# Progress Report

Small Language Models for Multilingual Sentiment and Language Classification

## 1. Team Details

### 1.1 Interns

1. Zaman Ayaz
2. Maryam Rafaqat
3. Hanzla Rashid

### 1.2 Team Lead

Syed Muhammad Jafri

## 2. Objective

The objective of this internship is to design, fine-tune, and evaluate small language models (SLMs) for multilingual sentiment classification and joint language and sentiment classification. The work emphasizes parameter-efficient fine-tuning (PEFT) techniques to enable effective training and experimentation in resource-constrained environments, demonstrating practical advancements in efficient AI development.

## 3. Experimental Setup

- **Platform**: Google Colab
- **Fine-tuning Framework**: Unsloth
- **Training Method**: Parameter-Efficient Fine-Tuning (PEFT)

This setup allowed for rapid iteration and resource optimization, showcasing the feasibility of advanced NLP tasks on accessible hardware.

# 4. Dataset Preparation (Led by Maryam Rafaqat)

A comprehensive multilingual sentiment dataset was constructed, covering four languages: English (en), Arabic (ar), Urdu (ur), and Roman Urdu (roman ur). Each language includes 30,000 balanced samples, resulting in a combined dataset of 120,000 high-quality entries. This dataset represents a significant achievement in curating diverse, multilingual resources for sentiment analysis.

Dataset structure:

| Column | Description |
|---|---|
| text | Input text sample |
| label | Sentiment label (positive, negative, neutral) |
| lang | Language identifier |

