Nordic AI Meet 2026

P3: LLMs, Language, Evaluation, Privacy & Trustworthy AI

9. Learning to Anonymise Without Seeing: Weakly Supervised PII Detection in Swedish Court Rulings

Authors: Aleksi Andreev1; Nils Alenäs1; Victor Lagerbring1, Mattias Tiger1
Affiliations: AI Academy, Division of Artificial Intelligence and Integrated Computer Systems (AIICS), Department of Computer and Information Science (IDA), Linköping University, Sweden.

Abstract: Public authorities must release the documents citizens request while protecting the individuals named in them, and the two duties pull against each other. Transparency and legal certainty argue for disclosing as much as possible, while data protection law requires exposing no more personal information than necessary, so redaction must remove identifiers without stripping the substance that made a document worth releasing. Manual redaction does not scale. However, automated de-identification models are conventionally trained on corpora containing the personal data they are built to remove, and such corpora cannot leave the institution holding them, which limits how widely the models can be shared.

We study Swedish Supreme Court rulings, a demanding instance because the material is both unusually sensitive and unusually important to publish in full. Swedish courts issue each ruling twice, as an unmasked original and as an officially redacted version. Aligning the two recovers the strings a court clerk judged identifying, supplying supervision at no annotation cost. We align 374 document pairs using fuzzy matching that tolerates OCR errors, resolves abbreviations and groups multi-line addresses, and against 20 hand-annotated rulings it reproduces manual annotation at a macro F1 of 0.885. We tune it toward recall, since an identifier that reaches publication cannot be withdrawn, whereas an over-redaction can be corrected.

We then examine whether a detector can be trained without retaining any real identifier. Each entity in the training corpus is replaced by its nearest neighbour in BERT embedding space, drawn from public Swedish registries of names, streets and companies. On real held-out rulings, a KB-BERT detector trained on this pseudonymised corpus reaches an entity-level F1 of 0.893 and leaves 4.2% of the identifiers a human annotator marked unmasked, against 0.926 and 7.3% for an equivalent detector trained on real identifiers. Substituting entities uniformly at random instead, holding corpus, hyperparameters and seed fixed, gives 0.809 and 21.6%. Precision is unchanged across the two conditions and the difference falls on recall, which suggests that the plausibility of the surrogate supplies part of the learning signal.

We then attack the surrogates under two threat models covering data and model access. Place names prove the most vulnerable. Of the eight location entities held in the candidate pool, an inversion attack recovers four at its first guess and six within ten. Personal names could not be tested, since the pool stored them in a form the attack could not reach. Training with DP-SGD at $\varepsilon = 8$ reduced the detector to an F1 of 0.0, although that run was short and does not establish a limit on what differential privacy could achieve. Semantic pseudonymisation appears to reduce disclosure risk without eliminating it, and the residual risk varies considerably by entity type.