**Computing Infrastructure (Operating System, hardware, etc )**

**Computing Infrastructure**

1. **Operating System**:
	* Local machine: Windows 11 Education (Version 23H2, OS Build 22631.4460)
2. **Hardware**:
	* Processor: 12th Gen Intel® Core™ i5-1235U CPU @ 1.30 GHz
	* Installed RAM: 8.00 GB (7.73 GB usable)
	* System type: 64-bit operating system, x64-based processor
3. **GPU**:
	* For training models: NVIDIA Tesla P100 (via Kaggle)
	* GPU Memory: Up to 16 GiB
4. **RAM**:
	* Kaggle environment: Maximum available: 29 GiB
	* Local machine: 8 GiB
5. **Storage**:
	* Kaggle Disk Space: Maximum 57.6 GiB

**Code Repository URL**

[**https://github.com/ezrauzair/ISEAR-Paper-Code**](https://github.com/ezrauzair/ISEAR-Paper-Code)

**Dataset URL**

[**https://huggingface.co/datasets/gsri-18/ISEAR-dataset-complete**](https://huggingface.co/datasets/gsri-18/ISEAR-dataset-complete)

**Selection Method**

The techniques chosen for this study were aimed at improving the accuracy and reliability of emotion classification, especially considering the challenges of the ISEAR dataset, such as limited data and class imbalance. The approach included the following steps

1. **Data Augmentation Process :**

To address the challenges of class imbalance and limited data in the ISEAR dataset, I implemented **data augmentation** using the **nlpaug library** and its **SynonymAug augmenter** (based on WordNet). Synonym replacement was applied to generate variations of the text, increasing the dataset size while preserving the emotional intent of the sentences.

**Steps Followed**:

* **Original Dataset**:
The ISEAR dataset initially contained the following emotion classes:
	+ Joy: **1079** sentences
	+ Fear: **1066** sentences
	+ Anger: **1082** sentences
	+ Sadness: **1071** sentences
	+ Disgust: **1076** sentences
	+ Shame: **1092** sentences
	+ Guilt: **1049** sentences
* **Augmentation Procedure**:
	+ Synonym replacement was applied to each emotion class.
	+ The augmentation ensured the linguistic and emotional integrity of the data.
* **Final Dataset**:
The dataset was expanded and balanced, with each class containing approximately 3,200 sentences:
	+ Joy: **3,237** sentences
	+ Fear: **3,198** sentences
	+ Anger: **3,246** sentences
	+ Sadness: **3,213** sentences
	+ Disgust: **3,228** sentences
	+ Shame: **3,276** sentences
	+ Guilt: **3,147** sentences

By combining the augmented data with the original dataset, the models were exposed to greater linguistic diversity, enabling better generalization and performance on unseen data.

**Repository**:
The augmented dataset and augmentation code are available at:
**GitHub Repository**: <https://github.com/ezrauzair/EmotionAugmentation-ISEAR>

1. **Transformer Models:**

State-of-the-art **transformer-based models** were chosen for their exceptional performance in Natural Language Processing tasks. These models were implemented to effectively capture the complexity of human emotions:

* **DeBERTa-v3-large**
* **Electra-base-discriminator**
* **RoBERTa-base**
* **T5-base**
* **XLNet-base-cased**

These transformers utilize advanced contextual embeddings and self-attention mechanisms, allowing them to model nuanced emotional expressions in text data.

1. **Hybrid Models (Transformer + CNN):**

To further enhance classification performance, **hybrid architectures** combining transformers with convolutional layers (CNNs) were implemented:

* **Transformers**: Provide rich contextual embeddings for textual inputs.
* **CNNs**: Extract spatial features from transformer embeddings, enabling finer detection of subtle emotional cues.

The hybrid models were tested using the following transformer architectures:

* DeBERTa-v3-large
* Electra-base-discriminator
* RoBERTa-base
* T5-base
* XLNet-base-cased

This integration leverages the strengths of transformers for contextual understanding and CNNs for capturing local patterns, resulting in more robust emotion classification.

4. **Evaluation and Comparison:**

* Training and evaluation of each model (both single transformers and their hybrids) was done to ascertain their level of performance in relation to precision, recall, F1-score.

**Assessment Metrics**

The evaluation metrics used—**accuracy**, **confusion matrix**, and **classification report** (precision, recall, F1-score)—are standard in emotion detection and NLP classification tasks.

1. **Accuracy**
	* **Definition**: Accuracy measures the proportion of correctly classified samples out of the total samples.
	* **Justification**: Accuracy provides a quick and high-level overview of model performance. However, it can be misleading in the case of class imbalance, as it may not reflect the model's performance for minority classes.
	* **Relevance**: In the ISEAR dataset, where class distribution was initially imbalanced, accuracy was used as a baseline metric to observe overall performance improvements after augmentation and model enhancements.
2. **Confusion Matrix**
	* **Definition**: The confusion matrix provides a detailed breakdown of the model's predictions by displaying the counts of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each class.
	* **Justification**: It helps analyze **class-specific performance** and identify where the model struggles, particularly in distinguishing between similar emotions (e.g., shame and guilt).
	* **Relevance**: Given that subtle emotions can be easily misclassified, the confusion matrix acts as a diagnostic tool to pinpoint areas for further model improvement.
3. **Classification Report**
The classification report consists of three key metrics: **Precision**, **Recall**, and **F1-Score**. Each of these metrics addresses specific evaluation needs:
	* **Precision**:
		+ **Definition**: Precision is the proportion of correctly predicted positive samples out of all predicted positive samples.
		+ **Justification**: High precision reduces false positives, ensuring that predictions for emotions (e.g., joy or anger) are reliable.
		+ **Relevance**: In tasks like emotion detection, where false alarms can compromise the system's credibility, precision is particularly important.
	* **Recall**:
		+ **Definition**: Recall (also known as sensitivity) measures the proportion of correctly predicted positive samples out of all actual positive samples.
		+ **Justification**: High recall reduces false negatives, ensuring that true emotional instances are not overlooked.
		+ **Relevance**: Recall is critical in emotion classification tasks where certain emotions (e.g., sadness or guilt) may be underrepresented and harder to detect.
	* **F1-Score**:
		+ **Definition**: The F1-Score is the harmonic mean of precision and recall, balancing both metrics into a single value.
		+ **Justification**: It is particularly useful for **imbalanced datasets** where accuracy alone may not suffice to evaluate performance.
		+ **Relevance**: For the ISEAR dataset, where emotions like shame and guilt had fewer samples initially, the F1-Score ensures fair assessment by accounting for both false positives and false negatives.

The combination of **accuracy**, **confusion matrix**, and **classification report** (precision, recall, and F1-Score) offers a comprehensive evaluation framework that is:

1. **Robust**: Each metric complements the others, ensuring that both overall performance (accuracy) and class-specific nuances (confusion matrix, F1-Score) are captured.
2. **Interpretative**: These metrics allow for detailed insights into the model’s strengths and limitations across individual emotions.
3. **Aligned to Task Requirements**: Emotion classification involves subtle and fine-grained distinctions between emotions. Precision and recall ensure that minority and overlapping classes are evaluated effectively.

By using these metrics, the study validates the improvements achieved through **data augmentation**, **transformer-based models**, and **hybrid architectures**, while providing a rigorous and interpretable performance analysis.