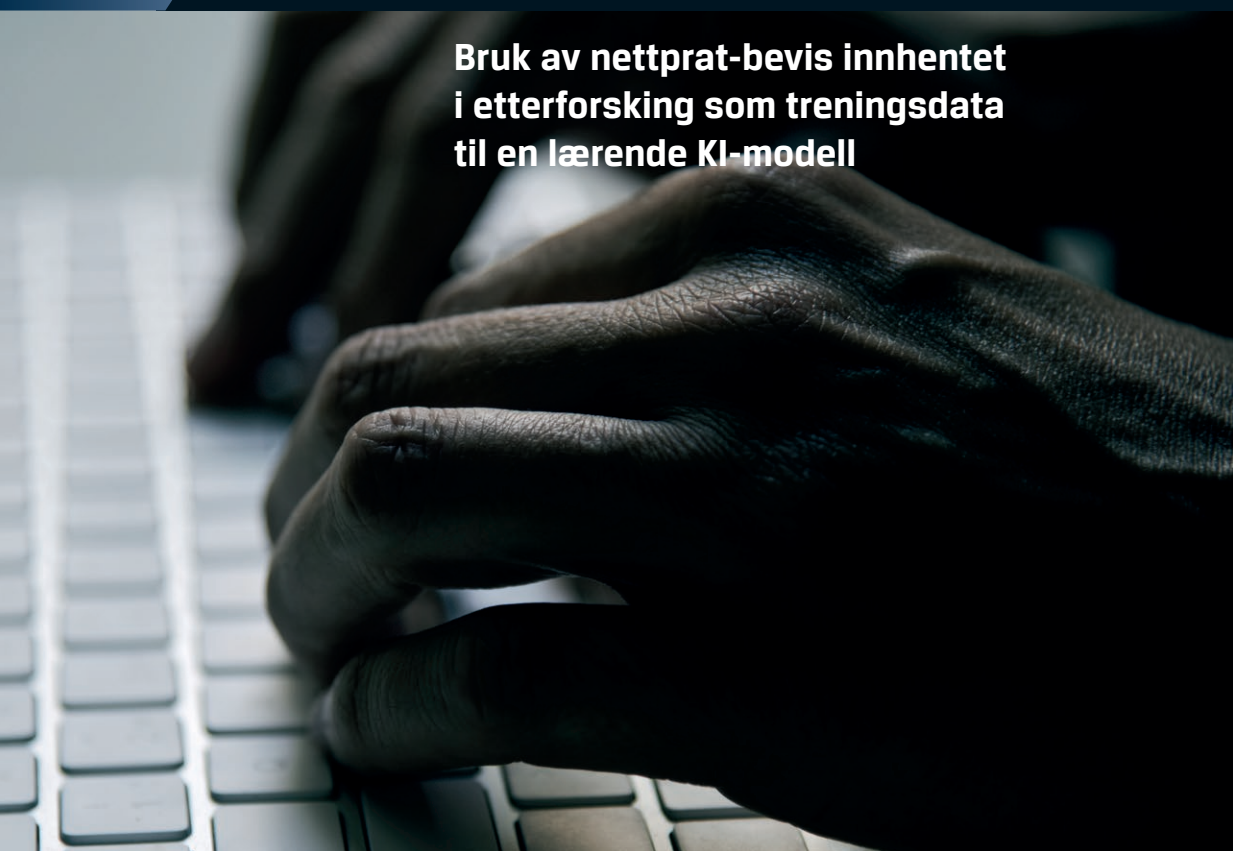
POLITIHØGSKOLEN

Nettprat-prosjektet

Bruk av nettprat-bevis innhentet i etterforsking som treningsdata til en lærende KI-modell



