

# Aidentifiering och pseudonymisering av svenska patientjournaltexter

Hercules Dalianis

## Sammanfattning

Inom hälsovården produceras patientjournaltexter idag som beskriver patientens symptom, diagnos, behandling och utfall och dessa sparas i digitalt format i en mycket stor mängd. Journalerna används sällan för forskning eftersom journalerna innehåller känslig information om patienterna. Ett sätt att använda journalerna är att aidentifiera journalerna och därefter pseudonymisera dem.

I vår studie har 100 svenska patientjournaltexter annoterats manuellt för PHI (Protected Health Information). PHI är information som kan identifiera patienten såsom namn, adress, telefonnummer, mm, denna annoterade mängd har därefter använts för att träna två olika maskininlärningsalgoritmer dels Conditional Random Field och dels Deep Learning LSTM, resultaten har evaluerats med relativt goda resultat. En regelbaserad ansats skulle kunna komplettera aidentifieringen och förbättra resultaten ytterligare. Aidentifieringen markerar upp den information som är känslig men pseudonymisering ersätter den uppmärskade information med pseudonymer för att eventuellt dölja missad aidentifiering, detta görs i vår studie med en regelbaserad ansats.