Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

THE LIMITS OF LARGE LANGUAGE MODELS IN FINE-GRAINED EMOTION DETECTION: A COMPARATIVE AND ERROR ANALYSIS STUDY

Mahrukh Rafique[*1], Ahmed Asja[2], Shahzad Babar[3], Muhammad Khan[4], Mukhtar Ali Soomro[5]

[*1]*IT intern at CPPA*
[2]*Student, University of Herefordshire*
[3,4,5]*Student, HITEC University Taxila*

[*1]mahrukhrafique786@gmail.com, [2]ahmedasjal86@gmail.com, [3]i.shahzadbabar@gmail.com,
[4]mkhan.edu01@gmail.com, [5]23sp-phd-cs-003@student.hitecuni.edu.pk

DOI: https://doi.org/10.5281/zenodo.20823378

Keywords
emotion recognition, NLP, fine-grained classification, transformer models, DistilBERT, RoBERTa, LoRA, multi-label classification, error analysis, GoEmotions.

Article History
Received: 25 April 2026
Accepted: 07 June 2026
Published: 24 June 2026

Corresponding Author: *
Mahrukh Rafique

Abstract
Emotion recognition in text is an increasingly important natural language processing task, yet the extent to which transformer-based models perform reliably on fine-grained, multi-label emotion classification remains poorly understood. This paper critically evaluates the effectiveness and failure modes of large language models applied to emotion detection, focusing specifically on how emotional granularity degrades classification performance and what structural error patterns emerge. Two benchmark datasets were used: the Emotion dataset (~20,000 Twitter posts across six coarse-grained categories) and GoEmotions (~58,000 Reddit comments across 28 fine-grained emotion categories). TF-IDF baselines with Logistic Regression and SVM were established first, followed by fine-tuning of DistilBERT on the Emotion dataset and DistilBERT, BERT-base, and RoBERTa with Low-Rank Adaptation (LoRA) on GoEmotions. On the coarse-grained task, DistilBERT reached 92.25% accuracy and macro-F1 of 0.87, well above the Logistic Regression baseline of 86.45% accuracy and macro-F1 of 0.80. On GoEmotions, RoBERTa+LoRA achieved micro-F1 0.61, macro-F1 0.55, and Hamming loss 0.0338 outperforming all baselines and DistilBERT by 8.9 macro-F1 points, yet substantially lower than coarse-grained performance, confirming that increased emotional granularity introduces structural difficulties that architecture alone cannot resolve. Structured error analysis identified four failure types: rare-class underperformance, universal semantic confusion across all 28 categories, over-prediction of dominant classes, and systematic under-detection of nuanced emotions. These findings argue for a diagnostic, failure-oriented evaluation framework as a professional and ethical requirement for emotion recognition research.

I. INTRODUCTION

Emotion recognition in text has become an important area within natural language processing (NLP) and affective computing. As digital communication increasingly occurs through social media, customer service platforms, healthcare systems, and automated decision-making pipelines, detecting emotion from written language has grown in practical importance. Applications include mental health monitoring,

customer experience analysis, social media analytics, human-computer interaction, and automated content moderation. Despite progress, the task remains fundamentally challenging because textual emotion is often implicit, context-dependent, culturally influenced, and shaped by ambiguity, sarcasm, and overlapping emotional states [1].

In sensitive applications, misclassifying distress as neutral or interpreting ironic disapproval as approval can carry serious real-world consequences. A system that performs well on aggregate metrics may still fail on the rare or nuanced emotions most critical to deployment contexts. This makes evaluation through headline metrics accuracy or micro-F1 insufficient as a basis for clinical or production decisions. The growing deployment of emotion recognition in high-stakes contexts therefore demands that researchers understand not only how well these models perform on average, but where, why, and how reliably they fail.

Traditional sentiment analysis classifies text broadly as positive, negative, or neutral. Fine-grained emotion recognition goes further, requiring identification of more specific and often overlapping emotional states. In multi-label settings, a single text may express several emotions simultaneously, adding further complexity. Although transformer architectures BERT [2-6], RoBERTa [7], DistilBERT [8-10] have substantially advanced NLP classification, their reliability specifically in fine-grained multi-label emotion recognition remains underexplored relative to the attention given to improving aggregate benchmark scores.

A. Research Objectives

This study investigates the limits of large language models in fine-grained emotion detection through a comparative and error-analytical framework. The core objectives are:

- Develop TF-IDF classical baseline models for both datasets as efficient reference systems.
- Systematically compare performance across coarse-grained (6-class) and fine-grained (28-class) classification.
- Examine the impact of class imbalance on macro-F1 scores for rare emotion categories.
- Fine-tune transformer models (DistilBERT, BERT, RoBERTa+LoRA) and evaluate against classical baselines.
- Conduct structured error analysis using confusion matrices, per-class FP/FN breakdowns, and semantic confusion mapping.
- Develop a four-type taxonomy of model failure modes and apply it across all 28 GoEmotions categories.

B. Research Questions

Three research questions guide the study: (1) Do LLMs perform significantly better than classical machine learning models in emotion detection? (2) Does performance degrade systematically as emotion categories become more fine-grained and multi-label? (3) What systematic failure patterns characterize transformer-based fine-grained emotion classification, and what do these imply for real-world deployment?

C. Novelty and Contribution

Although prior work has compared transformer models to classical baselines on emotion datasets, the present study distinguishes itself by placing structured failure analysis at its center rather than treating error examination as supplementary. A four-type error taxonomy is introduced rare-class underperformance, semantic confusion, over-prediction, and under-detection and applied systematically across all 28 GoEmotions categories. The finding that semantic confusion is universal across all categories, not just rare ones, is a novel quantitative observation. The study further demonstrates that RoBERTa+LoRA outperforms full fine-tuning of DistilBERT by 8.9 macro-F1 points while training only a small fraction of the model's total parameters, providing evidence that backbone architecture quality matters more than parameter count in this setting.

II. Related Work

A. Early Approaches: Lexicon-Based and Classical Methods

The earliest methods for text-based emotion recognition relied on manually constructed

affective lexicons, such as the NRC Emotion Lexicon [11] and the LIWC framework [12]. These approaches assign emotional valence to individual words and aggregate scores across a document. Their strengths are computational efficiency, interpretability, and the absence of a requirement for labeled training data. However, they are fundamentally context-insensitive: the word 'killing' carries opposite valence in 'he was killing it on stage' versus 'he was killing the mood,' yet a lexicon-based system assigns identical scores in both cases. Such methods systematically fail on negation, irony, and the implicit emotional expressions common in natural language.

Supervised machine learning represented a significant advance, introducing data-driven feature engineering with TF-IDF representations and classifiers such as Logistic Regression and SVMs. Researchers [13,12] demonstrated through the CARER system that classical supervised methods trained on large Twitter corpora could classify text into six basic emotion categories competitively, while also confirming that contextual nuance and sarcasm remained fundamental limitations. These insights motivated the use of classical baselines in the present study not to validate classical methods, but to establish a meaningful baseline against which the added value of contextual transformer representations can be precisely measured.

B. Deep Learning and Contextual Representations

Deep learning approaches first RNNs, then LSTMs, and subsequently attention-based architectures directly targeted the context-sensitivity limitations of lexicon-based and classical methods. RNN-based architectures process sequential input and can theoretically capture long-range dependencies within a sentence. Researchers [14] demonstrated that bidirectional LSTMs with attention mechanisms substantially outperformed classical baselines on fine-grained affect regression at SemEval-2018, establishing the value of contextual sequential modeling for emotion tasks. However, RNNs and LSTMs suffer from training instability, vanishing gradients, and difficulty parallelizing during training. The Transformer architecture introduced by researchers. [15] replaced recurrence entirely with self-attention, enabling efficient parallelization and the modeling of arbitrary-distance dependencies within a sequence simultaneously.

C. Transformer-Based Models: BERT, RoBERTa, and DistilBERT

BERT [16] represented a paradigm shift by pre-training a bidirectional encoder on masked language modeling and next-sentence prediction across large corpora, producing rich contextual embeddings that could be fine-tuned efficiently on downstream tasks. RoBERTa [17] improved upon BERT by training on larger mini-batches and more data, removing the next-sentence-prediction objective, and demonstrating substantially improved benchmark performance across a wide range of NLP tasks. DistilBERT [18] provided a distilled version retaining approximately 97% of BERT's performance while being 40% smaller and 60% faster through knowledge distillation making it an attractive option for resource-constrained deployment and as a lightweight fine-tuning baseline in comparative studies such as the present one.

Parameter-efficient fine-tuning, particularly LoRA [19], extended the reach of large pre-trained models by injecting trainable low-rank decomposition matrices into attention layers while keeping the pre-trained weights frozen. This reduces the trainable parameter count to under 1% of the full model while achieving performance comparable to full fine-tuning. The present study applies LoRA to RoBERTa on GoEmotions to test whether parameter-efficient adaptation of a stronger backbone outperforms full fine-tuning of a weaker one a question with direct implications for resource-constrained research and deployment.

D. GoEmotions and Fine-Grained Emotion Recognition

The GoEmotions dataset [20-22], introduced by researchers, marked a pivotal shift in emotion recognition research by providing 27 fine-grained emotion categories plus neutral, annotated across

approximately 58,000 Reddit comments with multi-label annotations. Prior datasets had relied on coarse categorical schemes; GoEmotions exposed the substantial performance degradation that accompanies granularity. Demszky et al. reported baseline BERT macro-F1 scores in the range of 0.46–0.54, markedly lower than coarse-grained results, and identified semantic label overlap and class imbalance as primary drivers. The present study uses GoEmotions as its fine-grained benchmark and extends its analysis with systematic quantitative characterization of failure modes.

E. Multi-Label Classification and Class Imbalance

Multi-label classification introduces complexity beyond single-label tasks, requiring independent binary predictions for each class and demanding careful choice of decision threshold and evaluation metric [24]. Micro-F1 is dominated by frequent classes and can be misleadingly high when rare classes perform poorly; Macro-F1 assigns equal weight to all classes and is therefore the primary metric in this study. Researchers [25] reported gaps of over 20 points between micro- and macro-F1 on GoEmotions, reflecting the severity of class imbalance with a maximum imbalance ratio near 185:1 between the most and least frequent categories. Researchers [26] surveyed class imbalance mitigation strategies including class-weighted loss, data augmentation, threshold adjustment, and cost-sensitive learning; this study employs a positive class weight of 2.0 to partially address imbalance without eliminating its diagnostic value as a variable.

F. Gaps in Existing Research

Despite extensive work on improving aggregate benchmark scores in emotion recognition, two important gaps remain. First, the overwhelming focus on maximizing performance metrics means that structured diagnostic analysis of failure modes particularly for fine-grained settings is underrepresented in the literature. Second, most studies compare a single classical baseline against a single transformer model; the graduated baseline development strategy used in this study provides a richer picture of where classical methods reach their performance ceiling and where transformer fine-tuning delivers genuine added value. Bostan and Klinger [27-29] highlighted annotation overlap between emotion categories as an upper-bound constraint on performance, a finding this study quantitatively extends through co-occurrence analysis of confusion pairs across all 28 GoEmotions categories.

III. Methodology

A. Research Design

This study adopts a diagnostic, comparative research design rather than one optimized purely for benchmark scores. The experimental pipeline has three stages: (1) establishing classical TF-IDF baselines with Logistic Regression and SVM; (2) fine-tuning and evaluating transformer models DistilBERT on the Emotion dataset, and DistilBERT, BERT-base, and RoBERTa with LoRA on GoEmotions; and (3) conducting systematic error analysis to identify failure patterns, rare-class behavior, and semantic confusion between emotionally proximate categories. This design directly addresses all three research questions by enabling both quantitative comparison of model types and qualitative characterization of failure modes.

B. Datasets

Two public benchmark datasets were selected to enable comparison between coarse-grained single-label and fine-grained multi-label emotion classification. The Emotion dataset [30] contains approximately 20,000 English-language Twitter posts, each labeled with one of six basic emotions anger, fear, joy, love, sadness, and surprise framing the task as single-label multi-class classification with a maximum imbalance ratio of roughly 9:1. GoEmotions [31] contains approximately 58,000 Reddit comments annotated with 27 fine-grained emotion categories plus neutral, permitting multiple labels per instance and a maximum imbalance ratio near 185:1. Both datasets are publicly available on Hugging Face with predefined train/validation/test splits.

Table I: Dataset Characteristics Comparison

Characteristic	Emotion Dataset	GoEmotions Dataset
Source	Twitter posts	Reddit comments
Year introduced	2018	2020
Instances	~20,000	~58,000
Emotion categories	6	27 + Neutral (28)
Label structure	Single-label	Multi-label
Avg. text length	Short (tweet)	Moderate (comment)
Max imbalance ratio	~9:1	~185:1
Annotation method	Distant supervision	Crowd-sourced

C. Data Preprocessing

For the Emotion dataset, all text was lowercased and excess whitespace removed. Standard stop words and emotive punctuation (e.g., exclamation marks) were retained, as these carry sentiment-relevant signal. TF-IDF Vectorizer with a maximum of 10,000 features generated inputs for classical models. For DistilBERT, the distilbert-base-uncased tokenizer was applied with truncation and dynamic padding via DataCollatorWithPadding.

For GoEmotions, label sets were converted to 28-dimensional multi-hot binary vectors to support multi-label training. Transformer inputs were tokenized to a maximum sequence length of 128 tokens with dynamic padding, and label tensors were cast to float32 for compatibility with the BCEWithLogitsLoss objective. The original predefined train, validation, and test splits were preserved for both datasets to ensure comparability with prior work and prevent data leakage.

D. Baseline Models

Two classical baselines were trained per dataset. The first applied TF-IDF vectorization with Logistic Regression (maximum 1,000 iterations). The second used TF-IDF with a linear-kernel Support Vector Machine. Both classifiers were used in standard multi-class form for the Emotion dataset. For GoEmotions, a one-vs-rest strategy was applied, training a binary classifier for each of the 28 emotion classes. Multiple SVM configurations were tested including calibrated SVMs with per-class threshold optimization and character-level TF-IDF fusion to provide a thorough picture of classical method performance rather than relying on a single default configuration.

E. Transformer Models and Training Configuration

DistilBERT was fine-tuned for six-class classification on the Emotion dataset for three epochs using cross-entropy loss. On GoEmotions, three transformer configurations were evaluated: DistilBERT (lightweight baseline), BERT-base (intermediate baseline), and RoBERTa with LoRA (parameter-efficient fine-tuning). RoBERTa was adapted using LoRA at rank 8 and alpha 16, leaving pre-trained weights frozen and updating only the low-rank adaptation matrices. All GoEmotions models used BCEWithLogitsLoss with a positive class weight of 2.0 to partially address class imbalance, and a fixed decision threshold of 0.5. All training was conducted on Google Colab using a T4 GPU.

Table II: DistilBERT Training Configuration Emotion Dataset

Hyperparameter	Value / Setting
Model checkpoint	distilbert-base-uncased
Number of labels	6 (single-label)
Loss function	Cross-Entropy (softmax)
Learning rate	2e-5
Batch size (train / eval)	16 / 16
Weight decay	0.01
Epochs	3
Optimizer	AdamW
Tokenizer padding	Dynamic (DataCollatorWithPadding)

Table III: Transformer Training Configurations GoEmotions

Parameter	DistilBERT	BERT-base	RoBERTa + LoRA
Checkpoint	distilbert-base-uncased	bert-base-uncased	roberta-base
Labels	28 (multi-label)	28 (multi-label)	28 (multi-label)
Loss function	BCEWithLogitsLoss	BCEWithLogitsLoss	BCEWithLogitsLoss
Learning rate	2e-5	2e-5	1e-4
Batch size	16	16	16
Epochs	20	3	5
LoRA rank	N/A	N/A	8
LoRA alpha	N/A	N/A	16
Pos. class weight	2.0	2.0	2.0
Decision threshold	0.5	0.5	0.5

F. Evaluation Metrics

For the Emotion dataset, standard accuracy, precision, recall, and weighted F1 were used alongside macro-F1. For GoEmotions, four metrics were employed:

- Micro-F1: aggregates TP, FP, FN across all classes; dominated by frequent categories.
- Macro-F1: computes F1 per class and averages without frequency weighting the primary metric in this study as it treats all classes equally regardless of support.
- Weighted-F1: per-class F1 averaged by class frequency.
- Hamming Loss: fraction of individual label predictions that are incorrect; lower is better.

G. Error Analysis Methodology

Structured error analysis was conducted at three levels. First, for the Emotion dataset, a confusion matrix identified all misclassification pairs, and class-specific error rates and top confusion pairs were ranked by frequency. Second, for GoEmotions, per-class binary confusion matrices were computed using scikit-learn's multilabel_confusion_matrix, yielding false-positive and false-negative rates for each of the 28 categories. False-negative rate (miss rate) was of

particular interest, as it quantifies failure to detect emotions that are genuinely present. Third, a co-occurrence analysis on confusion pairs was performed across all test instances: for each instance, the set of true emotions missed (false negatives) was paired with the set of incorrectly predicted emotions (false positives) to identify systematic semantic confusion patterns. This was supplemented by manual review of misclassified examples to contextualize quantitative findings [32].

