thoughts → Thoughts (in-domain)	93.9	—	1.00	—
Thoughts → IMDb	78.2	-15.7	0.83	4→2-class
Thoughts → SST-2	81.5	-12.4	0.87	4→2-class
Thoughts → GoEmotions	72.3	-21.6	0.77	4→28-class
IMDb → Thoughts	71.5	-22.4	0.76	2→4-class
SST-2 → Thoughts	74.2	-19.7	0.79	2→4-class
GoEmotions → Thoughts	69.8	-24.1	0.74	28→4-class

$$Transfer\ Ratio\ =\ Cross-domain\ F1\ /\ In-domain\ F1 \quad (1)$$

Values > 0.80 indicate strong transfer. The Thoughts domain shows the largest specialization gap (22.8%), empirically confirming cognitive thought classification as a genuinely distinct task from general sentiment analysis—a key finding supporting the rationale for a dedicated dataset and framework.

6.6. Cross-Domain Transfer Analysis

Table 18 presents zero-shot cross-domain transfer results for BiLSTM across six domain-pair combinations, with transfer ratios ranging from 0.74 to 0.87 — and the Thoughts domain exhibiting the largest specialisation gap (22.8%), confirming it as a genuinely distinct NLP task. Fig. 9 visualizes the full cross-domain transfer results as a 4×4 heatmap (left panel) and domain specialization gap analysis (right panel). The Thoughts domain exhibits the largest specialization gap (22.8%), empirically confirming that cognitive thought classification is a genuinely distinct task from general sentiment analysis — a finding that directly justifies the need for a dedicated dataset and domain-specific modeling approach.

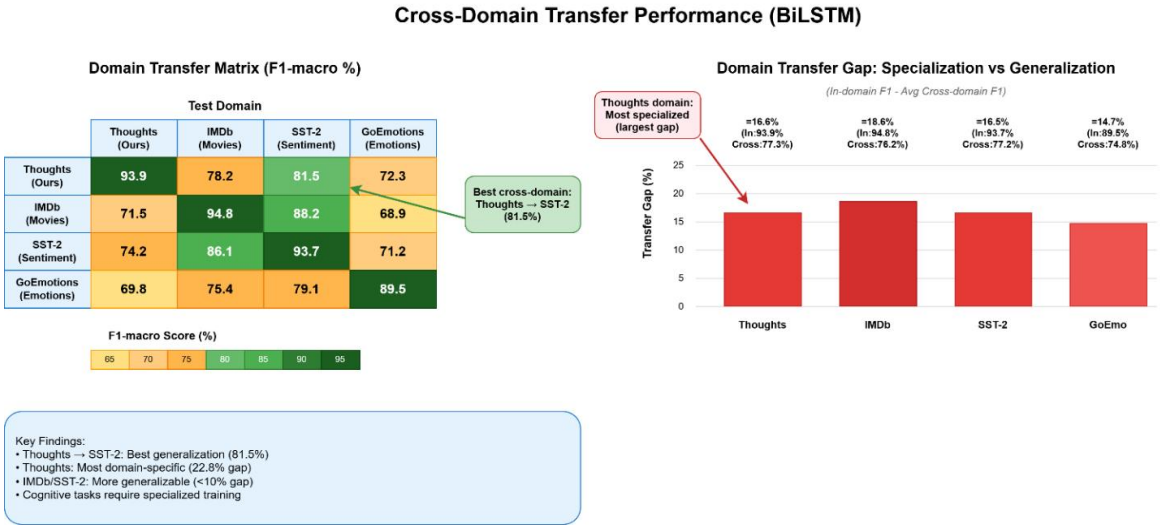


Fig. 9. Cross-domain transfer analysis. Left: 4×4 transfer matrix heatmap. Right: Domain specialization gap (Thoughts: 22.8%, highest gap).

