SYNTETISKE DATA

Dato: 23-01-2026

Status: Final

Version: 2

Formål

At dokumentere evalueringsmetoder og erfaringer fra brug af syntetiske data – herunder hvordan tilgangen har påvirket datakvalitet, modellens performance og generaliserbarhed. Dokumentet skal give indsigt i evalueringspraksis og fungere som grundlag for fremtidig kvalitetskontrol og udvikling.

Resume

Dokumentet gennemgår på et redegørende niveau vores tilgang til den syntetiske datagenerering.

Den syntetiske datagenerering er iterativt blevet forbedret via input fra fagligt personale, og løsningen til sygeplejefaglige udredningssamtaler er understøttet af en kategoriseringsmodel trænet på syntetiske samtaler, genereret via LLM.

Indholdsfortegnelse

1. Introduktion

Vi har i projektet arbejdet med 3 områder for værdiskabelse i sygeplejen, kaldet use cases, og har til hver af dem udviklet en løsning baseret på talegenkendelse, tekstklassificering og/eller generative tekst modeller. Løsningerne har det til fælles at de forsøger at afhjælpe byrden forbundet med at generere eller orientere sig i journaldata. Disse løsninger har vi gennem afprøvninger vist frem til sygeplejefagligt personale fra et bredt udsnit af kommuner.

Projektet har arbejdet med forskellige muligheder for værdiskabelse i syge- og hjemmeplejen samlet i 3 use cases (UC). Følgende er use cases overskrift, efterfulgt af en kort beskrivelse:

- **UC1: Sygeplejefaglig udredning**

Støtte til oprettelse af sygeplejetilstandsnotater på baggrund af en sygeplejefaglig udredningssamtale (SFU-samtale) mellem borger og sygeplejerske. Støtten ydes gennem udkast til notater baseret på samtalens indhold.

- **UC2: Opsummering af borgerjournal**

Opsummering af borgerjournal til støtte ved borgerbesøg udført af hhv. Sygeplejersker (SPL) og sundheds- og omsorgsassistenter og -hjælpere (SSA/SSH).

- **UC3: Indtaling af besøgsbeskrivelse**

Støtte til oprettelse og opdatering af besøgsbeskrivelser til hjælp med hjemmebesøg udført af SSA/SSH. Besøgsbeskrivelserne oprettelse på baggrund af en indtaling.

Vi har i projektet ikke haft adgang til ægte data, hverken journaldata, SFU-samtaler, eller besøgsbeskrivelses indtalinger. Det har derfor været afgørende for projektet at kunne generere syntetisk data til alle projektets formål. Vi har defineret følgende distinktion mellem forskellige typer af syntetisk data:

- **Manuelt syntetisk data**

Data genereret "manuelt", primært af fagligt personale i projektkommunerne og af afprøvningskommunerne i forbindelse med afprøvning af løsningerne. Dette data forventes at være det tætteste vi kommer på "ægte" data og bliver derfor primært benyttet til evaluering af modeller og samlede løsninger.

- **Automatisk syntetisk data**

Data genereret "automatisk", primært gennem den syntetiske datagenerering ved prompt af store sprogmodeller. Denne data beriges med validering og feedback fra fagligt personale. Denne data benyttes udelukkende til træning af modeller og udvikling af løsningerne.

Dette dokument beskriver den proces, vi har fulgt i arbejdet med at generere og validere syntetiske data via store sprogmodeller (automatisk syntetisk data). Denne data er brugt til udvikling af UC1 løsningen. Dokumentet gennemgår de teknikker og parametre, der er anvendt til at opnå data, som bedst muligt repræsenterer de reelle mønstre og variationer i den virkelige verden.

Afslutningsvis diskuteres nogle af de overvejelser vi har haft undervejs i forbindelse med den "automatisk" syntetiske datagenerering.

Du kan læse mere om tilgangen til den manuelle syntetiske datageneration, samt arbejdet med at definere, afgrænse og afprøve projektets use cases i det supplerende dokument "Talt: Uses cases og afprøvning fra et sundhedsfagligt perspektiv".

