Roadmap for BLP Subtask-1A: Bangla Hate & Offensive Speech Detection

# Objective

Build a leaderboard-competitive and research-ready system that neutralizes severe class imbalance, leverages Bangla-specialized and multilingual transformers, employs a two-step hierarchical setup, adds explainability for analysis, and follows a clean modular pipeline (data → models → post-processing).

# PHASE 1: Data Collection and Preprocessing

This phase focuses on preparing raw social-media-style Bangla text into a form usable by transformer models.

1. Normalize Unicode, punctuation, Bangla digits, and remove control characters.  
2. Preserve code-mix: keep English tokens; replace URLs/handles with <URL>, <USER>.  
3. Emoji→text mapping (বাংলা descriptors, e.g., 🙂→“হাসি”) to preserve sentiment cues.  
4. Minimal cleaning (avoid over-stemming/stopwording to keep semantic richness for transformers).  
5. Stratified splits to maintain class ratios across train/dev/test.  
6. Label hygiene and duplicate removal.

Why it helps: Ensures clean yet context-rich input, handles noisy social media data, and sets up for fair evaluation.

# PHASE 2: Model Baselines and Fine-Tuning

Multiple transformer models will be fine-tuned, no traditional models are used.

1. XLM-RoBERTa-base  
 - MaxLen=256–320, Batch=32, LR=2e-5, Epochs=3–5, Warmup=6%, WD=0.01, Dropout=0.1.  
 - Weighted CE for Step-1, Focal Loss for Step-2.  
  
2. BanglaBERT  
 - MaxLen=256–320, Batch=32, LR=2e-5, Epochs=3–5.  
 - Strong on monolingual Bangla.  
  
3. MuRIL  
 - MaxLen=256, Batch=32, LR=2e-5, Epochs=4.  
 - Effective for Indic code-mixed text.  
  
4. BanglaHateBERT  
 - MaxLen=256, Batch=32, LR=2e-5, Epochs=4–6.  
 - Specialized abusive corpora, improves minority-class detection.  
  
5. Distil-mBERT  
 - MaxLen=256, Batch=64, LR=3e-5, Epochs=5–8.  
 - Lightweight, for rapid ablations.

Why it helps: Transformer models provide contextual understanding and transfer learning from large corpora.

# PHASE 3: Imbalance Handling

1. Two-Step Hierarchy:  
 - Step-1: None vs Offensive (merge abusive, sexism, religious, political, profane).  
 - Step-2: Classify offensive into Abusive, Sexism, Religious, Political, Profane.  
2. Loss Functions:  
 - Step-1: Weighted Cross-Entropy.  
 - Step-2: Focal Loss (γ=2.0) with class-balanced α.  
3. Data Augmentation:  
 - Back-translation (BN→HI/UR→BN), synonym replacement, light EDA.  
4. Sampling:  
 - Oversample Sexism & Religious classes, undersample None moderately.  
5. Threshold Tuning & Calibration:  
 - Per-class decision thresholds optimized on dev set.

Why it helps: Ensures minority classes like Sexism and Religious Hate are fairly represented and learned.

# PHASE 4: Modifications & Novel Ideas

1. Curriculum by length: Train on shorter texts first, gradually add longer ones.  
2. Aspect-auxiliary head: Predict aspects (religion, politics, gender) jointly with class.  
3. Confidence-based self-training: Use pseudo-labeled high-confidence data to expand minority samples.  
4. Mixture-of-Encoders (MoE): Fuse XLM-R + MuRIL + BanglaHateBERT via a small gating MLP.  
5. Explainability: Token-level saliency/LRP visualizations to highlight why predictions occur.

Why it helps: Moves beyond standard fine-tuning by addressing dataset weaknesses and providing interpretable insights.

# PHASE 5: Evaluation

Metrics:  
- Primary: Macro-F1.  
- Secondary: Per-class F1, Confusion matrix, Calibration curves.  
  
Error Analysis:  
- Length buckets (e.g., 0–10, 50–100, 500+ tokens).  
- Two-step propagation errors.  
- Ablations: +Two-step, +Augmentation, +Curriculum, +MoE, +Calibration.  
- Human-interpretability: Evaluate faithfulness of explanations.

Why it helps: Macro-F1 ensures fairness across imbalanced classes. Error analysis highlights weaknesses.