KI og maskinlæring som metode i evaluering

Muligheter og eksempler

Jan Roar Beckstrøm

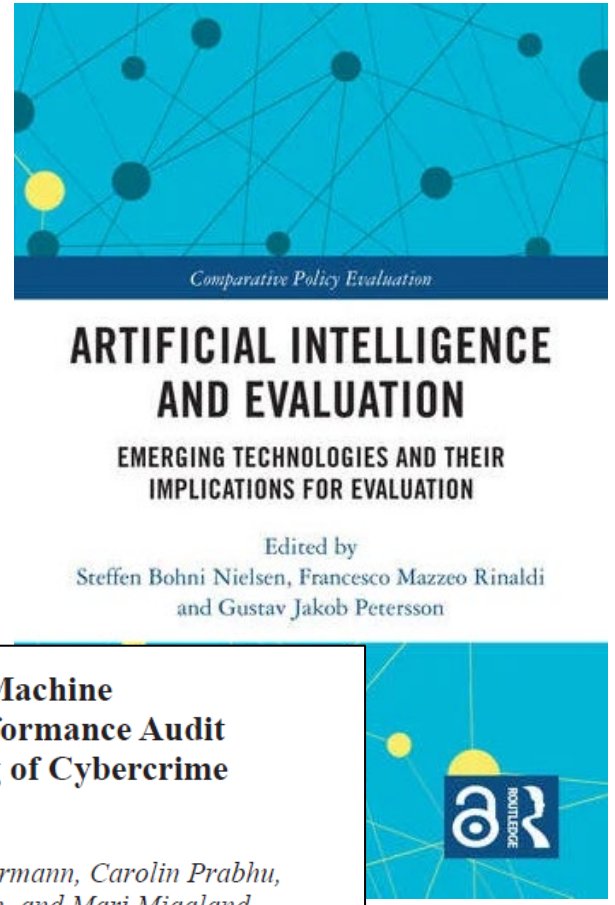
avdelingsdirektør

Riksrevisjonens innovasjonslab

Prompt: Lag en illustrasjon som kobler kunstig intelligens med evaluering. Stil skal være futuristisk cyberpunk. Bruk blå og grønntoner.

Hvor er metodefronten?

… har man tid og penger til å være der?
… og er det gøy der?



6 Text Mining and Machine Learning in a Performance Audit of Police Handling of Cybercrime in Norway

Tom Næss, Helge Holtermann, Carolin Prabhu, Lars Skage Engebretsen, and Mari Mjaaland

Introduction

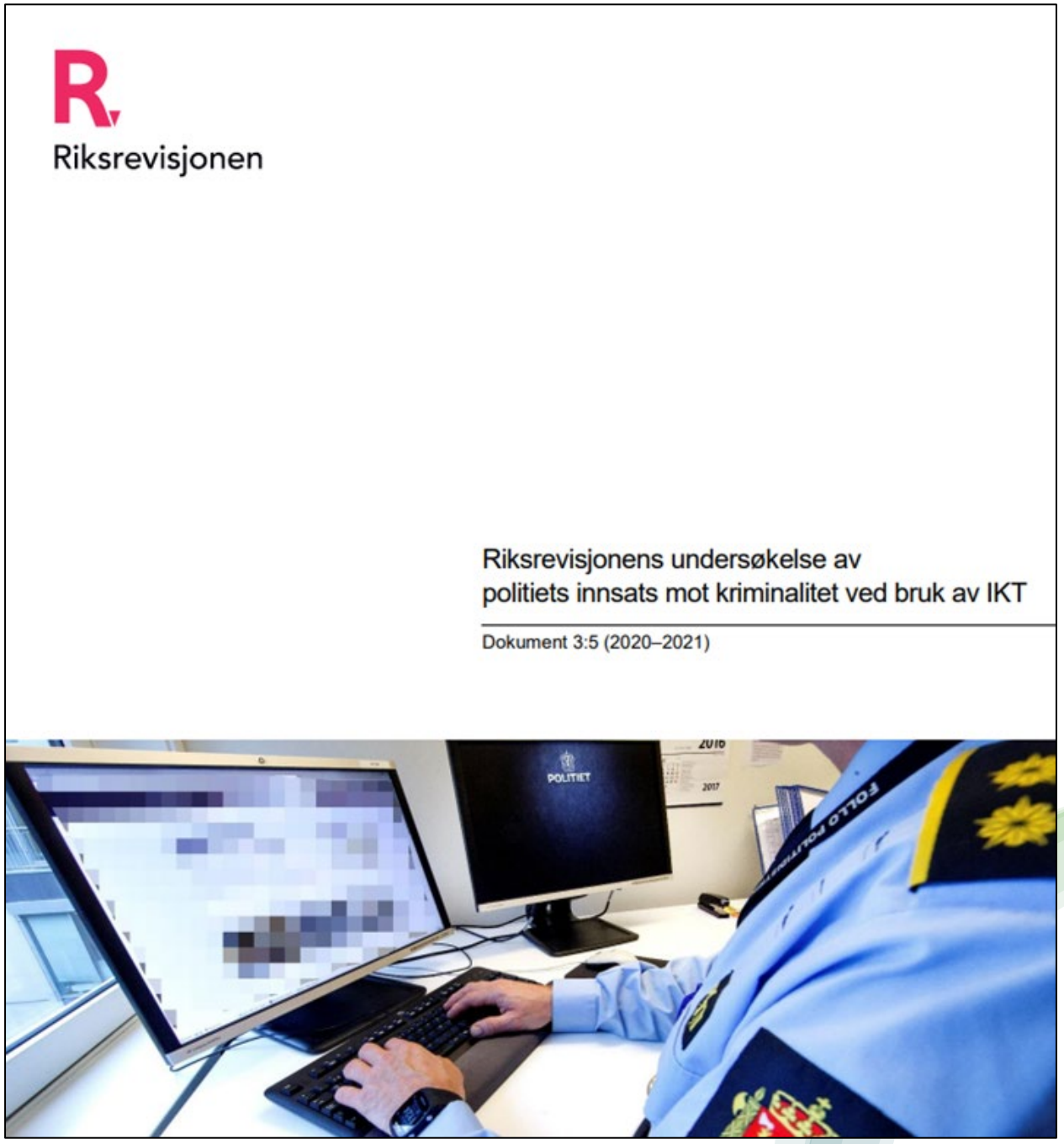
Technological development has opened new opportunities for criminal activity. While traditional theft has declined, crimes such as online fraud and identity theft have risen sharply in Norway. The shift in crime poses a challenge to national police authorities, as it requires new investigation methods. In 2021, we (the Office of the Auditor General of Norway) therefore conducted a performance audit[1] of the national police's efforts to combat cybercrime.[2,3] The aim was to assess whether the police had an overview of, investigated, and solved cases of cybercrime in accordance with the Police Act adopted by the Norwegian parliament.