Inger Marie Sunde

Prosjektleder

Jørgen Bendiksen

Nina Sunde

Oslo, 15. august 2022

SAMMENDRAG

Prosjektet ble gjennomført i et samarbeid mellom Politihøgskolen og Trøndelag politidistrikt. Formålet var å undersøke om data i straffesaker om seksuelle overgrep mot barn var egnet som treningsdata for å utvikle et forebyggende verktøy basert på maskinlæring (kunstig intelligens). Verktøyet hadde i så fall behov for data i nettprat mellom overgriper og fornærmede (barnet), som finnes i straffesakene. Prosjektet har avdekket flere problemer som står i veien for å realisere dette formålet;

– Størstedelen av data innsamlet i slike straffesaker krever mye behandling for å gjøres maskinlesbare i et konsistent format med tanke på innmating i en lærende algoritme. Ressursbruken ved å gjøre dette må følgelig veies mot verdien av annen innsats som kan gjøres uten bruk av verktøyet.

– Forskjellig praksis mellom politidistriktene i håndteringen av data sikret som bevis gjør at datatilfanget varierer mellom distriktene, både med hensyn til fullstendighet og kvalitet. Dette svekker kvaliteten til det totale datatilfanget med tanke på bruk som treningsdata for en lærende algoritme.

– Datatilfanget i straffesaker er utvilsomt en verdifull ressurs som kan utnyttes for å utvikle innovative verktøy som kan hjelpe politiet i å forebygge alvorlige seksuelle overgrep mot barn. Imidlertid tar verken datasystemene eller den praktiske bruken av dem, hensyn til behov som må ivaretas for å kunne utnytte dataene som ressurs i forebyggende arbeid. For å kunne ha et egnet datagrunnlag til forebyggingsformålet må politiet tenke nytt om behandlingen av slike data. Kravene som stilles til 'high-risk' KI-systemer og til treningsdataene kvalitet i den foreslåtte europeiske forordningen om kunstig intelligens bør være veiledende.

SUMMARY

The chatlog-project is a collaborative project between the Norwegian Police University College (Politihøgskolen) and Trøndelag Police District. The purpose was to investigate whether data collected during the criminal investigation of cases concerning sexual abuse of children, were suitable as training data for developing a preventative tool based on machine-learning (artificial intelligence). For the tool to be trained and realised, there was a need for data in chat-conversations between the perpetrator and the victim (the child). Such data were to be found in chat-logs collected as evidence during the criminal investigation of the cases. The project aimed at collecting such data from cases in all the Norwegian police districts. However, several problems that hinder access to and use of data for this purpose were detected:

– The main part of the data collected and secured in the investigation of such cases, requires much preparation to be converted into a machine-readable format suitable as input for training the learning model. The resources for doing this therefore need to be balanced against the possible gain of alternative approaches (without the tool) against the problem.

– The amount of data available for training a preventative tool varies substantially between the police districts. The disparity is caused by difference in practices between the police districts, when it comes to the handling of data secured as evidence. This impairs the data sets with respect to quality and completeness, and the overall quality of the total amount of data available as training data for a learning model.

– The data collected as evidence in criminal cases is potentially a valuable resource that can be used for developing innovative tools that can assist the police in the prevention of serious sexual abuse of children. However, neither the computer systems, nor the practical use of them, cater for the needs involved when the purpose is to use

the data as a resource in machine-learning for crime prevention. In order to build data sets suitable for preventive purposes, the police must rethink how they secure, process and store data. To this end, attention should be paid to the conditions relating to 'high-risk' AI systems and training data, as outlined in the proposed European Artificial Intelligence Act (AIA).

