CAPE: CONTEXT-AWARE PRIVATE EMBEDDINGS FOR PRIVATE LANGUAGE LEARNING

Richard Plant
r.plant@napier.ac.uk

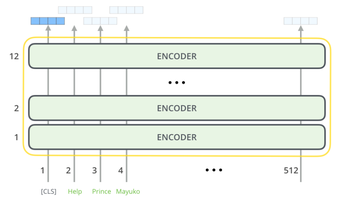
Valerio Giuffrida
v.giuffrida@napier.ac.uk

Dimitra Gkatzia
d.gkatzia@napier.ac.uk

Edinburgh Napier UNIVERSITY | School of Computing

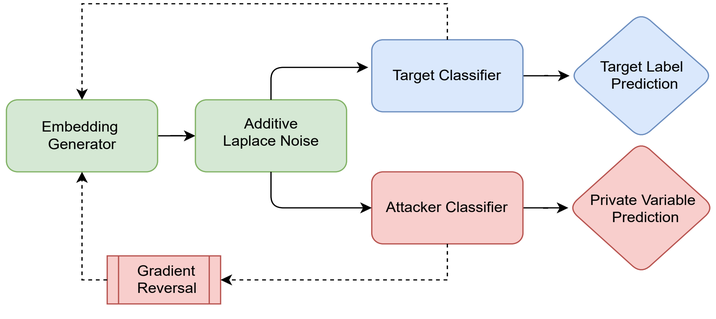
MOTIVATION AND CONTRIBUTION

- As creators of NLP models, we have a responsibility to make sure we use only essential information and prevent leaks of personal data from authors whose texts we utilise.
- We aim to obfuscate personal information (Preoțiuc-Pietro et al. 2015) while maintaining the desirable performance benefits of using pretrained embeddings.



- Previous attempts to mitigate this risk have relied on either adversarial learning (e.g. Coavoux 2018) which requires selection and annotation of specific characteristics to protect, or differential privacy (e.g. Lyu 2020) which can damage utility through application of noise.
- Our hybrid system combines both strategies, providing superior privacy outcomes.

PRIVATE EMBEDDING METHODOLOGY



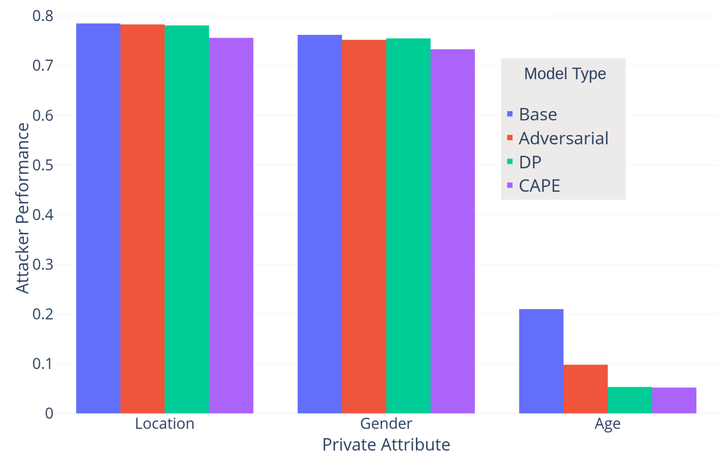
1. Generate embedding from input sequence: $x_e = f(x)$
2. Normalise representation: $x_e \leftarrow x_e - \min x_e / (\max x_e - \min x_e)$
3. Apply perturbation: $\tilde{x}_e = x_e + Lap(\frac{\Delta f}{\epsilon})$
4. Train classifiers: $\mathcal{L}(\tilde{x}_e, y, z; \theta_r, \theta_p) = -logP(y|\tilde{x}_e; \theta_r, \theta_p) - \lambda logP(\neg z|\tilde{x}_e; \theta_a)$

EXPERIMENTAL SETUP

We use English review texts from the Trustpilot dataset (Hovy et al. 2015), annotated with gender, age, and location private attributes. We compare our system against no privacy (base), and systems proposed by Coavoux (adversarial) and Lyu (DP).

We define two tasks:

- a legitimate sentiment prediction task from the input text sequence to the five-point review rating
- a simulated attacker task, attempting to learn a mapping from the embedding to the private attribute label



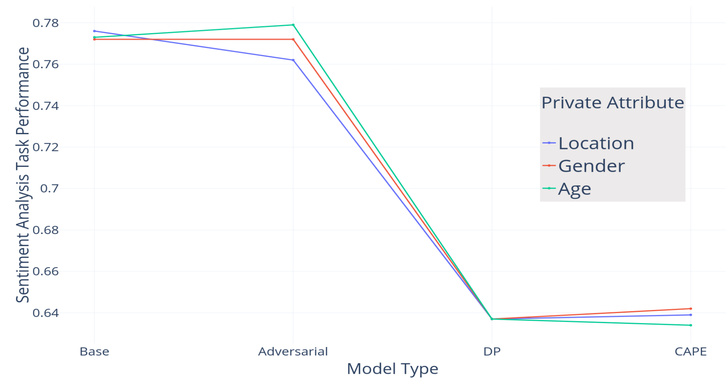
RESULTS

- We achieve a reduction in accurate predictions of private attribute values over the closest comparator
- Private attributes are predictable at differing rates, which may indicate variability in the semantic markers for each.
- Age is an extremely noisy variable, which we posit may be attributable to higher perceived privacy risk/greater willingness to lie by the user

UTILITY

We present results for our legitimate sentiment classifier task at a fixed privacy budget/noise level.

- Introducing DP-compliant noise causes a large drop in classifier accuracy
- For hybrid systems such as CAPE, relaxing the privacy budget to avoid performance degradation while targeting specific attributes could provide mitigation.



CONCLUSIONS

Our method provides ~3% reduction in attacker performance over the closest comparator.

Future research should consider:

- Multilingual datasets: In common with most research we consider only English texts, low resource languages may show different results.
- Better noise calibration: Traditional differential privacy causes utility degradation, relaxed definitions may be needed.
- Private by default: Our method still requires identification and annotation of sensitive data, building this into the embedding model itself would be the ideal solution.