Project: Private and Robust Text Data Linkage Models

COMP3850 Group Project Contribution - Macquarie University School of Computing (S1 2025)

Objective

To develop and evaluate privacy-preserving text data linkage models, specifically focusing on balancing linkage accuracy (utility) with robust data privacy guarantees using techniques like Local Differential Privacy (LDP). This project aimed to create algorithms suitable for securely linking unstructured text data, such as clinical notes, without exposing sensitive information.

My Contribution (XGBoost Prototype Focus)

As part of Group 46 for the COMP3850 project, my primary technical focus involved the development, integration, and evaluation of the Extreme Gradient Boosting (XGBoost) prototype model. This was one of several models explored by the group within the shared "SecureLink" web application platform.

Methodology & Key Achievements:

1. **Baseline Model Development:**
    - Developed and implemented a baseline XGBoost classifier trained on pre-processed (TF-IDF vectorization and Truncated SVD dimensionality reduction) unstructured text data to establish a non-private performance benchmark.
2. **Privacy-Preserving Solution (LDP):**
    - Engineered and integrated a Local Differential Privacy (LDP) mechanism by injecting controlled random noise into the feature vectors before training the XGBoost model. The level of privacy was controlled by the privacy budget parameter (ε).
3. **Privacy-Preserving Solution (Bloom Filter):**
    - Implemented and trained a separate XGBoost model using Bloom filter encoded representations, utilizing Dice similarity scores between record pairs as input features. This model was trained in a supervised manner using known match labels.
4. **Trade-Off Analysis:**
    - Conducted a comparative analysis of the privacy-utility trade-off across the three configurations (Raw, LDP, Bloom Filter) using standard metrics (Accuracy, Precision, Recall, F1 Score).

Quantified Results (Non-Confidential - Example using ε=7):

- **Baseline (Raw XGBoost):** Achieved F1 Score of **0.9118** (Accuracy: 0.9131, Precision: 0.9237, Recall: 0.9003).

- **LDP-Protected XGBoost (ε=7):** Maintained a high F1 Score of **0.8960** (Accuracy: 0.8980, Precision: 0.9118, Recall: 0.8808), demonstrating effective linkage despite the added noise.
- **Bloom Filter XGBoost (ε=7):** Achieved an F1 Score of **0.7252** (Accuracy: 0.8837, Precision: 0.8754, Recall: 0.6189), showing the impact of the encoding method on recall.
- **Conclusion:** Successfully quantified the trade-offs, showing LDP incurred only a minor ~1.7% F1 performance reduction compared to the raw model for this privacy budget, while significantly enhancing data privacy.

Confidentiality Notice

The source code, algorithms, and specific datasets associated with this group university research project are confidential. This summary outlines the overall project methodology, highlights my specific contribution related to the XGBoost model, and presents the non-confidential results derived from the group's project documentation.