Innhold

1 Prosjektets mandat og innretning

1.1 Mandat, bakgrunn og formål

Prosjektet har undersøkt om såkalte 'chattelogger' sikret som bevis i straffesaker om seksuelle overgrep mot barn, er egnet som treningsdata for å utvikle en maskinlæringsalgoritme til bruk i forebygging av nettovergrep mot barn. Bakgrunnen var at Politidirektoratet etter anmodning fra Justis- og beredskapsdepartementet, ba Politihøgskolen om å

> undersøke muligheten for å bruke chattelogger fra politiets etterforskinger til å utvikle forebyggende teknologi basert på maskinlæring.[1]

Oppdragsbrevet viser til et prosjektnotat basert på det forebyggende 'PrevBOT-konseptet', nærmere omtalt i punkt 1.4 nedenfor.[2] Oppdraget ble gitt i forbindelse med fremleggelse av Nasjonal strategi for samordnet innsats til forebygging og bekjempelse av internettrelaterte overgrep mot barn.[3] Det ble presisert at «Kompetanse fra Kripos og politidistriktene skal knyttes til arbeidet». Prosjektet har vært ledet av Politihøgskolen v/ professor Inger Marie Sunde.

I dialogen med Politidirektoratet høsten 2020 var det lagt opp til at både Kripos v/NC3 og Trøndelag politidistrikt skulle delta i prosjektet. På slutten av 2020 underrettet Kripos om at man likevel ikke hadde ressurser/kapasitet til å delta, heller ikke mot økonomisk kompensasjon.[4] Kripos har imidlertid, som ledd i sin alminnelige bistand til forskning

1 Jf. brev fra Politidirektoratet til Politihøgskolen 24. november 2020.

2 Prosjektnotat fra Politihøgskolen til Justisdepartementet, 27. oktober 2020.

3 Justis- og beredskapsdepartementet, Forebygging og bekjempelse av internettrelaterte overgrep mot barn. Nasjonal strategi for samordnet innsats (2021–2025). Strategi mot internettrelaterte overgrep mot barn (regjeringen.no)

4 I forbindelse med den nasjonale strategien ble det avsatt 1 million kroner til FoU i politiet.

som involverer politiets registre, bistått i kartleggingen av straffesaker som kan inneholde relevant materiale (se punkt 2.1.1).

Spesialetterforsker Jørgen Bendiksen ved Trøndelag politidistrikt har vært sentral i prosjektet. Høsten 2019 fullførte Bendiksen masteravhandlingen *'Automated detection of perpetrators in grooming conversations in Norwegian' ved NTNU.* [5] Avhandlingen viste lovende resultater for bruk av maskinlæring i analyse av chattelogger for å avdekke samtaler som ledet opp til (forsøk på) seksuelle overgrep mot barn. Analysen gjaldt logger på norsk fra norske straffesaker, slik også dette prosjektet gjelder. For Trøndelag politidistrikt føyer Nettprat-prosjektet seg inn i en innovativ satsing mot nettovergrep, blant annet gjennom SOBI-prosjektet i samarbeid med NTNU og St. Olavs Hospital.[6]

Det engelske uttrykket 'chat' som på norsk kan erstattes med 'nettprat', betyr direktemeldinger som foregår i tilnærmet sanntid mellom deltakerne, dvs. at samtaletjenesten legger opp til rask dialog mellom partene. Uttrykket 'chattelogger' får i tillegg frem at det er tale om samtaler som foreligger i skriftlig elektronisk form og følgelig er dokumenterbare. Selv om dette ikke fremkommer eksplisitt i uttrykket 'nettprat', har vi valgt likevel å holde oss til det norske uttrykket. Der det er nødvendig vil det bli presisert at det er tale om nettprat slik den foreligger i logger som er sikret i etterkant av samtalen.

1.2 Prosjektperiode, fritak for taushetsplikt, mm

Prosjektet hadde oppstartmøte mandag 3. august 2021, et halvt år senere enn opprinnelig forutsatt. Forsinkelsen skyldtes behovet for å få bekreftet finansiering av frikjøp av Bendiksen. Avtalen var at Bendiksen skulle frikjøpes i seks måneder for å innhente og analysere materialet. Tildelingsbrevet fra Justis- og beredskapsdepartementet til Politidirektoratet 17. mars 2021 omfattet Nettprat-prosjektet

5 Jørgen Bendiksen, Automated detection of perpetrators in grooming conversations in Norwegian, NTNU, 2019, masteravhandling.

6 Sektorsamarbeid om forskning på forebygging av seksuelle overgrep mot barn på internett (SOBI) - Institutt for psykisk helse - NTNU.

(kalt 'Teknologibasert forebygging av nettovergrep mot barn'), og nærmere sommeren bekreftet kontaktperson i Politidirektoratet at tildelingsbrevet til Politihøgskolen med tillegg og rettelser fra Politidirektoratet, ville inkludere Nettprat-prosjektet.[7] Politihøgskolen sendte deretter de nødvendige søknadene om fritak for taushetsplikt i forskning som beskrevet nedenfor. I oppstartmøtet 3. august ble det avklart at frikjøpet av Bendiksen først kunne skje fra 1. september, fordi han var midlertidig pålagt andre oppgaver på grunn av ferieavvikling ved politidistriktet. Frikjøpet gjaldt derfor ut februar 2022. Perioden ble skjøvet ut til medio mars fordi Bendiksen måtte bistå i noen straffesaker som alt var berammet. Deretter har det medgått tid til å ferdigstille rapporten.

Prosjektet hadde behov for logger over nettprat innhentet som bevismateriale i straffesaker. Siden slike opplysninger er taushetsbelagte, jf. politiregisterloven § 23, var gjennomføringen av prosjektet avhengig av at man fikk fritak fra taushetsplikten i medhold av § 33, som under visse betingelser åpner adgang til bruk av slike opplysninger i forskning. Riksadvokaten samtykket i brev 27. juli 2021 til at Bendiksen gis

> tilgang til det aktuelle materialet […], selv om dette til dels inneholder opplysninger av sensitiv karakter. Det legges vekt på at det kun er Bendiksen som skal ha tilgang til materialet, at det skal oppbevares skjermet i politinettet (Trøndelag politidistrikt) og at dataene skal anonymiseres før de benyttes videre.

Samtykket gjaldt straffesaker inntil fem år tilbake i tid, konkret saker hvor etterforsking ble åpnet 1. juli 2016 eller senere.[8]

I innhentingsfasen oppsto det behov for mer ressurser. Pensjonert politioverbetjent Ove Mule ved Trøndelag politidistrikt sa seg villig til å utføre arbeidet, og etter særskilt søknad ga Riksadvokaten samtykke

7 Rettelser og tillegg nr. 84 til resultatavtalene 2021- Forskningsmidler PHS, brev fra Politidirektoratet til Politihøgskolen, 1. oktober 2021.

8 Tidsrommet ble beregnet ut fra oppstart sommeren 2021, med skjæringspunkt ved halvåret 30/6-1/7.

til å utvide tilgangen til å omfatte Mule frem til utgangen av november 2021, jf. brev 13. oktober 2021.

For å identifisere straffesaker som kunne inneholde relevant materiale var det nødvendig først å ha et uttrekk fra STRASAK. Også tilgang til STRASAK for forskningsformål krever samtykke etter politiregisterloven § 33, og dette ble gitt av Politidirektoratet i brev 30. juli 2021. Det ble presisert at «selve uttrekket må skje etter nærmere avtale med Kripos».

1.3 Kompetansesammensetning

Prosjektet trengte kompetanse innen fenomenkunnskap, kriminalitetsforebyggende teori med digitale rom som kontekst, maskinlæring, og de rettslige rammene for politiets bruk av PrevBOT-teknologien i forebygging av nettovergrep. Kompetansebehovet ble dekket av spesialetterforsker Jørgen Bendiksen (maskinlæring), politioverbetjent/PhD-kandidat Nina Sunde ved Politihøgskolen (politietterforsking/kriminologi) og professor Inger Marie Sunde (rettslige rammer). Hver av dem har solid fenomenkunnskap.

1.4 Om PrevBOT-konseptet og behovet for data fra straffesaker

1.4.1 Hvordan PrevBOT er tenkt å fungere

'PrevBOT konseptet' nevnt i prosjektnotatet 27. oktober 2020, er beskrevet av Nina Sunde og Inger Marie Sunde i to vitenskapelige artikler. Artiklenes hovedtittel er '*Conceptualizing an AI based Police Robot for Preventing Online Child Sexual Exploitation and Abuse', med undertitlene:*

– 'Part 1 – The theoretical and technical foundations for PrevBOT', *Nordic Journal of Studies in Policing*, 2/2021.

– 'Part 2 – Legal Analysis of PrevBOT' (under fagfellevurdering).

Prosjektnotatet oppsummerte PrevBOT slik:

> PrevBOT er en intelligent chatbot som kan settes ut i offentlige pratekanaler (chatroom) på internett. Her vil den analysere åpent tilgjengelig informasjon (persondata) i form av chat-samtaler og opplysninger i brukerprofilen. På dette grunnlag kan den varsle om
>
> 1) Forekomst av seksualisert samtale mellom voksne og barn
>
> 2) Hvorvidt deltakere som opptrer som barn mest sannsynlig er voksen, og / eller med et annet kjønn enn oppgitt ('problematiske personer')
>
> 3) Tidligere domfelte som gjenopptar den ulovlige atferden på nett.

Funksjonenes formål ble begrunnet slik:

> PrevBOT kan dermed være et nyttig verktøy for å avdekke problematiske nettsteder/fora og personer som er spesielt risikable for barn […] Chatlogger og lagrede metadata kan også gi grunnlag for etterforsking, særlig dersom mistanken gjelder en serieovergriper.

I tillegg ble det forklart at PrevBOT kunne brukes til å intervenere overfor 'problematiske personer' ved å sende en forebyggende melding. PrevBOT-konseptet forutsetter menneskelig styring og kontroll, og agerer derfor ikke autonomt når det kommer til beslutning om hva som bør gjøres basert på prediksjonene som genereres. Dette skal utdypes i det følgende.

Artiklene konseptualiserer PrevBOT som et datasystem bestående av flere komponenter. *Input*-komponenten*/sensoren* er en chatbot som skal delta på ulike nettsteder rettet mot barn eller hvor barn har adgang. Chatbotens brukerprofil lages av operatøren (politibetjenten), f.eks. 'Lisa 12', 'Roger 14' eller annet. Chatboten er koblet til en *prosessor* med en *maskinlæringsalgoritme* (lærende modell) som analyserer nettpraten

chatboten observerer eller deltar i. Databehandlingen er basert på automatisert tekstanalyse (*NLP- Natural Language Processing*). Det er meningen at analysen skal utføres i sanntid.

Hovedformålet er å identifisere 'problematiske' nettsteder og personer, basert på forhåndsdefinerte risikoindikatorer. Risikoindikatorene lar seg avdekke ved analyse av de biometriske kjennetegn som kan trekkes ut av dataen, altså nettpraten mellom den potensielle overgriperen og et barn ('barnet' kan være PrevBOT med en brukerprofil som et barn, eller et annet barn som PrevBOT observerer at kontaktes av en voksen). De biometriske dataene er kjennetegn ved personens skrivestil, dvs. 'atferdsmessige egenskaper,' jf. definisjonen av 'biometrisk opplysning' i politiregisterloven § 2 nr. 16. Indikatorene gjelder en persons alder og kjønn, og samtaleinnholdets karakter, dvs. hvorvidt innleggene kan anses som seksualisert. Dersom seksualisert nettprat mellom voksen og barn avdekkes i tilknytning til et nettsted, kan det bli ansett som et 'problematisk sted' som bør prioriteres av politiet ved patruljering på nettet. I tillegg kan de medføre at en deltaker i den elektroniske samtalen anses å utgjøre en risiko for barn, noe artiklene kaller en 'problematisk person'. En person kan bli ansett som problematisk dersom vedkommende er voksen og gir uriktige opplysninger om alder og/eller kjønn i samtalen med et barn, eventuelt gir samtalen et seksuelt preg.

Analysen kan også ha som formål å identifisere personer som tidligere er domfelt for seksuelle overgrep mot barn, og som likevel kontakter barn på sosiale medier. Identifisering forutsetter at PrevBOT-systemet har en *referansedatabase* som inneholder domfeltes 'språklige fingeravtrykk' generert fra en tilstrekkelig mengde nettprat med vedkommende som deltaker.[9] Referansedatabasen vil være et identitetsregister på linje med slike som er etablert for DNA-profiler, foto og fingeravtrykk, jf. politiregisterloven §§ 12 og 13, jf. politiregisterforskriften kapittel 45 (DNA) og 46 (foto og fingeravtrykk). Tilstrekkelig samsvar (i henhold til en definert terskelverdi) mellom det språklige fingeravtrykket i nettpraten og et språklig fingeravtrykk i referansedatabasen, indikerer

9 Automatisert fremstilling av språklige fingeravtrykk er basert på den grenen av NLP som kalles 'Authorship Analysis' (PrevBOT Part I, kapittel 2.3, s. 5 og kapittel 4, s. 9-11).

at det er en person som tidligere er domfelt for seksuallovbrudd mot barn, som nettprater med et barn/PrevBOT.

Resultatet av maskinlæringsalgoritmens analyse er en *prediksjon*, dvs. at utdataene ('outputen') uttrykker sannsynlighet for at noe er tilfelle. Dette gjelder både om det er tale om å kategorisere nettprat som seksualisert, bestemme alder/kjønn til en deltaker i en samtale, eller identifisere en tidligere domfelt. Prediksjonen må vurderes av operatøren før man beslutter å iverksette tiltak på grunnlag av den.

Aktuelle personrettede tiltak er å sende en advarsel til en problematisk person, eller innlede etterforsking, dersom det er mistanke om at vedkommende er en serieovergriper. Etterforsking vil særlig være motivert av behovet for avklaring for å kunne stanse pågående overgrep mot barn. Det å sende en advarsel til den problematiske personen vil ha et forebyggende formål. For å kunne utføre denne funksjonen er PrevBOT utstyrt med en *standardisert meldingsfunksjon*.[10]

PrevBOT har *loggings- og lagringsfunksjoner* som sikrer elektroniske spor som kan være verdifulle i etterforsking for å spore opp en overgriper, men funksjonene er minst like viktige for å sikre kontrollmulighet (notoritet) og unngå feilbruk av systemet. Lagringsfunksjonen må settes opp med en korresponderende *funksjon for sletting* av unødvendige data, og data som er 'gått ut på tid ', jf. frister satt i politiregisterloven § 6 første ledd nr. 3, jf. § 50, og § 8.

Databehandlingen ved utvikling og bruk av PrevBOT er undergitt reglene i politiregisterlovgivningen. Bruken av PrevBOT suppleres videre med bestemmelser i politiloven når formålet er forebyggende (kategorisering, advarsel), og av straffeprosessuelle bestemmelser når formålet er etterforsking (identifisering, etterforsking).

Dataene som analyseres av PrevBOT er 'biometriske' i en vid betydning av ordet, fordi de refererer seg til en persons atferdsmessige egenskaper

10 Meldingsfunksjonen er beskrevet i PrevBOT Part I, kapittel 3.2, s. 9

ved deltakelse i nettprat. De er imidlertid ikke nødvendigvis 'biometriske opplysninger' slik dette defineres i politiregisterloven § 2 nr. 16. Denne definisjonen, som er basert på de likelydende definisjonene i GDPR artikkel 4(14) og Politidirektivet artikkel 3(13), snevrer inn de relevante dataene til å gjelde slike som 'muliggjør eller bekrefter en entydig identifikasjon' av personen.[11] Noe avhengig av hvor vidt man tolker 'muliggjør' innebærer dette at de biometriske dataene i vid forstand, bare er biometriske i politiregisterlovens mening når de behandles av PrevBOT for å *identifisere* en person (språklig fingeravtrykk), ikke når de gjelder kategorisering (alder, kjønn).