Eksempel 1:

Kategorisering av kriminalsaker

Se også:

https://www.riksrevisjonen.no/rapporter-mappe/no-2020-2021/undersokelse-av-politiets-innsats-mot-kriminalitet-ved-bruk-av-ikt /

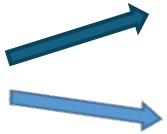


Problemstilling (en av dem)

Av alle anmeldte kriminalitetssaker

- hvor mange saker er relatert til IKT-kriminalitet?
- hvilke typer IKT-kriminalitet har vi?

- Et klassisk (binært) klassifiseringsproblem
  - 300 000+ tilfeller → trener en ML-algoritme



IKT- kriminalitet

Ikke IKT- kriminalitet

Alternative algoritmer testet

- Naive Bayes (dårlig resultat)
- Random Forest (sterkt overtrent)
- XGBoost (sterkt overtrent)
- **Support Vector Machine (SVM) – valgt**
  - SVM er en veiledet (supervised), svart boks ML-modell

Trene en modell ? Greit, men …

- Vi har ikke treningsdata!
- Vel, da må vi lage treningsdata…

Manuell klassifisering av 1072 saker

1. Samlet inn dokumenter fra totalt 334 544 straffesaker, totalt 396 917 documenter

2. Tilfeldig utvalg av 1072 saker, for manuell gjennomgang og klassifisering/merking

Resultat: 1072 manuelt merkede saker. Dette ble treningsdata

ML- klassifisering

- Data → politianmeldelsestekst , saksbeskrivelse mm.
- Noen begreper vil være spesifikke/hyppige for IKT-krimsaker
- Dermed er *ord* variablene som definerer prediksjonen

Valg av termer/variabler

- Noen ord er mer utbredt for IKT-kriminalitet enn for ikke-IKT-kriminalitet
- Vektet ved bruk av TF-IDF (basert på treningsdata)
  - (Termfrekvens – Invers dokumentfrekvens)
- Brukte de 150 ordene med høyest vekt fra hver klasse, fjernet vanlige ord (fra de 1072 treningssakene)
- Valgte de 70 ordene (variabler) med størst forskjell i vekt fra de to klassene

Treningsdata og variabler

Oppsummert så langt:

- Support Vector Machine
- 70 variabler (ord)
- 1072 manuelt merket saker som treningsdata

Den endelige, trente modellen

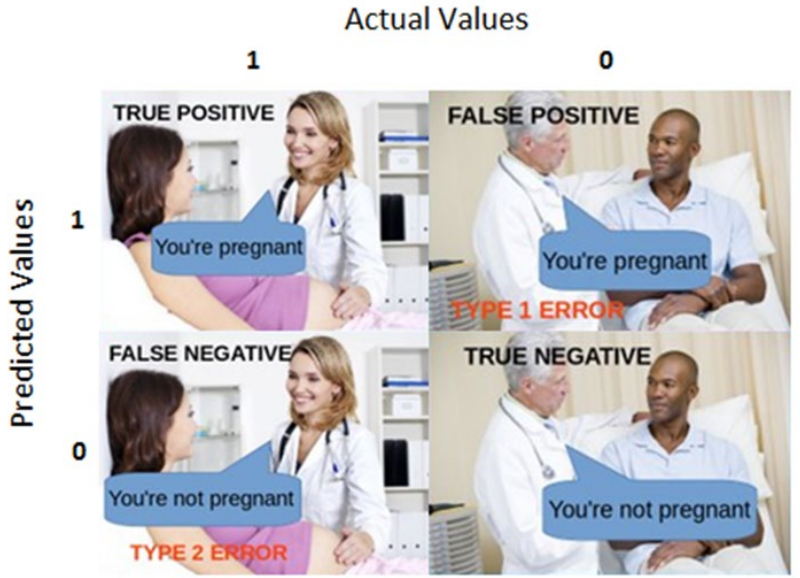
- Kjørt på hele populasjonen av saker (300 000+ tilfeller)
- Saker som tekstmessig mest ligner IKT-kriminalitet…
- … blir lagt i «IKT-kriminalitetsboksen» av modellen

Resultater…?

- Ulike mål på «model fit»
- - Ingen modell er perfekt (Ref. «modell»)

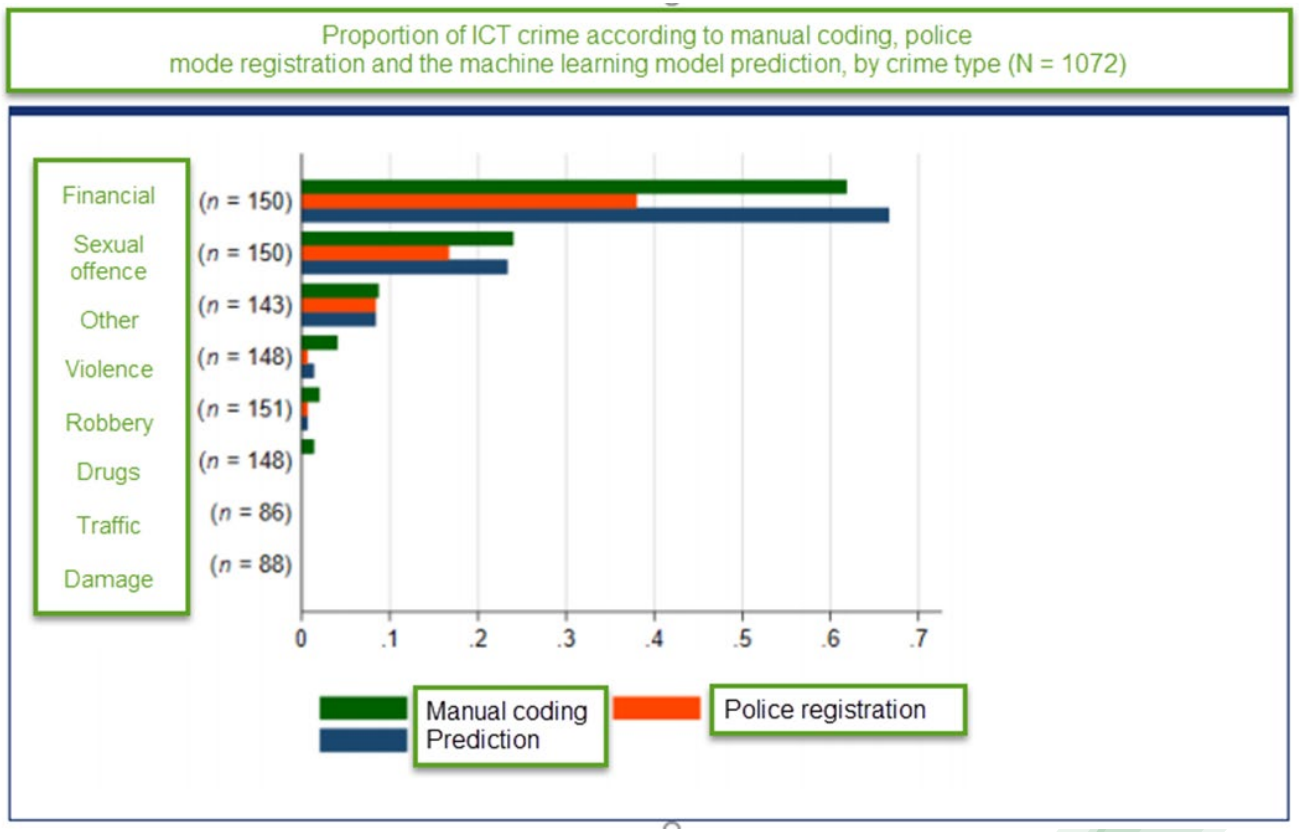
- F.eks . graviditet - hva er verst?
  - Å få budskapet om graviditet – når du faktisk ikke er det? (falsk positiv)
  - Å få beskjeden om ikke-graviditet – når du faktisk er gravid? (falsk negativ)

- Så: Hvilket mål er best i et bestemt tilfelle?



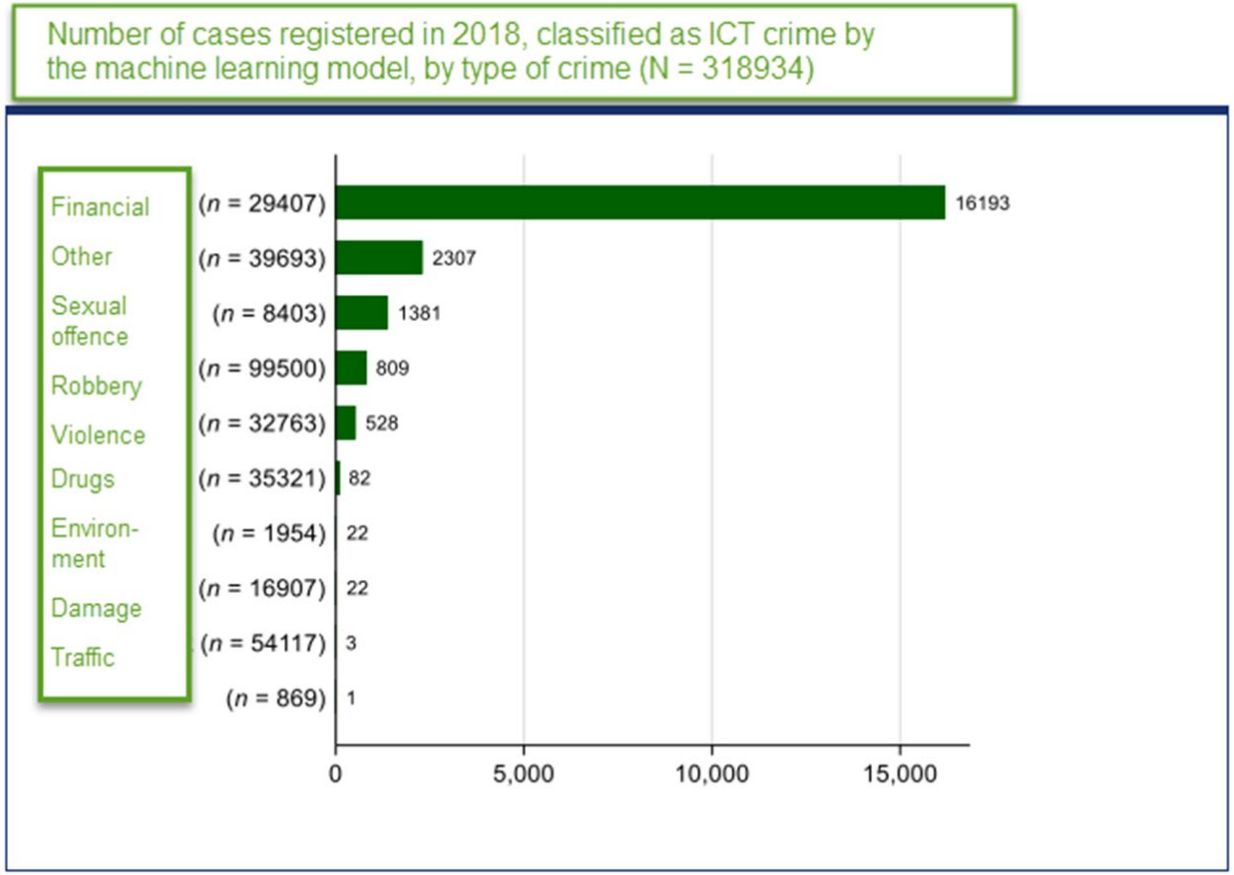
Forskjeller

Politiets registrering, manuell koding og prediksjon



Resultater…

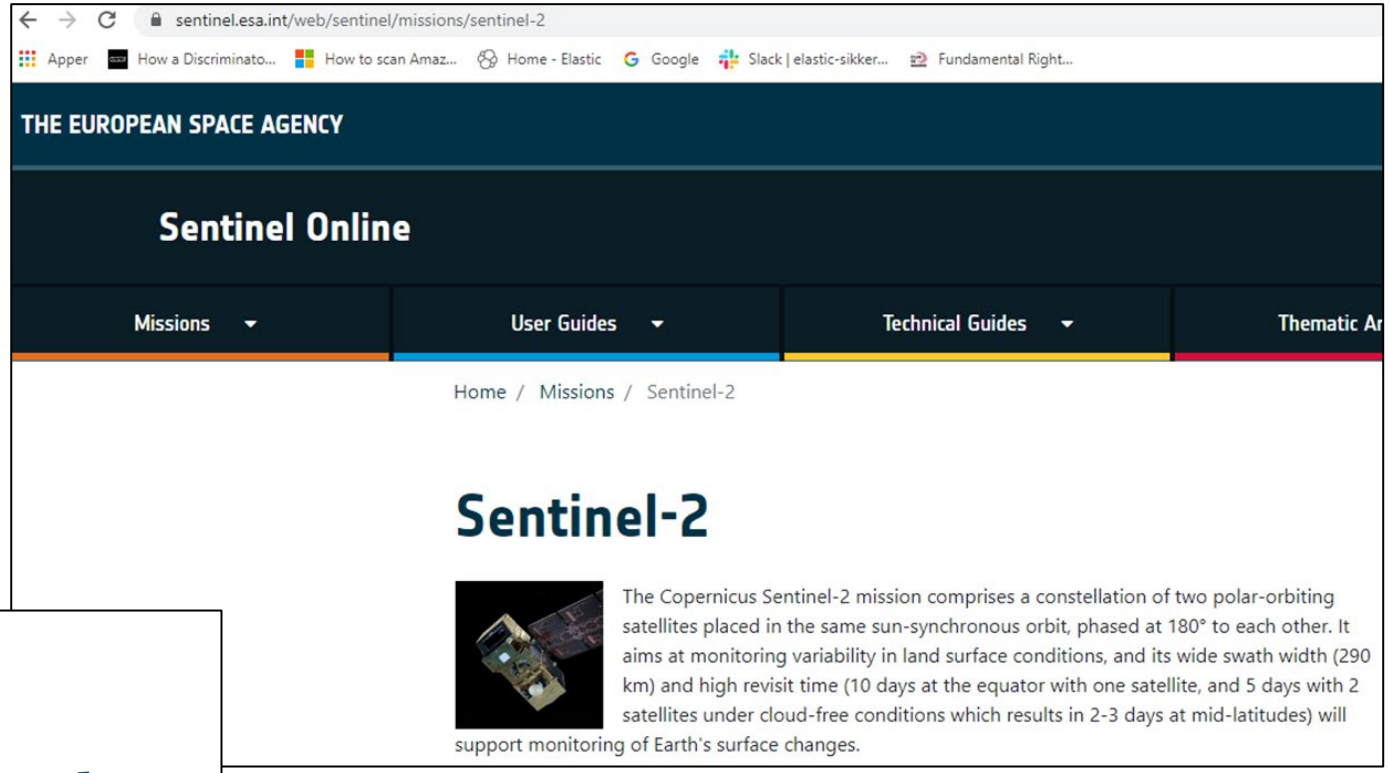
21 500 saker av til sammen 334 544 ble av modellen kategorisert som IKT-kriminalitet.



Eksempel 2:

ML og matsikkerhet

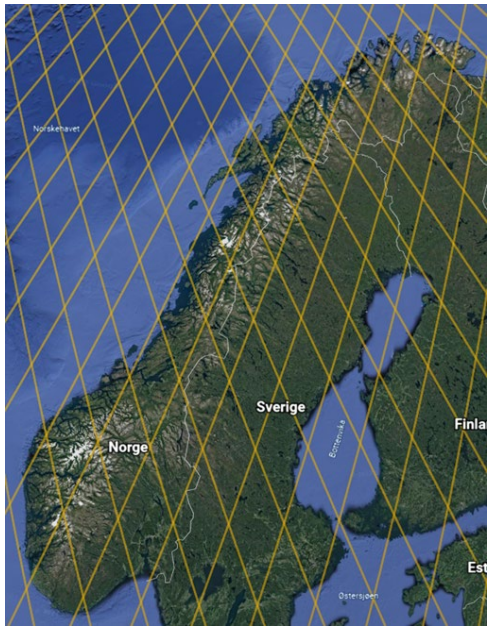
- Brukes faktisk jordbruksareal til matproduksjon?



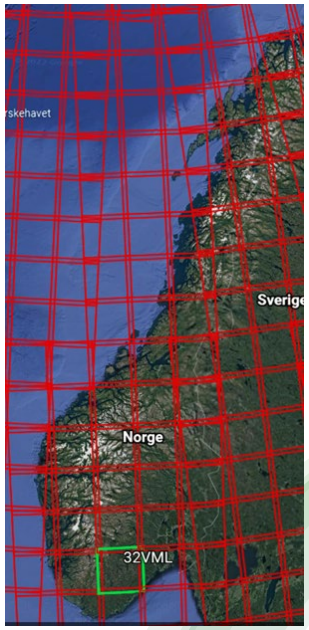
- Vi brukte ML på satellittbilder for å studere dette
- Trente en ML-modell for å gjenkjenne ulike typer avlinger
- For eksempel: Skyer var et problem…

