COMP3931 Individual Project — Assessor Progress Meeting Pack (Week 4)
Student: Nicolas Issa 201707349
Project Title: Fake News Detection on the LIAR Dataset (Research-led Investigation)
Supervisor: Dr Mohammad Ammar Alsalka
Date: ?

1) Aim and Research Framing (30 seconds)

This project investigates automated fake news / claim veracity classification using the LIAR dataset. The work is framed as a **research-led investigation**: establish strong baselines, introduce **one change at a time** with justification, and evaluate using consistent metrics and error analysis. The primary goal is to understand **what works, what fails, and why**, rather than building a large system.

2) Dataset + Evaluation Setup (30 seconds)

- **Dataset:** LIAR (Wang, 2017) with fixed train/valid/test split
    - Train: 10,240 | Valid: 1,284 | Test: 1,267
- **Task:** 6-class veracity prediction: pants-fire, false, barely-true, half-true, mostly-true, true
- **Primary metric: Macro-F1** (class imbalance + fair across labels)
- Outputs saved for reproducibility: metrics JSON, prediction CSVs, confusion matrices.

3) Baseline Results (Classical ML)

Model: TF-IDF + Logistic Regression

- **Unigram baseline (Valid):** Accuracy **0.215**, Macro-F1 **0.196**
- **Best classical variant (class_weight="balanced", Valid):** Accuracy **0.223**, Macro-F1 **0.224**
  Key observation: most errors are **adjacent-label confusions** (e.g., barely-true ↔ half-true), with “half-true” acting as a sink class.

4) Transformer Results (DistilBERT)

Model: DistilBERT (distilbert-base-uncased), 1 epoch, LR=2e-5, batch size 8 (eval 16)

- **Smoke test (max_steps=300):**

    - Valid: Acc **0.235**, Macro-F1 **0.136**
    - Test: Acc **0.233**, Macro-F1 **0.134**
      (pipeline validation only)
- **Full fine-tuning (max_length=64):**
    - Valid: Acc **0.245**, Macro-F1 **0.187**
    - Test: Acc **0.281**, Macro-F1 **0.223**
- **Controlled ablation (Variant A: max_length=128):**
    - Valid: Acc **0.261**, Macro-F1 **0.197**
    - Test: Acc **0.275**, Macro-F1 **0.211**
      Interpretation: longer context gives a small validation lift but **does not improve test generalisation** under the current setup.

5) Error Analysis (What fails)

Across baseline + transformer:

- Errors are dominated by **neighbouring-label drift** (false ↔ half-true ↔ mostly-true).
- The rare/extreme label **pants-fire** is consistently difficult.
- Overall, the main limitation appears to be **fine-grained label boundary ambiguity** using statement text alone.

6) Key Visuals

1. **Label distribution** (class imbalance motivation)

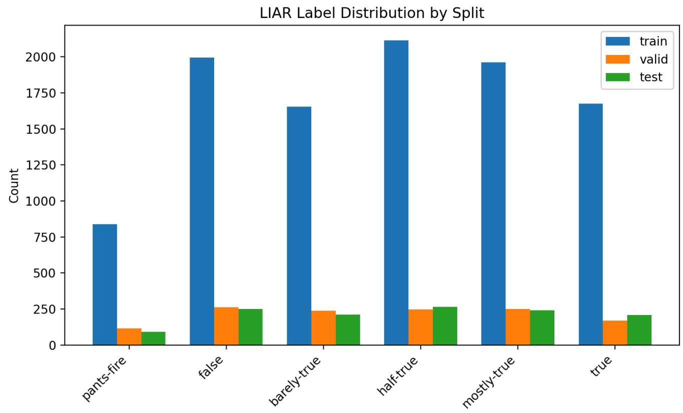

Figure: Label distribution across splits : Shows strong class imbalance across the six labels, motivating macro-F1 as the primary metric so minority classes are not ignored.