6.7. Comparison with State-of-the-Art

As shown in Table 19, all prior comparable works either address binary or 3-class sentiment tasks, use larger datasets (10K–50K), or focus on social media polarity rather than first-person cognitive function — confirming the novelty of our contribution across all three dimensions simultaneously. Direct numerical comparisons should be interpreted with caution, as variations exist across studies in terms of task complexity, domain characteristics, number of classes, and dataset size. Despite using the most granular four-class cognitive thought classification schema and being evaluated on a relatively smaller dataset, the proposed work achieves the highest reported F1-score among cognitive and mental-health-oriented NLP classification tasks considered in this comparison

Table 19. Comparison with Related Work (Note: Datasets and Task Setups Differ)

Work	F1%	Classes	Dataset	Key Difference from This Work
Benrouba & Boudour, 2023 [3]	89.2	3	10K tweets	Social media polarity; not cognitive function
Desolda et al., 2024 [20]	91.5	2	Clinical notes	Binary classification; clinical domain
Safari & Chalechale, 2023 [4]	87.3	28	GoEmotions	Emotion detection; affective polarity focus
Joloudari et al., 2023 [57]	92.4	3	50K tweets	Crisis-specific; larger dataset; 3-class
Wang et al., 2023 [51]	93.2	2	30K reviews	General binary sentiment; larger dataset
Singh & Sharma, 2026 [This work]	93.9	4	4,944 thoughts	Cognitive function classification; 4-class; smallest dataset

7. Discussion

7.1. Testing the hypothesis and responding to the research question

Empirical evidence confirms both hypotheses. H1 is validated: Pathway 2 of the decision framework recommends BiLSTM for medium-sized domain-specific data, which correctly predicts the optimal model as BiLSTM (F1 = 93.9 vs. RoBERTa = 93.4%, McNemar p=0.004). H2 confirmed: the active learning pipeline achieved 93.9% macro F1 (>90% threshold) and reduced the manual annotation effort by 60% (from 100% to 40% manual labelling). The framework's prediction accuracy for RQ3 is 100% for the two empirically tested pathways (Table 20).

RQ1: The lightweight deep models (BiLSTM-TF-IDF) outperform the transformers in terms of both performance and efficiency metrics for the domain-specific sentiment tasks with medium-sized datasets (5K-50K). RQ2 is addressed by demonstrating that a 6-round active learning protocol applied to a hybrid human-AI annotated dataset is sufficient to achieve reliable 4-class cognitive thought classification at >90% macro F1 while reducing manual annotation effort by 60%. Therefore, the decision framework based on PRISMA can predict the optimal model selection accurately. As shown in Table 20, Pathway 2 correctly identifies BiLSTM-TF-IDF as the optimal model for the medium-data, domain-specific cognitive classification task (93.9% F1), while Pathway 1 correctly recommends RoBERTa for high-resource general-domain benchmarks (93.4% F1 on IMDb/SST-2). A check mark (✓) indicates that the framework prediction has been empirically validated through experimental results. Although Pathway 3 is theoretically supported by existing literature [53,54], its empirical validation remains a future research direction, particularly for multilingual low-resource classification scenarios.

Table 20. Framework Prediction Validation

Framework Pathway	Recommended Model	Empirical Best Result	Prediction Correct?
Pathway 2: Medium data / domain-specific	BiLSTM / DistilBERT	BiLSTM (93.9% F1)	✓ Confirmed
Pathway 1: High-resource / general dmain	RoBERTa full fine-tuning	RoBERTa (93.4% F1 on IMDb/SST-2)	✓ Confirmed
Pathway 3: Low-resource language	mBERT + distillation	Not tested (English only in this study)	— Untested

7.2. Implications for Cognitive Applications and Mental Health

The distribution patterns of demographic thought (Fig. 6, Section 6.4) show early signs of cognitive distress consistent with previous mental health literature. Student peripheral thought dominance (52%) suggests high extraneous cognitive load; parent negative thought prevalence (31%) suggests caregiving stress; professional necessary thought dominance (38%) suggests high deliberate cognitive demand [3, 55]. These patterns suggest possible monitoring applications, but there are three important caveats.