Sentinel 2-fliser og baner

Mest relevante fliser og baner for vår analyse

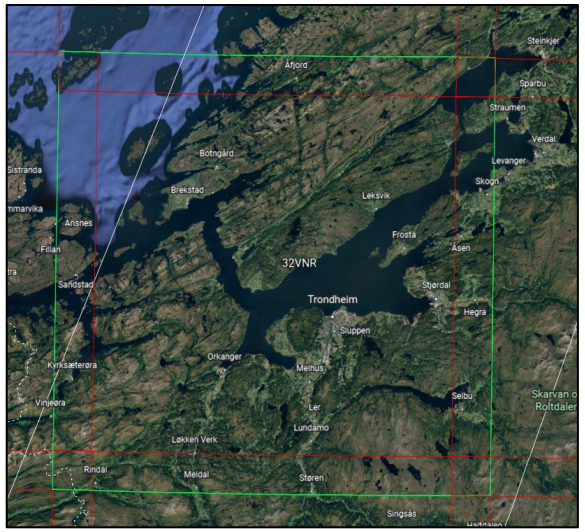


Flere baner



Flere fliser

Problemet med skyer – flis T32VNR



Google Earth view

S2 quicklook , 2019-08-07

(Algoritmen som ble brukt var convolutional neural net (CNN) - En type dyplæringsalgoritme)

remote sensing

MDPI

Article

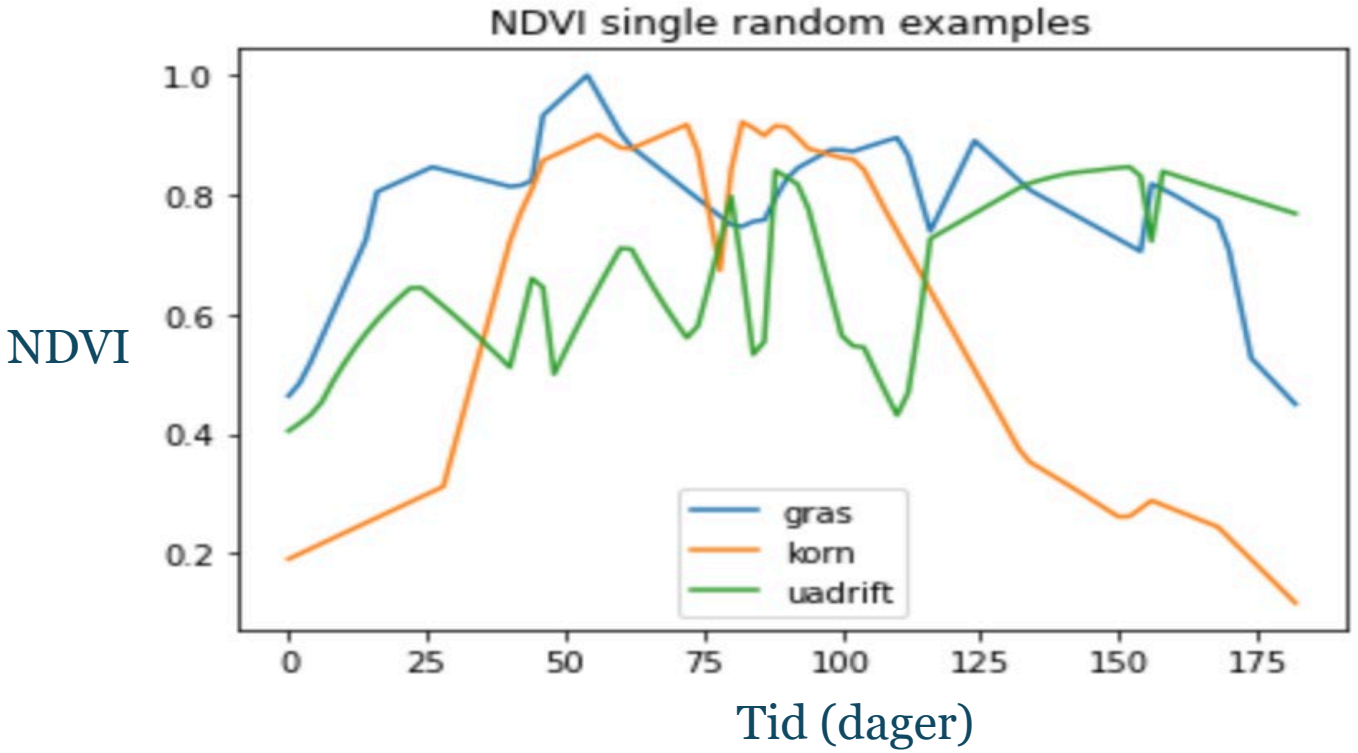
Mapping Seasonal Agricultural Land Use Types Using Deep Learning on Sentinel-2 Image Time Series

Misganu Debella-Gilo * and Arnt Kristian Gjertsen

Hvordan gjenkjenner maskinen ulike typer vekster?

NDVI: Normalized difference vegetation index

→ høy NDVI tyder på tett vegetasjon



NDVI i vekstsesongen:

- Korn er normalt bakkeformet
- Gress kan høstes flere ganger
- Ikke brukt: uregelmessig

- Plutselige fall tyder på skyer/skygger

Konklusjon matsikkerhet

- Funket ikke særlig bra…
- Som sagt: særlig pga. dårlig datakvalitet
- Ikke gode nok resultater til at det ble med i rapporten

- Kartverket har nå laget tidsserier av Sentinel-2 bilder *uten* skyer…

Eksempel 3:

Hvordan kan vi automatisere dokumentanalyse av kommuneplaner?

(courtesy of Siri Hellevik og Joachim Sandnes – forvaltningsrevisorer)

Formål

- Teste ut muligheter for å effektivisere dokumentanalyse i forvaltningsrevisjon
- Lese masse kommuneplaner, men det meste av innholdet i disse er ikke relevant?
- Hvordan spare tid på å finne relevant innhold?

