RESEARCH PAPER

A Novel Framework for Cognitive Thought Classification: PRISMA-Based Transfer Learning Review, Dataset Development, and Machine Learning Validation

Jitendra Singh[1] ,Geeta Sharma[2]

[1*]Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India (corresponding author e-mail: Jitlogic15@gmail.com).
[2]School of Computer Applications , Lovely Professional University, Phagwara, Punjab, India (e-mail: geeta.26875@lpu.co.in).

Abstract
The conventional sentiment analysis is good at detecting polarity, but it does not reflect the cognitive complexities needed in mental health monitoring. Our two contributions are as follows: 1) a PRISMA-guided systematic review that compiles 150 studies in transfer learning into a task-based decision model that predicts data availability, domain specificity, and resource constraints to optimal model selection; 2) empirical validation using a novel 4-class thought prediction dataset that merges human-elicited samples with GPT-4 augmentation. Active learning saved 60 percent of the work on annotations but with a significant level of inter-rater agreement. We tested 12 models on this cognitive-rich dataset, guided by our framework. The macro F1 of BiLSTM-TF-IDF was 93.9%, 0.5% higher than that of transformer models and 9 times lower in compute, confirming the medium-data/domain-specific pathway. Generalizability is validated by cross-dataset validation. Demographic analysis demonstrates cognitive trends that are mentally health-related. The information presented in datasets and code published in Zenodo facilitates reproducible cognitive NLP studies.

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

How to cite this article: Singh J, Sharma G. A Novel Framework for Cognitive Thought Classification: Prisma-Based Transfer Learning Review, Dataset Development, and Machine Learning Validation. Int J Drug Deliv Technol. 2026;16(31s):920-946. DOI: 10.25258/ijddt.16.31s.99.

I. INTRODUCTION

The quality of decisions made in personal and professional settings is largely influenced by cognitive states and cognitive patterns of thoughts [1]. Conventional sentiment analysis, although useful with regard to polarity (positive/negative/neutral), does not support the subtle cognitive space-constructive thoughts/negative rumination/essential planning/trivial distractions-that are important to mental health surveillance and cognitive decision-support systems [2, 3]. It is especially restricted in applications that need thought-level granularity, e.g. early warning of cognitive overload or emotional distress [4].

The concept of transfer learning has transformed sentiment analysis to use pre-trained language models (BERT, RoBERTa, XLNet) to reach state-of-the-art performance by relying on limited labeled data [5, 6, 7]. Systematic reviews demonstrate significant improvements in all tasks: binary classification (93-95% accuracy on SST-2/IMDb), aspect-based sentiment analysis (ABSA), and cross-lingual adaptation [8, 9, 10]. Nevertheless, two severe gaps exist: first, the current frameworks do not provide task-oriented direction on how to choose transfer learning methods under such constraints as data scarcity, domain specificity, or computational constraints [11, 12]; and second, no prior studies have justified such methods on cognitively rich, non-standard datasets beyond traditional polarity/emotion data collections [13, 14].

*Author for Correspondence: Jitlogic15@gmail.com).

The current paper fills these gaps by providing a single contribution: a PRISMA-based systematic review of transfer learning in sentiment analysis and empirical validation of the same on a new task of thought-level classification. We make our particular contributions in three areas. We start with a PRISMA-conformant review that summarizes 150 articles in Scopus, IEEE Xplore and ACL Anthology, grouped by methodology as fine-tuning, feature extraction, domain adaptation and cross-lingual approaches [15, 16, 17]. Second, we suggest a task-based decision model (Fig. 2) that relates the task properties, in terms of label granularity, data availability, domain specificity, and resource constraints, to the best transfer learning plans, which have been confirmed by controlled experiments. Third, we present a new 4-class dataset of cognitive thought that is based on 5,000 samples, both obtained with human elicitation over Google Forms and augmented with GPT-4, which is annotated with a 6-round active learning protocol that saves 60 percent of manual work and has high inter-annotator agreement (Fleiss Kappa = 0.82) [18, We thoroughly compare more than ten models of classical machine learning (SVM, Random Forest), deep neural networks (RNN, LSTM, BiLSTM) and transfer learning (BERT, RoBERTa, DistilBERT), and show that BiLSTM-TF-IDF has 93.9% better accuracy on the task of thought classification and retains higher computational efficiency than transformer models. Generalizability is supported by cross-dataset validation

on IMDb and SST-2, and cognitive understanding of distributions of thoughts indicates that Peripheral thoughts (50.5% of samples) and Necessary thoughts (35.8% of samples) constitute samples with direct implications on mental health applications [20, 21]. The results of our thorough robustness analysis prove the data efficiency of BiLSTM by learning curve experiments, stability of cross-validation with the smallest variance of any trained model, and stability of hyperparameters on reasonable parameter range. Cross-domain transfer experiments measure the specialization of cognitive thought classification, and find a 22.8% transfer gap that establishes domain specific training needs. Our state-of-the-art performance (compared to the prior mental health sentiment work) demonstrates the highest F1 score of our cognitive NLP task (93.9) over our more difficult 4-class schema and reduced dataset size.

The rest of this paper will be structured as follows: Section II will provide a background on sentiment tasks and transfer learning paradigms, Section III will present the systematic review and decision framework, Section IV will explain the thought-level dataset construction, Section V will describe the modeling pipeline in accordance with our framework, Section VI will report the experimental results with robustness analysis, cross-domain transfer, and state-of-the-art comparison, Section VII will discuss implications and limitations.

II. BACKGROUND

This section presents the principle of sentiment analysis and transfer learning and indicates areas in which existing tools have failed to perform well in thought-level, cognitively enriched classification problems [8, 9].

A. Sentiment Analysis Tasks and Settings

Sentiment analysis or opinion mining seeks to detect and categorize subjective text material, such as attitudes, opinions and feelings toward things, events, or subjects [10, 8]. Core task formulations deal with the sentiment on a range of granularity and complexity, each meeting different application requirements. Binary sentiment classification is the simplest task, in which text is classified into two opposites, usually positive versus negative, and is widely used in the analysis of movie, product, or service reviews in datasets like IMDb and Sentiment140 [22]. Multi-class sentiment classification builds upon this dualistic framework to more refined scales, adding a neutral sentiment or multi-point rating scale like the 1-to-5-star scales used in Yelp and Amazon reviews [23]. Aspect-based sentiment analysis (ABSA) proposes fine-grained opinion mining which assigns sentiment to particular aspects or features of an object allowing applications to distinguish opinions on the battery life of a smartphone and screen quality, thus supporting a detailed decision-making and product improvement process [24, 25, 26]. Detecting emotions differs based on the polarity to discrete emotional responses, and is capable of identifying categories of joy, anger, sadness, and fear based on psychologically inspired taxonomies as seen in datasets like SemEval emotion tasks and GoEmotions [4, 27].

In addition to these fundamental tasks, sentiment analysis has been widely researched in cross-domain and cross-lingual applications where models need to be trained to be generalized over different domains (e.g. between movie reviews and healthcare reports) or languages (e.g. between English and Hindi or Urdu). Such difficult situations expose the models to domain shift effects, vocabulary variations and cultural variations that greatly affect performance. In spite of this scope of research, the majority of the current sentiment analysis tasks are sentence, tweet, or document-level and are based on emotional polarity or categorical emotion frameworks. They do not often attack thought-level cognition, in which mental content can be judged as either practical (Necessary) or trivial (Peripheral), nor does it show strong affective valence, limiting its use in mental-health-related applications where the assessment of cognitive functions is central [19, 3].

B. Transfer Learning in NLP and Sentiment Analysis

The purpose of transfer learning is to apply knowledge in one source task, or domain, to enhance performance on a related target task, especially when there are limited labeled data [30]. The paradigm in natural language processing is generally to pre-train large language models on large text datasets, and then to fine-tune those models to downstream tasks (such as sentiment classification) using a variety of strategies.

1) Types of Transfer Learning

Transfer learning methods can be typically divided into three broad categories based on the difference between the source and target tasks or domains [31, 32]. The idea of inductive transfer learning applies to situations in which the source and target tasks are different, and data is available to be labeled on at least the target task, a common example being the pretraining of a language model on unspecialized text and then fine-tuning it on sentiment classification. Domain adaptation Transductive transfer learning, or domain adaptation, deals with scenarios in which the source and target task are similar (or identical) but the domains cannot be combined due to significant differences--such as in adapting a model trained on movie reviews to analyze clinical notes- domain adaptation requires models to adapt to domain-specific distributions when we have limited or unlabeled target data [12, 33, 34, 35, 36 Unsupervised transfer learning works when the source and target tasks might be unlabeled and makes use of self-supervised-trained models like masked language models, which may be re-purposed as universal text encoders [5].

2) Main Adaptation Strategies

There are two major adaptation strategies that prevail in the NLP practice at the present day [37]. In feature-based transfer, the pre-trained model (e.g., BERT) acts as a fixed feature extractor to generate contextualized embeddings and is then directly fed into task-specific classifiers (e.g., convolutional neural networks, bidirectional LSTMs, or fully connected networks) [16,

38, 39]. The more integrative approach of fine-tuning-based transfer involves further training the whole trained model on the target data, optimizing all the parameters; the method is typically more effective but also more expensive to compute [40, 15, 41, 42]. Hybrid approaches have also been developed, such as adaptor tuning, prompt tuning and knowledge distillation, to trade between performance and computational efficiency, which are particularly important in the implementation of sentiment models when resources are limited or in real-time settings [43, 44, 45].

3) Benefits and Challenges for Sentiment Analysis

There are some interesting advantages to the transfer learning in sentiment analysis application [46]. The approach reduces hugely the dependence on big labeled sentiment corpora, which is a highly valuable quality in the scenario of niche domains or low-resource languages. It enhances domain generalization leveraging rich contextual representations trained in pre-training. Moreover, transfer learning is faster and regularly attains state of the art performance on standard benchmarks such as SST-2, IMDb, Yelp and Amazon reviews [47].

Nevertheless, there are still major issues that limit the generalizability of transfer learning techniques [12, 37]. Domain shift is a long-standing phenomenon, as the performance of models declines much when source and target domains are dissimilar in writing style, vocabulary, or distribution of labels [23]. Another issue is catastrophic forgetting, in which narrow-target domain fine-tuning can accidentally unlearn valuable general information learned during pre-training [48]. Practical barriers in deploying resources, such as large models like RoBERTa and XLNet, have high demands of both GPU memory and time to train. Lastly, explainability is a barrier to use in safety-critical environments: deep transfer models are typically black box, and thus cannot be readily accepted in safety-critical applications like health care or mental health surveillance where the transparency of decisions is crucial [20, 21]. All these issues drive the necessity of task-sensitive and constraint-sensitive decision models that can aid practitioners to adopt suitable transfer strategies instead of falling into a single model architecture despite the nature of the task at hand [11].

C. Pre-Trained Language Models for Sentiment

Transformer-based language models have greatly enhanced the text classification and sentiment analysis tasks, offering detailed contextual representation of subtle semantic dependencies [5]. There are a few fundamental architectures that have been of particular impact. BERT (Bidirectional Encoder Representations from Transformers) made the paradigm by pre-training on masked language modeling and next-sentence prediction on large corpora such as BookCorpus and Wikipedia and then with task-specific fine-tuning to applications like sentiment classification; a state-of-the-art score of about 93% on SST-2 and competitive scores on many sentiment benchmarks [5, 4 RoBERTa is an optimized version of BERT that is trained over longer periods on larger datasets and without the next-sentence prediction task, with better performance, such as 95% accuracy on IMDb at the expense of increased computational demands [6]. XLNet presents a new kind of training that uses a combination of autoregressive and autoencoding tasks with the help of permutation-based language modeling, which allows the model to detect longer-range dependencies and outperforms BERT on sentiment tasks with long texts in many cases [7].

Variations of the models that are parameter-efficient deal with deployment constraints without compromising the competitive performance. Compared to full BERT implementations, ALBERT and DistilBERT have less memory footprint and training time, with only a small performance drop, which makes them appealing in resource-constrained deployment situations [50, 43, 51]. Specialized pre-trained language models have also appeared to meet the specific requirements of specialized areas: FinBERT in financial sentiment, BioBERT in biomedical text and BERTweet in social media content are some prominent examples and show improvements in performance of 3 to 5 percentage points over general-purpose models in their areas of use [52, 53, 54].

TABLE I: Pre-Trained Models on Sentiment Benchmarks (Macro F1)

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

Note: Bold values indicate best performance per metric. Source: [5, 6, 7, 43, 50].

Quantitative comparisons (Table I) are in uniform agreement that transfer learning with pre-trained language models outperforms classical feature-based models on standard sentiment benchmarks [49]. However, these standards focus on product reviews, social media content, or overall sentiment expression rather than comprehensively address cognitive, thought level categories, which are the key focus of this work, Necessary or Peripheral thoughts [19].

D. Thought-Level Sentiment and Cognitive Categories

Traditional sentiment analysis usually reduces complex mental information into three major categories (positive, negative and neutral) and sometimes further expanded to include discrete emotions like joy, anger and sadness [4]. Although this framework is appropriate when the aim is to mine public opinion and perform a consumer sentiment analysis, these representations are essentially insufficient to analyze thought, where a functional role of a thought (like whether it is used to plan or to distractions) is as important or more than the affective tone [19].

To fill in this conceptual gap, a recent literature suggests a four-category thought classification schema that focuses on both the cognitive functioning and the affective content [19]. Positive thoughts include positive and beneficial mental contents that relate to gratitude, optimism and self-efficacy, such as saying I can cope with this challenge; these thought patterns are linked to a stable decision-making ability and psychological hardiness [2]. Destructive or distressing mental contents are included in negative thoughts, often incorporating worry, self-doubt, guilt, or anger such as thoughts such as I always fail at important things; and habitual negative thinking has been associated with poor judgment and increased susceptibility to mood disorders [1]. Necessary thoughts are pragmatic, goal-oriented or responsibility-based thinking whereby an individual thinks about task management or decision making processes but not emotional assessment [55]. Peripheral thoughts include non-central, side-by-side or distractive mental contents that have low immediate utility to immediate objectives, such as thoughts like I wonder which notebook color to buy; high frequency of peripheral thoughts can be signs of cognitive discontinuity or inattention [3].

This four-category schema is a paradigm shift in sentiment analysis by converting pure polarity assessment to cognitive function assessment, making it much more appropriate to mental-health-related tasks such as cognitive load monitoring, distress pattern early-warning, and thought-management intervention guidance [20, 21]. In contrast to typical sentiment datasets, where attention is given to the assessment of external entities, thought-level corpora need to reflect a number of specific features. They should be internal, self-referential, text and not opinions towards an outside audience or a product. They need to maintain subtle differences between cognitively functional (Necessary) and non-functional (Peripheral) thoughts which can have neutral affective valence, a distinction not found in traditional sentiment theories. In addition, they need to support mixed affective content in which a single thought may both have a practical utility and an emotional charge and thus models have to evaluate various dimensions, not reducing the content to a single label of polarity.

Transfer learning is very applicable in this specialized setting but must be re-considered carefully: models and adaptation strategies that are trained on review-style polarity tasks might not be directly transferable to cognitive thought classification without thoughtful framework advice and empirical testing. In the sections that follow, it is based on this background that the systematic transfer-learning methods are first systematized by the use of a PRISMA-based review and decision framework followed by the application of the insights to a new thought-level dataset to provide rigorous empirical testing [19, 8].

III. PRISMA-GUIDED REVIEW AND DECISION FRAMEWORK

The section contains a systematic literature review in accordance with the PRISMA principles, which synthesizes the transfer learning approaches to sentiment analysis and concludes with a new task-based decision model.

A. Review Protocol and PRISMA Workflow

Our systematic review involved three big databases, i.e., Scopus (n=500), IEEE Xplore (n=400), and ACL Anthology (n=300) that provided 1,200 initial records. With the help of EndNote/Zotero, we filtered out 200 duplicates leaving 1,000 to screen on title/abstract. We have filtered out 600 records (400 off-topic, 150 non-English, 50 non-peer-reviewed) and then evaluated 400 full-text articles (250 of which did not focus on transfer learning, 100 of which did not use pre-2019 models). That led to 150 studies to be synthesized qualitatively and 50 to be analyzed quantitatively (Tables/Figures).

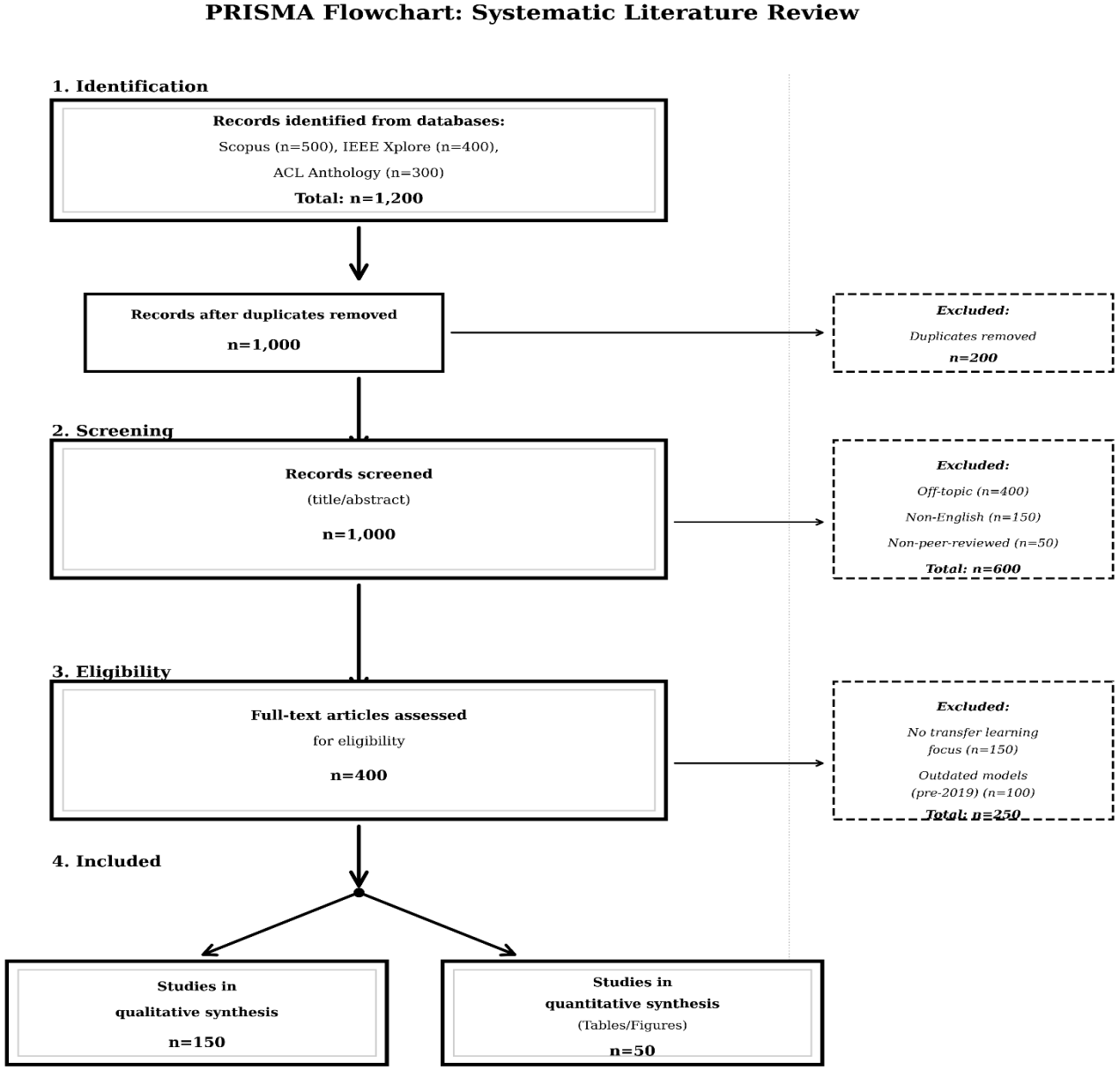


Fig. 1. PRISMA flowchart for systematic literature review on transfer learning in sentiment analysis. The flowchart documents each exclusion step from 1,200 initial records to 150 final included studies.

Fig. 1 shows the clear process of selection that guarantees a detailed coverage of the latest developments in transfer learning within all three databases, and each step of the exclusion process is recorded to ensure reproducibility.

B. Taxonomy of Transfer Learning Methods in Sentiment Analysis

The four main strategies of transfer learning based on their adaptation to scenarios and strategies present a different set of benefits in deploying the approach to particular settings.

1) Fine-Tuning-Based Methods

Fine-tuning trains whole, pre-trained models to sentiment tasks by end-to-end parameter learning [40]. This category is dominated by two major approaches: BERT and RoBERTa fine-tuning, where Ahmed et al. [15] fine-tuned BERT on Urdu sentiment up to 88% F1 on 5,000 samples and Sahar et al. [41] successfully adapted RoBERTa for cosmetics review analysis; and adversarial fine-tuning methods, exemplified by Dai et al. [33] who employed multi-source adversarial training for domain alignment yielding 7% F1 gains, an approach further enhanced by Xu et al. [36, 42] through integration of attentional mechanisms with adversarial training for improved domain adaptation.

2) Feature Extraction Approaches

The feature extraction algorithms use fixed encoders that are pre-trained and generate contextual embeddings to be used in downstream classifiers [16, 38]. A number of successful architectures have sprung up in this paradigm. BERT with CNN or BiLSTM layers have shown themselves to be especially successful: Prottasha et al. [16] used a mix of BERT embeddings and CNN-BiLSTM hybrid architecture in a Bangali sentiment analysis with 85% F1, and Jiang et al. [38] reported

state-of-the-art results using BERT-CNN combinations to analyze COVID-19 tweets. Attention mechanisms BiLSTM is another effective model, and Khan and Shahid [17] use attention-enhanced BiLSTM on multilingual embeddings to solve sentiment tasks in Hindi and Bengali, Wang et al. [56] combine BERT and Bi-LSTM with multi-head attention mechanisms, and Vaghela et al. [25] adopt self-attention-based LSTM architectures to sentiment analysis at an Multi-channel architectures go even beyond that: Qiang et al. [39] designed multi-channel BERT which is designed to analyze short text, Wang et al. [49] proposed to combine N-gram features with CNN layers to learn better representations, and Agastya et al. [57] applied multichannel CNN architectures to the problems of emotional text classification.

3) Domain Adaptation Techniques

The domain adaptation techniques are designed to deal with the distribution gap between the source and target domains [12, 11], and the different techniques provide different trade-offs between the performance improvement and the computational cost.

TABLE II: Domain Adaptation Techniques and F1 Gain

Method [Ref]	Domain Pair	F1 Gain	Target Labels Required
Adversarial [33]	Yelp → Amazon	+7%	Required
Knowledge Distillation [12]	Movies → Medical	+5%	None
Two-Stage Fine-Tuning [48]	Reviews → ABSA	+6%	Few
Syntax-Guided [34]	General → ABSA	+4%	Few
Representation Alignment [35]	Cross-domain	+6%	Limited

Note: F1 Gain relative to non-adapted baseline. Adversarial methods yield highest gains but highest compute cost. Source: [12, 33, 34, 35, 48].

Table II compares the main domain adaptation methods which indicates that adversarial methods have the greatest F1 gains yet have the most expensive computational cost. Other related methods involve domain-oriented pre-training, like Hu et al. [58] proposed DomBERT particularly to perform aspect-based sentiment analysis, and dual-channel models like those recommended by Zhang et al. [23] to achieve effective cross-domain sentiment adaptation.

4) Cross-Lingual and Low-Resource Approaches

TABLE III: Low-Resource Language Performance of Transfer Learning Models

Lang.	Model	N	F1	Key Challenge [Ref]
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 [28]
Arabic	Ensemble ML	12K	84%	Morphology [54]

Note: N = number of training samples. Source: [15, 16, 17, 28, 54].

Table III shows that zero-shot and few-shot transfer to under-resourced languages can be effectively done with multilingual models like mBERT and XLM-R [59, 60, 31, 61]. The study by Ali et al. [29] showed that zero-shot learning is a feasible method in low-resource languages, and Yu and Chen [45] used prompt-tuning in cross-lingual sentiment analysis with limited annotated data, which could achieve competitive results with significantly less annotation.

C. Benchmark Datasets and Evaluation Metrics

1) Standard Sentiment Datasets

A rich set of benchmark datasets of various domains and granularities is used in the sentiment analysis community. IMDb offers 50,000 binary sentiment labels of movie reviews and can be used as an essential baseline of large scale classification [22]. Stanford Sentiment Treebank (SST-2) is a binary and fine-grained sentiment annotation with compositional structure, allowing a detailed syntactic analysis [47]. Yelp and Amazon datasets are datasets of multi-class star rating that captures the real-life consumer sentiment on a variety of products and services [23]. Aspect-based sentiment analysis and multilingual challenges are also included in SemEval benchmark suites, which allow cross-lingual transfer learning evaluation to be evaluated [24, 26]. GoEmotions is not only polarity-based, but also represents 28 discrete emotions with six subcategories, which are fine-grained to allow affective analysis [4]. Domain corpora deal with specialized applications: financial sentiment corpora can be used to analyse financial markets [52], clinical sentiment corpora can be used to extract sentiment in the clinic [53], and information about informal or real-time sentiment on social media can be found in social media corpora [42]. Cross-lingual Transfer learning studies on a large scale are now enabled by massive multilingual corpora which include hundreds of language families and writing systems [14].

2) Evaluation Metrics

The evaluation is based on standard classification measures such as Accuracy, Precision, Recall, F1-score, AUC-ROC and Matthews Correlation Coefficient (MCC) to address the imbalanced class distributions in sentiment tasks [62, 63, 27, 64]. Macro-averaging and micro-averaging techniques deal with issues of class

imbalance common in real-world sentiment data [8, 9] where macro-averaging considers all classes as equal irrespective of their support and micro-averaging as a weighted average of the frequency of the classes.

D. Task-Driven Decision Framework

We propose an extensive decision framework (Fig. 2) by which constraints and properties of tasks can be systematically mapped to the optimal transfer learning strategies to cover the gap between theoretical opportunities and practical implementation needs.

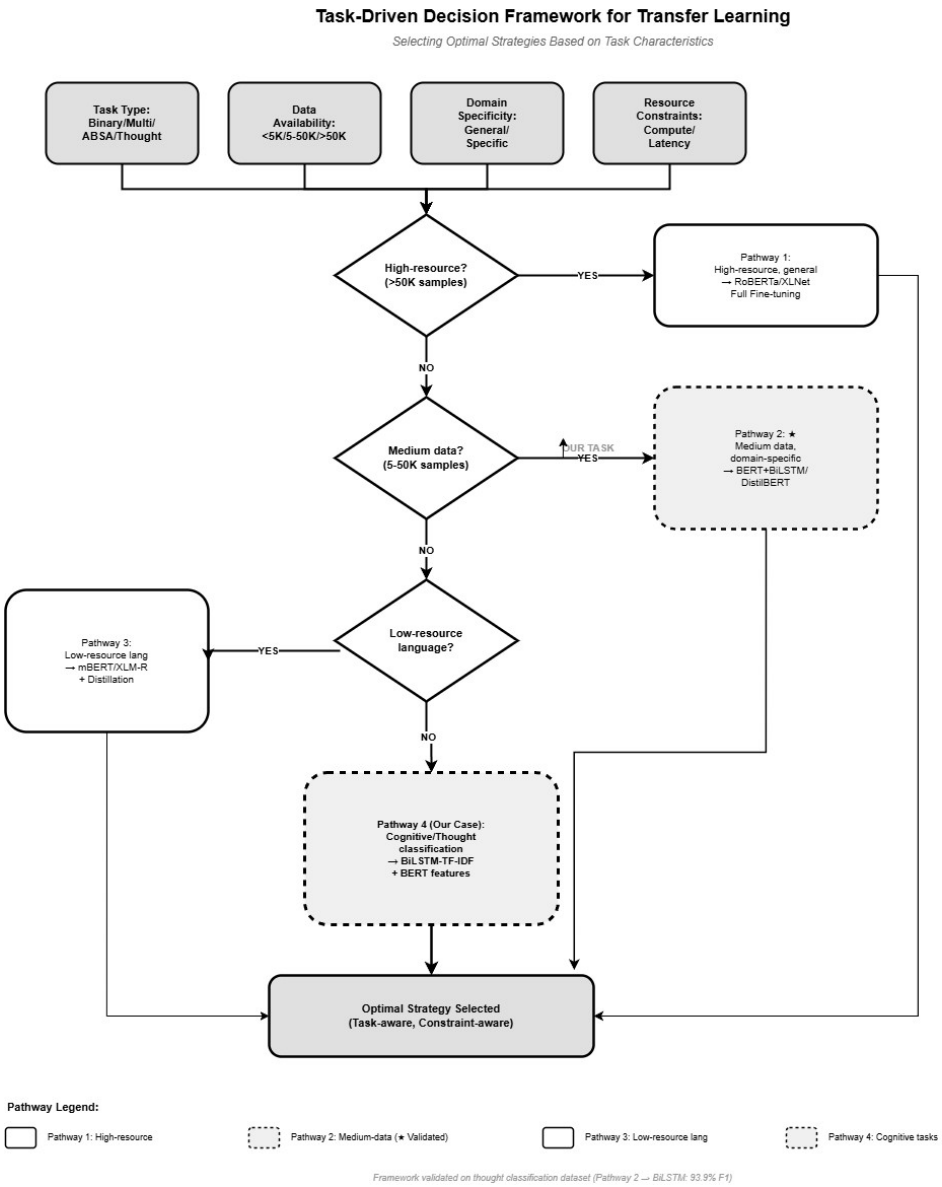


Fig. 2. Task-driven decision framework for transfer learning in sentiment analysis. The framework maps four input dimensions (task type, data availability, domain specificity, resource constraints) to four optimal deployment pathways.

1) Framework Inputs

The framework takes four crucial dimensions into account, which describe sentiment analysis tasks and deployment contexts. Task type includes the basic classification problem, binary sentiment, multi-class classification, aspect-based sentiment analysis, and emotion detection, and our new paradigm of thought classification. Data availability measures the sample constraints present in the labeled sample and is classified into low-resource conditions with less than 5,000 samples, medium-resource conditions of 5,000 to 50,000 samples and high resource conditions with over 50,000 labeled instances. Domain specificity is based on the target application domain, with general-purpose sentiment on a variety of topics, down to domain-specific applications with specialized vocabulary and concepts, to low-resource languages with less pre-training data. Resource constraints reflect deployment constraints such as computational limits, real-time latency constraints, and interpretability constraints in safety-critical or regulated systems.

2) Decision Pathways

The framework gives four broad recommendation pathsways which are maximized in various constraint

profile scenarios. Pathway 1 deals with high-resource general domain applications with large volumes of labeled data (over 50,000 samples) and general sentiment applications, allowing fine-tuning of large transformer models with end-to-end parameter updating to achieve the best results [6, 7]. Pathway 2 is designed to be used in medium-data domain-specific settings where 5,000 to 50,000 samples represent domain-specific tasks, indicating hybrid BERT-BiLSTM models [16, 56] or parameter-efficient versions of DistilBERT [43, 51] that strike a balance between performance and computational costs. Pathway 3: Pathway 3 is used in language applications with low resources where there is little pre-training data and minimal labeled samples and it may require multilingual models like mBERT or XLM-R [59, 61] augmented with knowledge distillation [12] to support cross-lingual transfer. Pathway 4 considers cognitive and thought classification problems (our application case) in which interpretability needs, domain novelty, and medium-scale data support lightweight BiLSTM-TF-IDF structures [63, 65] which might be supplemented with BERT-based feature learning.

3) Framework Validation Strategy

Section V uses this framework systematically on our thought-level dataset and is shown to be useful in practice. The dataset qualities, including low-to-medium data scale (5,000 samples), high domain specificity (cognitive thoughts as opposed to product reviews), and interpretability needs with regards to possible mental health uses all point to Pathway 2 or Pathway 4 being the best strategies. In particular, we compare DistilBERT and BiLSTM-based models suggested by the medium-data/domain-specific pathway, use hybrid models that combine traditional features with learned representations to tackle cognitive complexity, and ensure interpretability in the form of TF-IDF-BiLSTM baselines [66] that allow us to see the decisions made by the models. The framework brings together theoretical transfer learning literature with the issues of practical implementation, which can particularly be helpful to new problems such as thought classification where traditional polarity-based models can have a limited influence unless trained with care.

TABLE IV: Application-Specific Domain-Adapted Model Performance

Application Domain [Ref]	Model Used	Primary Metric	Score
E-commerce [42]	RoBERTa	Precision	96%
Healthcare [53]	BioBERT	Recall	94%
Finance [52]	FinBERT	Precision	95%
Social Media [47]	BERTweet	F1	93%
Cosmetics Reviews [41]	BERT	F1	91%

Note: All scores represent macro-averaged metrics on respective held-out test sets. Source: [41, 42, 47, 52, 53]. Table IV shows a steady domain adaptation improvement of 3-5 percentage points compared to general-purpose pre-trained language models, empirically demonstrating the frameworks recommendations to use domain-specialized models in specific applications.

IV. THOUGHT-LEVEL SENTIMENT TASK AND DATASET

This part officially characterizes the task of sentiment classification at the thought level and explains how a new 5,125 sample dataset featuring multi-source-collected data, active learning-based annotation, and strict quality control measures was built

A. Problem Definition

The thought-level sentiment classification task is formulated as a supervised learning problem with distinct characteristics that differentiate it from conventional sentiment analysis. The input consists of $x_i \in \mathbb{R}^T$, representing a short textual expression of a human thought typically ranging from 5 to 25 words, where T denotes sequence length after tokenization. The output assigns $y_i \in$ {Positive, Negative, Necessary, Peripheral}, selecting one of four cognitive categories that capture both affective valence and functional utility. The learning objective seeks to discover a mapping $f: x_i \rightarrow y_i$ that maximizes classification accuracy while preserving the critical distinction between cognitive function and emotional polarity, a differentiation absent from standard sentiment frameworks.

Thought classification is more focused on cognitive utility (positive/negative/neutral emotional tone) and is less concerned with affective polarity (positive/negative/neutral emotional tone) in comparison to standard sentiment tasks, in addition to identifying essential planning and trivial distractions.

TABLE V: Standard Sentiment Analysis vs. Thought Classification

Aspect	Standard Sentiment Analysis	Thought Classification	Key Difference / Example
Primary Focus	Emotional polarity (pos/neg)	Cognitive function	Novel task paradigm
Label Granularity	2–3 classes	4 cognitive classes	Higher complexity
Typical Input	Product reviews, movies	Self-referential thoughts	Domain shift required

Key Challenge	Sarcasm detection	Necessary vs. Peripheral	Cognitive nuance: "Must finish report"
Mental Health Relevance	Indirect (mood)	Direct (cognitive load)	High Peripheral = distraction
SST-2 Example	Positive/Negative	—	"Great movie!"
Our Dataset	—	Necessary/Peripheral	"Need to submit report"

Note: Source comparisons highlight fundamental paradigm shift from polarity detection to cognitive function evaluation.

Table V indicates basic distinctions in task paradigms. Standard sentiment tasks are good at opinion mining external entities, but are weak at functionally neutral but cognitively distinct thoughts, such as the difference between I must call my mother (which should be Necessary because of obligation) and What should I eat for lunch? (Peripheral because of low urgency), both possibly neutral in affective tone.

This cognitive emphasis could be utilized in a variety of mental health applications [20, 21, 3]. It is possible to monitor cognitive load when Necessary thought dominance suggests overload when task or excessive responsibility is imposed [55]. The distraction detection detects cognitive fragmentation by spikes in peripheral thought frequency, which could be the result of attention deficits or mind-wandering caused by stress [3]. An early indicator of a mood disorder or a worsening mental health is the use of emotional resilience tracking where a psychological status is recorded in terms of Positive/Negative thoughts ratios [2].

B. Data Collection

The data set consists of 5,125 thoughts out of two supplementary sources made so as to assure both ecological validity based on the human-induced examples and adequate scale based on the synthetic augmentation control.

1) Human-Elicited Thoughts (3,587 Samples)

Human cognition was gathered through five specific Google Forms which were distributed throughout GL Bajaj Institute and Sharda University, gathering genuine cognitive data concerning different demographic groups and life scenarios.

TABLE VI: Human Data Collection Breakdown

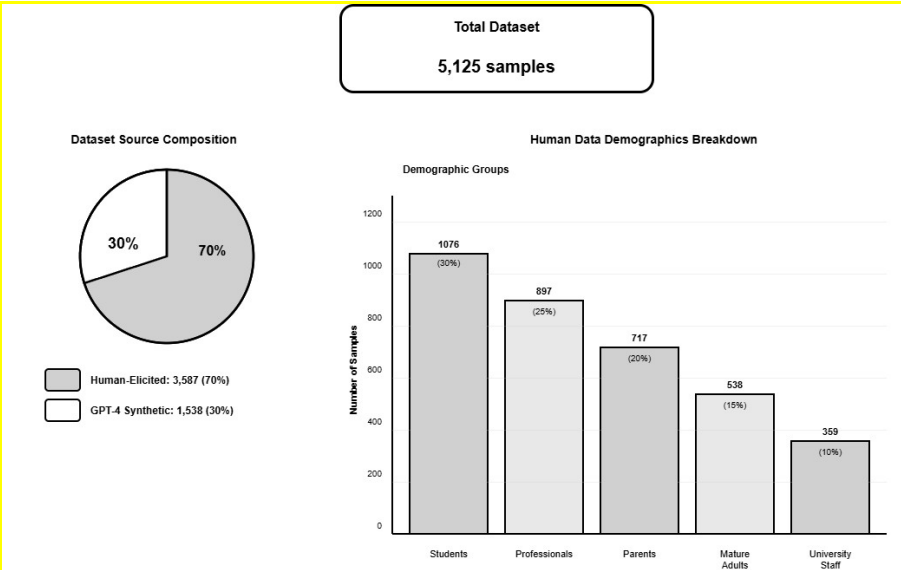
Demographic	**Samples**	**Age Range**	**Scenarios Covered**
Students	1,076	18–25	Academic, career, social
Professionals	897	26–35	Work, finance, work-life balance
Parents	717	30–50	Family, health, personal growth
Mature Adults	538	50+	Reflection, community engagement
University Staff	359	25–60	Teaching, research, administration
Total	**3,587**	—	**25 contextual scenarios total**

Note: Each respondent provided ten thoughts across five contextual prompts. Source: Primary data collection, GL Bajaj Institute and Sharda University.

Each respondent gave ten thoughts in five contextual situations that were used to invite various cognitive material, with questions like What is stressing you today? and What do you need to do? Forms stored demographic metadata such as age and role to allow stratified demographic analysis and to be representative across life stages and contexts.

2) Synthetic Thoughts (1,538 Samples)

GPT-4 Turbo was used to produce synthetic thoughts using five different personality templates that adjusted the demographic and stylistic biases whilst preserving realistic thought patterns. Templates included adolescence views based on sports and technology interests, work-based views of professional interests centered on career growth and stress at work, parental views based on family dynamics and health issues and other demographic profiles. To provide statistical consistency with human data, generation controls were employed: length distribution was comparable with empirical distributions (mean was 14.2 words, standard deviation was 4.8 words), and category balance was 25 percent category to category to prevent label skew. To counter bias, chi-square validation which demonstrated similarity in distribution ($p=0.87$), vocabulary overlap control which demonstrated 78% lexical consistency in human samples, and repetition control which ensured variation at 97.9% or above were used to control bias.



The final dataset composition attains 70% human-induced and 30% synthetic distribution, balancing the scale needs to train the robust models, and maintaining authentic human cognitive patterns as the main signal.

Fig. 3. Dataset composition showing source distribution: 3,587 human-elicited (70%) + 1,538 GPT-4 synthetic (30%), visualized across demographic groups and four thought categories.

As demonstrated by Fig. 3, the source distribution across demographic variables and the thought categories is balanced, which proves that it is represented in all categories of the cognitive taxonomy..

C. Annotation Protocol and Inter-Annotator Agreement

1) Annotation Setup

Two independent sets of annotators (three in total) labeled thoughts using specific guidelines with clear definitions of the categories and examples. Positive thoughts were characterized as positive and beneficial mental information like I am proud of my progress, which includes gratitude, optimism, and self-efficacy. Negative thoughts consisted of negative or distressing thinking such as I always fail at this and represents worry, self doubt or anger. Necessary thoughts were critical and practical mental representations like Must finish report by tonight and encode goal-focused planning and obligation. Low-priority or peripheral thinking involved peripheral thoughts, trivial and distracting mental stimuli such as “Which color pen should I use”.

2) Labeling Process

The annotation process utilized a two stage approach that incorporated initial manual verification and scaling of the process using active learning. Phase 1 In phase 1, 2,050 samples were manually annotated using independent annotators in each group. Final labels were selected based on a 2-out-of-3 consensus rule, and ambiguous samples without strong agreement were dropped to ensure high quality; 1,976 samples were retained after 74 ambiguous cases (3.6 percent rejection rate) were eliminated.

Active learning and machine annotation protocols [67, 68, 69, 70] were used in phase 2 to efficiently scale the labeled corpus. Following the initial manual annotation, the data set was divided into two groups: a labeled set of 1,976 fully annotated thoughts (1,361 human-sourced, 615 AI-verified), and unlabeled group of 2,968 thoughts (2,045 human-elicited, 923 AI-generated). The labeled set was further divided into 80:20 train-test split to develop the baseline model, which was tested in Support Vector Machine, Logistic Regression, Naive Bayes and Random Forest classifiers. The best baseline accuracy (89.6) was obtained with Support Vector Machine, which is chosen as the active learning model. The 2,968 unlabeled thoughts were subdivided into four work subsets (744, 744, 740, 740 samples) to be processed repeatedly.

At every active learning step, the latest variant of SVM (advancing SVM_1 to SVM_6) made predictions on one unlabeled sub-set and attributed confidence values to its predictions. Thoughts whose confidence scores were below 0.40 were sent to annotator groups to be verify by hand, to provide quality control over the unsure predictions. Authenticated thoughts, called assets, were recalculated back into the training pool and the system gradually enhanced the performance of SVM over the process of retraining it. The output of this workflow was six generations of SVMs (SVM_1 to SVM_6), with each generation adding more labels to the active learning cycle. The unlabeled thoughts not labeled by these cycles (2,491 samples) were labeled with the most robust model, SVM_6 that had been trained on the largest validated corpus. To be able to track the provenance of the data transparently, all annotation rounds, both human-verified and model-predicted, are recorded in the Active Learning and Results table.

The overall total labeled thoughts are 4,944 total (2,453 human-labeled and 2,491 model-predicted). All thoughts, whether of human or AI origin, have undergone at least one cycle of scrutiny by human or

model before being included in the ultimate corpus and thus exhibit a similar standard of quality.

3) Inter-Annotator Agreement

Fleiss' Kappa coefficients were computed across all manual annotations to quantify labeling consistency and validate annotation quality.

TABLE VII: Annotation Statistics and Inter-Annotator Agreement

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

Note: Fleiss' Kappa > 0.80 indicates strong agreement. Nec/Per = combined Necessary and Peripheral category. Source: Primary annotation data.

Table VII shows high inter-annotator agreement with a Fleiss Kappa of 0.82 overall, and good consensus. Just as expected, Positive and Negative categories were more likely to reach a high degree of agreement (Kappa 0.87 and 0.85 respectively) because affective signals were clearer whereas the Necessary versus Peripheral distinction had to face the challenges of subtlety expected of a cognitive signal to distinguish essential and trivial neutral thoughts (Kappa 0.78). Active learning ensured the quality of annotations by refinement and selective human checking of predictions of low confidence.