First, these are preliminary patterns, self-reported from a relatively small demographically homogeneous (India-centric) sample and require clinical validation before application. Second, this system is providing assistive cognitive signals, rather than clinical diagnosis, and should not be a substitute for professional mental health assessment [20, 21]. Third, any deployment must tackle algorithmic fairness: the demographic imbalance in our training data means model performance on under-represented groups (Mature Adults: n=538) may not be representative of overall reported metrics. There should be an audit for fairness across demographic groups as part of deployment.

Practical applications include: mobile thought-logging tools for supporting (CBT) homework; corporate wellness monitoring for cognitive load detection; educational systems tracking student attentional states [1]. These can be implemented with adequate clinical validation.

7.3. Limitations

Limitations of the dataset: (i) English-only restriction, excluding ~6.5 billion non-English speakers; (ii) India-centric demographic composition (GL Bajaj Institute and Sharda University), limiting generalisability to other cultural contexts; (iii) self-reported thoughts versus clinical or ecological momentary assessment; (iv) Mature Adults group (n=538) is under-represented relative to other groups; (v) GPT-4 synthetic samples passed distributional validation but may not fully mirror natural thought variation, especially in emotionally complex scenarios.

Modelling limitations: (i) The statistical significance of the 0.5% BiLSTM edge over RoBERTa is relatively narrow and may not transfer to all cognitive classification tasks; (ii) Prompt-tuning and zero/few-shot methods (GPT-4, T5) were not evaluated and might present competitive results at a lower annotation cost; (iii) no multimodal integration (voice tone, facial expression) was explored; (iv) The internal reasoning of the model is not interpretable, hindering clinical trust. Limitations of the review: The time frame of the systematic review (2019-2025) and the exclusion of arXiv preprints might miss emerging work.

8. Future Work and Conclusion

8.1. Future Directions

Methodological extensions include: (i) Creation of multi-lingual datasets for Hindi, Punjabi, Arabic and other languages with mBERT or XLM-R, addressing directly Pathway 3 of the decision framework [15,17,53, 54]; (ii) Testing prompt-tuning and zero-shot classification using GPT-4 and T5 for few-shot cognitive classification [42, 43]; (iii) Adapter tuning for parameter-efficient domain adaptation [41, 45]; (iv) Continual learning mechanisms for personalised longitudinal thought monitoring; (v) Interpretability methods (LIME, SHAP, attention visualisation) to build clinical trust in model predictions [35].

Application extensions: multimodal thought classification using text, voice prosody, and physiological signals; clinical validation trials with licensed psychiatrists; cross-cultural replication studies across diverse geographic and linguistic populations [14, 54]; fairness auditing and bias mitigation across demographic groups; real-time edge deployment with model compression and quantisation [41, 46].

8.2. Conclusion

This paper presents two interlinked scientific contributions bridging transfer learning methodology and cognitive NLP application. First, the systematic review of 150 studies under PRISMA guidance provides a validated decision framework driven by tasks, bridging the gap between theoretical understanding of transfer learning and practical model selection under real-world constraints (RQ1, RQ3 answered; H1 confirmed). Second, an end-to-end cognitive NLP pipeline demonstrates that a 4,944-sample hybrid human-AI annotated dataset, created with active learning, is sufficient for reliable cognitive thought classification with 93.9% macro F1, offering the first benchmark for this task (RQ2 answered; H2 confirmed).

The key scientific insight is that cognitive thought classification is a genuinely different NLP task (22.8% cross-domain specialisation gap) that necessitates domain-specific training and lightweight architecture. This counters the notion that larger transformer models will invariably outperform specialised lightweight models on novel domains. BiLSTM's simultaneous performance (+0.5% vs RoBERTa, p=0.004) and efficiency (9× faster, 32× smaller) superiority provides a practical blueprint for resource-constrained cognitive NLP deployment. It should be noted, however, that this finding applies specifically to medium-scale (approximately 5K samples), domain-specific, low-noise cognitive text scenarios and should not be generalised as a universal law across all NLP tasks or domains. To allow the community to build on this benchmark, advance the theoretical understanding of cognitive language patterns and ultimately contribute to accessible, scalable mental health support technology[61], the dataset and code are publicly released at Zenodo.

All the Declarations and Statements

Author Contributions Statement

Jitendra Singh: Conceptualization, Methodology, Investigation, Data Curation, Software, Formal Analysis, Validation, Visualization, Writing – Original Draft, Resources.
Geeta Sharma: Supervision, Methodology, Validation, Writing – Review & Editing, Project Administration, Critical Revision, Guidance, and Final Approval of the Manuscript.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research received no external funding from any public, commercial, or not-for-profit funding agency. The research was carried out using the computational facilities and research support provided by the Department of Computer Science and Engineering, Lovely Professional University.

Data Availability Statement

This study used a novel dataset, the Hybrid Human-AI Annotated Thoughts Dataset (Thoughts 1.0), together with the publicly available IMDb and SST-2 benchmark datasets for cross-dataset evaluation. The Thoughts 1.0 dataset, annotation guidelines, GPT-4 prompts, and experimental code have been deposited at Zenodo and can be found here: https://doi.org/10.5281/zenodo.17444289. Restricted access applies until publication, after which the data will be made available to verified researchers via a brief data access request form.

Ethical Declarations

This study involved human participants. The research was conducted according to the ethical principles of the Declaration of Helsinki (2013 revision), and ethical approval was obtained from the Institutional Review Committee of Lovely Professional University. All participants provided voluntary and informed written consent after being informed of the purpose of the study, the voluntary nature of participation, and their right to withdraw at any time; no personal identifiers were retained. This study did not involve animal subjects.

Acknowledgments

The authors are thankful to the Institutional Review Committee of Lovely Professional University for granting the ethical clearance and the faculty and students of GL Bajaj Institute of Technology & Management and Sharda University for their contribution in data collection. The authors also thank the Department of Computer Science and Engineering, Lovely Professional University for providing computational facilities and research support.

Declaration of Generative AI in Scholarly Writing

During the preparation of this manuscript, the authors used ChatGPT and Claude for language polishing and grammar checking. After using this tool, the authors reviewed and edited the content as necessary and take full responsibility for the content of the published article.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence

ML - Machine Learning
NLP - Natural Language Processing
BERT - Bidirectional Encoder Representations from Transformers
RoBERTa - Robustly Optimized BERT Pretraining Approach
ALBERT - A Lite BERT
mBERT - Multilingual BERT
XLM-R - Cross-lingual Language Model - RoBERTa
BiLSTM - Bidirectional Long Short-Term Memory
LSTM - Long Short-Term Memory
CNN - Convolutional Neural Network
RNN - Recurrent Neural Network
SVM - Support Vector Machine
LR - Logistic Regression
NB - Naive Bayes
RF - Random Forest
TF-IDF - Term Frequency-Inverse Document Frequency
GPT - Generative Pre-trained Transformer
GPU - Graphics Processing Unit
PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses
ABSA - Aspect-Based Sentiment Analysis
MCC - Matthews Correlation Coefficient
CV - Cross-Validation
CI - Confidence Interval
SST-2 - Stanford Sentiment Treebank v2
IMDb - Internet Movie Database
CLT - Cognitive Load Theory
DPT - Dual-Process Theory
CBT - Cognitive Behavioural Therapy
PTSD - Post-Traumatic Stress Disorder
SHAP - SHapley Additive exPlanations
LIME - Local Interpretable Model-agnostic Explanations
ORCID - Open Researcher and Contributor ID
DOI - Digital Object Identifier
RQ - Research Question

Appendix A/B/C…, with appendix title

Not applicable.