1.4.2 PrevBOT-artiklenes konklusjoner

Den første artikkelen beskriver PrevBOT-konseptet, teknologien og hva de forebyggende teoriene innebærer for forebygging utført av politiet i det digitale rom. Artikkelen konkluderer med at PrevBOT vil være en velegnet teknologi for å støtte politiets arbeid mot seksuelle overgrep på nett. Den andre artikkelen belyser de rettslige rammene for PrevBOT. Analysen gjelder metodens forhold til retten til privatliv og rettferdig rettergang (provokasjon og selvinkriminering), jf. EMK artikkel 8 og 6. Konklusjonen er at PrevBOT kan benyttes slik det forebyggende konseptet foreslår, innenfor velkjente skranker som beskytter mot provokasjon og selvinkriminering.

I tillegg behandles personopplysningsvernet, herunder bruk av biometriske opplysninger. Siden målgruppen er et internasjonalt publikum gjelder drøftelsen det europeiske Politidirektivet (2016/680), ikke spesifikt politiregisterloven. Formålet er å avklare om lovgiver på nasjonalt nivå har handlingsrom til å vedta lovgivning som gir adgang til å bruke PrevBOT. Konklusjonen er at Politidirektivet gir lovgiver slikt handlingsrom.

Endelig er PrevBOT analysert i forhold til AIA (Rådets kompromisstekst).[12] Konklusjonen er at PrevBOT klassifiseres som et 'high-risk'

11 GDPR og Politidirektivet er henholdsvis EU forordning 2016/679 og direktiv 2016/680.

12 Se fn. 15.

KI-system. Det innebærer at systemet kan utvikles og brukes, forutsatt at kravene til kvalitet, risikovurdering og kontroll overholdes, jf. AIA del III kapittel 2.

Oppsummert er konklusjonen at de rettslige rammene ikke er til hinder for at norsk politi kan benytte PrevBOT til å forebygge nettovergrep mot barn. Det må imidlertid tas hensyn til at forhandlingene rundt AIA fremdeles pågår, hvor både definisjonen av 'biometriske data' og 'AI-system' stadig er gjenstand for diskusjon. Dette behandles derfor litt nærmere i punktet nedenfor. Uansett utfall er det på det rene at det vil være behov for noe spesialregulering i politiregisterlovgivningen.

1.4.3 Betydningen av den europeiske forordningen om kunstig intelligens (AIA)

PrevBOT vil bli undergitt den foreslåtte europeiske forordningen om kunstig intelligens *(Artificial Intelligence Act ('AIA'))* dersom denne blir vedtatt.13 AIA inneholder en definisjon av 'biometriske data' i artikkel 3(33) som i Kommisjonens forslag var likelydende med de nevnte i GDPR artikkel 14(4) og Politidirektivet artikkel 3(13).[14] Rådet har imidlertid foreslått å slette vilkåret om at dataene må muliggjøre eller bekrefte personens unike identitet.[15] Hvis dette blir stående vil også databehandlingen som inngår i PrevBOTs kategoriseringsfunksjon anses å gjelde biometriske data i AIAs forstand. Europaparlamentet ved den sammensatte saksførende komiteen IMCO/LIBE, har på sin side foreslått at AIA beholder en definisjon av biometriske data som er likelydende med GDPR og Politidirektivets, sml. Kommisjonens forslag.[16] I tillegg foreslås en ny definisjon av '*biometrics-based data*' som

13 European Commission: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts, 21 April 2021, COM(2021) 206 final.

14 *Ibid.*

15 European Council: Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 14278/21, 29 November 2021. Presidency of the Council of the European Union. Compromise Text.