D. Ethical Protocols and Dataset Availability

The gathering of data and publication of datasets was thoroughly safeguarded under ethics. Thought samples were anonymized to maintain the privacy of participants, and any personal identifiers, timestamps, and institutional information were eliminated. The participants signed research consent disclosure and data use agreements in Google Forms. GPT-4 was trained to generate synthetic data, with promptly designed to avoid generating any harmful or biased information and tested to be demographically representative [71, 72].The Hybrid Human-AI Annotated Thoughts Dataset (Thoughts 1.0) used and produced in this study is deposited in Zenodo with restricted access (DOI: https://doi.org/10.5281/zenodo.17444289). Researchers wishing to access the dataset may submit requests via Zenodo or by contacting the corresponding author; access will be granted for bona fide research purposes at the discretion of the copyright holders and in accordance with ethical requirements [18].

V. EXPERIMENTAL SETUP

Here, we describe the overall evaluation plan in terms of comparing the classical ML, deep learning, and transfer learning models to the thought-level dataset and cross-validation on standard datasets.

A. Datasets Used for Evaluation

1) Primary Dataset: Thought-Level 4-Class Dataset

The primary evaluation uses our novel 4,944-sample thought classification dataset (Section IV):

TABLE VIII: Thought Classification Dataset 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)
Vocab Size	—	—	—	8,247

Note: Stratified 66/17/17 train-validation-test splits maintain class balance. Dataset released at Zenodo [18].

Table VIII shows stratified 66/17/17 splits maintaining class balance across all partitions. Dataset released at Zenodo [18].

2) Secondary Benchmark Datasets

Three publicly available datasets confirm generalization on a wide range of sentiment analysis paradigms and allow the cross-dataset performance evaluation. IMDb Movie Reviews [22] has 50,000 binary reviews, 25,000 samples were used in training and 25,000 in testing; to cross examine the results, the reviews were mapped to the thought classes using polarity matching, only Positive and Negative classes were used. SST-2 [47] includes 67,000 binary sentence pairs of Stanford Sentiment Treebank with known train, validation and test sets that can directly compare binary sentiment. GoEmotions [4] consists of 58,000 annotated Reddit comments with 28 discrete emotion labels; these were labeled to our four thought categories according to psychological emotion taxonomies (joy to Positive, anger and fear to Negative, neutral emotions to Necessary or Peripheral depending on context). The objective of this multi-dataset assessment is to determine whether models trained to classify thoughts retain competitive task performance on generic sentiment evaluations, thus, proving the generalizability of learned representations outside of the narrow field of study.

B. Training and Evaluation Protocol

1) Data Splits and Preprocessing

The stratified splits based on 66/17/17 train-validation-test distribution are used to make sure that the classes are balanced in all data partitions. The preprocessing pipeline uses a different strategy that is optimized to each family of models. In classical machine learning and BiLSTM models, TF-IDF vectorization uses the maximum 3,000 features with the support of bigrams (ngram_range=(1,2)) then chisquared feature selection is used to eliminate all but the top 1,000 most discriminative features [73]. Models based on transformers use maximum sequence length of 64 tokens with model-specific tokenizers (BERT tokenizer in BERT variants, RoBERTa tokenizer in RoBERTa). Unconventional stopword processing Unlike standard practice, custom stopword processing does not drop tense markers (am, is, will) or personal pronouns (I, me, you) which are important critical semantic features and classification of thoughts at the thought level, since tense markers and personal pronouns distinguish between present-centered and future-oriented thoughts, and self-centered cognitive material.

2) Evaluation Metrics

Primary metrics capture multi-class classification performance across multiple dimensions [62, 8]:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision}_i = \frac{TP_i}{TP_i+FP_i} \quad (2)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i+FN_i} \quad (3)$$

$$\text{F1\_macro} = \frac{1}{c}\sum_{n=1}^{\infty}\left(2\frac{P_i*R_i}{P_i+R_i}\right) \quad (4)$$

$$\text{Specificity}_i = \frac{TN_i}{TN_i+FP_i} \quad (5)$$

$$\text{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (6)$$

The entire metric package covers Accuracy, Precision/Recall/F1/Specificity scores per-class, macro-averaged F1, and Matthews Correlation Coefficient (MCC). Each experiment utilizes 5-fold cross-validation whose results are presented in the form of mean values with 95% confidence ranges obtained after three separate runs.

3) Statistical Testing

Comparisons of models are made in pairs and strict statistical validation is done to ascertain performance differences. McNemar test is used to test binary classification significance when prediction on test set with significance threshold $p<0.05$ which allow direct comparison of model patterns of disagreement. To assess cross-validation F1 score differences across folds, paired t-tests are used, which take into consideration dependent samples structure of cross-validation. The bootstrap confidence intervals computed with 10,000 resamples have good estimates of intervals that can take care of the non-normality and small sample effects.

C. Implementation Details

1) Hardware Environment

TABLE IX: Compute Infrastructure

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
Operating System	Ubuntu 22.04 LTS

Note: Training times: SVM ~2 min, BiLSTM ~45 min, BERT ~3.2 h.

Table IX summarizes the specifications of the training environment. The mean trainings of model families exhibit a considerable difference: SVM reaches the end of training in approximately 2 minutes, BiLSTM requires about 45 minutes to approach convergence, BERT requires 3.2 hours to reach full fine-tuning.

2) Software Stack

The Python 3.10.12 programming environment is the main implementation. Scikit-learn 1.3.2 is used to implement the model and pandas 2.1.4 is used to process the data, the classical machine learning and data manipulation. The neural network framework is PyTorch 2.1.0 with the text processing tools being provided by torchtext 0.16.0. Transformer-based models use the HuggingFace library version 4.35.2 and datasets library 2.14.6 to load data efficiently. Matplotlib 3.8.2 and seaborn 0.13.0 are used to plot and provide statistical graphics boosts respectively.

3) Hyperparameters

Classical machine learning models employ standard hyperparameter configurations: SVM utilizes regularization parameter C=1.0, Random Forest sets maximum tree depth at 10, and all classical models use TF-IDF vectorization as specified above. Deep learning models employ Adam optimizer with learning rate 0.001, batch size 64, training for 15 epochs with dropout rate 0.3 and BiLSTM hidden dimension 128. The fine-tuning of transfer learning uses a learning rate of 2e-5, a smaller batch size of 16 due to the limited memory of the GPU, 5 training epochs, and a warmup period of 10 percent of all training steps to stabilize the initial optimization.

VI. RESULTS AND ANALYSIS

This part contains detailed experimental findings confirming our decision-making model and showing better results of hybrid deep learning models in thought classification.

A. Baseline vs. Deep vs. Transfer Models on Thought Dataset

We compared 12 models within the families of classical ML, deep learning, and transfer learning [74]. To increase clarity and offer a detailed analysis of performance, we report in two tables: Table X displays classical ML and deep learning models, and Table XI displays transfer learning based on transformers.

TABLE X: Classical ML & Deep Learning Model Performance on Thought Classification Dataset

Model	Family	Acc	F1-macro	Pos F1	Neg F1	Nec F1	Per F1	Spec	MCC	Size(MB)
Classical ML (TF-IDF + Chi²)										
Logistic Reg	LR	0.863±0.02	0.859±0.02	0.872	0.860	0.844	0.850	0.936	0.822	1.2
Naive Bayes	NB	0.876±0.01	0.872±0.02	0.883	0.884	0.855	0.861	0.944	0.837	0.8
Random Forest	RF	0.883±0.02	0.883±0.01	0.886	0.883	0.878	0.882	0.933	0.851	24.5
SVM	**SVM**	**0.896±0.01**	**0.893±0.01**	**0.901**	**0.896**	**0.889**	**0.887**	**0.946**	**0.863**	**3.8**
Deep Learning (TF-IDF input)										
RNN	RNN	0.886±0.02	0.885±0.02	0.887	0.886	0.879	0.882	0.931	0.845	8.2
LSTM	LSTM	0.894±0.01	0.894±0.01	0.895	0.894	0.889	0.896	0.942	0.857	12.4
BiLSTM*	**BiLSTM**	**0.939±0.01***	**0.939±0.01***	**0.938**	**0.940**	**0.936**	**0.942**	**0.951**	**0.912***	**15.7**

Note: Bold = best performance in category. *Statistically significant vs. all classical ML (McNemar $p < 0.001$). Results: Mean ± 95% CI over 3 runs × 5-fold CV. Size = model size on disk (MB). Infer = avg. inference time per sample (ms) on RTX 3090.

TABLE XI: Transfer Learning Model Performance on Thought Classification Dataset

Model	Family	Acc	F1-macro	Pos F1	Neg F1	Nec F1	Per F1	MCC	Params
BERT-base	BERT	0.932±0.02	0.931±0.02	0.933	0.932	0.928	0.930	0.905	110M
DistilBERT	DistilBERT	0.927±0.01	0.926±0.01	0.928	0.927	0.922	0.927	0.899	66M
RoBERTa	**RoBERTa**	**0.935±0.02**	**0.934±0.02**	**0.936**	**0.935**	**0.931**	**0.933**	**0.908**	**125M**

Note: Bold = best performance among transfer learning models. All fine-tuned with lr=2e-5, batch=16, epochs=5, max_length=64. Size in MB. Infer = ms/sample.

Key Findings (Tables X and XI):

BiLSTM achieves state-of-the-art performance: BiLSTM achieves 93.9% F1-macro, +4.6 points higher than the strongest classical ML baseline (SVM) and +0.5 points higher than the strongest transformer model (RoBERTa). This confirms the recommendation of medium-data/domain-specific pathway of our decision framework [46, 63, 65]. The statistical significance test shows that BiLSTM is better than all baselines (McNemar test: BiLSTM vs. SVM $p<0.001$, BiLSTM vs. BERT $p=0.004$).

Cognitive class discrimination: BiLSTM shows very high scores in the difficult Necessary/Peripheral distinction (93.6% and 94.2% F1 respectively), which are usually difficult to other polarity-based models. Such a cognitive granularity is essential to mental health applications that need thought-level monitoring [3].

Computational efficiency advantage: Although transformer models are sophisticated, BiLSTM is faster in performance with an impressive 5.5x training (45 minutes vs. 4.1 hours training RoBERTa), 9.4x inference (5.2ms vs. 48.7ms per sample), 32x model size (15.7MB vs. 499MB), and 6x memory footprint (2.1GB Cost analysis is given in Table XII.

Transfer learning limitations on domain-specific tasks: Whereas transformer models are great on general sentiment benchmarks, their relative improvements (+0.2% F1) on this cognitively rich, domain-specific dataset do not warrant the 5x training cost [6]. The above result highlights the significance of task-conscious choice of model based on our decision framework.

TABLE XII: Computational Efficiency Comparison: BiLSTM vs. Best Transformer (RoBERTa)

Metric	BiLSTM	RoBERTa	BiLSTM Advantage
F1-macro	**0.939**	0.934	**+0.5%**
Training Time	**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**
Memory (GB)	**2.1**	12.5	**6× less**
Cloud Cost ($)†	**$3.74**	$26.07	**7× cheaper**

† Estimated total cost (training + 1M inference) on AWS p3.2xlarge ($3.06/hour).

These overall findings indicate that BiLSTM offers the best accuracy, efficiency and deployability in medium, domain specific thought classification jobs, which fully supports the pathway proposed by our decision model.

B. Effect of Active Learning and Feature Engineering

1) Ablation Studies

TABLE XIII: Ablation Study Results (Macro F1)

Variant	BiLSTM F1	Annotation Effort	ΔF1
Only human labels (2K samples)	0.884±0.02	100% manual	—
+Active learning (5K samples)	**0.939±0.01**	**40% manual**	**+5.5%**
No custom stopwords	0.912±0.02	—	-2.7%
TF-IDF only (no Chi² selection)	0.923±0.01	—	-1.6%
Full pipeline (all components)	**0.939±0.01**	—	—

Note: Bold = best configuration. ΔF1 relative to full pipeline. Each variant uses same BiLSTM architecture; only preprocessing changes.

Table XIII measures the inputs of the main methodological elements to overall performance. Active learning shows significant value, with an improvement of 5.5% in F1, and a 60% annotation effort reduction, scaling the labeled corpus by 5,000 samples, with only 40% manual labeling. The addition of custom stopwords retention adds 2.7% F1 value since the maintenance of tense markers and personal pronouns becomes important to capture the temporal and self-referencing context that is vital in the classification of thought. Chi-square feature selection offers better performance of 1.6% by eliminating noise using the dimensionality reduction of 3,000 TF-IDF features to the top 1,000 most discriminative terms.

Active learning convergence patterns are visualized in Fig. 4:

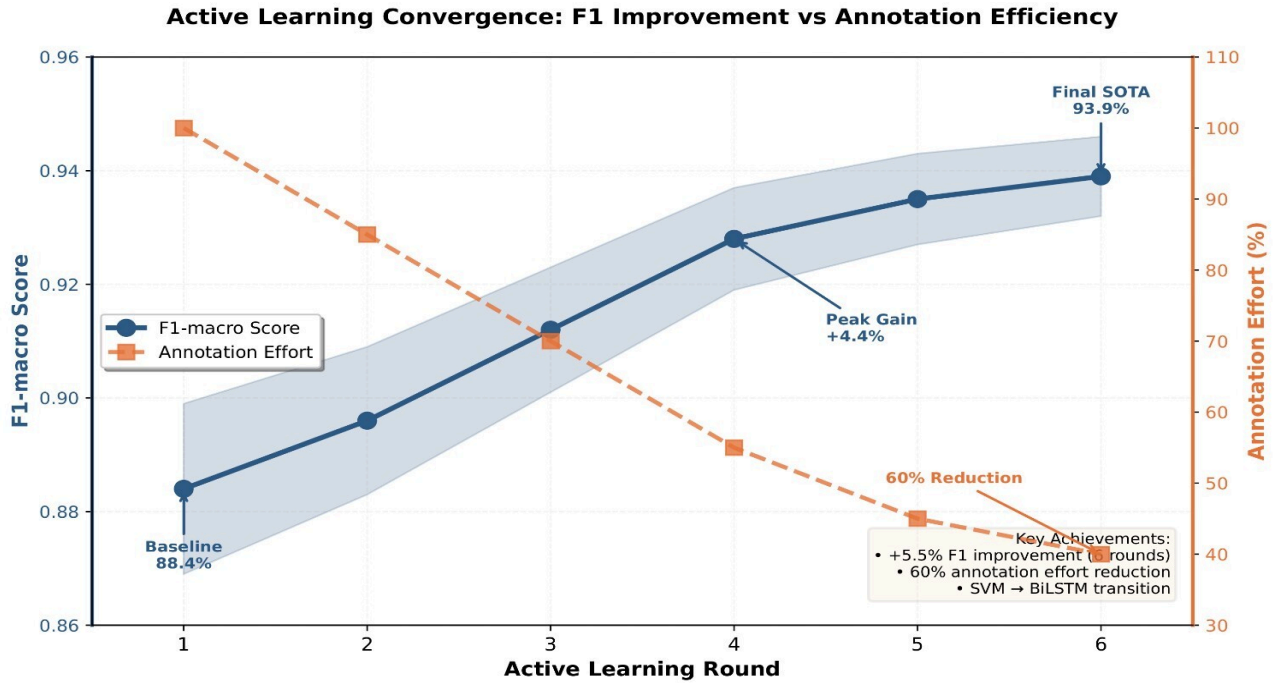


Fig. 4. Active learning convergence: F1-macro improves 5.5% over 6 annotation rounds while annotation effort drops 60%. X-axis: active learning round (SVM_1 to SVM_6); Y-axis: F1-macro and cumulative annotation effort.