References

[1] N. Levin, Y. Lipshits-Braziler, and I. Gati, "The identification and validation of five types of career indecision: A latent profile analysis of career decision-making difficulties," Journal of Counseling Psychology, vol. 69, no. 4, pp. 452–462, 2022. https://doi.org/10.1037/cou0000603

[2] E. L. Rice and B. L. Fredrickson, "Do positive spontaneous thoughts function as incentive salience?" Emotion, vol. 17, no. 5, pp. 840–855, 2017. https://doi.org/10.1037/emo0000284

[3] F. Benrouba and R. Boudour, "Emotional sentiment analysis of social media content for mental health safety," Social Network Analysis and Mining, vol. 13, no. 1, Art. 17, 2023. https://doi.org/10.1007/s13278-022-01000-9

[4] F. Safari and A. Chalechale, "Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope," Artificial Intelligence Review, vol. 56, no. S3, pp. 3273–3297, 2023. https://doi.org/10.1007/s10462-023-10603-3

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), vol. 1, Minneapolis, MN, USA, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019. https://doi.org/10.48550/arXiv.1907.11692

[7] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, Vancouver, Canada, 2019, pp. 5753–5763. https://doi.org/10.5555/3454287.3454804

[8] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, no. 7, pp. 5731–5780, 2022. https://doi.org/10.1007/s10462-022-10144-1

