Empirical Evaluation of Membership Inference Attacks on NLP Text Classifiers: A Baseline Study on SST-2

William Novak and Muhammad Abusaqer

Trustworthy Language Intelligence Lab (TLI Lab), Department of Math, Data & Technology, Minot State University, Minot, ND
muhammad.abusaqer@minotstateu.edu

Why membership inference matters

A membership inference attack (MIA) tries to answer one question: was this record used to train the model? A correct answer can reveal that a person's text was part of a sensitive dataset, and black-box access to prediction outputs is often enough to mount the attack [1].

Text classifiers are trained on user-generated content and served through prediction APIs, and NLP classifiers are known MIA targets [3]. Yet most text-classification studies, including comparative benchmarks of traditional and transformer models, report accuracy and rarely measure privacy [7]. Leakage is linked to overfitting: a model that fits its training data too tightly assigns members systematically lower loss than non-members [2].

This study. We benchmark a TF-IDF + Logistic Regression pipeline against a fine-tuned DistilBERT classifier on SST-2 under one attack and one evaluation protocol, then test a single lightweight mitigation per model.

RQ1. How much membership information can be inferred from a traditional text classifier compared with a compact transformer on SST-2?

RQ2. How closely is privacy leakage associated with overfitting-related behavior in this setting?

RQ3. To what extent can simple mitigation choices reduce leakage while preserving utility?

Dataset: SST-2

What it is. The Stanford Sentiment Treebank, binary version [5]: sentences and phrases from movie reviews, each labeled positive or negative. The GLUE release [4] has 67,349 training and 872 validation items (55.8 and 50.9 percent positive, respectively).

Why it fits this study. A standardized benchmark, so results are comparable to prior work; an official, clean validation split that serves as the non-member pool; and small enough to train both pipelines under one identical protocol.

Examples. “enigma is well-made” (positive); “contains no wit , only labored gags” (negative); “the clumsy cliché of the ugly american” (negative).

Threat model and attack

Attack set. 872 members from target-train and 872 non-members from the GLUE validation split (balanced).

↓

Query the target model (black-box). The attacker sees only $p_\theta(y \mid x)$, the predicted probability of the true label y for sentence x; no parameters, no training data [1], [2].

↓

Per-example loss. $\text{loss}(x, y) = -\log p_\theta(y \mid x)$

↓

Attack score. score = −loss. A higher score means the example is more likely a training member.

↓

Decision and metrics. Predict member if score ≥ τ. Report Attack AUC over all τ and Attack Advantage = $\max_\tau [\text{TPR}(\tau) - \text{FPR}(\tau)]$.

Experimental setup

Dataset	SST-2 from GLUE [4], [5] (see Dataset block)
Splits	Official training split partitioned by stratified sampling (seed 42) into target-train (80%, 53,879 sentences) and target-dev (20%, 13,470). The official validation split (872 sentences) is held out as the clean non-member pool.
TF-IDF + LR	Lowercase, English stopwords removed, unigrams and bigrams, min_df = 2, max_features = 20,000; liblinear solver, max_iter = 1000; baseline C = 1.0
DistilBERT	distilbert-base-uncased [6]; max length 128 tokens, batch size 16, learning rate 2×10^{-5}, weight decay 0.01; baseline 3 fine-tuning epochs
Mitigations	LR: stronger regularization, C = 0.3. DistilBERT: fewer fine-tuning epochs, 2 instead of 3
Metrics	Utility: development accuracy, macro F1-score, and mean per-example loss on the target-train, target-dev, and validation splits. Privacy: Attack AUC and Attack Advantage (best TPR minus FPR over all thresholds)
Tools	Python, scikit-learn, Hugging Face Transformers, PyTorch; Jupyter workflow with saved probabilities, per-example losses, and attack summaries

Results

All four runs (baseline = Table 1; mitigation runs from Section 4.2 of the paper). Losses are mean per-example losses.

Metric	TF-IDF + LR C = 1.0 baseline	TF-IDF + LR C = 0.3	DistilBERT 3 epochs baseline	DistilBERT 2 epochs
Dev accuracy	0.8756	0.8569	0.9466	0.9465
Macro F1	0.8727	0.8525	0.9460	0.9458
Train loss	0.3105	0.4151	0.0510	0.0771
Dev loss	0.3501	0.4358	0.2341	0.2005
Validation loss	0.4553	0.4894	0.4471	0.3777
Attack AUC	0.5615	0.5385	0.5800	**0.5654**
Attack Advantage	0.1697	0.1330	0.1365	**0.1170**

Change from baseline to mitigation within each model family (Table 2 of the paper).