H. Tools and Technologies

Table IV: Tools and Technologies Used

Tool / Library	Version	Role in Project
Python	3.12	Core programming language
PyTorch	Colab default	Tensor operations, GPU training, BCEWithLogitsLoss
Hugging Face Transformers	Latest stable	DistilBERT, BERT, RoBERTa models and Trainer API
Hugging Face PEFT	Latest stable	LoRA configuration and adapter injection for RoBERTa
Hugging Face Datasets	Latest stable	Emotion and GoEmotions dataset loading
scikit-learn	Latest stable	TF-IDF, Logistic Regression, LinearSVC, metrics
pandas / numpy	Latest stable	Data manipulation and array operations
matplotlib / seaborn	Latest stable	Confusion matrices and performance visualizations
Google Colab	T4 GPU	Cloud GPU compute environment

IV. Results and Analysis

A. Emotion Dataset Baseline Performance

On the Emotion dataset test set, TF-IDF with Logistic Regression achieved 86.45% accuracy and weighted F1 of 0.859. These are competitive results that confirm the effectiveness of classical approaches for relatively straightforward coarse-grained emotion classification, providing a meaningful reference point for measuring transformer model improvement.

Table V: Baseline Performance Emotion Dataset (Logistic Regression)

Metric	Score
Accuracy	0.8645
Precision (Weighted)	0.8648
Recall (Weighted)	0.8645
F1-Score (Weighted)	0.8594
Macro-F1	0.80

Table VI: Per-Class Performance Logistic Regression on Emotion Dataset

Emotion	Precision	Recall	F1-Score	Support
Sadness	0.89	0.93	0.91	581
Joy	0.84	0.96	0.89	695
Love	0.82	0.58	0.68	159
Anger	0.90	0.81	0.85	275
Fear	0.88	0.79	0.83	224
Surprise	0.84	0.48	0.62	66

The most frequent classes in the test set sadness and joy are also the best performing, while the worst-performing classes are surprise and love. Surprise has very low support (66 instances), while love suffers from semantic proximity to joy, contributing to its comparatively low recall of 0.58.

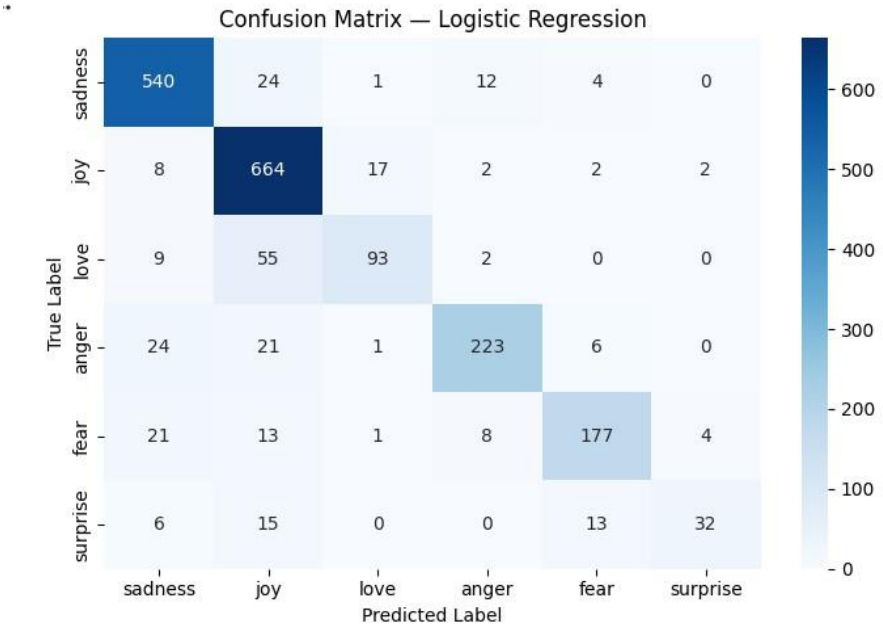

Fig. 1. Confusion Matrix Logistic Regression on Emotion Dataset.

B. Emotion Dataset DistilBERT Performance

Fine-tuning DistilBERT on the Emotion dataset produced a substantial improvement across all metrics, achieving 92.25% accuracy, weighted F1 of 0.9215, and macro-F1 of 0.87 a 6.2-point weighted F1 improvement and 7-point macro F1 improvement over Logistic Regression. This affirmatively answers the first research question: transformer models meaningfully outperform classical baselines on coarse-grained emotion classification.

Table VII: Comparative Summary Emotion Dataset

Metric	Logistic Regression	DistilBERT	Improvement
Accuracy	0.8645	0.9225	+0.0580
Precision (Weighted)	0.8648	0.9212	+0.0564
Recall (Weighted)	0.8645	0.9225	+0.0580
F1-Score (Weighted)	0.8594	0.9215	+0.0621
Macro-F1	0.80	0.87	+0.07

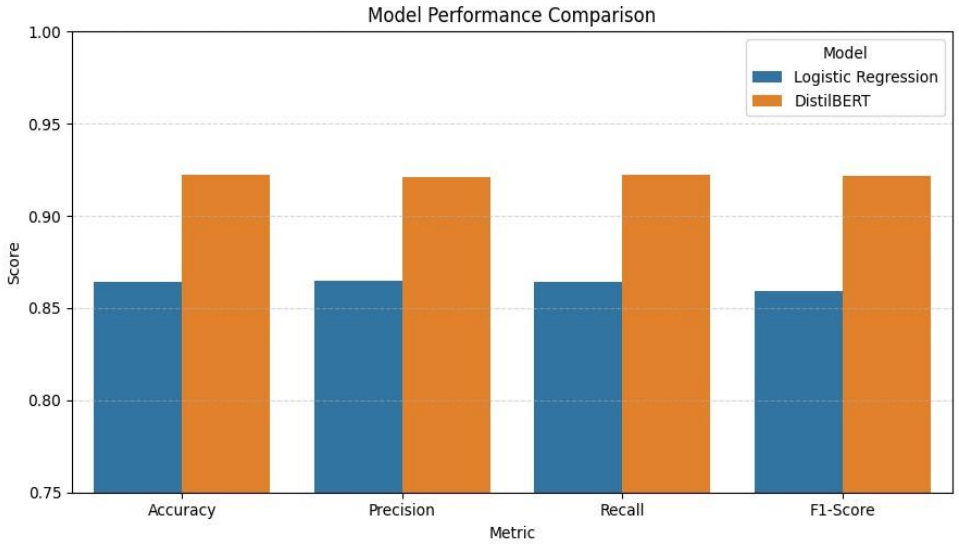

Fig. 2. Model Performance Comparison Logistic Regression vs. DistilBERT on Emotion Dataset.

C. Emotion Dataset Error Analysis

DistilBERT misclassified 155 out of 2,000 test examples (7.75% error rate). The single most common confusion pair was love predicted as joy (29 instances) and joy predicted as love (23 instances), consistent with semantic proximity between these categories. Anger/sadness and fear/surprise were the next most common confusion pairs, also reflecting genuine semantic overlap.

Table VIII: Top Confusion Pairs DistilBERT on Emotion Dataset

True Label	Predicted Label	Count
Love	Joy	29
Joy	Love	23
Anger	Sadness	15
Surprise	Fear	14
Fear	Surprise	10
Anger	Fear	9
Sadness	Anger	8
Surprise	Joy	8

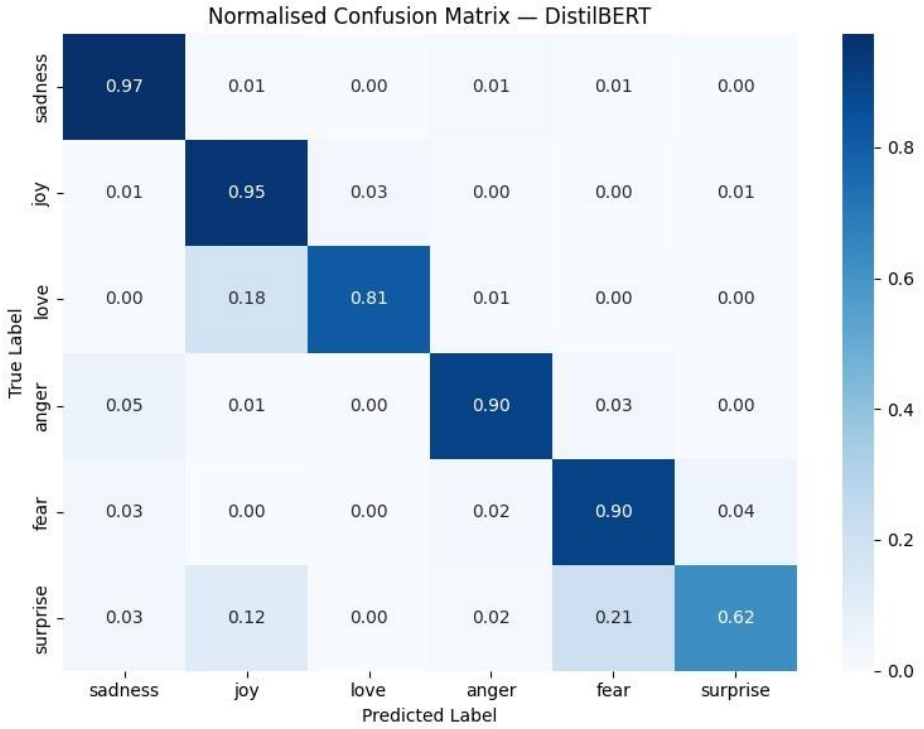

Fig. 3. Confusion Matrix — DistilBERT on Emotion Dataset.