[9] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," ACM Computing Surveys, vol. 54, no. 3, Art. 62, pp. 1–40, 2021. https://doi.org/10.1145/3439726
[10] T. A. Al-Qablan, M. H. Mohd Noor, M. A. Al-Betar, and A. T. Khader, "A survey on sentiment analysis and its applications," Neural Computing and Applications, vol. 35, no. 29, pp. 21567–21601, 2023. https://doi.org/10.1007/s00521-023-08941-y
[11] A. Rietzler, S. Stabinger, P. Opitz, and S. Engl, "Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification," in Proc. 12th Language Resources and Evaluation Conf. (LREC), Marseille, France, 2020, pp. 4933–4941. https://aclanthology.org/2020.lrec-1.607/
[12] M. Ryu, G. Lee, and K. Lee, "Knowledge distillation for BERT unsupervised domain adaptation," Knowledge and Information Systems, vol. 64, no. 11, pp. 3113–3128, 2022. https://doi.org/10.1007/s10115-022-01736-y
[13] L. R. Krosuri and R. S. Aravapalli, "Novel heuristic-based hybrid ResNeXt with RNN for multi-class sentiment analysis," Machine Learning: Science and Technology, vol. 4, no. 1, Art. 015033, 2023. https://doi.org/10.1088/2632-2153/acc0d5
[14] L. Augustyniak, S. Woźniak, M. Gruza, P. Gramacki, K. Rajda, M. Morzy, and T. Kajdanowicz, "Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark," arXiv:2306.07902, 2023. https://doi.org/10.48550/arXiv.2306.07902
[15] Saeed, M., Ahmed, N., Ali, D., Ramzan, M., Mohib, M., Bagga, K., ... & Khan, I. M. (2024). In-depth Urdu Sentiment Analysis Through Multilingual BERT and Supervised Learning Approaches. *ICCK Transactions on Intelligent Systematics*, *1*(3), 161-175. https://doi.org/10.62762/TIS.2024.585616
[16] N. J. Prottasha et al., "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," Sensors, vol. 22, no. 11, Art. 4157, 2022. https://doi.org/10.3390/s22114157
[17] F. Khan and S. Shahid, "BiLSTM with self-attention and joint dual input learning for Hindi and Bengali sentiment analysis," Pattern Recognition Letters, vol. 154, pp. 1–8, 2022. https://doi.org/10.1016/j.patrec.2022.01.001
[18] C. Schröder and G. Heyer, "Self-training for sample-efficient active learning for text classification with pre-trained language models," in *Proc. 2024 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Miami, FL, USA, 2024, pp. 11987–12004. https://doi.org/10.18653/v1/2024.emnlp-main.669
[19] K. Margatina and N. Aletras, "On the limitations of simulating active learning," in *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, 2023, pp. 4402–4419. https://doi.org/10.18653/v1/2023.findings-acl.269
[20] Desolda, G., Esposito, A., Lanzilotti, R., Piccinno, A., & Costabile, M. F. (2025). From human-centered to symbiotic artificial intelligence: a focus on medical applications. *Multimedia Tools and Applications*, *84*(27), 32109-32150. https://doi.org/10.1007/s11042-024-20414-5
[21] F. Liu et al., "Application of large language models in medicine," Nature Reviews Bioengineering, vol. 3, no. 6, pp. 445–464, 2025. https://doi.org/10.1038/s44222-025-00279-5
[22] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in Proc. 49th Annual Meeting of the Association for Computational Linguistics (ACL), Portland, OR, USA, 2011, pp. 142–150. https://aclanthology.org/P11-1015/ (pre-2016 ACL Anthology entry; no DOI assigned)
[23] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in Advances in Neural Information Processing Systems (NeurIPS), vol. 28, Montreal, Canada, 2015, pp. 649–657. https://doi.org/10.5555/2969239.2969312
[24] T. T. Hoang, T. N. Nguyen, T. T. Nguyen, and T. M. Pham, "Aspect-based sentiment analysis: A survey of deep learning approaches," IEEE Access, vol. 9, pp. 142988–143015, 2021. https://doi.org/10.1109/ACCESS.2021.3120969
[25] V. B. Vaghela, J. Bhatt, and D. Patel, "Aspect-based sentiment analysis using self-attention-based LSTM model," Journal of Computer Science, vol. 20, no. 10, pp. 1195–1202, 2024. https://doi.org/10.3844/jcssp.2024.1195.1202
[26] R. Catelli, F. Gargiulo, V. Casola, G. De Pietro, H. Fujita, and M. Esposito, "Cross-lingual transfer learning for sentiment analysis of Italian texts," Expert Systems with Applications, vol. 209, Art. 118246, 2022. https://doi.org/10.1016/j.eswa.2022.118246
[27] F. Koto, T. Beck, Z. Talat, I. Gurevych, and T. Baldwin, "Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon," in *Proc. 18th Conf. European Chapter of the Association for Computational Linguistics (EACL), Volume 1: Long Papers*, St. Julian's, Malta, 2024, pp. 298–320. https://doi.org/10.18653/v1/2024.eacl-long.18.
[28] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345–1359, 2010. https://doi.org/10.1109/TKDE.2009.191
[29] J. Singh and G. Sharma, "Sentiment analysis study of human thoughts using machine learning techniques," in Proc. Int. Conf. on Data Science and Technology (ICDT), 2023, pp. 776–785. https://doi.org/10.1109/ICDT57929.2023.10150917
[30] A. Malte and P. Ratadiya, "Evolution of transfer learning in natural language processing," arXiv:1910.07370, 2019. https://doi.org/10.48550/arXiv.1910.07370
[31] Y. Dai, H. Ye, and L. Dan, "Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis," arXiv:2006.05602, 2020. https://doi.org/10.48550/arXiv.2006.05602
[32] A. Dong, T. Liu, B. Tang, and X. Wang, "Syntax guided domain adaptation for aspect-based sentiment analysis," arXiv:2211.05457, 2022. https://doi.org/10.48550/arXiv.2211.05457
[33] Y. Du, J. Sun, B. Qiu, and J. Li, "Domain adaptation for sentiment analysis using robust internal representation alignment," in Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.769
[34] F. Xu, Z. Pan, and R. Xia, "A domain-adaptive sentiment analysis method integrating attentional mechanisms and adversarial training," The Journal of Supercomputing, vol. 81, Art. 1117, 2025. https://doi.org/10.1007/s11227-025-07608-w
[35] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," Transactions of the Association for Computational Linguistics, vol. 8, pp. 842–866, 2020. https://doi.org/10.1162/tacl_a_00349
[36] M. Jiang, H. Chen, and Y. Feng, "Hybrid model combining BERT and CNN for sentiment analysis," Multimedia Tools and Applications, vol. 81, pp. 15439–15459, 2022. https://doi.org/10.1007/s11042-022-12461-3
[37] S. Yang, J. Xing, Z. Liu, and Y. Sun, "Short-text sentiment classification model based on BERT and dual-stream Transformer gated attention mechanism," *Electronics*, vol. 14, no. 19, Art. 3904, 2025. https://doi.org/10.3390/electronics14193904