- Bruke chatGPTs API til automatiserte spørringer

Utvalg for uttesting

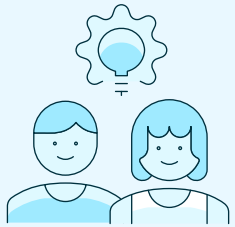
- 14 planer fra kommuner, som delvis handler om helse

- Utvalgskriterier:
  - Kommuner med mange korttidsinnleggelser på sykehus
  - Kommuner med mange utskrivningsklare pasienter (= pasienter som kommunene ikke klarer å ta imot)

Innledende fase

1

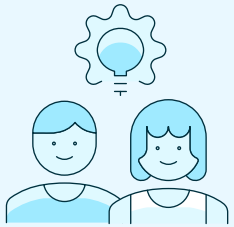
Utforskning og eksperimentering



Lese, forstå og node 6 utvalgte planer

2

Designe et prompt



Finne en god formulering og spesifisere format på svar

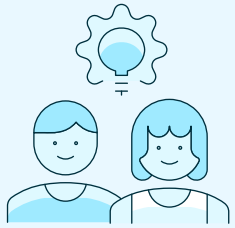
3

Maskinen svarer

Se raskt om output ser fornuftig ut

4

Validering



Manuell gjennomgang av stikkprøver

Prompten de endte opp med:

Du er en hjelpsom AI-assistent som forholder deg til følgende regler:

- Du skal ikke komme med oppsummerende beskrivelser av dokumentet.
- Jeg kommer til å gi deg en liste med temaer. Du skal lete etter innhold som er relevant for disse temaene.
- For hvert tema, skal du vurdere hvor relevant teksten er ved å velge mellom én av følgende kategorier: **Ikke i det hele tatt, I liten grad, I noen grad, I ganske stor grad, I svært stor grad.**

- Temaene du skal se etter er:
  1. **Samhandling i helsesektoren**
  2. **Samarbeid med spesialisthelsetjenesten**
  3. **Utskrivningsklare pasienter**
  4. **Pasientforløp**
  5. **Samhandlingsreformen**

- Du skal vurdere hvert tema for seg, og strukturere vurderingen av hvert tema slik, i markdown-format:
  - **Tema [nummer]: NAVN PÅ TEMA:**
  - **RANGERING AV RELEVANS:**
  - **HVORFOR TEKSTEN ER RELEVANT (maks 30 ord):**

Her kommer innholdet du skal vurdere: *[Her gjengis innholdet fra én A4-side]*

Eksempel på svar fra én spesifikk A4-side

- **Tema 5:**
Samhandlingsreformen

- **RANGERING AV RELEVANS:**
I svært stor grad

- **HVORFOR TEKSTEN ER RELEVANT:**
Teksten nevner direkte at samhandlingsreformen vil kreve ytterligere satsing på rehabilitering fra kommunenes side.