D. GoEmotions Classical Baseline Results

The GoEmotions baselines reveal a stark picture. Even under best case classical configurations, performance on the 28 class multi label task is substantially lower than on the Emotion dataset, reflecting the combined effects of fine-grained labels, extreme class imbalance (up to 185:1), and multi-label complexity. Multiple SVM configurations were tested to avoid a misleadingly pessimistic classical baseline.

Table IX: GoEmotions Classical Baseline Results

Model	Micro-F1	Macro-F1
LR — Unigrams (20k, default)	0.420	0.238
LR — Bigrams, balanced (50k)	0.494	0.433
SVM Calibrated — Per-class threshold	0.518	0.425
SVM — Word + Char TF-IDF (80k)	0.513	0.348

E. GoEmotions Transformer Fine-Tuning Results

Initial transformer fine tuning improved on classical baselines: DistilBERT achieved micro-F1 0.54 and macro-F1 0.46, and BERT-base reached micro-F1 0.52 and macro-F1 0.46. RoBERTa with LoRA was the clear best performer across all metrics, with an 8.9 point macro-F1 advantage over DistilBERT despite training only a fraction of the model's total parameters.

Table X: RoBERTa + LoRA Evaluation Metrics GoEmotions

Metric	Score
Micro-F1	0.6133
Macro-F1 (Primary Metric)	0.5493
Weighted-F1	0.6101
Hamming Loss	0.0338
Exact-Match Accuracy	0.4479

Table XI: Top-5 and Bottom-5 Per-Class Performance RoBERTa + LoRA

Emotion	Precision	Recall	F1	Support
Gratitude	0.903	0.920	0.911	352
Amusement	0.751	0.913	0.824	264
Love	0.750	0.870	0.805	238
Fear	0.667	0.795	0.725	78
Admiration	0.617	0.823	0.705	504
...	...	...	...	...
Caring	0.414	0.444	0.429	135
Grief	0.500	0.333	0.400	6
Annoyance	0.378	0.422	0.399	320
Disappointment	0.382	0.331	0.355	151
Realization	0.397	0.214	0.278	145

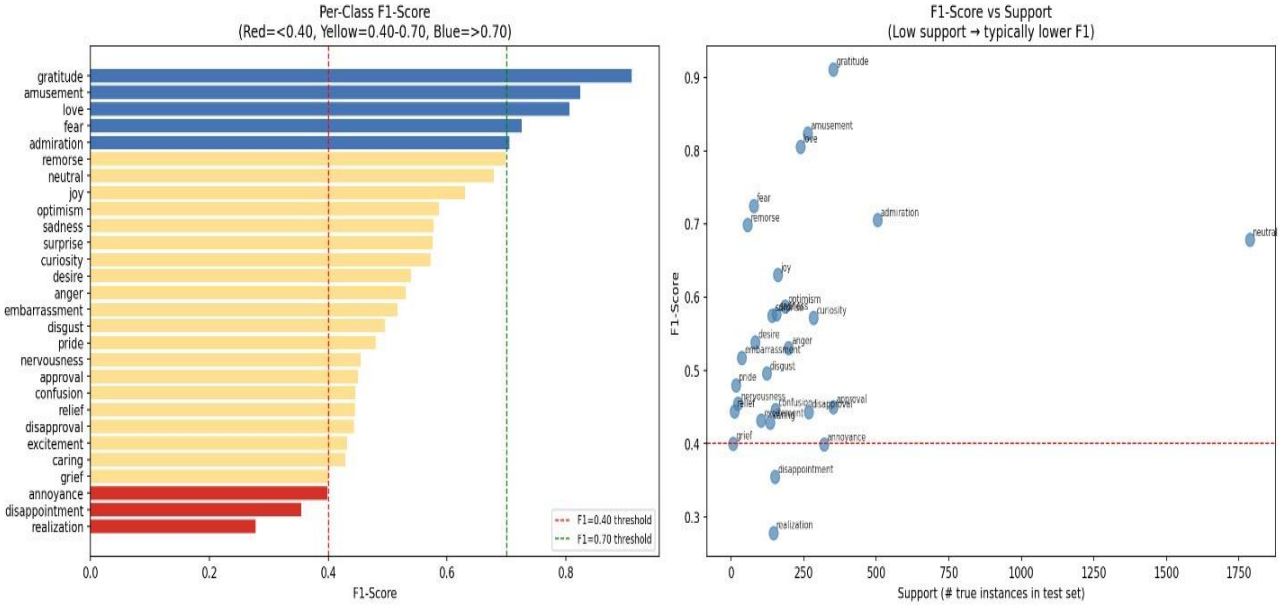

Fig. 4. Per-class F1-Score and F1 vs. Support across 28 GoEmotions categories RoBERTa+LoRA.

F. Rare vs. Frequent Class Performance Gap

Rare categories (bottom 30% by support, average 46 examples) scored an average F1 of 0.521, versus 0.643 for frequent categories (top 30%, average 485 examples) a 12.2-point gap. Crucially, several frequently occurring categories (such as annoyance with 320 test instances) still performed poorly, indicating that semantic ambiguity, not solely data scarcity, drives much of the performance deficit in rare categories.

Table XII: Rare vs. Frequent Class Performance Gap

Group	Avg Support	Avg F1	Sample Classes
Rare classes (bottom 30%)	46	0.521	grief (6), relief (11), pride (16), nervousness (23)
Frequent classes (top 30%)	485	0.643	neutral (1787), admiration (504), gratitude (352)
Performance gap	—	0.122	Frequent classes score 12.2 points higher

V. Error Analysis and Failure Taxonomy

A. Semantic Confusion Patterns

The co-occurrence analysis of confusion pairs across the GoEmotions test set revealed strong systematic patterns. Of the top 15 confusion pairs, 12 involve the neutral class either as the true label (where the model incorrectly predicts an emotion) or as the predicted label (where the model collapses a genuine emotion into neutrality). This reflects the dominance of neutral in the training data (14,219 examples), creating a strong default tendency toward neutrality when emotional signal is ambiguous or subdued.

Table XIII: Top Semantic Confusion Pairs GoEmotions Test Set

True Emotion	Predicted Instead	Count
Approval	Neutral	73
Neutral	Admiration	63
Annoyance	Neutral	61
Disapproval	Neutral	60
Neutral	Curiosity	56
Neutral	Annoyance	54
Neutral	Approval	48
Realization	Neutral	43
Neutral	Confusion	37
Confusion	Curiosity	29
Anger	Neutral	27

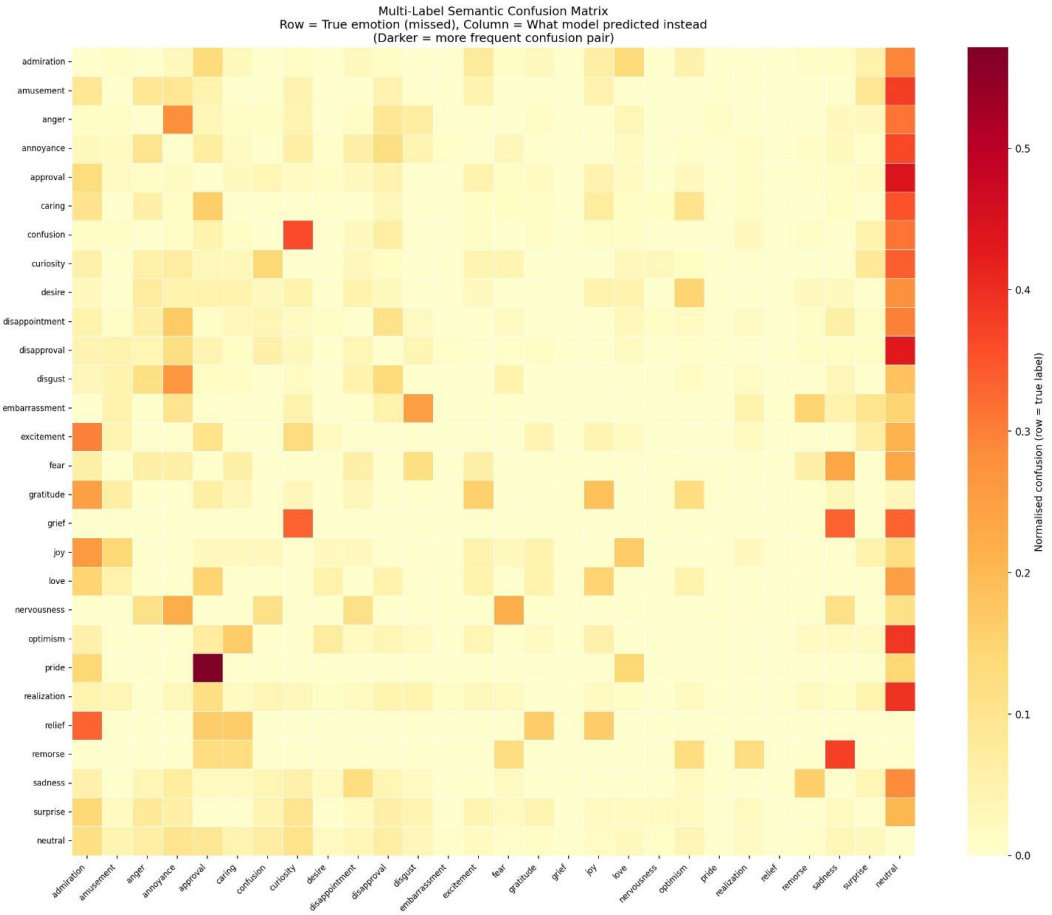

Fig. 5. Semantic Confusion Network — Top confusion pairs across GoEmotions test set.