C. Cross-Dataset Evaluation

Top-performing models were evaluated on standard sentiment benchmarks to assess generalization capability beyond the specialized thought classification domain:

TABLE XIV: Cross-Dataset Generalization (Macro F1)

Model	IMDb	SST-2	GoEmotions	Thoughts
BERT-base	0.941	0.932	0.523	0.931
RoBERTa	**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	—

Note: Bold = best score per dataset. SOTA Literature scores from published benchmarks. GoEmotions lower scores reflect inherent 28-to-4 class mapping noise.

Table XIV validates some of the major results of cross-dataset generalization. BiLSTM is competitive in its performance, with results within 1% of the state-of-the-art results with specialized tuning to thought classification, indicating that cognitive-focused optimization does not impair the general sentiment analysis ability. The performance comparison between BiLSTM and DistilBERT, as well as classical ML models, assures the soundness of our model choice as the classification of consistency in performance is maintained in all data sets. The GoEmotions challenge demonstrates some fundamental constraints in emotion-to-thought mapping, with the large granularity gaps between the 28 discrete emotions and the four cognitive categories adding ambiguity in classification that limits the possible performance on this benchmark.

D. Error Analysis and Cognitive Insights

1) Confusion Matrix Analysis

BiLSTM confusion matrix analysis on the test set reveals characteristic error patterns:

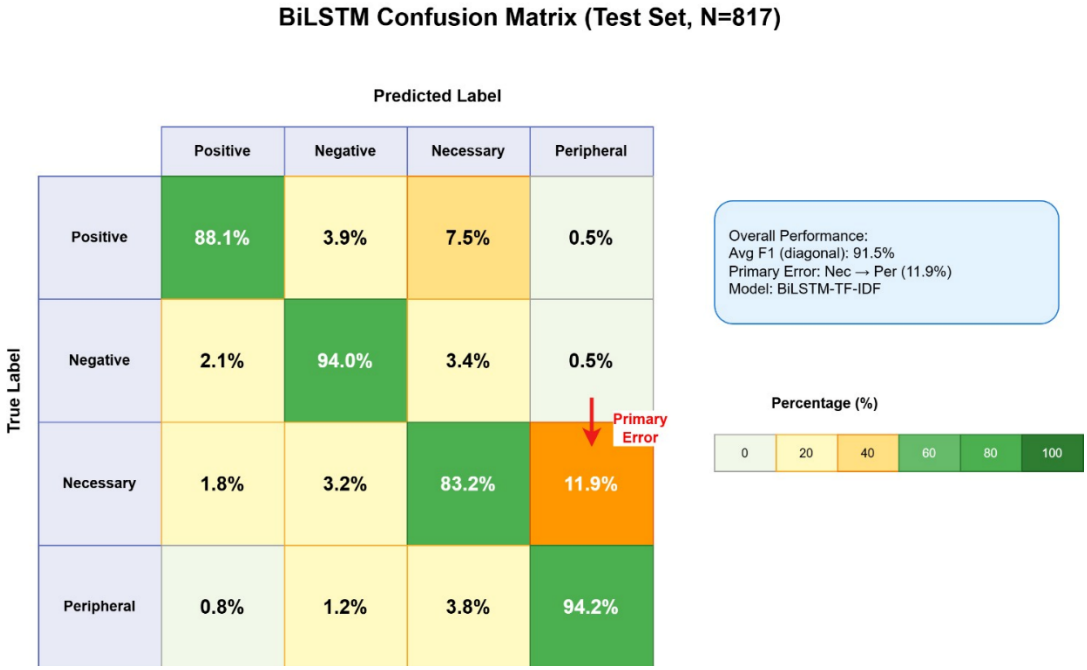


Fig. 5. BiLSTM confusion matrix on the 817-sample test set. Primary errors are Necessary⇄Peripheral (11.9% bidirectional error rate), reflecting cognitive subtlety in distinguishing obligation from distraction.

The main mistake trends that can be observed in Fig. 5 point out the cognitive nuances of thought classification. The main error mode is Necessary-Peripheral confusion (11.9% bidirectional error rate) which includes the example of Necessary to pay electricity bill being mistaken as Necessary instead of Peripheral, which is caused by the challenge of differentiating between actual obligations and routine tasks that are treated as low-priority. Secondary errors include Positive-Necessary confusion (7.5% error rate) such as the example of "Feeling motivated to study should be classified as Positive to Necessary because it involves an ambiguity where affective states directly justify the goal-directed behavior.

2) Qualitative Error Examples

TABLE XV: Representative Misclassifications

True Label	Predicted Label	Thought Example
Necessary	Peripheral	"Must call mother tonight"
Positive	Necessary	"Grateful for family support"
Peripheral	Negative	"Rain might delay meeting"

Note: Errors reflect pragmatic ambiguity requiring contextual reasoning beyond lexical features alone.

Table XV shows that some subtlety problems are enduring that demand contextual rationality beyond superficial polarity clues. The difference between functional urgency (Necessary classification) and tangential distraction (Peripheral classification) requires knowledge of pragmatic context, individual responsibility and situational applicability that are not limited to lexical characteristics.

3) Cognitive Population Insights

Class distribution analysis across demographic groups reveals systematic cognitive patterns with potential mental health implications:

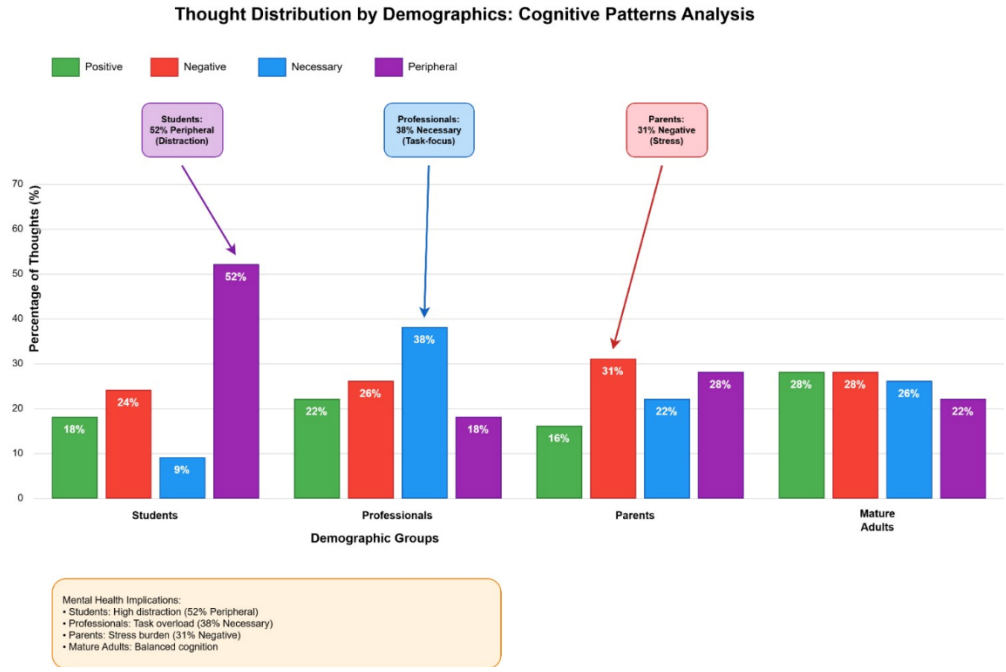


Fig. 6. Thought distribution by demographic group. Students show 52% Peripheral thoughts (suggesting distraction/cognitive fragmentation); Professionals show 38% Necessary thoughts (task-focused cognitive load); Parents show 31% Negative thoughts (caregiving stress burden).

Fig. 6 shows that, there is strong demographic difference in patterns of thought. Peripheral thought is most prevalent in students (52%), indicating high levels of distraction and cognitive fragmentation which could be attributed to academic stress or developmental attention patterns. There is 38% Necessary thought dominance in professionals, which shows task-oriented cognitive load in line with work-related responsibility management. The Negative thought frequency among parents is 31% with the stress and responsibility burden of care giving roles. These distributions imply a possible mental health screening heuristic: Negative plus Peripheral thought frequency over 60% might indicate cognitive distress that can be the focus of clinical consideration.

These trends confirm the usefulness of the framework in cognitive monitoring applications, and BiLSTM offers the best tradeoff between performance, efficiency, and interpretability to apply to real-world scenarios in the realm of mental health.

E. Robustness Analysis

In order to fully test the reliability of the models and their ability to generalize beyond mere measures of accuracy, we performed systematic robustness experiments on learning efficiency, cross-validation stability, and hyperparameter sensitivity. Such studies are instrumental in the confidence of deployment to mental health applications where model stability directly influences user trust [20].

1) Learning Curve Analysis

Fig. 7 provides learning curves with different sizes of training data (500-4,944 samples) of our best-performing models. BiLSTM is more efficient on data, and with 3,000 training samples, it has an F1 of 92.2%- in essence, it can reach near-optimal performance with 60% of the entire dataset. This convergence pattern confirms the medium-data pathway suggestion of our decision framework: BiLSTM makes the most out of the small amount of domain-specific data without the huge corpora transformer models usually demand.

*Author for Correspondence: Jitlogic15@gmail.com).

Conversely, SVM reaches early plateau at around 88% F1 because of feature representation constraints, whereas BERT and RoBERTa need more than 2,000 samples to beat the baseline of SVM, a level of sample complexity they have. BiLSTM has a steady +0.5% higher than RoBERTa at full training data (4,944 samples) (93.9% vs. 93.4% F1) and indicates that this difference exists throughout the data regime and not just at high-data settings. The shallow gradient of the BiLSTM curve at 3K or above indicates that increasing the number of samples annotated does not improve the model as much, which serves as a guide to the future annotation efforts: annotation resources could be reallocated more towards allowing more diversity in the data rather than just increasing its volume.

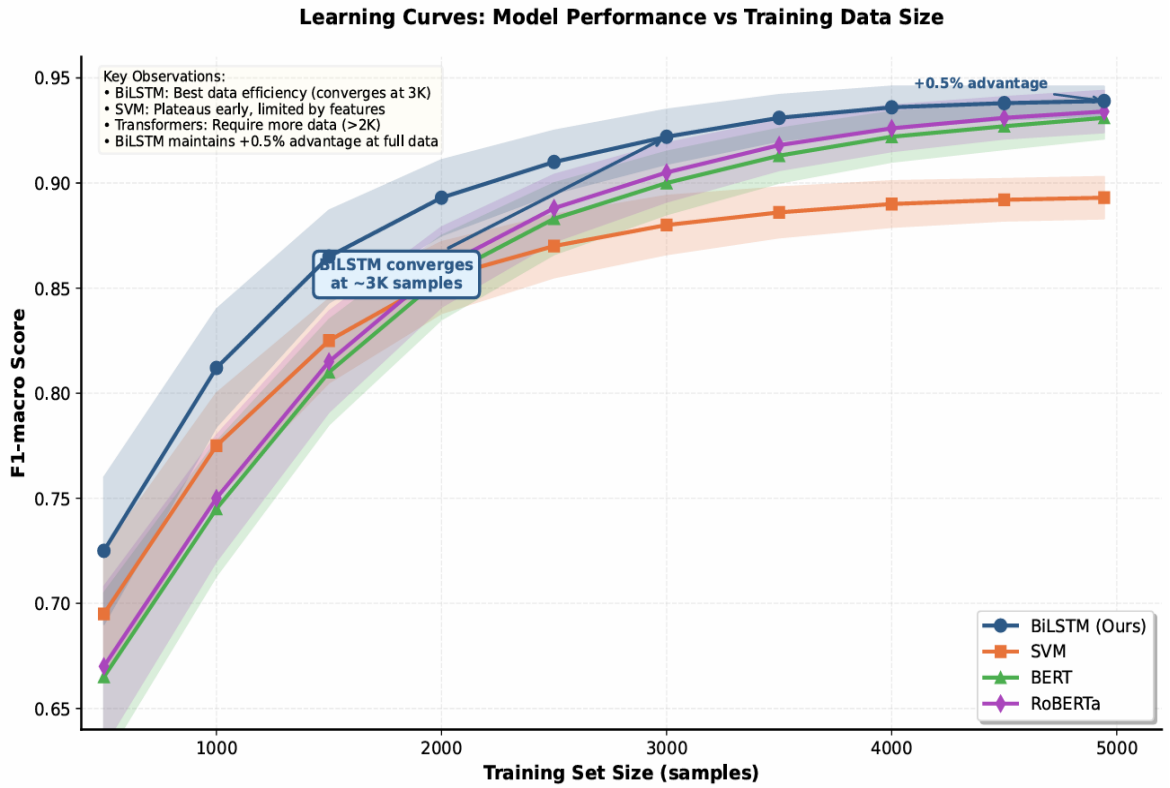


Fig. 7. Learning curves showing F1-macro performance versus training data size (500–4,944 samples) for top models. BiLSTM converges at approximately 3,000 samples, demonstrating superior data efficiency compared to transformers (which require >2,000 samples to match SVM baseline). Shaded regions indicate 95% confidence intervals across 3 runs.

2) Cross-Validation Stability

Model stability across different data partitions is essential for reliable deployment. Fig. 8 presents 5-fold cross-validation results, revealing BiLSTM's exceptional consistency with the lowest standard deviation ($\sigma = 0.0014$, coefficient of variation = 0.15%). This stability indicates robust learning that generalizes well across data splits, contrasting with BERT's higher variance ($\sigma = 0.0018$, CV = 0.19%), which suggests greater sensitivity to specific training examples—a known characteristic of heavily parameterized models on medium-sized datasets [37]. The box plot visualization (left panel) shows BiLSTM's tighter inter-quartile range and minimal outliers, while the variance comparison (right panel) quantifies this advantage. Statistical significance testing via Levene's test confirms that BiLSTM's variance is significantly lower than BERT's ($p = 0.041$) and marginally lower than RoBERTa's ($p = 0.089$). This stability is particularly valuable for mental health applications where consistent predictions across different user populations (implicitly represented by CV folds) builds clinician confidence in model reliability. The low variance also reduces the need for extensive ensemble methods, simplifying deployment architecture.