Skalering

5

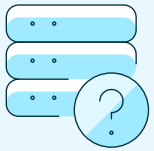
Pre-prosessering av dokumenter

RStudio

Lese inn 14 dokumenter og stykke dem opp i 450 enkeltsider

6

Maskinell innlesing

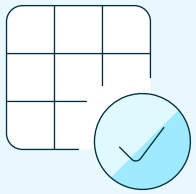


x 450

450 automatiske spørringer

7

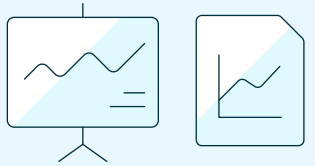
Strukturering av data



Svarene struktureres i et datasett i R

8

Visualisering

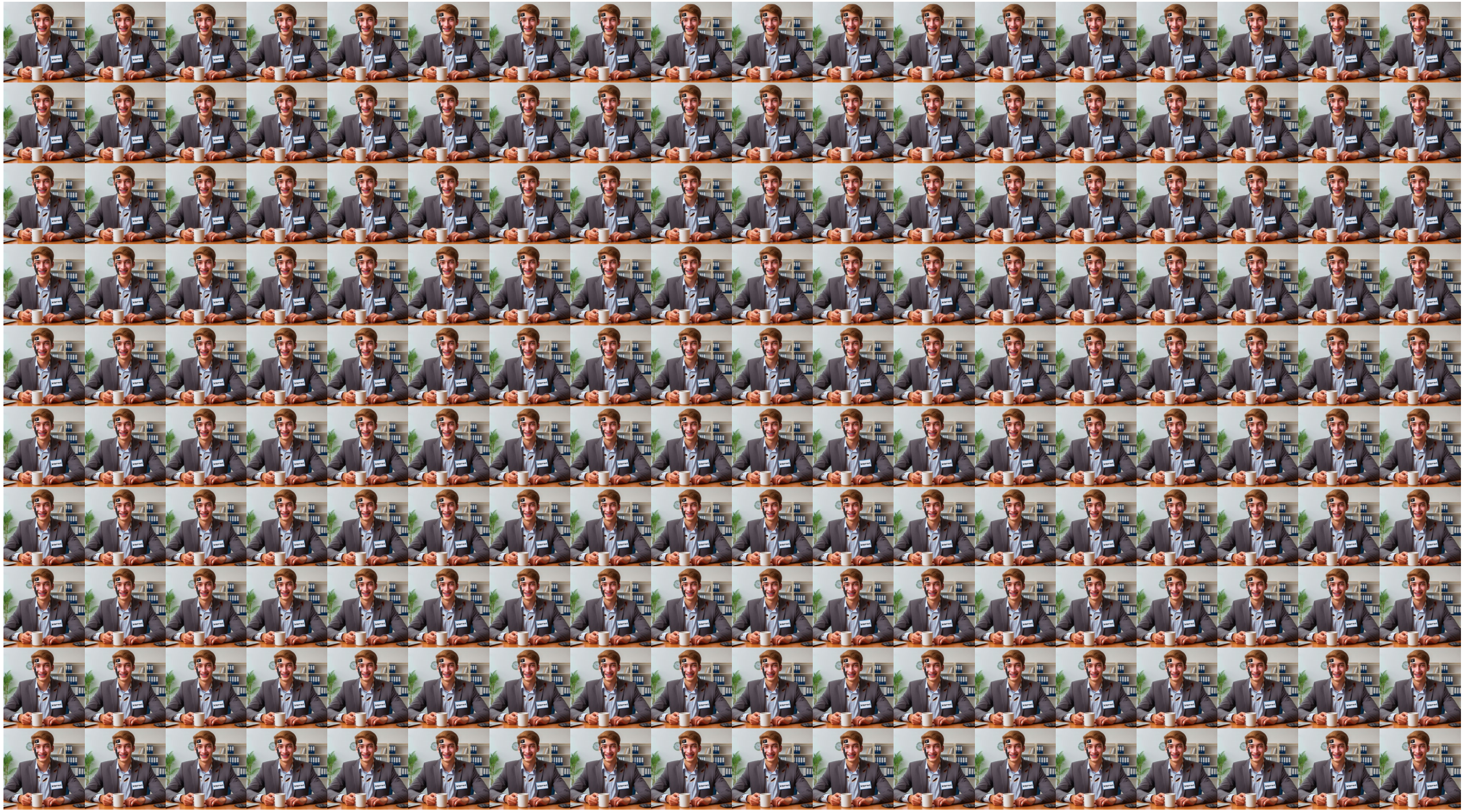


Resultatet visualiseres i en interaktiv figur

«Å bruke en språkmodell er som å ha tilgang til 1000 ivrige, smått inkompetente forskningsassistenter»

(Sitat: Joachim Sandnes)


Interrent



Fredrikstads kommunedelplan
for helse og velferd 2016-2027

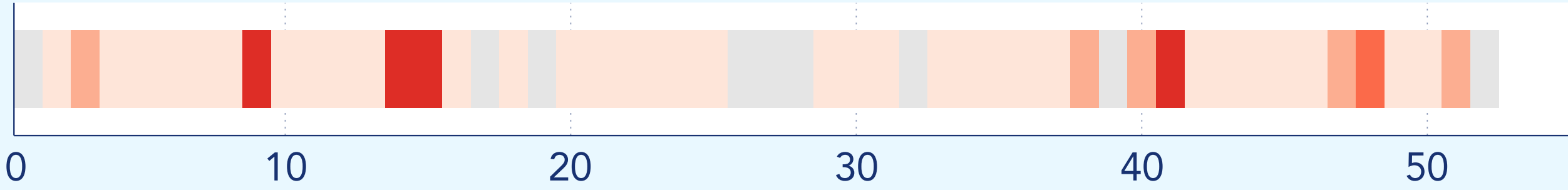
Eksempel:
Visualisering for én rapport

Resultater for temaet «Utskrivningsklare pasienter»