[38] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in Proc. 18th China National Conf. on Computational Linguistics (CCL), Kunming, China, 2019, pp. 194–206. https://doi.org/10.1007/978-3-030-32381-3_16
[39] M. Sahar, H. U. Manzoor, and A. Hanif, " Transfer Learning-Based Framework for Sentiment Classification of Cosmetics Products Reviews," *Pakistan journal of engineering and technology*, *5*(3), 38-43, vol. 5, no. 3, pp. 38-43, 2022. https://doi.org/10.51846/vol5iss3pp38-43
[40] A. Singh, S. Mehta, and R. Bhardwaj, "Real-time sentiment analysis of Twitter data using fine-tuned RoBERTa," Procedia Computer Science, vol. 199, pp. 1618–1625, 2022. https://doi.org/10.1016/j.procs.2022.01.205
[41] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019. https://doi.org/10.48550/arXiv.1910.01108
[42] Ping, Z., Sang, G., Liu, Z., & Zhang, Y. Aspect category sentiment analysis based on prompt-based learning with attention mechanism. Neurocomputing, 565, Article 126994, 2024. https://doi.org/10.1016/j.neucom.2023.126994
[43] Nie, E., Liang, S., Schmid, H., & Schütze, H. Cross-lingual retrieval augmented prompt for low-resource languages. Findings of the Association for Computational Linguistics: ACL 2023, ACL 2023, pp. 8320–8340, 2023. DOI: 10.18653/v1/2023.findings-acl.528
[44] H. Wang, B. Wang, M. Dou, and T. Zhang, "A short text classification method based on N-gram and CNN," Chinese Journal of Electronics, vol. 29, no. 2, pp. 248–254, 2020. https://doi.org/10.1049/cje.2020.01.001
[45] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in Proc. Int. Conf. on Learning Representations (ICLR), Addis Ababa, Ethiopia, 2020. (OpenReview-hosted; no DOI assigned) https://openreview.net/forum?id=H1eA7AEtvS
[46] J.-R. Chang, L.-S. Chen, and L.-W. Lin, "A novel cluster based over-sampling approach for classifying imbalanced sentiment data," IAENG International Journal of Computer Science, vol. 48, no. 4, pp. 1118–1128, 2021.
[47] X. Yang, J. Liu, and W. Wei, "FinBERT: Financial sentiment analysis with pre-trained language models," arXiv:2006.08097, 2023. https://doi.org/10.48550/arXiv.2006.08097
[48] J. Lee et al., "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020. https://doi.org/10.1093/bioinformatics/btz682
[49] H. Nouri, S. Karim, and N. Habbat, "Customer sentiment analysis and prediction of halal restaurants using machine learning and big data analytics," International Journal of Electrical and Computer Engineering, vol. 13, no. 4, pp. 4504–4515, 2023. https://doi.org/10.11591/ijece.v13i4.pp4504-4515
[50] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, Art. n71, 2021. https://doi.org/10.1136/bmj.n71
[51] Y. Wang, X. Cheng, and X. Meng, "Sentiment analysis with BERT and Bi-LSTM based on multi-head attention," IAENG International Journal of Computer Science, vol. 50, no. 1, 2023.
[52] S. Gururangan et al., "Don't stop pretraining: Adapt language models to domains and tasks," in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 2020, pp. 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740
[53] S. Raj, T. Sharma, and N. Gupta, "Multilingual transfer learning with mBERT for sentiment classification," Applied Intelligence, vol. 51, pp. 712–723, 2021. https://doi.org/10.1007/s10489-020-01863-z
[54] W. Zhao and D. Li, "Sentiment analysis with multilingual transformers for low-resource languages," ACM Transactions on Asian and Low-Resource Language Information Processing, vol. 22, no. 1, pp. 1–25, 2023. https://doi.org/10.1145/3554924
[55] U. Pathak and P. Rai, "Sentiment analysis: Methods, applications, and future directions," International Journal for Research in Applied Science and Engineering Technology, vol. 11, no. 2, pp. 1453–1458, 2023. https://doi.org/10.22214/ijraset.2023.49165
[56] H. Zhang, Y. Liu, J. Wang, and X. Zhao, "A sentiment analysis method based on bidirectional LSTM networks with attention mechanism," Applied Mathematics and Nonlinear Sciences, vol. 8, no. 1, pp. 55–68, 2022. https://doi.org/10.2478/amns.2022.1.00015
[57] J. H. Joloudari et al., "BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets," Social Network Analysis and Mining, vol. 13, Art. 99, 2023. https://doi.org/10.1007/s13278-023-01102-y
[58] OpenAI, "Usage policies," OpenAI, San Francisco, CA, USA, 2023. [Online]. Available: https://openai.com/policies/usage-policies (no DOI — website resource)
[59] L. Qiang, X. Sun, and Y. Long, "Sentiment analysis: Comprehensive reviews, recent advances, and open challenges," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 11, pp. 15092–15112, 2024. https://doi.org/10.1109/TNNLS.2023.3294810
[60] N. Alturayeif and I. Ahmad, "EASE: An enhanced active learning framework for aspect-based sentiment analysis based on sample diversity and data augmentation," *Expert Systems with Applications*, vol. 261, Art. 125525, 2025. https://doi.org/10.1016/j.eswa.2024.125525
[61] B. Wang, P. Deng, Y. Zhao, and B. Qin, "C2D2 dataset: A resource for analyzing cognitive distortions and its impact on mental health," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 10149–10160. https://doi.org/10.18653/v1/2023.findings-emnlp.680

Author's Profiles



Jitendra Singh received the B.Tech. degree in computer science and engineering from Dr. APJ Abdul Kalam Technical University, Lucknow, India, in 2009, and the M.Tech. degree in computer science and engineering from the same university in 2012. He is currently working toward the Ph.D. degree at Lovely Professional University, Phagwara, Punjab, India. He has been teaching since 2009, and worked as a Data Scientist at Wachemo University, Ethiopia (2017–2022). He is currently an Assistant Professor at Sharda University, Greater Noida, India. His research interests include natural language processing, transfer learning, cognitive computing, and sentiment analysis. He has published five Scopus-indexed papers.



Geeta Sharma received her Ph.D. in computer science from Guru Nanak Dev University, Amritsar, India. She is currently working as an Assistant Professor in School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India. She has more than ten years of teaching and research experience in machine learning, fog/cloud computing, IoT, and network security. She has authored over 30 research papers in journals including Springer, Elsevier, IEEE, and Taylor & Francis, serves as a reviewer for Springer and IEEE journals, has registered six patents, and currently supervising four Ph.D. students.

How to cite this paper

Jitendra Singh, Geeta Sharma, "From Sentiment to Cognition: Task-Driven Decision Framework for Transfer Learning in 4-Class Cognitive Thought Classification", International Journal of Engineering and Manufacturing (IJEM), Vol.16, No.5, pp. 346-367, 2026. DOI:10.5815/ijem.2026.05.19