Chatbot Training for Healthcare Application - Privacy and Anonymization

Suryalaxmi Ravianandan

Seminar: Data Governance Application
Degree: M.Sc. Data Science
September 2025

1 Privacy Risks

The primary privacy risk is the potential for re-identification of individuals based on a combination of quasi-identifiers (QIs). The QIs considered in medical student mental health dataset are the following:

- age_group
- year_group
- sex
- glang_gen

While individually these attributes may not uniquely identify someone, their combination can narrow down the possibilities significantly, potentially linking records back to specific individuals, especially when combined with external data. Because the data includes sensitive attributes (health, cesd, stai_t, etc.) , the danger is even greater—if someone is re-identified, their private personal information could be exposed.

2 Anonymization Techniques

The implementation part mainly focuses on the following anonymization techniques:

1. **Tokenization:** The original id column was replaced with a randomly generated id_token. This process replaces direct identifiers with a randomly generated value, breaking the direct link to the original identity while still allowing for internal tracking or linking of records within the anonymized dataset if necessary.
2. **Generalization:** The age and glang columns were generalized into broader categories (age_ group and glang_gen). This reduces the specificity of these attributes, making it harder to identify individuals based on these values. The year and sex columns were also mapped to broader categories, serving a similar purpose.
3. **Suppression:** The k_l_anonymize function suppresses (sets to NaN) the QI values for groups of records that do not meet the specified k-anonymity and l-diversity thresholds. This directly addresses the identified risk of re-identification by ensuring that combinations of QI values below the thresholds are not present in the anonymized dataset.

3 Legal Obligations - GDPR Principles

The anonymization process aligns with core principles of the *General Data Protection Regulation (GDPR, 2016)* [3]. Under GDPR, there is an important legal distinction:

- **Anonymous data** is considered *outside the scope of GDPR* because it can no longer be linked to an identifiable individual by any reasonably likely means.
- **Pseudonymized data**, on the other hand, remains *personal data* under GDPR, since re-identification is still possible if additional information is available. Techniques such as tokenization are therefore pseudonymization measures, not full anonymization.

In the present dataset, the applied methods (tokenization, generalization, suppression) aim to reduce the likelihood of re-identification and thereby support GDPR principles, including:

- **Data Minimization:** Generalization and suppression reduce the specificity of quasi-identifiers, ensuring only data strictly necessary for research purposes is retained.
- **Purpose Limitation:** By anonymizing or pseudonymizing the data, its potential misuse beyond the intended research context is reduced.
- **Integrity and Confidentiality:** Removing direct identifiers, aggregating quasi-identifiers, and suppressing high-risk records protect individual confidentiality.

It is crucial to note that achieving full GDPR compliance requires more than technical anonymization: a comprehensive governance framework is needed, including consent management, access control, audit trails, and organizational safeguards.

4 Conclusion

Anonymization techniques are crucial for protecting individual privacy in sensitive datasets like medical student mental health data, primarily by mitigating re-identification risks through quasi-identifiers. While methods such as tokenization, generalization, and suppression align with GDPR principles like data minimization, they inherently involve a trade-off with data utility. Higher privacy (e.g., through increased k and l values) leads to greater data loss and reduced analytical precision. Therefore, a balanced approach is essential, carefully selecting anonymization parameters to preserve research validity while upholding robust privacy standards.

5 Recommendations and Limitations

1.Recommendations

- **Layered Protections:** Combine anonymization with organizational safeguards (e.g., restricted access, consent management, encryption) to strengthen compliance with GDPR and reduce residual risks.
- **Continuous Monitoring:** Regularly assess the effect of anonymization on analytical accuracy to ensure research validity.

2. Limitations

- Optimal choices for anonymization parameters (k, l) vary by dataset size, sensitivity, and intended research use.
- Anonymization reduces but does not eliminate re-identification risks—especially if adversaries can link with external data sources.
- High privacy thresholds may result in excessive data suppression, rendering parts of the dataset unusable.

References

[1] Anna Leschanowsky, “Data Anonymization and Privacy-Enhancing Technologies (PETs),” Data governance Application, FAU Erlangen-Nürnberg, 26 May 2025. Lecture.

[2] Anna Leschanowsky, Dr. Birgit Popp, “Legal Requirements and Compliance,” Data governance Foundation, FAU Erlangen-Nürnberg, 4 Nov 2024. Lecture.

[3] European Parliament and Council. (2016). *General Data Protection Regulation (GDPR), Regulation (EU) 2016/679*. Official Journal of the European Union. Includes Recital 26, which clarifies that **anonymous data** falls outside the scope of GDPR, while **pseudonymized data** remains personal data and thus within scope.

[4] Sweeney, L. (2002). k-anonymity:A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557–570.

[5] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 3–es.

[6] Wiest, I. C., Leßmann, M.-E., Wolf, F., Ferber, D., Van Treeck, M., Zhu, J., Ebert, M. P., Westphalen, C. B., Wermke, M., & Kather, J. N. (2025). Deidentifying medical documents with local, privacy-preserving large language models: The LLM-Anonymizer. *Journal of Medical Internet Research.*