The confusion→curiosity pair (29 instances, normalized rate 0.363) reflects genuine linguistic overlap: both categories involve questions, uncertainty, and orientation toward seeking information, and their surface forms are often indistinguishable. The anger→neutral confusion (27 instances) suggests that passive or moderate expressions of anger are not reliably distinguished from the absence of emotion.

B. Per-Class False Positive and False Negative Analysis

The false-positive and false-negative analysis reveals asymmetric error profiles across emotion categories. Some emotions are primarily under-detected (high false-negative rate) while others are predominantly over-predicted (high false-positive rate). Realization, disappointment, grief, and relief stand out as persistently under-detected. Annoyance is particularly noteworthy: despite moderate support (320 test instances), the model misses 185 (57.8% miss rate) while simultaneously generating 222 false positives indicating high semantic confusion rather than simple class imbalance.

Table XIV: Per-Class FP/FN Breakdown Selected Emotions

Emotion	TP	FP	FN	FP Rate	FN Rate	F1
Realization	31	47	114	0.009	0.786	0.278
Disappointment	50	81	101	0.015	0.669	0.355
Grief	2	2	4	0.000	0.667	0.400
Relief	4	3	7	0.001	0.636	0.444
Pride	6	3	10	0.001	0.625	0.480
Annoyance	135	222	185	0.043	0.578	0.399
Nervousness	10	11	13	0.002	0.565	0.455

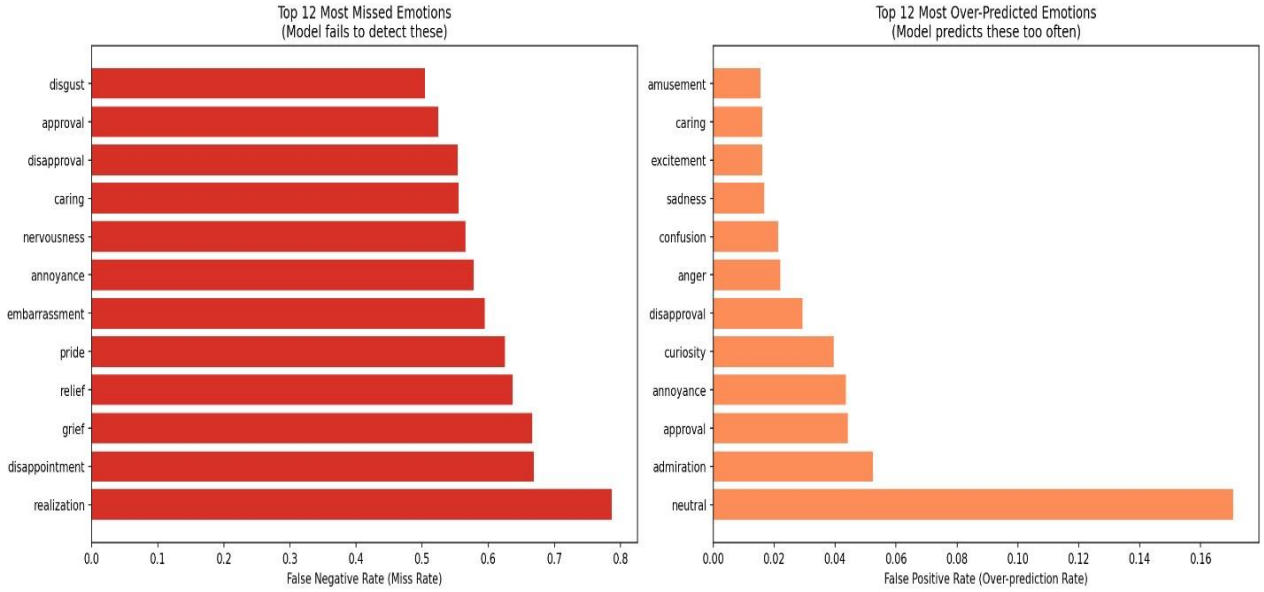

Fig. 6. Top 12 most missed (false negative) and most over-predicted (false positive) emotion categories.

C. Qualitative Error Examination

Manual inspection of misclassified examples revealed four recurring error archetypes corresponding to the quantitative patterns identified above.

Table XV: Qualitative Error Examples

Text Example	True Label(s)	Predicted	Error Type	Explanation
OP is just a kid. Let's just move on.	Approval	Neutral	Semantic confusion	Approval is pragmatic, not explicit. Model defaults to neutral.
Crap. I need more Excedrin. STAT.	Disappointment	Annoyance, Desire	Ambiguous multi-label	Text contains frustration, need, and mild distress. Predicted labels are partially reasonable.
Boomers ruined the world.	Neutral	Disappointment	Reverse semantic error	Model prediction is linguistically reasonable; neutral ground truth reflects annotator judgment.
Eff your video - love Canada Stupid geolock.	Anger, Annoyance	Love	Sarcasm / lexical cue error	Model relies on 'love' surface form; fails to recognize sarcastic usage.

D. Error Taxonomy

Based on the quantitative and qualitative analysis, a four-type failure taxonomy is introduced and applied across all 28 GoEmotions categories:

Table XVI: Error Taxonomy Summary Applied to 28 GoEmotions Categories

Failure Type	Definition	Count (of 28)	Example Classes
TYPE-1: Rare Class Underperformance	Low support + F1 < 0.40	0 in isolation	Compounded with TYPE-2 & TYPE-4
TYPE-2: Semantic Confusion	High confusion rate with ≥1 other label	28 (all)	All 28 categories
TYPE-3: Over-Prediction	Elevated FP rate (top 30% of classes)	9	neutral, admiration, annoyance, approval, anger
TYPE-4: Under-Detection	Elevated FN rate (top 30% of classes)	9	realization, disappointment, grief, relief, pride

TYPE-1 (rare-class underperformance) does not appear in isolation: every rare class also exhibits TYPE-2 (semantic confusion) or TYPE-4 (under-detection), demonstrating that rarity and semantic overlap compound rather than act independently. TYPE-2 (semantic confusion) is universal all 28 categories exhibit meaningful confusion with at least one other label indicating that this is a structural property of the GoEmotions taxonomy rather than a weakness affecting only certain classes. TYPE-3 and TYPE-4 each affect nine categories, often co-occurring in the same category (annoyance exhibits both simultaneously).