Starten av rapporten

Rødere farge = høyere relevans

Slutten av rapporten



Sidetall i rapporten

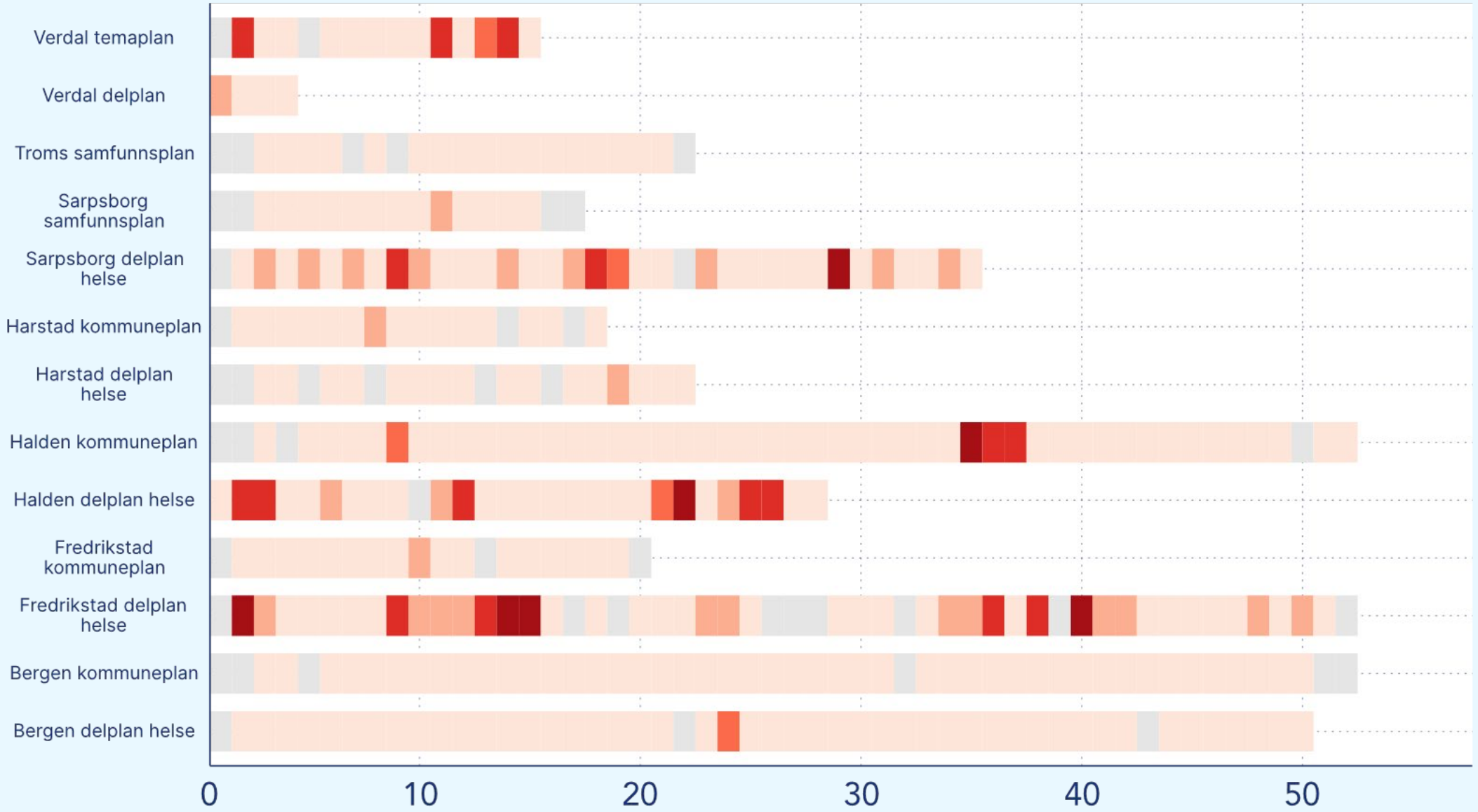
Ikke vurdert

Ikke i det hele tatt

I liten grad

I noen grad

I svært stor grad



Tema: Samarbeid med spesialisthelsetjenesten

Ikke vurdert

Ikke i det hele tatt

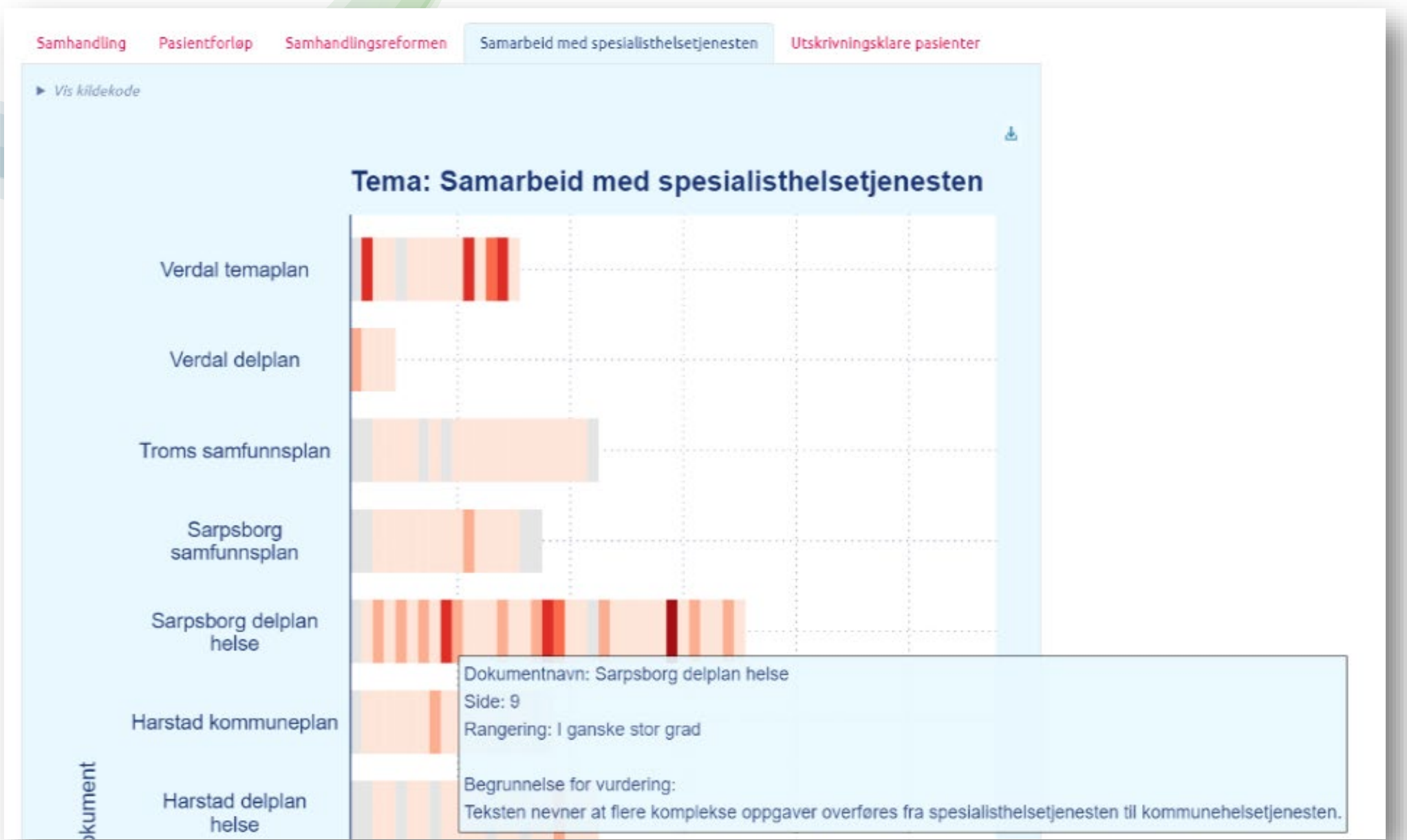
I liten grad

I noen grad

I ganske stor grad

I svært stor grad

Laget et interaktivt dokument for å vite hvor man bør begynne å lese



Men kan vi stole på resultatet?

Konklusjon: Ja, stort sett.

Språkmodellen finner alt det relevante innholdet, men også noe som er irrelevant.

Med andre ord: Flere falske positive enn falske negative, som er bra!

Eksempel 4:

Et dokument på 100 sider…

… som en 10 min. podcast?

Regjeringen.no

Søk

Søk

Tema Dokument Aktuelt Departement Regjering

Forsiden • Tema • Statlig forvaltning • IT-politikk

Del/tips Skriv ut

Ny nasjonal digitaliseringsstrategi

Med denne strategien setter vi retningen for det digitale Norge frem mot 2030 og for å gjøre Norge til det mest digitaliserte landet i verden.

Digitaliserings- og forvaltningsdepartementet

Norge skal bli verdens mest digitaliserte land

Forutsetningene for digitaliseringen

Samordning

Infrastruktur

Sikkerhet

Personvern

Kompetanse

Innsats på prioriterte områder

Data og datadeling

Kunstig intelligens

Grønn omstilling

Innovativt næringsliv

Inkludering

NotebookLM EXPERIMENT

Try NotebookLM

Do your best creating

NotebookLM is your personalized AI research assistant powered by Google's most capable model, Gemini 1.5 Pro.

Try NotebookLM

Collaborate with a virtual research assistant

When you upload the documents that are central to your projects, NotebookLM instantly becomes an expert in the information that matters most to you.

Ting jeg ikke rakk å snakke om…

- RAG pipelines for skreddersøm av LLMs
  - https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1
- Våre planer fremover (vi har store planer ☺)
- KI-revisjonen
- **… men takk for meg!**