Methodology

1 Dataset and Preprocessing

The dataset consists of Reddit comments categorized according to the 16 Myers-Briggs Type Indicator (MBTI) personality types. The comments are divided into 16 separate tables, each corresponding to one personality type.

Understanding how federated learning (FL) handles data distribution is crucial for evaluating its effectiveness. To ensure a thorough comparison, the dataset will be distributed among clients using three different strategies, each designed to test different aspects of FL's performance:

1.1 Balanced Split

Each client receives an equal amount of data from all personality types, but with different samples. This ensures that every client has a diverse representation of personality types while avoiding artificially duplicated data.

1.2 Real-World Distribution

Personality types are distributed according to their real-world prevalence, based on statistical data from 40 million MBTI test results (e.g., INTJs constitute 2% of the population). Each client receives an independent subset following the same statistical distribution. This setup tests FL's ability to work under real-world conditions where data is naturally imbalanced while ensuring that every client receives a unique but proportional dataset.

1.3 Random Split

Each client is assigned a random subset of personality types, with between 1 and 7 personality types per client. This setup simulates real-world scenarios where data availability is inconsistent across devices. This allows us to analyze whether FL can generalize well when clients only have access to a fraction of the full dataset.

2 Models and Training

The models used for training and evaluation will be **DistilBERT** and **TinyBERT**, selected for their balance between efficiency and performance in NLP tasks.

Each model will be trained under two conditions:

- **Traditional Centralized Learning (CL):** The model is trained on a centralized dataset.
- **Federated Learning (FL):** The model is trained across multiple clients, following the dataset splits described earlier. This setup allows us to compare FL's performance against a traditional centralized approach before applying any optimizations.

3 Experiment Setup

The federated learning process will be tested on varying numbers of clients to analyze how scalability impacts model performance. Three different client sizes will be used:

- **Small** (e.g., 10 clients)
- **Medium** (e.g., 50 clients)
- **Large** (e.g., 150 clients)

For each of the three dataset splits (Balanced, Real-World, Random), all three client sizes will be tested. Training will run for multiple rounds, and results will be averaged over three runs to ensure consistency.

4 Experimentation Stages

1. **Baseline Comparison:**
   - Train both FL and CL models under identical initial conditions (without hyperparameter tuning).
   - Compare FL and CL across all dataset splits and client sizes to evaluate raw performance.
   - Measure performance using accuracy, precision, recall, and F1-score.
2. **Model Selection:**
   - Based on FL and CL results, determine whether DistilBERT or TinyBERT performs better overall.
3. **Fine-Tuning Stage:**

- Fine-tune the best FL model: Optimize parameters such as learning rate, batch size, and apply different FL strategies like **FedAvg** and **FedOpt**.
- Fine-tune the best CL model: Optimize for best performance within centralized learning.

4. **Final Comparison:**
   - Compare fine-tuned FL vs. fine-tuned CL to assess whether FL can match centralized training in performance.
   - Analyze trade-offs in accuracy, computational efficiency, and scalability.

5 Evaluation Metrics

To assess model performance, the following classification metrics will be used:

- Accuracy
- Precision
- Recall
- F1-score

The evaluation will compare results across the three dataset splits and three client sizes. After averaging results across different runs, the best-performing FL model will be selected for comparison with the fine-tuned CL model.

6 Final Considerations

The random split inherently acts as an additional test case, examining how FL models handle incomplete data distribution across clients. Unlike an explicit ablation study, this setup naturally assesses whether FL can generalize well when trained on fragmented datasets.

7 Research Questions

This study aims to answer the following key research questions:

1. Can federated learning achieve results comparable to centralized learning while maintaining data security advantages?
2. What are the benefits of federated learning when applied to MBTI personality-based text classification?