Final Project Outline: YouTube Engagement Prediction

Project Title

"What Drives Engagement in Climate Change Videos? A Machine Learning Analysis of YouTube Video Performance"

1. Data Collection & Preparation

- Use YouTube Data API v3 with keywords like 'climate change', 'global warming', etc.
- Fields: video_id, title, description, tags, viewCount, likeCount, commentCount, publish date, channel info, duration, category, captions, thumbnail presence
- Handle missing fields, quota management, nested JSON structure

2. Exploratory Data Analysis (EDA)

- Cluster videos based on content topics using TF-IDF + KMeans or HDBSCAN
- Use t-SNE or UMAP for dimensionality reduction and visualizations
- Explore outliers, patterns in engagement, and temporal patterns

3. Feature Engineering & Preprocessing

- Textual: Title sentiment (VADER), TF-IDF features, clickbait detection, use of numerics, punctuation
- Temporal: Hour of post, weekday, month, weekend/holiday flags
- Engagement Ratios: likes/view, comments/view, likes/subscriber
- Channel and content flags, normalized and one-hot encoded
- Dimensionality reduction if needed

4. Model Development

- Classification (high vs. low engagement) or Regression (predict views/likes)
- Models: Logistic/Linear Regression, Random Forest, XGBoost, Neural Net, optional BERT+tabular fusion
- Metrics: F1, AUC, MAE, RMSE, R^2
- Use stratified cross-validation, SHAP for interpretability

5. Model Comparison & Selection

Final Project Outline: YouTube Engagement Prediction

- Compare using accuracy and interpretability
- Validate with curves and SHAP analysis
- Retrain final model and justify selection

6. Communication & Reporting

- Deliverables: Clear narrative, strong visuals (SHAP, clusters, engagement plots)
- Final report covers end-to-end workflow
- Optional dashboard: Input video features -> predict engagement

7. Creativity & Extensions

- Compare climate videos vs. other types
- Analyze captions vs. engagement (accessibility)
- Cluster content types: doom, activism, science, lifestyle
- Time-based forecasting of future engagement