2. Data generering via LLM

Vi ønskede en syntetisk datagenerering implementeret i et Python modul, hvor vi iterativt kunne forbedre promptingen og hurtigt kunne generere nye samtaler baseret på kvaliteten af samtalerne og ekstern feedback. Samtidig ville vi også gerne beholde historikken for vores prompting-strategier, for dels at dokumentere historikken og udviklingen i datagenereringen, men også for altid at kunne gå tilbage til en tidligere konfiguration. Koden skulle ligeledes være modulært, hvor små prompt-moduler med specifikke formål kunne genbruges og indgå på tværs i forskellige endelige prompt-metoder. Vi har derfor udviklet et modul til syntetisk datagenerering. Modulet tager udgangspunkt i et konfigurerbart script. Her kan forskellige parametre varieres via en config-fil, med formålet at opnå et varieret datasæt, genereret på tværs af flere kørsler.

Følgende parametre er konfigurerbare:

- Antallet af samtaler
- Længden på samtalerne
- Antallet af sygeplejetilstande der snakkes om per samtale
- Hvilke sygeplejetilstande man vil have med i samtalerne
- Hvilken prædefineret promptmetode, der skal bruges
- Hvilken sprogmodel, der skal generere det

Når man kører scriptet, tager den disse indstillinger, genererer samtalerne, nedbryder samtalen linje for linje og skriver dem til data science forretnings-database med relevante tags og kommentarer, der ligeledes angives i config-filen.

Vores overordnede prompt-metoder, som man vælger i forbindelse med genereringen, er oprettet som kombinationer af mindre prompt-funktioner, der hver især opfylder deres eget specifikke formål, som fx beskrivelse af output-formatet eller indhentning af FSIII-dokumentation for de relevante sygeplejetilstande.