2. **Baseline confusion matrix** (adjacent-label confusions)

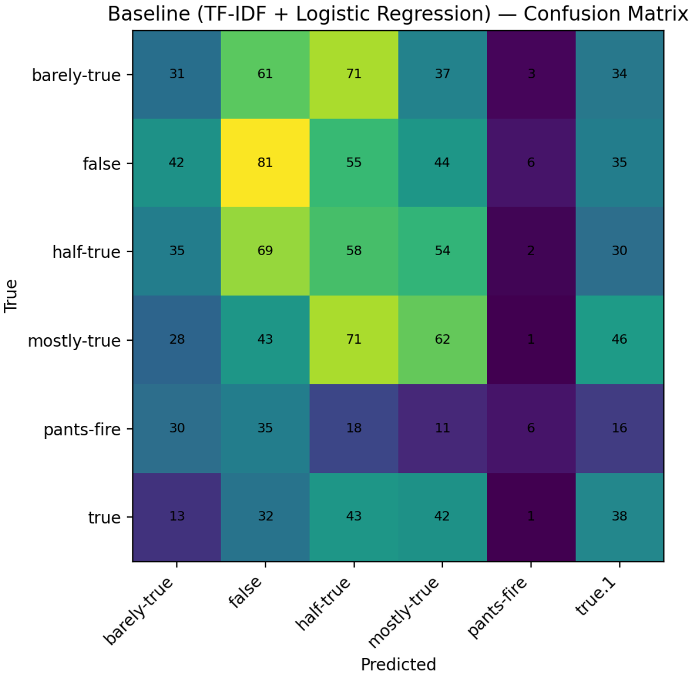

Figure : Baseline confusion matrix (validation) : Errors are concentrated between adjacent labels (e.g. barely-true/half-true/mostly-true), indicating bag-of-words features struggle with fine-grained veracity boundaries.

3. **Transformer full confusion matrix** (main benchmark)

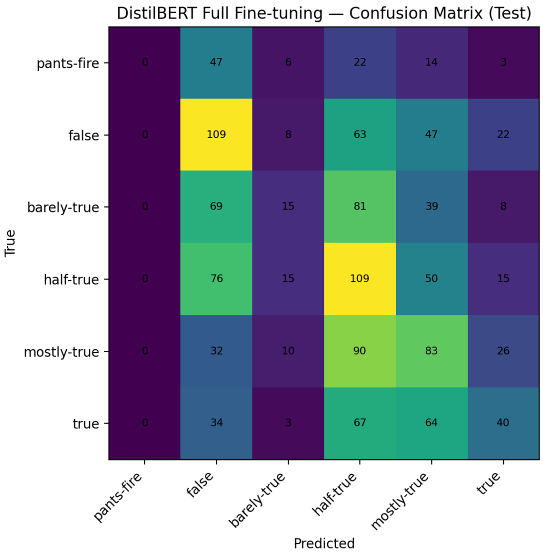

Figure : DistilBERT full confusion matrix test (test) : Compared with the baseline, predictions are more spread across classes and test performance improves, but adjacent-label confusion remains a dominant failure mode.

7) What I will do next (Week 5+ plan)

- Extend transformer evaluation with deeper, structured error analysis (per-class metrics + representative errors).

- If time permits: one additional controlled ablation (e.g., small LR change) or limited metadata integration, evaluated consistently.

- Finalise report Chapters 3–5 with clean presentation, consistent figure/table references, and reproducibility details.

My questions

- Are my evaluation choices (macro-F1 primary, fixed split) appropriate for LIAR's imbalance?
    - His answer:

- Is the scope/novelty sufficient for a research-led investigation (baselines → controlled improvements → analysis)?
    - His answer:

- For Chapter 3, how much implementation detail do you expect (pipeline + reproducibility vs deep engineering)?
    - His answer:

- Would you recommend one more controlled ablation (e.g., LR) or shifting effort to deeper error analysis + discussion?
    - His answer:

- Is it worth including limited LIAR metadata as a small extension, or should I keep it text-only and focus on interpretation?
    - His answer :

Notes: