Statement of Purpose (SoP)
DSL501: Machine Learning Project

Name: Rohit Raghuwanshi
Roll No.: 12341820

1. Project Details

- **Project Title:** Enhancing Generalization in Imbalanced and Low-Resource NLP Tasks using Teaching Regularization
- **Code Repo Link :https://github.com/rohitraghuwanshi07/Machine-Learning-Project**.
- **If Own Idea:** Yes.
  This project builds upon Liu et al. (2024) "Learning from Teaching Regularization" (NeurIPS 2024) but extends it to a novel setting: class-imbalanced and low-resource Natural Language Processing (NLP) tasks, which were not explicitly addressed in the original paper.

2. Problem Statement

Recent advancements in deep learning have shown the power of teacher-student paradigms and knowledge distillation. The work of Liu et al. (2024) introduced *Teaching Regularization*, a method where the teacher model is trained jointly with the student to encourage smoother, more generalizable predictions. While the original paper demonstrated strong performance gains on standard supervised datasets, a critical gap remains: it did not test the method on **imbalanced or low-resource scenarios**, which are very common in real-world NLP applications such as hate speech detection, abusive content moderation, and underrepresented language sentiment analysis.

This project aims to bridge that gap. Specifically:

- Investigate whether teaching regularization can improve model generalization in settings with limited labeled data (low-resource).
- Evaluate its effectiveness in imbalanced data distributions, where the majority class dominates (e.g., non-hate tweets in hate speech datasets).
- Compare against standard baselines (fine-tuning BERT/DistilBERT without teaching regularization).

Importance: In ML, class imbalance and data scarcity significantly reduce model robustness and fairness. Demonstrating that teaching regularization helps in these domains would open a new direction in applying this method to real-world problems.

3. Methodology

The methodology will involve the following steps:

1. **Baseline Setup:**

- Start with pre-trained models such as BERT-base and DistilBERT.
- Fine-tune on benchmark datasets (IMDB for sentiment, Davidson hate speech dataset for toxic content classification).

2. **Teacher-Student Setup with Teaching Regularization:**
   - Teacher: A fine-tuned BERT-base model.
   - Student: A smaller model such as DistilBERT or BERT-tiny.
   - Apply teaching regularization (from Liu et al., 2024), where the teacher is regularized to avoid overfitting and guide the student with softened probability distributions.
3. **Experiment Design:**
   - **Balanced Data Scenario:** Train on full IMDB dataset (50K reviews).
   - **Imbalanced Scenario:** Train on Davidson et al. (2017) hate speech dataset, which is skewed towards non-hate labels.
   - **Low-Resource Scenario:** Subsample IMDB and Davidson datasets to small sizes (1k–5k examples).
4. **Preprocessing:**
   - Tokenization using BERT tokenizer.
   - Truncation/padding to 128 tokens per input.
   - Lowercasing text, removing URLs, emojis, and non-ASCII characters.
5. **Evaluation Metrics:**
   - Accuracy and F1-score (macro) for balanced datasets.
   - Precision, Recall, and Macro-F1 for imbalanced datasets.
   - Robustness measures for low-resource settings.
6. **Tools & Frameworks:**
   - Hugging Face Transformers library.
   - PyTorch for implementation of training loops.
   - Google Colab Pro for GPU support.

4. Dataset Details

- **IMDB Movie Reviews Dataset** (50K labeled reviews, balanced sentiment classification).
  Source: `https://ai.stanford.edu/~amaas/data/sentiment/`
- **Davidson Hate Speech Dataset** (25K tweets labeled as hate, offensive, or neutral; highly imbalanced).
  Source: `https://github.com/t-davidson/hate-speech-and-offensive-language`
- **Synthetic Low-Resource Splits:** Sub-sampling IMDB and Davidson datasets to 1K–5K examples to simulate data-scarce settings.

5. Required Resources

- **Hardware:** Training will primarily use GPU resources on Google Colab Pro (Tesla T4 or P100). For small-scale experiments, MacBook Air M2 (8GB RAM, 256GB SSD) will be used locally.
- **Software:** Python 3.10, PyTorch, Hugging Face Transformers, Scikit-learn, Pandas, Matplotlib.
- **Other Tools:** GitHub for version control, Overleaf for LaTeX documentation.

6. Novelty of Approach

The novelty lies in extending teaching regularization to:

- **Imbalanced NLP tasks:** Demonstrating that teacher-student training can reduce bias towards majority classes.
- **Low-resource NLP tasks:** Showing improved generalization when very few examples are available, which is crucial for underrepresented languages or domains.
- **Comparative Analysis:** Providing detailed empirical results against standard baselines like BERT fine-tuning and knowledge distillation.

7. Individual Contribution

- **Name:** Rohit Raghuwanshi
- **Roll No.:** 12341820
- **Contribution:** Responsible for the complete project lifecycle including literature review, dataset preprocessing, teacher-student setup implementation, experiments, evaluation, and final documentation. This is an individual project.

8. Expected Outcomes

- **Quantitative:** Demonstrate that teaching regularization improves Macro-F1 and Recall in imbalanced datasets, and prevents performance degradation in low-resource splits.
- **Qualitative:** Provide insights into why teaching regularization helps in scarce/imbalanced data situations.
- **Deliverables:**
  - Final trained teacher-student models.
  - Detailed comparative analysis with baselines.
  - Well-documented project report and GitHub repository.

9. References

- Liu et al., 2024. Learning from Teaching Regularization. *NeurIPS 2024*. https://papers.nips.cc/paper_files/paper/2024/file/01ce1ae7f94d139e4917f9e4425a4f38-Paper-Conference.pdf
- Hinton et al., 2015. Distilling the Knowledge in a Neural Network. *NeurIPS*. https://arxiv.org/abs/1503.02531
- Xu et al., 2023. Addressing Class Imbalance in NLP: A Survey. *ACL Anthology*. https://aclanthology.org/2023.acl-long.120/
- Davidson et al., 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *ICWSM*. https://ojs.aaai.org/index.php/ICWSM/article/view/14955