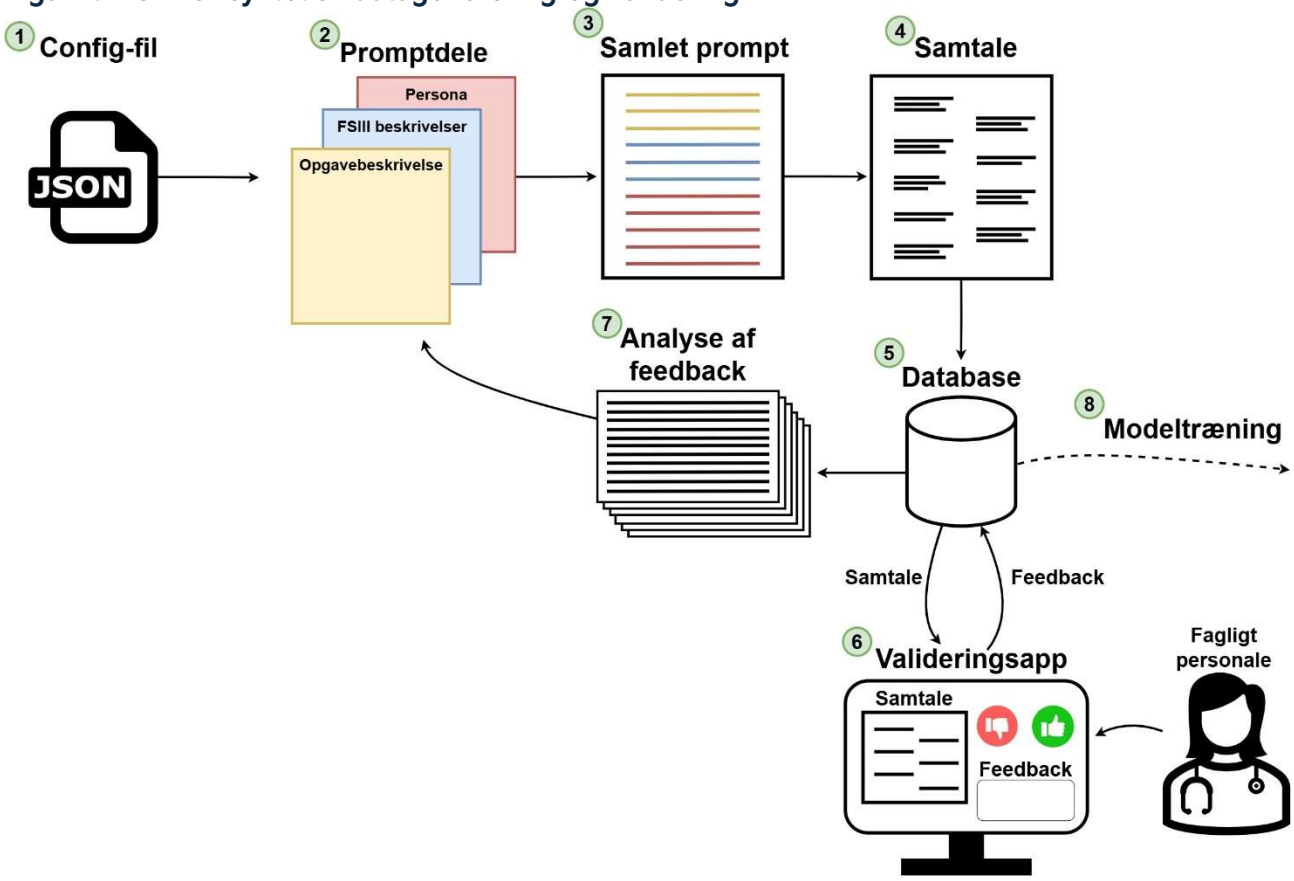
Projektet har gennemgået forskellige iterationer af den syntetiske datagenerering, hvor der i begyndelsen blev arbejdet med generering af mindre samtalebidder, blev der senere i projektet forsøgt at generere fulde samtaler. De fulde samtaler har gennemgået flere faser; det er bl.a. forsøgt at inkludere nogle af de mindre samtalebidder - som var blevet valideret af fagligt personale - i genereringen af de fulde samtaler. Information om sygeplejetilstandene fra FSIII-materialet er også forsøgt inkluderet i genereringen og effekten af at generere og inkludere varierende borgerpersonaer er også i omfattende grad blevet afsøgt. Hvor samtalebidderne og FSIII-materialet blev trukket fra et fast datamateriale og tilføjet for at give højere kvalitet og større faglighed af samtalerne, blev personaerne genereret med en LLM med henblik på at skabe en mere realistisk variation i samtalerne ud fra informationer som borgeres alder, personlighedstræk, tætte relationer og (tidligere) job.

Flowet i den syntetiske datagenerering og dertilhørende validering er illustreret med 7 skridt i figur 1:

1. Konfigurationer indlæses fra en config-fil i json-format.
2. Ud fra den valgte promptmetode, genereres de tilhørende promptdele.
3. Promptdelene samles til en samlet prompt.
4. Samtalen genereres med den valgte sprogmodel.
5. Samtalen gemmes i databasen.

6. Samtalen sendes til valideringsappen, hvor fagligt personale evaluerer samtalen. Feedbacken sendes tilbage til databasen.
7. Kommentarerne fra evalueringerne analyseres og giver anledning til ændringer i nogle af promptmetoderne.
8. Validerede samtaler, som er blevet godkendt, sendes til træning af klassifikationsmodellen.

Figur 1: Flow for syntetisk datagenerering og validering



3. Validering af samtaler

De syntetiske SFU-samtaler blev valideret af sygeplejefagligt personale. Indsamling af denne validering blev understøttet af en valideringsapp[1] udviklet til formålet. Det faglige personale blev instrueret i at give den enkelte samtale en samlet binær vurdering (god nok / ikke god nok) på baggrund af en række forskellige kriterier, blandt andet realisme og sygeplejefaglig kvalitet. En begrundelse blev tilkoblet vurderingen og her blev personalet opfordret til at give konkrete eksempler på hvorfor samtalen opnåede den konkrete vurdering og hvor i samtalen eksempler der underbyggede begrundelsen kunne findes. Godkendte samtaler blev efterfølgende brugt til modeltræning og begrundelser, både negative og positive, blev brugt til forbedring af den syntetiske datagenerering.

