Phase 1: Foundation Setup — Model and Data

Objective: Establish a working BERT-based sentiment classifier for Hindi text.

What to Do:

1. **Choose a pre-trained model** that supports Hindi (IndicBERT, Multilingual BERT, or HindiBERT).
2. **Collect or prepare a dataset** containing Hindi sentences labeled with sentiments — positive, negative, and neutral.
3. **Preprocess the data**:
    - Clean text (remove emojis, special characters if needed).
    - Tokenize Hindi text using the model's tokenizer.
4. **Build your sentiment classifier:**
    - Either fine-tune BERT directly for classification, or
    - Use BERT just to generate embeddings and train a separate classifier (like a small neural network).
5. **Train and evaluate** your model to make sure it reaches acceptable accuracy before adding explainability.

Outcome: You now have a trained sentiment model that takes Hindi text as input and outputs the predicted sentiment.

Phase 2: Visual Explainability — Attention Analysis

Objective: Understand what parts of the Hindi sentence BERT focuses on while making predictions.

What to Do:

1. **Enable attention outputs** in the model to access its attention maps.
2. **Visualize attention patterns** for a few test sentences:
    - See which words (tokens) BERT pays most attention to.
    - Identify how attention changes across layers (lower = structure, higher = meaning).
3. **Interpret the patterns**:
    - For positive sentences, BERT might focus on adjectives like "□□□□□" or "□□□□□□".
    - For negative ones, it might highlight "□□□□", "□□□□", etc.

Outcome: You gain an intuitive understanding of how the model internally relates Hindi words to each other. This is your first explainability layer.

Phase 3: Local Explainability — Per Prediction Insights

Objective: Explain why the model made a specific prediction for a given Hindi sentence.

What to Do:

1. **Pick representative samples** — one positive, one neutral, one negative.
2. **Apply a word-importance method** such as:
    - **SHAP** (most consistent): to calculate how much each word contributed to the final prediction.
    - **LIME** (simpler): to approximate how removing or changing a word affects the outcome.
3. **Interpret and record the explanations:**
    - Highlight which words increased or decreased the sentiment score.
    - Note patterns (e.g., intensifiers like "□□□□" make positive words stronger).

Outcome: You can now point to each word in a sentence and explain how it influenced the model's decision.

□ Phase 4: Gradient-based Explainability — Understanding Neural Influence

Objective: Explore how deeply the model's internal activations depend on specific input tokens.

What to Do:

1. Use gradient-based techniques (like Integrated Gradients conceptually).
2. Compare results with your SHAP or LIME outputs to see if they align.
3. Identify words that consistently have strong influence across multiple examples.

Outcome: You now understand not just which words mattered, but *how strongly* they affected the neural computations inside BERT.

□ Phase 5: Counterfactual & Bias Analysis

Objective: Test how sensitive or biased the model is toward certain Hindi words or structures.

What to Do:

1. **Create counterfactual sentences** by making small, meaningful changes:
    - Add or remove negation words like "□□□□".
    - Replace adjectives with their opposites ("□□□□□" → "□□□□").
2. **Observe how the prediction changes**:
    - If the model correctly flips sentiment, it's behaving logically.
    - If not, you may have discovered a bias or a gap in understanding.

3. **Record these insights** for model improvement and fairness reporting.

Outcome: You'll know whether your model is fair, logical, and consistent — or whether it over-relies on specific words or patterns.

Phase 6: Visualization and Presentation

Objective: Build a way to present your XAI findings clearly and interactively.

What to Do:

1. Create a simple interface (web or notebook-based) that:
    - Accepts a Hindi sentence.
    - Displays the model's sentiment prediction.
    - Highlights important words in colors (e.g., red for negative, green for positive).
2. Include attention visualizations and SHAP explanations side by side.
3. Use clear legends, color gradients, and Hindi text-friendly fonts.

Outcome: A viewer can visually understand *what the model predicted* and *why it did so* — in an intuitive, human-friendly way.

Phase 7: Documentation and Reporting

Objective: Summarize your process, findings, and insights for presentation or submission.

What to Include:

1. **Introduction:** Why explainability matters in sentiment analysis.
2. **Model Overview:** Your Hindi BERT setup and dataset.
3. **Explainability Layers:**
    - Attention visualization
    - Local word importance (SHAP/LIME)
    - Gradient-based reasoning
    - Counterfactual testing
4. **Key Insights:**
    - Which Hindi words or structures most affect sentiment predictions.
    - How well BERT's attention aligns with human reasoning.
    - Any discovered biases or weaknesses.
5. **Conclusion:**
    - Discuss how explainability improved your understanding of the model.
    - Suggest future improvements (e.g., data balancing, multilingual fine-tuning).

Outcome: A polished report or presentation showing not just a working AI system — but an *explainable* and *trustworthy* one.

Suggested Timeline

Phase	Description	Time Estimate
1	Build & train BERT-based Hindi sentiment model	2 days
2	Visualize BERT’s attention	1 day
3	Add SHAP/LIME explanations	1 day
4	Explore gradient-based interpretability	0.5 day
5	Perform counterfactual/bias analysis	0.5 day
6	Create visualization interface	1 day
7	Documentation and final polish	1 day

Total: Around 6–7 days for a complete, presentation-ready explainable sentiment model.

Final Deliverable

By the end, your project will include:

- A **Hindi sentiment classifier** built on BERT.
- Multi-layer explainability:
  - **Attention visualization** (what BERT focused on)
  - **Word importance heatmaps** (why it predicted what it did)
  - **Gradient-based influence** (how deep learning layers respond)
  - **Counterfactual tests** (logical and fairness verification)
- A clean, interactive **visual interface** showing explanations.
- A concise, well-written **report** summarizing insights.