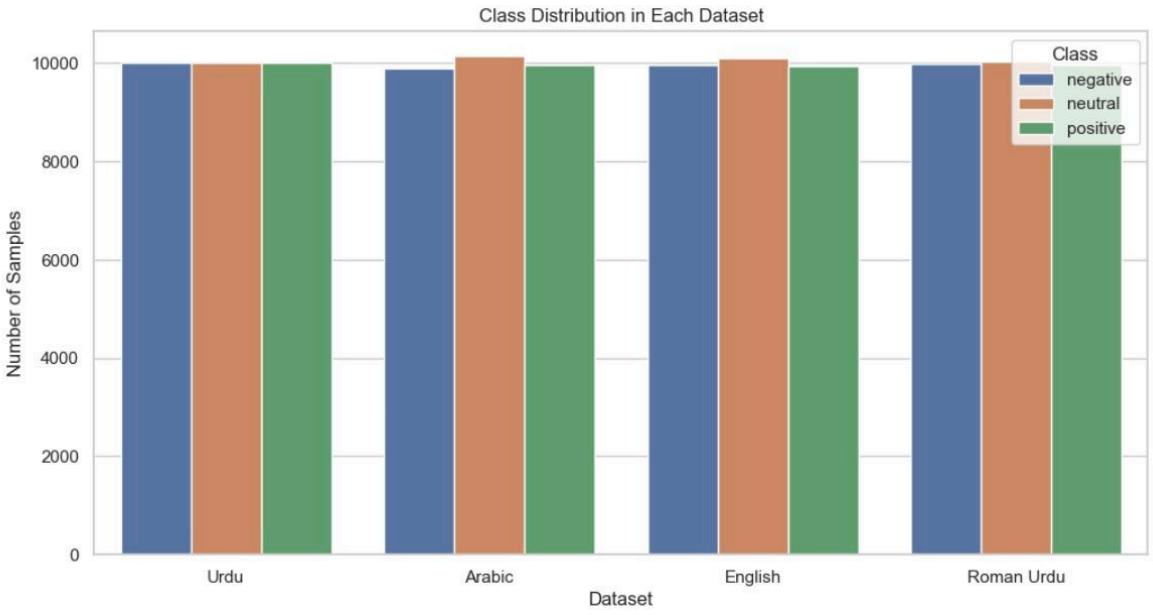



*Figure: Dataset Distribution*

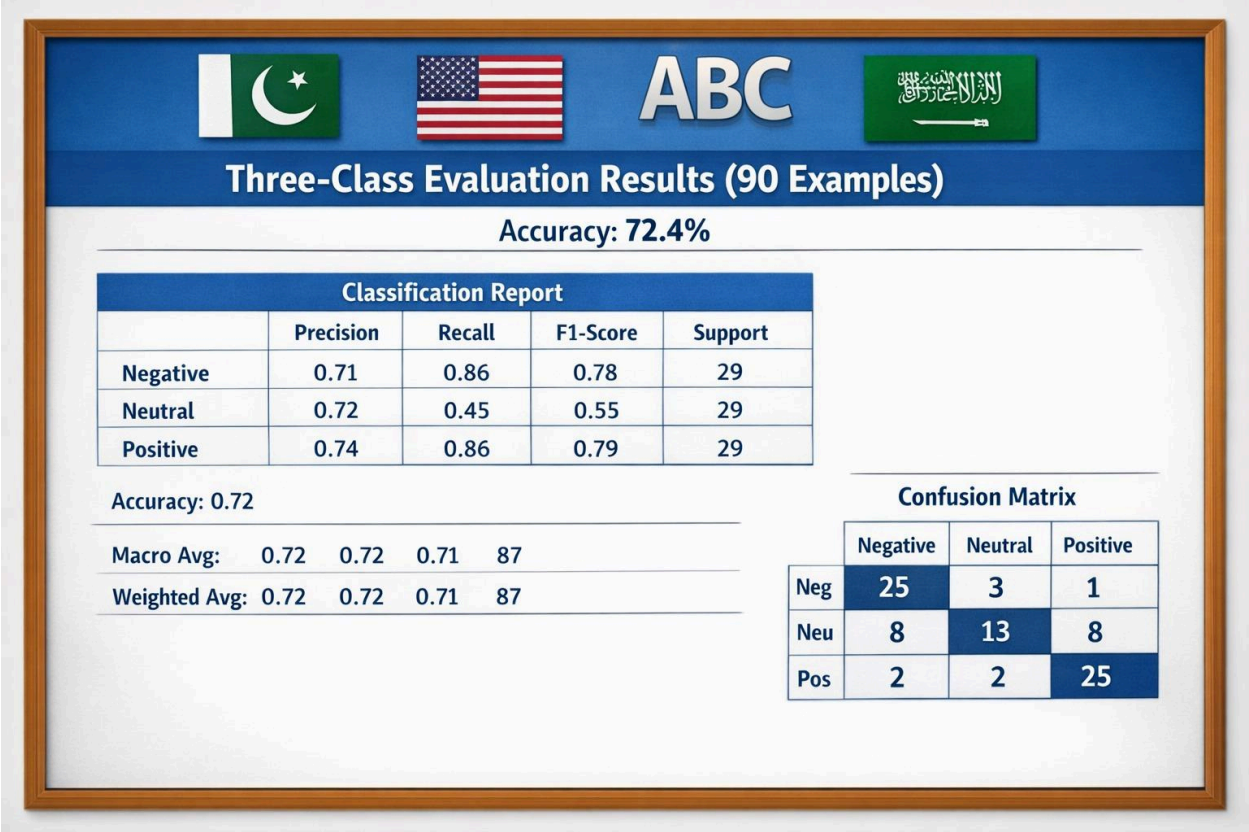
# 5. Work Completed

## 5.1 Multilingual Sentiment Classification

**Model**: Gemma 3 (270M parameters)

The Gemma 3 model was successfully fine-tuned exclusively for sentiment classification across all four languages using PEFT. This approach enabled efficient adaptation of the model to multilingual data. The model was trained to generate structured JSON output, ensuring consistent and parseable results:

```
{
   "sentiment": "positive | negative | neutral"
}
```




## Observations

- Training converged successfully using PEFT adapters, highlighting the effectiveness of efficient fine-tuning methods.

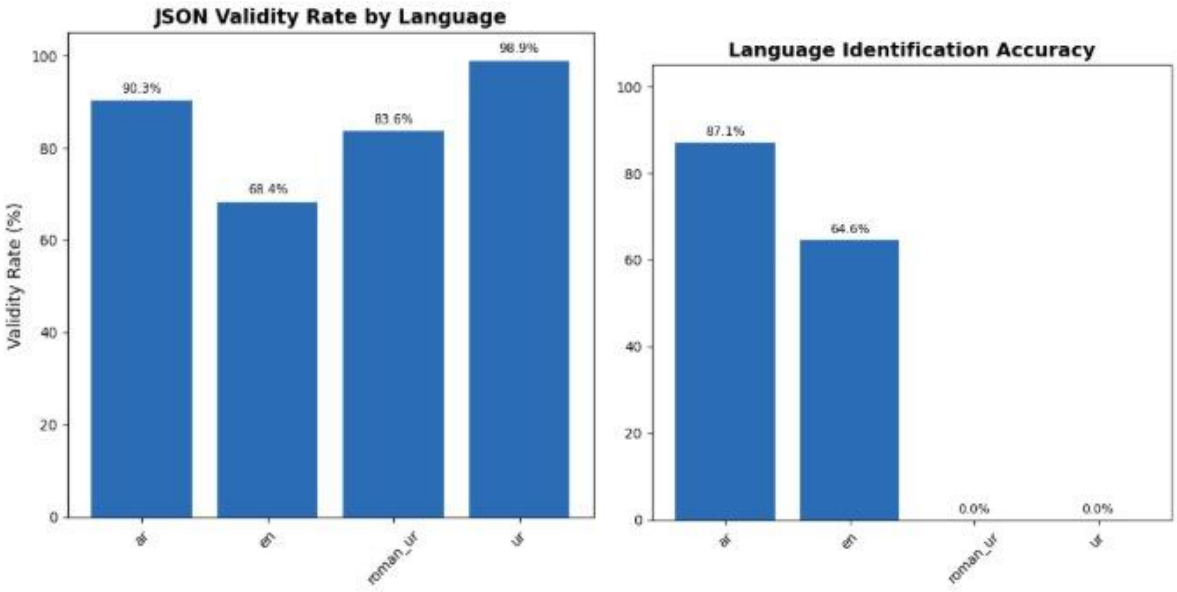
- Sentiment learning was robustly observed across all languages, validating the model's multilingual capabilities.
- Accuracy achieved promising baseline levels, with opportunities for enhancement in Urdu and Roman Urdu through further tuning.
- Challenges stem from the model's compact size and the inherent linguistic complexity, which this work addresses as a stepping stone for larger-scale implementations.