[1] Se 'Kode og udviklingsmiljø'-dokumentet for yderligere information on valideringsapp'en

4. Overvejelser, anbefalinger og konklusioner

I den ideelle verden ville vi have adgang til ægte samtaler mellem borgere og sygeplejersker (og deres rettede transskriberinger) i et så stort omfang, at alle sygeplejetilstande og deres variationer i emner var berørt. Da det ikke har været tilfældet, har den syntetiske generering haft som mål at lave så realistiske samtaler som muligt og med så stor variation som mulig. Vores primære værktøj til at teste denne validering kommer fra input af det faglige personale gennem valideringsklienten. Processen har været iterativ, hvor de indledende samtaler var meget upersonlige og urealistiske. Efter implementering af konkrete forbedringstiltag, såsom borgerpersona'er, blev samtalerne bedre. Undervejs med indsamling af feedback til de syntetiske samtaler blev en voksende bunke af godkendte samtaler indhentet. Denne proces garanterer en eller anden grad af realisme for den enkelte samtale, men garanterer ikke en realistisk variation i den samlede bunke af godkendte samtaler. Dette bør videre arbejde med syntetisk data håndtere.

En udvidelse, der potentielt kunne have gavnet datagenereringen, var analyser af ægte samtaler, dette kunne underbygge en forbedret realisme af den enkelte samtale og give indsigt i den samlede forventede variation. Det kunne være fx. være fordelinger af, hvor meget der bliver snakket om de forskellige sygeplejetilstande før man går videre, hvor ofte en sygeplejetilstand bliver taget op igen senere i samtalen og en analyse af den samlede liste af sygdomme og deres håndtering, der bliver berørt til hver sygeplejetilstand. Viden om fordelingen af disse sygdomme gør, at vi tilfældigt kunne trække fra denne fordeling, og indhente baggrundsviden om den pågældende sygdom til prompten ved genereringen af en samtale. Denne viden ville derfor kunne hjælpe med at styre, at både formen og indholdet af samtalerne bedre afspejlede virkeligheden og var fagligt korrekte.

Den primære benyttelse af de automatisk generede syntetiske SFU-samtaler er træning af klassifikationsmodellen. Her er det værd at bemærke at vi har valgt en modelarkitektur[i] der benytter en præ-trænet tekst embedding model, finetunet via kontrastive learning, hvor både positive og negative træningspar dannes på tværs af alle kategorier. Dette er designet til at opnå god performance på en relativt lille mængde data. Antallet af træningspar der dannes ved træning, vokser hastigt med mængden af træningsdata som funktion af antallet af kategorier, hvilket betyder at træningskompleksiteten stiger med større mængder data og vi erfarer at det har negativ effekt på den endelige performance. Vi har en kraftig mistanke om at dette skyldes at den samlede semantiske variation i data, er meget lille, hvorfor større mængder data i nogen grad hjælper til at nedbryde grænserne mellem kategorierne. Specielt kategorien 'andet', har store semantiske overlap med mere eller mindre alle kategorier og (dimensionsreducerede) visualiseringer af data viser et udfordrende klassifikationsrum. Vi har derfor set størst gevinst ved træning på mindre udsnit af data, bedst, når semantisk centrale træningseksempler udvælges.

Vi har i et begrænset omfang afsøgt muligheden for at bruge andre modelarkitekturer, der i bedre grad kunne gøre brug af den voksende mængde syntetisk data tilgængelig. Denne afsøgning ledte ikke til bedre kategoriseringsmodeller.

Hele modeludvælgelsesprocessen er i stor grad begrænset af den relativt lille mængde evalueringsdata. Dette skyldes primært at alt manuelt syntetisk data indsamlet fra afprøvning af løsningen ikke blev beriget med annotering af kategorier på linjeniveau. Derfor har vi haft begrænset mulighed for at udvælge modeller på en statisk signifikant måde.

[i] Se 'Data og modeller'-dokumentet for uddybende information om modeller