COMP7015 Artificial Intelligence – Group Project Rubrics

Overview

The assessment of the course project is based on the implementation, the final report, and the in-person presentation that each group submits. Each group member will be given an individual score based on his/her contribution to the project and the quality of their presentation.

- **Implementation:** 40%
- **Final Report:** 30%
- **Presentation:** 30%

1. Implementation Rubrics (40%)

(A) Topic 1: Human Action Recognition (HMDB51)

Criterion (Max)	Excellent (Advanced)	Satisfactory (Minimum)	Unsatisfactory
Data prep & preprocessing (12)	Robust, reproducible splits; thoughtful class selection/balance; resilient 3–4 frame extraction across fps/length; efficient caching/IO; well-justified augmentations and normalization. (9–12)	Correct train/val/test split; extract 3–4 frames; composite into one image; basic normalization/augmentation; functional dataloader. (5–8)	Split/leakage issues; extraction/composition incorrect or missing; pipeline not functional. (0–4)
Modeling: baseline + transfer (12)	Well-implemented custom 2D CNN and strong transfer model with sound fine-tuning (freeze/unfreeze, schedulers); plus	Working custom 2D CNN and correctly adapted pretrained model (e.g., ResNet18/34) with proper head; trains to reasonable performance. (5–8)	One or both models missing or incorrect; training unstable/incorrect. (0–4)

	temporal/3D modeling or other substantive improvements. (9–12)		
Experimentation & evaluation (12)	Systematic tuning and ablations; per-class metrics, F1, confusion matrix; error analysis; trade-offs vs compute; optionally more classes or UCF101 with justified protocol. (9–12)	Basic tuning of LR/batch/regularization; reports accuracy and at least one additional metric; confusion matrix; clear test protocol. (5–8)	Little/no tuning; incorrect metrics/protocol; no meaningful analysis. (0–4)
Code quality & lab readiness (4)	Modular, documented, config-driven; deterministic; efficient; clear README; proper acknowledgements; runs in FSC 8/F. (3–4)	Runnable in lab; README with commands; seeds set; paths configurable. (2)	Not runnable, unclear, or missing documentation. (0–1)

(B) Topic 2: Sentiment Analysis (IMDb)

Criterion (Max)	**Excellent (Advanced)**	**Satisfactory (Minimum)**	**Unsatisfactory**
Data pipeline & vectorization (12)	Strong pipeline with subword tokenization/bucketing, caching/prefetch; sequence length and OOV analysis; robust loaders. (9–12)	Proper split with seeds; clean text; tokenize; vocab from training only; OOV handling; pad/truncate; efficient loaders. (5–8)	Split issues/leakage; weak/incorrect vocab or padding; pipeline not functional. (0–4)
Modeling & RNN embeddings (12)	BiLSTM/attention or hybrid models; rigorous pretrained LM fine-tuning (e.g., BERT) compared to	LSTM/GRU with trainable embedding; pretrained embeddings (GloVe/Word2Vec) used correctly; frozen	One or both embedding setups missing/incorrect; unstable/incorrect training. (0–4)

	RNN baselines; clear training strategies. (9–12)	vs fine-tuned compared fairly. (5–8)	
Experimentation & evaluation (12)	Systematic search/ablations; accuracy, precision, recall, F1; calibration/thresholds; error buckets; robustness (e.g., negation/sarcasm); optional fine-grained ratings. (9–12)	Tune LR, batch, dropout, units/layers; report accuracy plus precision/recall/F1; learning curves; compare trainable vs pretrained. (5–8)	Minimal tuning; metrics misused; no fair comparison or analysis. (0–4)
Code quality & lab readiness (4)	Modular, documented, config-driven; deterministic; clear README; acknowledgements; runs in FSC 8/F. (3–4)	Runnable in lab; README with commands; seeds set; paths configurable. (2)	Not runnable, unclear, or missing documentation. (0–1)

(C) Topic 3: Open Topic

Criterion (Max)	**Excellent (Advanced)**	**Satisfactory (Minimum)**	**Unsatisfactory**
Problem formulation & algorithm choice (8)	Well-justified, ambitious, course-aligned choices; discusses alternatives and trade-offs. (6–8)	Clear problem, dataset, and reasonable baseline algorithm fit. (4–5)	Poor fit or unjustified choices; unclear problem. (0–3)
Implementation completeness & engineering (14)	Sophisticated methods and solid engineering (schedulers, checkpoints, mixed precision as relevant); stable and efficient training. (11–14)	End-to-end baseline pipeline; produces results; sensible regularization/training loop. (7–10)	Partial/buggy pipeline; missing key components or results. (0–6)
Experimentation & evaluation (14)	Strong baselines; rigorous ablations; proper metrics/protocol; robustness/efficiency analyses; limitations/trade-offs. (11–14)	Basic correct protocol and metrics; baseline comparison; clear test split. (7–10)	Weak/incorrect evaluation; no meaningful analysis. (0–6)
Code quality & lab readiness (4)	Modular, documented, config-driven; deterministic; clear README; runs in FSC 8/F. (3–4)	Runnable in lab; README with commands; seeds set; paths configurable. (2)	Not runnable, unclear, or missing documentation. (0–1)

2. Report and In-person Presentation Rubric (60%)

Criterion (Max)	**Excellent (Advanced)**	**Satisfactory (Minimum)**	**Unsatisfactory**
Organization & storytelling (12)	Strong narrative arc; logical flow; high-impact visuals; within 5-page limit and ~8 minutes; smooth transitions among members. (9–12)	Clear structure; visuals present and legible; within limits; each member presents. (5–8)	Disorganized; poor visuals or overrun; missing member participation. (0–4)
Technical correctness & rationale (14)	Accurate explanations; well-argued design choices; trade-offs and constraints tied to course concepts; limitations stated. (11–14)	Methods correctly described; basic rationale; minimal errors. (7–10)	Major inaccuracies; weak or missing rationale. (0–6)
Results, visuals & insight (24)	Comprehensive metrics; clear tables/figures; ablations/robustness; error analysis; per-class or task-suitable analyses; evidence-backed conclusions with compute-performance discussion. (18–24)	Proper metrics reported; readable visuals; required comparisons (e.g., baseline vs transfer; trainable vs pretrained) and basic error analysis. (12–17)	Missing/incorrect metrics; unclear visuals; no meaningful comparisons/insights. (0–11)
Contributions, delivery & Q&A (10)	Specific, balanced member contributions in report; confident delivery; accurate, concise Q&A across the team. (8–10)	Contributions listed; all members present; basic Q&A handled. (5–7)	Contributions missing; uneven participation; poor Q&A. (0–4)