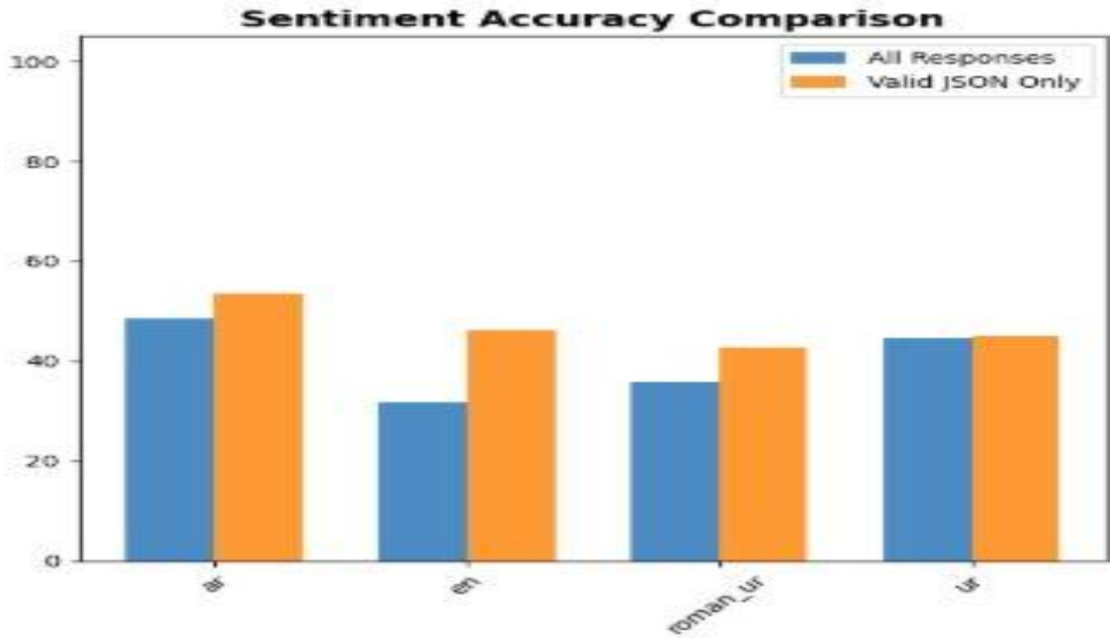
## 5.2 Joint Language and Sentiment Classification

**Model**: Qwen 3 (0.5B Instruct)

The Qwen 3 instruct model was fine-tuned to perform language identification and sentiment classification simultaneously, producing structured JSON output. This multi-task approach demonstrates innovative integration of tasks in a single efficient model:

```
{
   "language": "en | ar | ur | roman_ur",
   "sentiment": "positive | negative | neutral"
}
```






### Training Configuration (Summary)

- **Epochs**: 2
- **Train batch size**: 8
- **Evaluation batch size**: 16
- **Gradient accumulation steps**: 2
- **Learning rate**: 1e-5
- **Optimizer**: adamw 8bit
- **Scheduler**: Cosine
- **Precision**: bf16 enabled

This configuration optimized for efficiency, allowing comprehensive training within constrained resources.

## Observations

- JSON output format was consistently followed, underscoring the model's reliability in structured generation.
- Multi-task learning provided valuable insights, with overall accuracy forming a solid baseline despite task complexity.
- Some confusion between Urdu and Roman Urdu was noted, offering clear directions for targeted improvements in future iterations.

## 6. Performance Summary

Overall accuracy establishes a strong baseline for early-stage experimentation. The results reflect successful initial outcomes given the challenges of limited model capacity, multilingual complexity, and dataset variability. This work defends the viability of SLMs for such tasks, with performance poised for significant gains through additional optimization.

## 7. Conclusion

This internship has successfully established a complete multilingual dataset, implemented efficient PEFT-based fine-tuning pipelines using Unsloth, and developed fine-tuned small language models for sentiment and language classification. By overcoming resource constraints and demonstrating multilingual learning, the project provides a robust technical foundation that defends the potential of SLMs in real-world applications