Membership Inference Attack on a DistilBERT AG News Model Using Shadow Models and Supervised Attack Networks

Apu Kumar Chakroborti S. Vignesh Kumar Pandian

Abstract

This report documents the implementation and evaluation of a shadow-model–based membership inference attack (MIA) against a pretrained DistilBERT classifier fine-tuned on AG News. The attack pipeline follows the classical approach: constructing shadow datasets, training shadow models, extracting per-example features from model outputs, building a labeled attack dataset, and training a small neural attack classifier. The system outputs membership predictions for evaluation samples, and validation metrics are reported.

1 Problem Setup

Victim model: A fine-tuned DistilBERT classifier stored in `victim_model_distilbert_agnews`. **Shadow pool:** `sampled.csv`. **Validation input:** `validation_samples.csv`. **Validation ground-truth:** `validation_results.txt`. **Required output:** `mia_lm_results.txt`, containing:

$$\text{id} \quad \text{membership\_prediction} \in \{0, 1\}.$$

2 Attack Method: Shadow-Model Supervised MIA

The implemented pipeline uses only the supervised shadow-model method.

2.1 Shadow Dataset Construction

For each shadow model:

- Randomly split `sampled.csv` into a shadow *train set* (treated as members) and a shadow *out set* (treated as non-members).
- Multiple shadows (e.g., 5) are used to generate a more robust attack dataset.

2.2 Shadow Model Training

Each shadow model is:

- initialized from the same DistilBERT checkpoint as the victim,
- trained for a small number of epochs (10),
- optimized using AdamW,
- trained with a special *NoisyLabelCrossEntropy* loss.

2.3 Feature Extraction

For each sample, the code extracts the following features from the model forward pass:

- cross-entropy loss (per example),
- prediction confidence (max softmax probability),
- entropy of output distribution,
- correctness indicator,
- logit margin (top1–top2),
- optional embedding/logit norms.

These are concatenated into a feature vector $x \in \mathbb{R}^d$.

2.4 Attack Dataset Construction

For each shadow:

$$X_{\text{members}} \leftarrow \text{features from shadow-train},$$
$$X_{\text{nonmembers}} \leftarrow \text{features from shadow-out}.$$

Stacking across shadows yields a supervised binary classification dataset:

$$(X, y), \quad y \in \{0, 1\}.$$

2.5 Attack Model Training

The attack model is a small MLP:

$$\text{AttackMLP}: \quad 64 \to 32 \to 16 \to 1$$

trained using:

- Adam optimizer for a long schedule (500 epochs),
- followed by LBFGS full-batch refinement (10 steps).

2.6 Final Membership Prediction

For each evaluation sample:

$$x = \text{extract\_features(victim, sample)}$$

$$p = \sigma(f_{\text{attack}}(x))$$

$$\text{predict member} = \begin{cases} 1, & p \geq 0.5 \\ 0, & p < 0.5. \end{cases}$$

3 Implementation Summary

The following key components are used (and fully implemented in code):

- `AttackMLP`
- `NoisyLabelCrossEntropy`
- `extract_features`
- `train_shadow_model`
- `build_attack_dataset`
- `train_attack_model`
- `attack_victim`

4 Validation Results

The code reports the following validation metrics (based on `validation_samples.csv`):

Metric	Value
ROC AUC	0.5212
Accuracy	0.381
Precision	0.1984
Recall	0.734375
F1	0.3129
Confusion matrix	$\begin{bmatrix} 240 & 568 \\ 51 & 141 \end{bmatrix}$

Shadow-model MIA results.

Observations:

- ROC AUC slightly above 0.5 $\rightarrow$ weak but nonzero MIA signal.
- High recall but low precision $\rightarrow$ the model predicts many positives.
- Accuracy modest $\rightarrow$ shadow models may not mimic victim well.

Conclusion

This report describes the implemented membership inference attack, which is exclusively a *shadow-model supervised attack*. The current results show limited but present membership leakage, and the system is fully extensible for improvement.