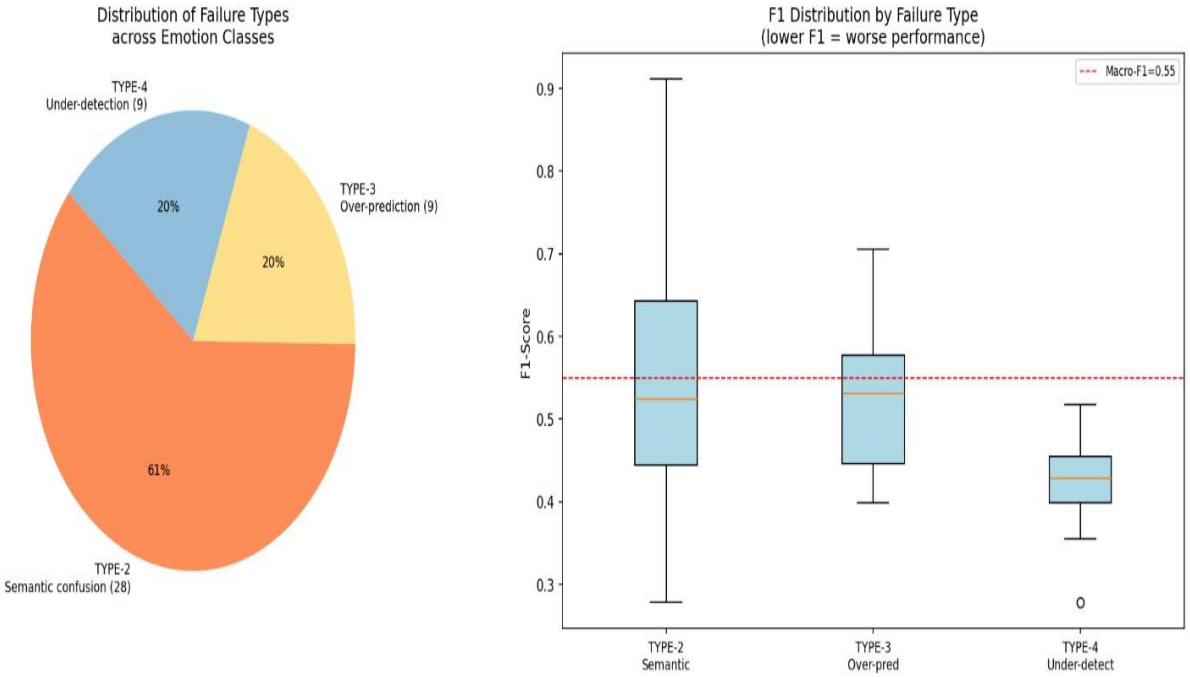

Fig. 7. Distribution of failure types across emotion classes.

VI. Discussion

A. Interpretation of Results

The results of this study provide clear, quantitative confirmation that transformer-based models offer a meaningful advantage over classical baselines on coarse-grained emotion classification DistilBERT's 6.2-point weighted F1 improvement over Logistic Regression on the Emotion dataset is unambiguous. This advantage is consistent with findings in the broader NLP literature that contextual embeddings outperform static TF-IDF features on text classification tasks [6], [8].

However, the picture changes substantially on the fine-grained GoEmotions task. Even the best-performing model RoBERTa+LoRA at macro-F1 0.549 operates at a level that would be inadequate for many deployment contexts, with systematic failure to detect nuanced emotions and universal semantic confusion across all 28 categories. This confirms the second research question affirmatively: performance degrades substantially and structurally as emotional granularity increases, and this degradation cannot be resolved by improving architecture alone.

The 8.9-point macro-F1 advantage of RoBERTa+LoRA over fully fine-tuned DistilBERT is particularly instructive. RoBERTa+LoRA trains only a small fraction of the model's total parameters yet outperforms full fine-tuning of a smaller model, consistent with Hu et al.'s [10] finding that backbone architecture quality matters more than training parameter count. This has practical implications for resource-constrained research: investing in a stronger pretrained backbone with parameter-efficient adaptation may be more effective than exhaustive full fine-tuning of a weaker one.

B. Comparison to Prior Literature

The study's results align with prior literature while extending it in key respects. The DistilBERT vs. Logistic Regression comparison on the Emotion dataset (macro-F1 0.87 vs. 0.80) echoes established findings from Devlin et al. [6] and Sanh et al. [8]. The RoBERTa+LoRA macro-F1 of 0.549 on GoEmotions exceeds the BERT-based range of 0.46–0.54 reported by Demszky et al. [9] while extending their largely qualitative observations with systematic quantitative confusion analysis.

The dominant role of the neutral class in confusion patterns corroborates class imbalance concerns from He and Garcia [11] with specific instance-level evidence. The universality of semantic confusion across all 28 categories resonates with Bostan and Klinger's [12] annotation overlap analysis and provides a new quantitative characterization of it.

The novel finding that no category exhibits TYPE-1 failure in isolation that rare-class underperformance always co-occurs with semantic confusion or under-detection challenges the common assumption that collecting more training examples for rare categories is the primary remedy. Semantic distinctiveness, not frequency alone, appears to determine per-class F1 in this setting.

C. Technical Challenges

Several technical challenges were encountered during implementation. The most persistent was managing the multi-label classification pipeline for GoEmotions, where a subtle type mismatch in label tensor formats caused ineffective early training runs, resolved by implementing a custom data collator. Computational constraints also shaped methodological choices: training full RoBERTa on GoEmotions without LoRA would have been substantially more expensive, making LoRA not merely a research choice but a practical necessity. Threshold selection for multi-label prediction introduced additional complexity, with per-class threshold tuning identified as a direction for future work.

VII. Conclusion and Future Work

A. Summary of Findings

This study set out to investigate the limits of large language models in fine-grained emotion detection, motivated by the observation that aggregate benchmark metrics provide an incomplete and potentially misleading picture of model capability. The experimental program spanning classical TF-IDF baselines, fine-tuned DistilBERT, BERT-base, and RoBERTa+LoRA across two datasets at different levels of emotional granularity provides a detailed and critically grounded answer.

On the coarse-grained Emotion dataset, DistilBERT achieves macro-F1 of 0.87 a solid and reliable result that reflects the relative tractability of six-class single-label classification. On GoEmotions, even the strongest model (RoBERTa+LoRA, macro-F1 0.549) exhibits systematic failure modes that structural improvements alone cannot resolve: semantic confusion is universal across all 28 categories, the neutral class exerts a gravitational pull on ambiguous predictions, and rare categories are under-detected at rates that would be unacceptable in applied settings.

The four-type error taxonomy introduced here rare-class underperformance, semantic confusion, over prediction, and under-detection provides a more actionable framework than aggregate metrics for understanding and communicating model limitations. The finding that semantic distinctiveness, not frequency alone, governs per-class performance has direct implications for how practitioners approach rare-class improvement in emotion recognition.

B. Future Work

Five directions for future investigation emerge from this study. First, per-class threshold calibration: the SVM baseline demonstrated that per-class threshold optimization substantially improves macro-F1; applying the same strategy to RoBERTa+LoRA predictions — independently tuning thresholds for each of the 28 categories on the validation set — is straightforward and would likely yield a meaningful improvement, particularly for underconfident rare categories.

Second, inter-label relationship modelling: the semantic confusion analysis demonstrates that the model conflates emotionally proximate categories in systematic and predictable ways. Architectures that explicitly model label co-occurrence — label embedding networks or graph neural networks applied to the label space — could address TYPE-2 failures by learning inter-label relationships during training rather than treating each class as an independent binary prediction.

Third, data augmentation for rare classes: techniques such as synonym substitution, back-translation, or GPT-based paraphrase generation

could increase training examples for the least frequent categories. Given that failure appears driven more by semantic ambiguity than pure data scarcity, augmentation would need to generate semantically distinguishable rather than merely additional instances. Fourth, cross-domain generalization: both datasets draw from social media platforms with distinctive registers characterized by informality, sarcasm, and abbreviation; whether the failure modes identified here generalize to clinical notes, literary texts, or customer reviews is an important open question. Fifth, broader application of the error taxonomy: testing the four failure types across different architectures, datasets, and training configurations would establish its generalizability as an audit framework for emotion recognition systems.

C. Ethical Considerations

Both datasets used in this study are publicly available from the Hugging Face repository and were collected from public social media platforms. No personally identifying information was involved, and no institutional ethics approval was required. The study was conducted in compliance with GDPR principles and the Data Protection Act 2018. The failure modes documented here — particularly systematic under-detection of distress-related emotions such as grief and disappointment — carry direct implications for deployers of emotion recognition systems in healthcare or moderation contexts. As these systems are increasingly used in contexts with genuine human consequences, failure-oriented evaluation is not optional but a professional and ethical responsibility.

REFERENCES

[1] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[2] J. W. Pennebaker, M. E. Francis, and R. J. Booth, Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Lawrence Erlbaum Associates, 2001.

[3] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, "CARER: Contextualized affect representations for emotion recognition," in Proc. 2018 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 3687–3697.

[4] C. Baziotis, N. Pelekis, and C. Doulkeridis, "NTUA-SLP at SemEval-2018 Task 1: Predicting affective content in tweets with deep attentive RNNs and transfer learning," in Proc. 12th Int. Workshop on Semantic Evaluation (SemEval-2018), 2018, pp. 245–255.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 5998–6008.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North American Chapter of the Assoc. for Computational Linguistics (NAACL-HLT), 2019, pp. 4171–4186.

[7] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.

[9] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proc. 58th Annual Meeting of the Assoc. for Computational Linguistics (ACL), 2020, pp. 4040–4054.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in Proc. 10th Int. Conf. on Learning Representations (ICLR), 2022.

[11] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[12] L. A. M. Bostan and R. Klinger, "An analysis of annotated corpora for emotion classification in text," in Proc. 27th Int. Conf. on Computational Linguistics (COLING), 2018, pp. 2104–2119.

xAhmed, D., Dillshad, V., Danish, A. S., Jahangir, F., Kashif, H., & Shahbaz, T. (n.d.). *Enhancing Home Automation through Brain-Computer Interface Technology*. Retrieved http://xisdxjxsu.asia

Bint-E-Asim, H., Iqbal, S., Danish, A. S., Shahzad, A., Huzaifa, M., & Khan, Z. (n.d.). *Exploring Interactive STEM in Online Education through Robotic Kits for Playful Learning* (Vol. 19). Retrieved http://xisdxjxsu.asia

Danish, A. S., Khan, Z., Jahangir, F., Malik, A., Tariq, W., Muhammad, A., & Khan, Y. (n.d.). *Exploring the Effectiveness of Augmented Reality based E-Learning Application on Learning Outcomes in Pakistan: A Study Utilizing VARK Analysis and Hybrid Pedagogy*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Malik, A., Lashari, T. A., Javed, M. A., Asim, H. B., Muhammad, A., & Khan, Y. (n.d.). *Evaluating the User Experience of an Augmented Reality E-Learning Application for the Chapter on Work and Energy using the System Usability Scale*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Waheed, Z., Sajid, U., Warah, U., Muhammad, A., Khan, Y., & Akram, H. (n.d.). *Exploring Temporal Complexities: Time Constraints in Augmented Reality-Based Hybrid Pedagogies for Physics Energy Topic in Secondary Schools*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-a). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-b). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Khan, J., Jalil, Z., Ali, S., & Samad Danish, A. (2019). Implementation of Smart Aquarium System Supporting Remote Monitoring and Controlling of Functions using Internet of Things. In *Journal of Multidisciplinary Approaches in Science* (Vol. 9). JMAS.

Lashari, T. A., Danish, A. S., Lashari, S. A., Sajid, U., Khan, Z., & Saare, M. A. (n.d.). *Impact of custom built videogame simulators on learning in Pakistan using Universal Design for Learning*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Aizaz, F., Bilal, H., Shahrose, A., Asad, R. M., & Rafiq, M. T. (n.d.). *Navigating Cybersecurity in Global Software Development: Insights, Challenges, and Practices*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Haider, I., Batool, S., Javed, M. A., & Tariq, W. (n.d.). *Enhancing Social Media Text Analysis: Investigating Advanced Preprocessing, Model Performance, and Multilingual Contexts*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-a). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-b). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Attique Khan, M., Ijaz, S., & -UR-Rehman, O. (n.d.). *Exploring Student Morale through Technology Acceptance Model in Higher Education: A Study of AR-Based E-Learning Application*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Noor, N., Hamid, Y., Ali Khan, H., Muneeb Asad, R., & Muhammad Yar Khan, A. (n.d.). *Augmented Narratives: Unveiling the Efficacy of Storytelling in Augmented Reality Environments*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Warah, U., Sajid, U., Adnan Javed, M., & Muhammad Yar Khan, A. (n.d.). *Evaluating the Feasibility and Resource Implications of an Augmented Reality-Based E-Learning Application: A Comprehensive Research Analysis*. Retrieved http://xisdxjxsu.asia

Spectrum of Engineering Sciences ISSN (e) 3007-3138 (p) 3007-312X. (n.d.). https://doi.org/10.5281/zenodo.17365636

Talha Rafiq, M., Osama Habib Gilani, S., Hina Habib Gilani, S., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Augmented Reality Interface for Seamless Control and Management of IoT Devices in Unity Engine*. Retrieved http://xisdxjxsu.asia

Tariq, W., Ali, I., Naeem, H., Batool, S., Faizan Hassan, M., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Enhancing Educational Outcomes through Augmented Reality: A Case Study on Newton's Laws of Motion*. Retrieved http://xisdxjxsu.asia

Yar Khan, A., Samad Danish, A., Hamid, Y., Khan, F., & Ali Kiani, S. (n.d.). *Unraveling Pakistan's Network Landscape-Legacy Structures vs. SDN Paradigms in the Internet Age in IoT Architecture*. Retrieved http://xisdxjxsu.asia

Ahmed, D., Dillshad, V., Danish, A. S., Jahangir, F., Kashif, H., & Shahbaz, T. (n.d.). *Enhancing Home Automation through Brain-Computer Interface Technology*. Retrieved http://xisdxjxsu.asia

Bint-E-Asim, H., Iqbal, S., Danish, A. S., Shahzad, A., Huzaifa, M., & Khan, Z. (n.d.). *Exploring Interactive STEM in Online Education through Robotic Kits for Playful Learning* (Vol. 19). Retrieved http://xisdxjxsu.asia

Danish, A. S., Khan, Z., Jahangir, F., Malik, A., Tariq, W., Muhammad, A., & Khan, Y. (n.d.). *Exploring the Effectiveness of Augmented Reality based E-Learning Application on Learning Outcomes in Pakistan: A Study Utilizing VARK Analysis and Hybrid Pedagogy*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Malik, A., Lashari, T. A., Javed, M. A., Asim, H. B., Muhammad, A., & Khan, Y. (n.d.). *Evaluating the User Experience of an Augmented Reality E-Learning Application for the Chapter on Work and Energy using the System Usability Scale*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Waheed, Z., Sajid, U., Warah, U., Muhammad, A., Khan, Y., & Akram, H. (n.d.). *Exploring Temporal Complexities: Time Constraints in Augmented Reality-Based Hybrid Pedagogies for Physics Energy Topic in Secondary Schools*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-a). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-b). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Khan, J., Jalil, Z., Ali, S., & Samad Danish, A. (2019). Implementation of Smart Aquarium System Supporting Remote Monitoring and Controlling of Functions using Internet of Things. In *Journal of Multidisciplinary Approaches in Science* (Vol. 9). JMAS.

Lashari, T. A., Danish, A. S., Lashari, S. A., Sajid, U., Khan, Z., & Saare, M. A. (n.d.). *Impact of custom built videogame simulators on learning in Pakistan using Universal Design for Learning*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Aizaz, F., Bilal, H., Shahrose, A., Asad, R. M., & Rafiq, M. T. (n.d.). *Navigating Cybersecurity in Global Software Development: Insights, Challenges, and Practices*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Haider, I., Batool, S., Javed, M. A., & Tariq, W. (n.d.). *Enhancing Social Media Text Analysis: Investigating Advanced Preprocessing, Model Performance, and Multilingual Contexts*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-a). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-b). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Attique Khan, M., Ijaz, S., & -UR-Rehman, O. (n.d.). *Exploring Student Morale through Technology Acceptance Model in Higher Education: A Study of AR-Based E-Learning Application*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Noor, N., Hamid, Y., Ali Khan, H., Muneeb Asad, R., & Muhammad Yar Khan, A. (n.d.). *Augmented Narratives: Unveiling the Efficacy of Storytelling in Augmented Reality Environments*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Warah, U., Sajid, U., Adnan Javed, M., & Muhammad Yar Khan, A. (n.d.). *Evaluating the Feasibility and Resource Implications of an Augmented Reality-Based E-Learning Application: A Comprehensive Research Analysis*. Retrieved http://xisdxjxsu.asia

Spectrum of Engineering Sciences ISSN (e) 3007-3138 (p) 3007-312X. (n.d.). https://doi.org/10.5281/zenodo.17365636

Talha Rafiq, M., Osama Habib Gilani, S., Hina Habib Gilani, S., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Augmented Reality Interface for Seamless Control and Management of IoT Devices in Unity Engine*. Retrieved http://xisdxjxsu.asia

Tariq, W., Ali, I., Naeem, H., Batool, S., Faizan Hassan, M., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Enhancing Educational Outcomes through Augmented Reality: A Case Study on Newton's Laws of Motion*. Retrieved http://xisdxjxsu.asia

Yar Khan, A., Samad Danish, A., Hamid, Y., Khan, F., & Ali Kiani, S. (n.d.). *Unraveling Pakistan's Network Landscape-Legacy Structures vs. SDN Paradigms in the Internet Age in IoT Architecture*. Retrieved http://xisdxjxsu.asia

Ahmed, D., Dillshad, V., Danish, A. S., Jahangir, F., Kashif, H., & Shahbaz, T. (n.d.). *Enhancing Home Automation through Brain-Computer Interface Technology*. Retrieved http://xisdxjxsu.asia

Bint-E-Asim, H., Iqbal, S., Danish, A. S., Shahzad, A., Huzaifa, M., & Khan, Z. (n.d.). *Exploring Interactive STEM in Online Education through Robotic Kits for Playful Learning* (Vol. 19). Retrieved http://xisdxjxsu.asia

Danish, A. S., Khan, Z., Jahangir, F., Malik, A., Tariq, W., Muhammad, A., & Khan, Y. (n.d.). *Exploring the Effectiveness of Augmented Reality based E-Learning Application on Learning Outcomes in Pakistan: A Study Utilizing VARK Analysis and Hybrid Pedagogy*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Malik, A., Lashari, T. A., Javed, M. A., Asim, H. B., Muhammad, A., & Khan, Y. (n.d.). *Evaluating the User Experience of an Augmented Reality E-Learning Application for the Chapter on Work and Energy using the System Usability Scale*. Retrieved http://xisdxjxsu.asia

Danish, A. S., Waheed, Z., Sajid, U., Warah, U., Muhammad, A., Khan, Y., & Akram, H. (n.d.). *Exploring Temporal Complexities: Time Constraints in Augmented Reality-Based Hybrid Pedagogies for Physics Energy Topic in Secondary Schools*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-a). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-b). *Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning*. Retrieved http://xisdxjxsu.asia

Khan, J., Jalil, Z., Ali, S., & Samad Danish, A. (2019). Implementation of Smart Aquarium System Supporting Remote Monitoring and Controlling of Functions using Internet of Things. In *Journal of Multidisciplinary Approaches in Science* (Vol. 9). JMAS.

Lashari, T. A., Danish, A. S., Lashari, S. A., Sajid, U., Khan, Z., & Saare, M. A. (n.d.). *Impact of custom built videogame simulators on learning in Pakistan using Universal Design for Learning*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Aizaz, F., Bilal, H., Shahrose, A., Asad, R. M., & Rafiq, M. T. (n.d.). *Navigating Cybersecurity in Global Software Development: Insights, Challenges, and Practices*. Retrieved http://xisdxjxsu.asia

Muhammad, A., Khan, Y., Danish, A. S., Haider, I., Batool, S., Javed, M. A., & Tariq, W. (n.d.). *Enhancing Social Media Text Analysis: Investigating Advanced Preprocessing, Model Performance, and Multilingual Contexts*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-a). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-b). *CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Attique Khan, M., Ijaz, S., & -UR-Rehman, O. (n.d.). *Exploring Student Morale through Technology Acceptance Model in Higher Education: A Study of AR-Based E-Learning Application*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Noor, N., Hamid, Y., Ali Khan, H., Muneeb Asad, R., & Muhammad Yar Khan, A. (n.d.). *Augmented Narratives: Unveiling the Efficacy of Storytelling in Augmented Reality Environments*. Retrieved http://xisdxjxsu.asia

Samad Danish, A., Warah, U., Sajid, U., Adnan Javed, M., & Muhammad Yar Khan, A. (n.d.). *Evaluating the Feasibility and Resource Implications of an Augmented Reality-Based E-Learning Application: A Comprehensive Research Analysis*. Retrieved http://xisdxjxsu.asia

Spectrum of Engineering Sciences ISSN (e) 3007-3138 (p) 3007-312X. (n.d.). https://doi.org/10.5281/zenodo.17365636

Talha Rafiq, M., Osama Habib Gilani, S., Hina Habib Gilani, S., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Augmented Reality Interface for Seamless Control and Management of IoT Devices in Unity Engine*. Retrieved http://xisdxjxsu.asia

Tariq, W., Ali, I., Naeem, H., Batool, S., Faizan Hassan, M., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). *Enhancing Educational Outcomes through Augmented Reality: A Case Study on Newton's Laws of Motion*. Retrieved http://xisdxjxsu.asia

Yar Khan, A., Samad Danish, A., Hamid, Y., Khan, F., & Ali Kiani, S. (n.d.). *Unraveling Pakistan's Network Landscape-Legacy Structures vs. SDN Paradigms in the Internet Age in IoT Architecture*. Retrieved http://xisdxjxsu.asia

[13] Ahmed, D., Dillshad, V., Danish, A. S., Jahangir, F., Kashif, H., & Shahbaz, T. (n.d.). Enhancing Home Automation through Brain-Computer Interface Technology. Retrieved http://xisdxjxsu.asia

[14] Bint-E-Asim, H., Iqbal, S., Danish, A. S., Shahzad, A., Huzaifa, M., & Khan, Z. (n.d.). Exploring Interactive STEM in Online Education through Robotic Kits for Playful Learning (Vol. 19). Retrieved http://xisdxjxsu.asia

[15] Danish, A. S., Khan, Z., Jahangir, F., Malik, A., Tariq, W., Muhammad, A., & Khan, Y. (n.d.). Exploring the Effectiveness of Augmented Reality based E-Learning Application on Learning Outcomes in Pakistan: A Study Utilizing VARK Analysis and Hybrid Pedagogy. Retrieved http://xisdxjxsu.asia

[16] Danish, A. S., Malik, A., Lashari, T. A., Javed, M. A., Asim, H. B., Muhammad, A., & Khan, Y. (n.d.). Evaluating the User Experience of an Augmented Reality E-Learning Application for the Chapter on Work and Energy using the System Usability Scale. Retrieved http://xisdxjxsu.asia

[17] Danish, A. S., Waheed, Z., Sajid, U., Warah, U., Muhammad, A., Khan, Y., & Akram, H. (n.d.). Exploring Temporal Complexities: Time Constraints in Augmented Reality-Based Hybrid Pedagogies for Physics Energy Topic in Secondary Schools. Retrieved http://xisdxjxsu.asia

[18] Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-a). Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning. Retrieved http://xisdxjxsu.asia

[19] Faizan Hassan, M., Mehmood, U., Samad Danish, A., Khan, Z., Muhammad Yar Khan, A., & Muneeb Asad, R. (n.d.-b). Harnessing Augmented Reality for Enhanced Computer Hardware Visualization for Learning. Retrieved http://xisdxjxsu.asia

[20] Khan, J., Jalil, Z., Ali, S., & Samad Danish, A. (2019). Implementation of Smart Aquarium System Supporting Remote Monitoring and Controlling of Functions using Internet of Things. In Journal of Multidisciplinary Approaches in Science (Vol. 9). JMAS

[21] Lashari, T. A., Danish, A. S., Lashari, S. A., Sajid, U., Khan, Z., & Saare, M. A. (n.d.). Impact of custom built videogame simulators on learning in Pakistan using Universal Design for Learning. Retrieved http://xisdxjxsu.asia

[22] Muhammad, A., Khan, Y., Danish, A. S., Aizaz, F., Bilal, H., Shahrose, A., Asad, R. M., & Rafiq, M. T. (n.d.). Navigating Cybersecurity in Global Software Development: Insights, Challenges, and Practices. Retrieved http://xisdxjxsu.asia

[23] Muhammad, A., Khan, Y., Danish, A. S., Haider, I., Batool, S., Javed, M. A., & Tariq, W. (n.d.). Enhancing Social Media Text Analysis: Investigating Advanced Preprocessing, Model Performance, and Multilingual Contexts. Retrieved http://xisdxjxsu.asia

[24] Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-a). CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN. Retrieved http://xisdxjxsu.asia

[25] Muhammad Yar Khan, A., Shaheen, S., Samad Danish, A., Warah, U., -UR-Rehman, O., & Faizan Hassan, M. (n.d.-b). CISCO Packet Tracer Enterprise Level Architecture using the Concept of SDN. Retrieved http://xisdxjxsu.asia

[26] Samad Danish, A., Attique Khan, M., Ijaz, S., & -UR-Rehman, O. (n.d.). Exploring Student Morale through Technology Acceptance Model in Higher Education: A Study of AR-Based E-Learning Application. Retrieved http://xisdxjxsu.asia

[27] Samad Danish, A., Noor, N., Hamid, Y., Ali Khan, H., Muneeb Asad, R., & Muhammad Yar Khan, A. (n.d.). Augmented Narratives: Unveiling the Efficacy of Storytelling in Augmented Reality Environments. Retrieved http://xisdxjxsu.asia

[28] Samad Danish, A., Warah, U., Sajid, U., Adnan Javed, M., & Muhammad Yar Khan, A. (n.d.). Evaluating the Feasibility and Resource Implications of an Augmented Reality-Based E-Learning Application: A Comprehensive Research Analysis. Retrieved http://xisdxjxsu.asia

[29] Spectrum of Engineering Sciences ISSN (e) 3007-3138 (p) 3007-312X. (n.d.). https://doi.org/10.5281/zenodo.17365636

[30] Talha Rafiq, M., Osama Habib Gilani, S., Hina Habib Gilani, S., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). Augmented Reality Interface for Seamless Control and Management of IoT Devices in Unity Engine. Retrieved http://xisdxjxsu.asia

[31] Tariq, W., Ali, I., Naeem, H., Batool, S., Faizan Hassan, M., Samad Danish, A., & Muhammad Yar Khan, A. (n.d.). Enhancing Educational Outcomes through Augmented Reality: A Case Study on Newton's Laws of Motion. Retrieved http://xisdxjxsu.asia

[32] Yar Khan, A., Samad Danish, A., Hamid, Y., Khan, F., & Ali Kiani, S. (n.d.). Unraveling Pakistan's Network Landscape-Legacy Structures vs. SDN Paradigms in the Internet Age in IoT Architecture. Retrieved http://xisdxjxsu.asia