Model family	Δ Dev acc.	Δ Macro F1	Δ Attack AUC	Δ Attack Adv.
Logistic Regression (C = 1.0 to 0.3)	−0.0187	−0.0202	−0.0230	−0.0367
DistilBERT (3 to 2 epochs)	−0.0001	−0.0002	−0.0145	−0.0195

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Security and Privacy (SP), 2017, pp. 3 to 18.

[2] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in Proc. IEEE 31st Computer Security Foundations Symp. (CSF), 2018, pp. 268 to 282.

[3] V. Shejwalkar, H. A. Inan, A. Houmansadr, and R. Sim, "Membership inference attacks against NLP classification models," in NeurIPS 2021 Workshop on Privacy in Machine Learning, 2021.

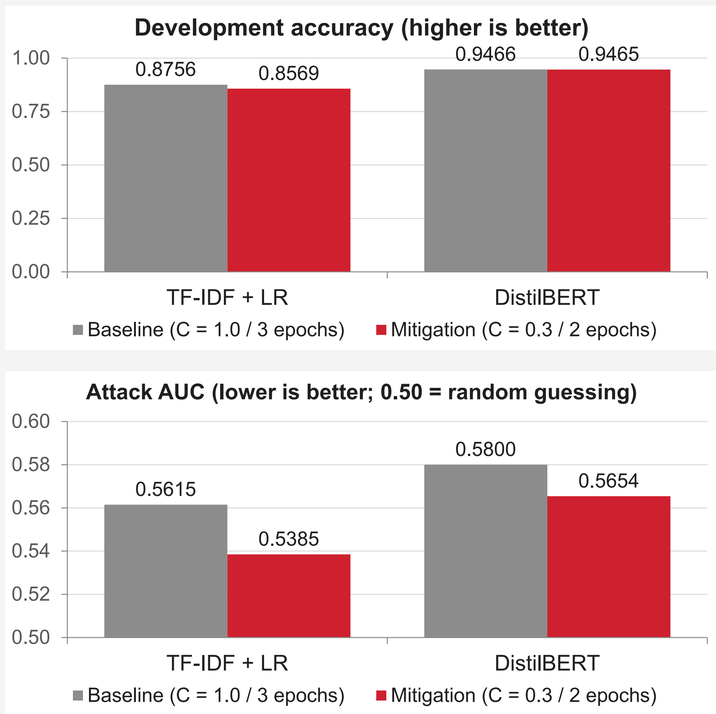
[4] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A multi-task benchmark and analysis platform for natural language understanding," in Proc. ICLR, 2019.

[5] R. Socher et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in Proc. EMNLP, 2013, pp. 1631 to 1642.

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," arXiv:1910.01108, 2019.

[7] M. Abusaqer and J. Saquer, "A comparative analysis of transformer and traditional ML models for cyberbullying detection on Twitter (now X)," in Proc. IEEE 49th COMPSAC, 2025, pp. 1607 to 1612.

Baseline vs. mitigation



Key findings

Higher utility did not buy privacy. DistilBERT was far more accurate than the baseline (0.9466 vs. 0.8756) and still leaked slightly more on average (Attack AUC 0.5800 vs. 0.5615). Logistic Regression had the larger Attack Advantage (0.1697 vs. 0.1365), so the two privacy metrics rank the models differently. Do not judge risk from one metric.

Leakage was measurable but moderate. Every Attack AUC stayed in the 0.54 to 0.58 range: above chance, far from reliable extraction of membership.

Leakage tracked the generalization gap. Members received lower mean loss than non-members in both families (LR 0.3105 vs. 0.4553; DistilBERT 0.0510 vs. 0.4471), the exact signal the attack exploits.

Cheap mitigations worked, at model-dependent cost. C = 0.3 cut LR's Attack AUC by 0.0230 but cost 0.0187 dev accuracy. Two epochs cut DistilBERT's Attack AUC by 0.0145 with accuracy essentially unchanged (0.9466 to 0.9465) and lower dev and validation losses, consistent with less overfitting: the best privacy-utility trade-off of the four runs.

Scope and next steps. One dataset, one attack family, two mitigation settings: a compact, reproducible baseline rather than a general claim. Next: more datasets, shadow-model and confidence-based attacks, stronger weight decay, calibration, differential privacy training, and cybersecurity-oriented text tasks where leakage matters operationally.

Trustworthy Language Intelligence Lab

Minot State University. Lab page: msaqer.github.io/lab.html

Reproducible Jupyter notebooks (seed 42) are available from the authors.

Presented at the 58th Midwest Instruction and Computing Symposium (MICS 2026), Eau Claire, WI, March 27 to 28, 2026.

BE SEEN. BE HEARD. BE YOU.