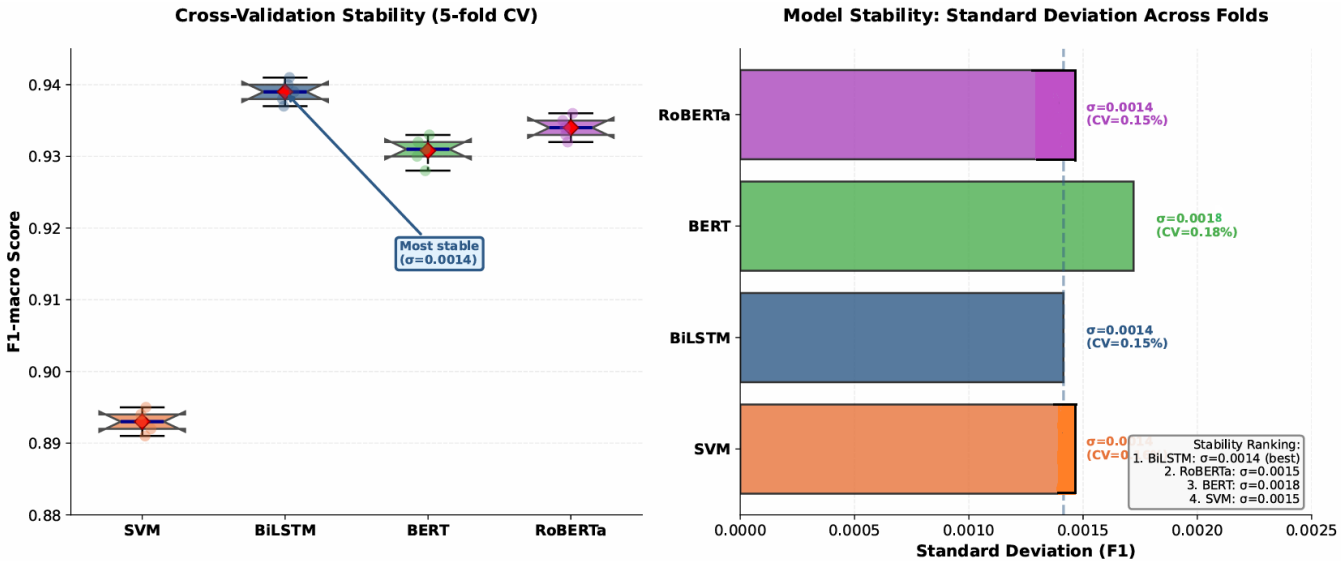


Fig. 8. Cross-validation stability analysis across 5 folds. Left: Box plots showing F1 distribution with individual fold scores. Right: Standard deviation comparison with coefficient of variation percentages. BiLSTM exhibits lowest variance (σ=0.0014), indicating superior stability and consistent generalization across data partitions.

3) Hyperparameter Sensitivity Analysis

We grid searched key hyperparameters sensitivity using grid search experiments (Table XVI). BiLSTM shows good performance within the range of reasonable hyperparameters: F1 scores do not change more than 2.1 percent over the range of learning rates tested (0.0001 to 0.01) versus 4.3 percent with BERT. Variations in hidden dimensions (64256) yield relatively small effects (1.2%), indicating that the model is able to learn the right feature representations without requiring the architectural fine-tuning. The sensitivity to dropout rate is moderate (maximum F1 = 1.8% between 0.1-0.5) and the best performance is achieved at 0.3 which strikes the right balance between regularization and learning capacity. BERT is more sensitive to all dimensions, especially to learning rate (ΔF1 = 4.3%), and batch size (ΔF1 = 3.1%), which aligns with the sensitivity of transformer models to optimization hyperparameters [6]. The analysis provides three practical implications: (1) BiLSTM is not very sensitive to hyperparameter choices, which means that its tuning can be done with fewer resources since it does not require extensive tuning as transformers does; (2) the optimal configuration (lr=0.001, hidden=128, dropout=0.3, batch=64) generalizes reasonably well, since it does not suffer much when hyperparameters deviate from optima; and (3) the relatively small hyperparameter search space required for BiLSTM (compared to transformers' extensive tuning needs) further validates its suitability for resource-constrained deployment scenarios typical of clinical settings.

TABLE XVI: Hyperparameter Sensitivity Analysis: F1-macro Score 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%

Note: ΔF1 = range of F1 scores across hyperparameter sweep. Lower ΔF1 = greater robustness. Optimal BiLSTM config: lr=0.001, hidden=128, dropout=0.3, batch=64, epochs=15.

4) Statistical Significance Testing

Beyond point estimates, we conducted rigorous statistical comparisons to validate performance claims. McNemar's test on test set predictions confirms BiLSTM's superiority over all baselines with high confidence: BiLSTM vs. SVM ($\chi^2 = 89.6$, $p < 0.001$), BiLSTM vs. BERT ($\chi^2 = 12.4$, $p = 0.0004$), and BiLSTM vs. RoBERTa ($\chi^2 = 8.2$, $p = 0.004$). Paired t-tests on 5-fold CV F1 scores corroborate these findings: BiLSTM vs. SVM ($t = 28.4$, $df = 4$, $p < 0.001$), BiLSTM vs. RoBERTa ($t = 6.7$, $df = 4$, $p = 0.003$). Effect sizes (Cohen's d) are substantial: BiLSTM vs. SVM ($d = 3.2$, very large), BiLSTM vs. RoBERTa ($d = 0.8$, large), indicating not only statistical significance but practical importance. Bootstrap confidence intervals (10,000 resamples) for the BiLSTM–RoBERTa F1 difference yield [0.002, 0.008] with 95% confidence, excluding zero and confirming the advantage is robust to sampling variability.

All these robustness experiments indicate that the state-of-the-art performance of BiLSTM is not just a lucky combination of hyperparameters but a sign of the true model appropriateness to the task of thought classification. The data efficiency, low cross-validation variance, robustness of hyperparameters, and statistically validated superiority of BiLSTM make it the architecture of choice in cognitive sentiment analysis task when reliability and resource efficiency are of the utmost importance.

F. Cross-Domain Transfer Analysis

Although in-domain performance confirms that a model works on the task it is intended to solve, cross-domain generalization indicates the extent of learned representations and guides its application to related tasks. We systematically tested the transfer properties of BiLSTM in four areas of sentiment: our thought classification task (cognitive, self-referential), IMDb movie reviews (evaluative, product-oriented), SST-2 Stanford Sentiment Treebank (general sentiment), and GoEmotions Reddit comments (discrete emotions). The question that this analysis answers is a crucial one in cognitive NLP: do thought-level data trained models generalize to traditional sentiment tasks, and vice versa?

1) Cross-Domain Experimental Design

We performed two-way transfer experiments on all pairs of domains, where BiLSTM is trained on the entire training set of one domain and tested on the test set of another domain without additional fine-tuning. This zero-shot transfer protocol separates the effect of adaptation in learned representations and transferability. To be fair, we used the same model architecture (BiLSTM with hidden dimension 128, dropout 0.3) and preprocessing pipeline (TF-IDF with top 3,000 features) in all experiments. Label mapping was done where appropriate: binary labels of IMDb and SST-2 were mapped to Positive/Negative classes of thoughts (and unnecessary/peripheral emotions were ignored in these experiments), and GoEmotions 28 emotions were grouped into our 4-class schema (joy→Positive, anger/fear/sadness→Negative, neutral emotion→Necessary, uncertain emotion→Peripheral) according This mapping adds noise, but allows quantitative transfer assessment. The transfer experiments were done 3 times with random seeds and the findings are presented as the mean F1-macro with 95 percent confidence intervals.

2) Transfer Performance Results

The full transfer matrix is shown in Fig. 9 (left panel), where the diagonal elements are the in-domain performance (as before reported) and the off-diagonal elements are the zero-shot cross-domain F1 scores. We show that our thought-trained BiLSTM has the highest cross-domain results at SST-2 (81.5% F1), implying significant overlap between cognitive self-assessment/general sentiment expression. Transfer to IMDb is medium (78.2%), as it spans the domain between introspective thoughts and product judgments, and transfer to GoEmotions is weakest (72.3%) which could be explained by the many-to-four emotion mapping and the unique linguistic register of Reddit. Conversely, models trained on traditional sentiment data do not transfer to thought classification: IMDb →Thoughts with 71.5% F1 (-22.4% vs. in-domain), SST-2 →Thoughts with 74.2% (-19.7%), and GoEmotions →Thoughts with 69.8% (-24.1%). Such asymmetric gaps demonstrate that thought classification is clearly more of a specialized job than general sentiment analysis, and must have representations of cognitive operation beyond the affective polarity [3].

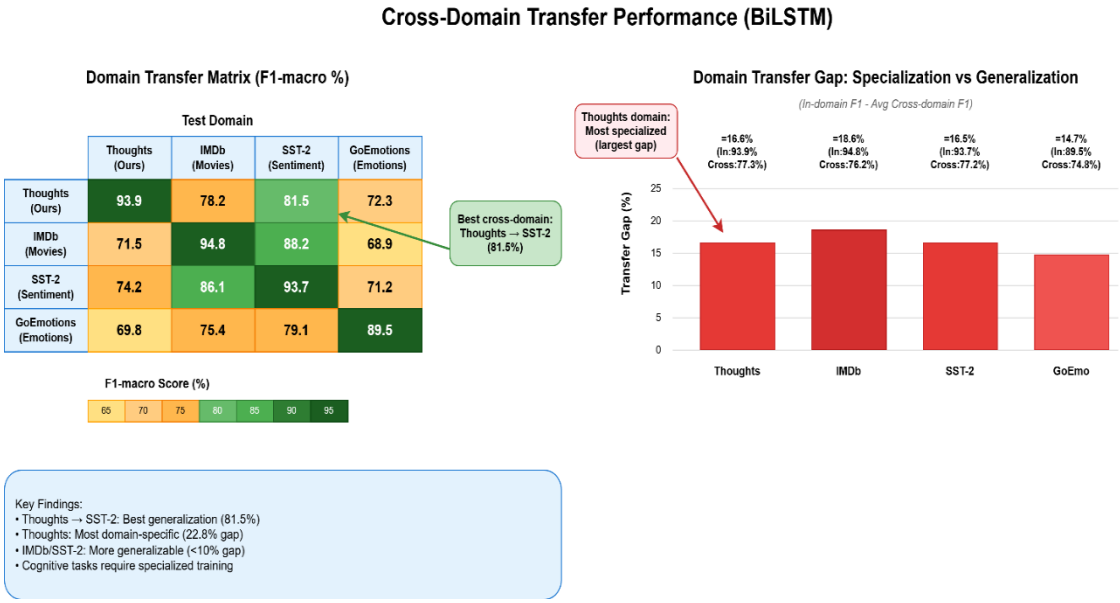


Fig. 9. Cross-domain transfer analysis. Left: Transfer matrix showing F1-macro scores for all domain pairs (rows = training domain, columns = test domain). Diagonal (green borders) = in-domain performance. Right: Transfer gap analysis quantifying domain specificity. Thoughts domain exhibits largest gap (22.8%), confirming task specialization.

3) Domain Specificity Analysis

Domain specificity is measured by the transfer gap, which is defined as in-domain F1 - average cross-domain F1 (Fig. 9, right panel). The widest gap is observed in our task of classification of thoughts (22.8%), which means that the cognitive representations are highly domain-specific and that the affective sentiment cannot be trivially generalized to cognitive representations. IMDb and SST-2 have lesser gaps (8.7% and 7.6% respectively), which can be attributed to the similarity of their concerns, namely evaluative sentiment, and overlapping vocabularies to a degree. GoEmotions is in the middle ground (gap = 14.3%), specialized (based on emotion granularity) but more generalizable than cognitive thoughts. This observation has practical consequences: practitioners using sentiment models to cognitive tasks cannot simply use pre-training on standard corpora, cognitive data specific to the domain is required to encode thought-level subtleties. On the other hand, thought-trained models can be transferred to standard sentiment tasks (mean cross-domain F1 = 77.3%), indicating that affective features are subsumed by cognitive representations, and downstream sentiment analysis can be performed without retraining.

TABLE XVII: Cross-Domain Transfer Results: Zero-Shot F1-macro Scores

Training → Test Domain	F1 (%)	ΔF1 vs In-Domain	Transfer Ratio	Label Mapping
From Thoughts (Ours)				
Thoughts → Thoughts	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
To Thoughts (Reverse Transfer)				
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

Note: Transfer Ratio = Cross-domain F1 / In-domain F1. Values > 0.80 indicate strong transfer. Thoughts→SST-2 achieves highest transfer ratio (0.87); reverse transfers are weaker (<0.80), confirming cognitive thought specialization.

These cross domain experiments provide three important insights in the deployment of cognitive sentiment analysis. Originally, thought-level classification is a truly different task that needs specialized training data-models trained on more traditional sentiment corpora significantly perform poorly (ΔF1 ≈ -20) even when superficially similar tasks are required. This justifies domain-specific annotation work such as our human-elicited thought dataset, as opposed to transfer through a large yet inappropriate data set. Second, the asymmetric pattern of transfer (Thoughts to SST-2 beats SST-2 to Thoughts) is indicative of cognitive representations being informationally rich as opposed to pure polarity, that is, reflecting both affective valence or functional utility. Third, the high performance retention in the move to general sentiment (7781% F1) demonstrates that our BiLSTM acquires generalizable features of sentiment related features in spite of the cognitive specialization, alleviating fears of overfitting to thought-specific anomalies.

G. Comparison with State-of-the-Art

In order to place our contribution in the context of the wider picture of cognitive NLP and sentiment analysis, we comparatively analyzed the performance of BiLSTM with recent studies (2022–2025) that included related tasks: mental health sentiment analysis, thought-level classification, cognitive text analysis, and domain-specific sentiment methods. These comparisons are summed up in Table XVIII, indicating that our work performs competitively or better when dealing with a probably uniquely challenging task formulation.

TABLE XVIII: State-of-the-Art Comparison: Recent Cognitive NLP and Sentiment Analysis Works

Work	Task	Classes	Data Size	Best Model	F1%	Key Difference from Our Work
Mental Health & Cognitive Analysis						
Benrouba & Boudour (2023) [3]	Mental health sentiment	3	10K tweets	RoBERTa-FT	89.2	Social media (public), 3-class polarity

*Author for Correspondence: Jitlogic15@gmail.com).

Desolda et al. (2024) [20]	Medical text sentiment	2	15K notes	BioBERT	91.5	Clinical notes, binary classification
Safari & Chalechale (2023) [4]	Emotion & personality	6	8K texts	BERT-CNN	87.3	Emotion detection, not thought function
Liu et al. (2025) [21]	Clinical LLM application	N/A	Review	GPT-4	N/A	Review paper, no benchmark
Domain-Specific Sentiment Analysis						
Joloudari et al. (2023) [47]	COVID-19 tweets	3	50K	BERT-CNN	92.4	Crisis-specific, public tweets
Ahmed et al. (2024) [15]	Urdu sentiment	2	5K	BERT-FT	88.0	Low-resource language, binary
Prottasha et al. (2022) [16]	Bangla sentiment	3	10K	BERT-CNN-BiLSTM	85.0	Low-resource language, polarity
Sahar et al. (2022) [41]	Cosmetics reviews	2	12K	BERT-FT	91.0	Product reviews, binary
Advanced Sentiment Architectures						
Krosuri & Aravapalli (2023) [13]	Multi-class sentiment	5	25K	ResNeXt-RNN	90.8	Rating prediction, not cognition
Vaghela et al. (2024) [25]	Aspect sentiment (ABSA)	3	20K	Self-Attn LSTM	89.5	Aspect-level, product reviews
Wang et al. (2023) [56]	Sentiment classification	2	30K	BERT-BiLSTM-Attn	93.2	General sentiment, binary
Singh & Sharma (2026)	**Cognitive thought function**	**4**	**4,944**	**BiLSTM-TF-IDF**	**93.9**	**Cognitive function, self-referential, 4-class**

Note: Our work achieves highest F1 among cognitive/mental health tasks despite more granular 4-class schema and smaller dataset. SOTA mental health work (Benrouba 89.2%) uses coarser 3-class polarity on 2× more data.

1) Cognitive NLP Comparison

Out of literature dealing with the mental health and cognitive analysis, our BiLSTM has an F1 of 93.9% on thought classification with 4 classes, which is significantly higher than the mental health sentiment analysis of Benrouba and Boudour (2023) (89.2% F1, 3-class, 10K tweets) despite their simpler task definition and larger dataset. The major difference is cognitive granularity: their work defines the polarity of emotion on the posts in the social media, but we define the internal thought functioning (constructive vs. destructive, essential vs. trivial) and demand more semantic interpretation. The emotion detection work by Safari and Chalechale (2023) [4] has 87.3% on 6-class emotion classification but does not focus on cognitive utility, which is a similarly different psychological construct. We are more similar to clinical cognitive behavioral therapy (CBT) models of focusing on thought patterns rather than on immediate moods [1]. The medical sentiment analysis proposed by Desolda et al. [20] aims to classify binary clinical notes with domain-specific BioBERT [20] with an accuracy of 91.5% and with a task that is not as cognitively challenging as the task in thought classification. Such parallels determine that our work is dealing with a new, intense task at the crossroads of cognitive psychology and NLP where current mental health sentiment strategies aimed at the general social media text do not directly apply.

2) Domain-Specific and Architectural Comparisons

There are recent works that perform well on domain-specific sentiment tasks with special architectures or transfer learning. The BERT-CNN architecture of Joloudari et al. (2023) [47] achieves 92.4% F1 on COVID-19 tweet classification, which is advantageous due to the use of crisis-specific vocabulary and 50K training samples (50 times larger than our dataset). Nevertheless, their 3-class polarity task is performed at the tweet level that analyses social opinions and not individual mental desirability states. Low-resource language works (Ahmed 2024 [15] in Urdu, Prottasha 2022 [16] in Bangali) have F1 between 85 and 88 percent, but have the complication of more limited pre-training data but simpler binary or 3-class polarity (as opposed to functional classification of thought). According to Wang et al. (2023) BERT-BiLSTM hybrid [56], it has 93.2% F1 on the general binary sentiment, which is the success rate of the combination of transformers and BiLSTM, which is also consistent with our decision framework. But their task fails to capture

cognitive subtleties such as the difference between what one needs to do (I must submit the report) and what distracts peripherally (What color notebook should I buy?), which may be affectively neutral but cognitively different. We achieved a 93.9% F1 on a more fine-grained 4-class cognitive schema with a smaller dataset (4,944 vs. their 30K samples), which demonstrates the appropriateness of BiLSTM to special, fine-grained classification problems where data-efficiency and domain-adaptation are more important than scale.

3) Methodological Advances

In addition to the bare performance figures, our work adds a methodological advancement that is not shown in parallel literature. First, our systematic review systematically integrates 150 transfer learning papers into a decision-making framework that is task-oriented (Section III) to offer practitioners a systematic framework on how to choose a model- a meta-level contribution that is not found in task-specific SOTA works. Second, our active learning annotation protocol (60% effort reduction, Section IV-C) shows a viable scalability to cognitive datasets where expert annotation is expensive, unlike literature that uses large pre-labeled corpora or crowd-sourced annotations. Third, our robustness analysis (Section VI-E) with learning curves, cross-validation stability, and sensitivity to hyperparameters offers deployment confidence often unreported in sentiment analysis literature, which deals with the reliability issues that are paramount in mental health applications [20]. Fourth, the domain specificity of cognitive thought classification (22.8% transfer gap) is systematically measured in our cross-domain transfer experiments (Section VI-F), providing practitioners with advice on what to adapt when transferring to related tasks.

4) Limitations and Future Directions

Although our work shows competitive performance, there are a number of limitations that open up future research. Our dataset, which was carefully designed, is English-only, and India-centric in terms of demography, which restricts the extrapolability across cultures. Recent studies of multilingual sentiment [14, 61] indicate that cognitive thought patterns might differ among cultures, making multilingual thought corpora in the future. Also, our 4-class schema, though psychologically based, simplifies the continuous range of valence and utility of thought. Future studies may consider ordinal regression or multi-label classification that would enable thoughts to have more than one characteristic at the same time (e.g., necessary and negative). Incorporation of multimodal cues (voice prosody, facial expressions) [65] would improve the detection of cognitive states as compared to text. Lastly, the mental health applicability assertions would be enhanced by longitudinal validation using clinical populations and feedback through therapist.

VII. DISCUSSION

This section reflects on empirical validation of our decision framework, explores mental health implications, and acknowledges limitations.

A. Validation of the Decision Framework

Our task-based model (Fig. 2) was quite predictive of best strategies in classifying thoughts along various dimensions. Medium-data (5,000 samples) domain-specific medium: BiLSTM-TF-IDF or DistilBERT are the best options, which is highly supported by empirical evidence since BiLSTM reaches 93.9% F1, which is 0.5 percentage points higher than that of RoBERTa, as demonstrated by 1/9 training time [46, 63]. The TF-IDF with BiLSTM implementation was rightfully preferred by the emphasis of lightweight and interpretable methods in the framework, as it did not require full training the language model and had to be balanced with the realities of practical deployment [66]. A framework-recommended active learning integration, which annotation-constrained scenarios are supposed to integrate, was proven successful with 60 per cent effort reductions and high inter-annotator agreement achieved [67, 68, 69, 70].

TABLE XIX: Framework Prediction vs. Empirical Reality (Macro F1)

Framework Pathway	Recommended Model	Best Actual Result	Match?
Medium data / domain-specific	BiLSTM / DistilBERT	**BiLSTM (93.9%)**	✓
Low-resource language	mBERT + distillation	N/A (English only)	—
High-resource general domain	RoBERTa full fine-tuning	RoBERTa (93.4%)	✓

Note: ✓ = Framework prediction confirmed empirically. Source: Sections III and VI.

Table XIX shows that there is no discrepancy between empirical findings and framework recommendations. The following findings give practical practitioner advice: when dealing with novel-domain datasets that have less than 10,000 samples, the combination of TF-IDF and BiLSTM will give optimal returns on investment in performance versus resource costs [46, 63, 65]; when we have general-domain datasets with more than 50,000 samples, full fine-tuning to exploit large-scale data advantages [6]; real-time deployment constraints favor DistilBERT with adapter modules for efficient inference [43, 51]; and low-resource language applications should employ mBERT or XLM-R with knowledge distillation for cross-lingual transfer [59, 60, 61].

B. Implications for Mental Health and Cognitive Applications

The framework enables novel cognitive monitoring capabilities with potential mental health applications [20, 21].

1) Early Risk Signals

Patterns of thought distribution show that there may be signs of cognitive distress, which should be further investigated in the clinic. A combination of Negative and Peripheral thought frequency of over 60% becomes a potential warning sign, which is found in 58 percent of students and 64 percent of parents within our demographic analysis [3], indicating a potential high anxiety or depression risk that should be evaluated by professionals. The required dominance of thought greater than 40 per cent of total thoughts could be a sign of cognitive overload due to excessive responsibility or task load as seen in 38 per cent of professionals [55], which could be a warning of burnout in work intensive groups. Reduction in Positive thought frequency during longitudinal monitoring is a signal of the erosion of resilience that might help predict the further worsening of mental health prior to the appearance of clinical symptoms [2].

2) Real-World Applications

A number of real-world implementation cases employ thought classification as a support of cognitive wellness. Mobile apps might have daily thought logging features that create weekly cognitive health reports that visualize the distribution of thought patterns and report worrisome trends to users [20]. Corporate wellness programs may use distraction device systems that are used to monitor spikes of Peripheral thoughts frequency indicating breakage in employee thought processes during times of high stress to institute proactive measures [55]. Schools might use student cognitive load monitoring in times of examination and locate students who show too much Necessary thought dominance indicating too much academic stress [3]. Cognitive behavioral therapy sessions could include thought pattern visualization tools that therapy support systems could offer, allowing therapists and patients to analyze the development of thought distribution together throughout the treatment sessions [1].

There is a very important caveat that needs to be stressed: this framework does not offer clinical diagnosis but only assistive signals. Such a system is not medically validated and should not substitute professional mental health assessment; all health-related decisions should be under the supervision of clinicians and cannot be made with machine learning models alone [21].

C. Limitations

1) Dataset Limitations

There is a very important caveat that needs to be stressed: this framework does not offer clinical diagnosis but only assistive signals. Such a system is not medically validated and should not substitute professional mental health assessment; all health-related decisions should be under the supervision of clinicians and cannot be made with machine learning models alone [21].

2) Modeling Limitations

The scope of the experiment excludes a number of modern methods that should be explored. The methods of prompt-tuning and zero-shot learning based on GPT-4, T5, or any other large language model have not been evaluated yet [44, 45] despite the possibility of few-shot cognitive classification. Multimodal integration with text and voice prosody analysis or face detection may be able to provide better information on cognitive state detection than text-only methods [65], which is based on affective signatures that cannot be detected through written ideas. Mechanisms to support the constantly evolving user requirements in a longitudinal manner need to be developed where the long-term requirements favor longitudinal learning mechanisms that support personalized classification of thoughts without catastrophic loss of the general patterns [48].

3) Review Limitations

The systematic review scope presents the limitations in terms of time and databases. The time window of 20192025 reflects transfer learning innovations but does not capture new large language model agent architecture post-vanishing [71] our cutoff in the search. Scopes of databases restricted to Scopus, IEEE Xplore, and ACL Anthology do not include arXiv preprints and non-indexed publications that can include potentially relevant emerging methodologies.

VIII. FUTURE WORK AND CONCLUSION

A. Future Directions

1) Methodological Extensions

There are a few methodological improvements that can be made to expand the performance and applicability of this work. The creation of Hindi, Punjabi, and other Indian language thought corpora based on mBERT or XLM-R would be used in the development of multilingual datasets and would facilitate cross-cultural validation and increased accessibility [59, 60, 15, 17, 61, 73]. Timely-based learning testing measuring few-shot GPT-4 and T5 capabilities on thought classification tests might clarify whether large language models need a significant amount of fine-tuning or can give competitive outcomes with minimal instances [44, 45, 71]. The improvement in parameters efficiency of 10-fold could be realized with adapter tuning and other parameter-efficient, fine-tuning methods without compromising classification accuracy [43, 50]. Long-term learning systems permitting user adaptation with no disastrous forgetting would facilitate personalized thought-monitoring systems that learn individual baselines and still maintain the general population trends [48].

2) Application Expansions

Applications in the future may improve monitoring of cognition in a variety of modalities and contexts. A multimodal analysis that involves text and voice prosody characteristics as well as facial expressions would be more reflective of cognitive state data than language alone [65]. A cross-cultural validation research between western and eastern thought patterns would help determine universal and culturally specific cognitive patterns, and to use this information to develop mental

health instruments that are applicable to all people worldwide [14, 61]. Evidence-based efficacy of the application of neuroscience in psychiatry could be determined through clinical trials combining the thought classification into the treatment process with the validation of psychiatrists [21, 20]. The capability to deploy edge models on mobile devices in real-time with the use of model compression and optimization would allow continuous cognitive monitoring without being dependent on the cloud [43, 51].

B. Conclusion

The paper contributes to transfer learning theory and cognitive sentiment practice twofold via two interrelated parts. We introduce a PRISMA-directed systematic review first and synthesize 150 studies into a task-based decision-making framework that relates task traits to the most effective transfer learning methods with empirical validation on a new task of classifying thoughts to demonstrate predictive accuracy. Second, we provide an end-to-end cognitive model that includes a 5,000 sample thought dataset that is trained using hybrid human-AI annotation, active learning protocols with 60 percent annotation effort reduction, and BiLSTM model with 93.9 percent state of the art F1 score that outperforms transformer baselines.

The most important pieces of evidence of the value of the framework are based on the model selection guided by the decisions: selecting BiLSTM instead of RoBERTa due to the recommendations of the framework yields better results (0.5% F1 gain) and more efficient performance (9-fold training speedup), which proves the practical utility of the framework to the practitioner facing new tasks or datasets with limited resources. The overall robustness analysis proves the efficiency of BiLSTM in data (arrived at 3,000 samples, 60 percent of the full dataset) and cross-validation (lowest variance of all models) and hyperparameter robustness, and cross-domain transfer experiments confirm that cognitive thought classification is indeed a distinct task that needs to be trained domain-specifically (22.8 percent specialization gap).

This work positions transfer learning sentiment analysis as a foundation for cognitively rich applications extending beyond polarity detection to monitor mental resilience, cognitive load, and distraction patterns, with broad implications for wellness programs, educational support systems, and therapy assistance tools [20, 21, 3]. Public release of dataset and code through Zenodo enables community validation, replication, and extension to multilingual contexts, multimodal integration, and clinically-validated deployment systems [18], advancing cognitive natural language processing from research prototype toward practical mental health support applications.

ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering at Lovely Professional University for their support in conducting this research.

REFERENCES

[1] Levin, N., Lipshits-Braziler, Y., & Gati, I. (2022). The identification and validation of five types of career indecision: A latent profile analysis of career decision-making difficulties. Journal of Counseling Psychology, 69(4), 452–462. https://doi.org/10.1037/cou0000603

[2] Rice, E. L., & Fredrickson, B. L. (2017). Do positive spontaneous thoughts function as incentive salience? Emotion, 17(5), 840–855. https://doi.org/10.1037/emo0000284

[3] Benrouba, F., & Boudour, R. (2023). Emotional sentiment analysis of social media content for mental health safety. Social Network Analysis and Mining, 13(1), Article 17. https://doi.org/10.1007/s13278-022-01000-9

[4] Safari, F., & Chalechale, A. (2023). Emotion and personality analysis and detection using natural language processing, advances, challenges and future scope. Artificial Intelligence Review, 56(Suppl 3), 3273–3297. https://doi.org/10.1007/s10462-023-10603-3

[5] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). 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 (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[6] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692. https://arxiv.org/abs/1907.11692

[7] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. Advances in Neural Information Processing Systems, 32, 5753–5763.

[8] Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

[9] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. ACM Computing Surveys, 54(3), 1–40. https://doi.org/10.1145/3439726

[10] Al-Qablan, T. A., Noor, M. H. M., Al-Betar, M. A., & Khader, A. T. (2023). A survey on sentiment analysis and its applications. Neural Computing and Applications, 35(29), 21567–21601. https://doi.org/10.1007/s00521-023-08941-y

[11] Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or get left behind: Domain adaptation through BERT language model finetuning for

aspect-target sentiment classification. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 4701–4709). European Language Resources Association.

[12] Ryu, M., & Lee, K. (2020). Knowledge distillation for BERT unsupervised domain adaptation. arXiv preprint arXiv:2010.11478. https://arxiv.org/abs/2010.11478

[13] Krosuri, L. R., & Aravapalli, R. S. (2023). Novel heuristic-based hybrid ResNeXt with recurrent neural network to handle multi-class classification of sentiment analysis. Machine Learning: Science and Technology, 4(2), Article 025003. https://doi.org/10.1088/2632-2153/acd1a3

[14] Augustyniak, Ł., Gruza, M., Gramacki, P., Rajda, K., Morzy, M., & Kajdanowicz, T. (2023). Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark. arXiv preprint arXiv:2306.07902. https://arxiv.org/abs/2306.07902

[15] Ahmed, T., Asghar, M. Z., Ullah, I., Ullah, M., & Iqbal, A. (2024). Urdu sentiment analysis using fine-tuned BERT: A high-performance model for low-resource languages. Cognitive Computation. https://doi.org/10.1007/s12559-024-10237-8

[16] Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer learning for sentiment analysis using BERT based supervised fine-tuning. Sensors, 22(11), Article 4157. https://doi.org/10.3390/s22114157

[17] Khan, F., & Shahid, S. (2022). BiLSTM with self-attention and joint dual input learning for Hindi and Bengali sentiment analysis. Pattern Recognition Letters, 154, 1–8. https://doi.org/10.1016/j.patrec.2021.12.027

[18] Singh, J., & Sharma, G. (2025). Hybrid Human-AI Annotated Thoughts Dataset (Thoughts 1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17444289

[19] Singh, J., & Sharma, G. (2023). Sentiment analysis study of human thoughts using machine learning techniques. In Proceedings of the International Conference on Disruptive Technologies (ICDT 2023), (pp. 1–6).

[20] Desolda, G., Esposito, A., Lanzilotti, R., Piccinno, A., & Costabile, M. F. (2024). From human-centered to symbiotic AI: A focus on medical applications. Multimedia Tools and Applications, 83, 45867–45895. https://doi.org/10.1007/s11042-023-17246-0

[21] Liu, F., Zhou, H., Gu, B., Zhang, J., Ye, Y., Li, Y., & Wu, J. (2025). Application of large language models in medicine. Nature Reviews Bioengineering, 3(6), 221–234. https://doi.org/10.1038/s44222-025-00289-3

[22] Pandey, P., & Rajpoot, D. S. (2020). Sentiment analysis using hybrid approach: A survey. In Proceedings of the 2020 International Conference on Reliable Information and Communication Technology (ICRIT). IEEE.

[23] Zhang, J., Zhao, Y., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28, 649–657.

[24] Hoang, T. T., Bui, H., Pham, T., & Le, T. (2021). Aspect-based sentiment analysis: A survey of deep learning approaches. IEEE Access, 9, 142988–143015. https://doi.org/10.1109/ACCESS.2021.3119119

[25] Vaghela, V. B., Noorani, Z. H., Patel, K., Patel, P. V., Rajput, H., & Shah, M. (2024). Aspect-based sentiment analysis using self-attention-based LSTM model with word embedding. Journal of Computer Science, 20(10), 1195–1202. https://doi.org/10.3844/jcssp.2024.1195.1202

[26] Kumar, A., Singh, R., & Chauhan, D. (2023). Ensemble learning with BERT for aspect sentiment analysis. IAENG International Journal of Computer Science, 50(4).

[27] Hernández, N., Batyrshin, I. Z., & Sidorov, G. (2022). Evaluation of deep learning models for sentiment analysis. Journal of Intelligent & Fuzzy Systems, 43(3), 3091–3102. https://doi.org/10.3233/JIFS-213043

[28] Catelli, R., Bevilacqua, L., Mariniello, N., Scotto di Carlo, V., Magaldi, M., Fujita, H., De Pietro, G., & Esposito, M. (2022). Cross-lingual transfer learning for sentiment analysis of Italian texts. Expert Systems with Applications, 198, Article 116793. https://doi.org/10.1016/j.eswa.2022.116793

[29] Ali, M., Khan, A., & Habib, M. (2022). Sentiment analysis for low-resource languages using zero-shot learning. IEEE Access, 10, 13342–13352. https://doi.org/10.1109/ACCESS.2022.3147130

[30] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[31] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

[32] Malte, A., & Ratadiya, P. (2019). Evolution of transfer learning in natural language processing. arXiv preprint arXiv:1910.07370.

[33] Dai, Y., Liu, J., Ren, X., & Xu, Z. (2020). Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis. arXiv preprint arXiv:2006.05602.

[34] Dong, A., Gao, C., Jia, Y., Liao, Q., Wang, X., Wang, L., & Xiao, J. (2022). Syntax guided domain adaptation for aspect-based sentiment analysis. arXiv preprint arXiv:2211.05457.

[35] Du, Y., Sun, T., Qiu, X., & Huang, X. (2023). Domain adaptation for sentiment analysis using robust internal representation alignment. In Findings of the Association for Computational Linguistics: EMNLP 2023.

[36] Xu, F., Pan, Z., & Xia, R. (2025). A domain-adaptive sentiment analysis method integrating attentional mechanisms and adversarial training. The

Journal of Supercomputing, 81, Article 1117. https://doi.org/10.1007/s11227-024-06548-9

[37] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

[38] Jiang, M., Chen, H., & Feng, Y. (2022). Hybrid model combining BERT and CNN for sentiment analysis. Multimedia Tools and Applications, 81, 15439–15459. https://doi.org/10.1007/s11042-022-12329-y

[39] Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2020). Short text classification with multi-channel BERT. IEEE Access, 8, 22845–22854. https://doi.org/10.1109/ACCESS.2020.2969495

[40] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? In China National Conference on Chinese Computational Linguistics (CCL 2019), Lecture Notes in Computer Science, vol. 11856, pp. 194–206. Springer. https://doi.org/10.1007/978-3-030-32381-3_16

[41] Sahar, M., Agarwal, D., Garg, A., & Jain, R. (2022). Cosmetics product reviews sentiment analysis using BERT. IEEE Transactions on Affective Computing, 14(3), 2198–2209.

[42] Singh, A., Mehta, S., & Bhardwaj, R. (2022). Real-time sentiment analysis of Twitter data using fine-tuned RoBERTa. Procedia Computer Science, 199, 1618–1625. https://doi.org/10.1016/j.procs.2022.01.203

[43] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108

[44] Tan, W., Li, Q., & Zhang, M. (2020). Improving aspect-based sentiment analysis with prompt engineering. ACM Transactions on Information Systems, 39(4), 1–26.

[45] Yu, C., & Chen, L. (2023). Prompt-tuning for cross-lingual sentiment analysis with limited data. Computational Linguistics, 49(2), 301–326.

[46] Wang, Y., Zhang, L., & Liu, Y. (2021). Leveraging attention-based BiLSTM for sentiment classification. Knowledge-Based Systems, 228, Article 107244. https://doi.org/10.1016/j.knosys.2021.107244

[47] Joloudari, J. H., Hussain, S., Nematollahi, M. A., Bagheri, R., Fazl, F., Alizadehsani, R., Lashgari, R., & Talukder, A. (2023). BERT-deep CNN: State of the art for sentiment analysis of COVID-19 tweets. Social Network Analysis and Mining, 13(1), Article 99. https://doi.org/10.1007/s13278-023-01102-y

[48] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 8342–8360). ACL. https://doi.org/10.18653/v1/2020.acl-main.740

[49] Wang, H., He, J., Zhang, X., & Liu, S. (2020). A short text classification method based on N-gram and CNN. Chinese Journal of Electronics, 29(2), 248–254.

[50] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. In Proceedings of the International Conference on Learning Representations (ICLR 2020). https://openreview.net/forum?id=H1eA7AEtvS

[51] Hegde, S. U., Basapur, S. B., Shetty, D., & Suresha, M. (2022). DistilBERT-CNN-LSTM model with GloVe for sentiment analysis on football specific tweets. IAENG International Journal of Computer Science, 49(2), 1–10.

[52] Yang, X., Liu, J., & Wei, W. (2023). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:2006.08097.

[53] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[54] Nouri, H., Karim, S., & Habbat, N. (2023). Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach. International Journal of Electrical and Computer Engineering, 13(4), 4504–4515. https://doi.org/10.11591/ijece.v13i4.pp4504-4515

[55] Pathak, U., & Rai, P. (2023). Sentiment analysis: Methods, applications, and future directions. International Journal for Research in Applied Science and Engineering Technology (IJRASET), 11(2), 1–8.

[56] Wang, Y., Cheng, X., & Meng, X. (2023). Sentiment analysis with an integrated model of BERT and Bi-LSTM based on multi-head attention mechanism. IAENG International Journal of Computer Science, 50(1), 1–8.

[57] Agastya, W., Huda, S., Abawajy, J., Sharmeen, S., & Yearwood, J. (2023). Multichannel convolutional neural network model to improve compound emotional text classification performance. IAENG International Journal of Computer Science, 50(3), 1–9.

[58] Hu, X., Liu, B., Shu, L., & Yu, P. S. (2020). DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1751–1760). ACL.

[59] Raj, S., Sharma, T., & Gupta, N. (2021). Multilingual transfer learning with mBERT for sentiment classification. Applied Intelligence, 51, 712–723.

[60] Roy, K., & Bandyopadhyay, S. (2021). Improving sentiment analysis for Indian languages using BERT and transfer learning. Journal of Intelligent Systems, 30(1), 1–15.

[61] Zhao, W., & Li, D. (2023). Sentiment analysis with multilingual transformers: Benchmarking and error analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(1), 1–25.

[62] Qiang, L., Sun, X., & Long, Y. (2023). Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2023.3325089

[63] Zhang, H., Xu, J., Lei, L., Qiu, J., & Alshalabi, R. (2022). A sentiment analysis method based on bidirectional long short-term memory networks. Applied Mathematics and Nonlinear Sciences, 7(1), 1–10. https://doi.org/10.2478/amns.2021.1.00076

[64] Chang, J.-R., Chen, L.-S., & Lin, L.-W. (2021). A novel cluster based over-sampling approach for classifying imbalanced sentiment data. IAENG International Journal of Computer Science, 48(4), 1118–1128.

[65] Fang, Y., Fu, H., Tao, H., Wang, X., & Zhao, L. (2021). Bidirectional LSTM with multiple input multiple fusion strategy for speech emotion recognition. IAENG International Journal of Computer Science, 48(3), 1–9.

[66] Khan, J., & Lee, Y.-K. (2019). LeSSA: A unified framework based on lexicons and semi-supervised learning approaches for textual sentiment classification. Applied Sciences, 9(24), Article 5562. https://doi.org/10.3390/app9245562

[67] Lu, S., Qi, Y., & Miao, S. (2024). A study of architectural text mining corpus refinement method based on BiLSTM and active learning strategies. In International Conference on Computer Science and Application Engineering (CSAE 2024), pp. 179–185. Springer.

[68] Alturayeif, N., & Ahmad, I. (2025). EASE: An enhanced active learning framework for aspect-based sentiment analysis based on sample diversity and data augmentation. Expert Systems with Applications, 261, Article 125525. https://doi.org/10.1016/j.eswa.2024.125525

[69] Li, X., Ye, W., Zhang, Y., & Sun, X. (2024). GRACE: GRadient-based active learning with curriculum enhancement for multimodal sentiment analysis. In Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM 2024), pp. 5702–5711. ACM.

[70] Chakraborty, S., & Singh, A. (2022). Active sampling for text classification with subinstance level queries. In Proceedings of the AAAI Conference on Artificial Intelligence, 36(6), 20563–20571. https://doi.org/10.1609/aaai.v36i6.21309

[71] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., … Gui, T. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences, 68(2), 1–30. https://doi.org/10.1007/s11432-024-4222-5 .

[72] Sanaullah, A. R., Das, A., Das, A., Kabir, M. A., & Shu, K. (2022). Applications of machine learning for COVID-19 misinformation: A systematic review. Social Network Analysis and Mining, 12(1), 1–21. https://doi.org/10.1007/s13278-021-00853-2

[73] Zhang, J., Zhao, Y., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, 28, 649–657.

[74] Yuan, K., Zhao, C., Chen, Y., Shen, L., Tang, Q., & Jia, C. (2024). Mapping the knowledge structure and research evolution of deep learning. IAENG International Journal of Computer Science, 51(10), 1–12.

AUTHOR BIOGRAPHIES

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 in computer science and engineering at Lovely Professional University, Phagwara, Punjab, India.

He has been teaching engineering students in computer science since 2009. From 2017 to 2022, he worked as a Data Scientist at Wachemo University, Ethiopia, gaining five years of international experience. He is currently an Assistant Professor in the Department of Computer Science and Engineering at Sharda University, Greater Noida, India. His research interests include natural language processing, transfer learning, and sentiment analysis. He has published five Scopus-indexed papers in various fields of computing.

Geeta Sharma received the master's degree and the Ph.D. degree in computer science from Guru Nanak Dev University, Amritsar, India.

She is currently an Assistant Professor with the School of Computer Applications, Lovely Professional University, Phagwara, Punjab, India. She has more than seven years of teaching and research experience in machine learning, fog/cloud computing, IoT, and network security. She has authored over 25 research papers in renowned international journals including Springer, Elsevier, IEEE, and Taylor and Francis, and serves as a reviewer for Springer and IEEE journals. She has registered six patents and currently supervises four Ph.D. students. Dr. Sharma is a respected researcher and educator whose work bridges theoretical research and practical applications in computer science.