16 European Parliament - IMCO/LIBE: Draft report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM2021/0206 – C9-0146/2021 – 2021/0106(COD)). Committee on the Internal Market and Consumer Protection, Committee on Civil Liberties, Justice and Home Affairs, 20 April 2022.

kan utledes ved teknisk behandling av en persons fysiske, fysiologiske eller atferdsmessige '*signaler*', som '*may or may not*' muliggjøre eller bekrefte unik identifikasjon av personen.[17] Som det fremgår gjelder den supplerende definisjonen databehandling av 'signaler' til forskjell fra 'egenskaper', og fjerner vilkåret om at dataene må være identifiserende.

Nytten av den supplerende definisjonen kan umiddelbart være vanskelig å få øye på, men uansett peker den i samme retning som Rådets forslag, nemlig at forståelsen av 'biometriske data' gjøres videre enn det som følger av GDPR og Politidirektivet, med den følge at også formål utover det å identifisere en person omfattes. Videre synes det å være enighet blant EU-organene om at AIA både bør regulere biometriske identifikasjonssystemer (definert i artikkel 3(36)) og biometriske kategoriseringssystemer (artikkel 3(35)). Rettslig sett må PrevBOT følgelig være å anse som et sammensatt system som må ta hensyn til regler for politiets bruk av KI-systemer som kan kategorisere og identifisere personer.

I skrivende stund er definisjonene fortsatt gjenstand for diskusjon og forhandlinger på EU-nivå, og behandles ikke nærmere her. Et siste forhold skal imidlertid nevnes, og det er at definisjonen av 'AI-system' i artikkel 3(1), jf. Annex I særlig bokstav c, er så vid at det for politiets del kan resultere i at bruk av *automatisert beslutningsstøtte generelt* faller inn under AIA s regime, såfremt beslutningen retter seg mot en person eller gruppe personer. Det vises til AIA artikkel 5(1)(d) om sanntids biometrisk identifikasjon på offentlig sted; Annex III(1) om bruk av biometriske identifikasjonssystemer annet enn på offentlig sted; og, Annex III(6) som nevner spesifikke politimessige anvendelser av KI rettet mot personer. Mens politiets behandling av personopplysninger for politimessige formål bestandig må ha rettslig grunnlag i politiregisterlovgivningen (jf. Politidirektivet), medfører AIA at databehandling som er lovlig etter disse reglene, likevel kan bli forbudt. Databehandlingen vil uansett bli ansett som 'high-risk' og dermed undergitt strenge krav til hvordan datasystemene utvikles,

17 *Ibid, endring 64, s. 49.*

rigges og kontrolleres, foruten krav til operatørenes kompetanse og ferdigheter. Siden AIA ikke er endelig utformet, kan det vanskelig sies noe sikkert om den rettslige risikoen for bruk av PrevBOT, annet enn at systemet vil anses som 'high-risk'.[18]

1.4.4 Behov for treningsdata fra norske straffesaker

En maskinlæringsalgoritme er autonom i den forstand at den kan anvende 'kunnskap' eller 'erfaring' ervervet fra tidligere tilfeller til å løse nye tilfeller av lignende art. Kunnskapen/erfaringen som er en forutsetning for at PrevBOT-algoritmen skal kunne tas i bruk på sosiale medier, må erverves gjennom trening på relevante data i tilstrekkelig mengde og kvalitet. For PrevBOT er relevante treningsdata logger over nettprat, dvs. direktemeldingene mellom potensiell voksen overgriper og et barn. Treningsdataene må oppfylle flere krav: Kommunikasjonen må for det første være *reell*. Dette er nødvendig for å unngå feilkilder. Et eksempel på en feilkilde er hvor voksne som utgir seg for å være barn på internett for å tiltrekke seg og avsløre overgripere, og overdriver den antatte stilen til et barn, f.eks. for mange emojis, for mye bruk av slang o.l.[19]

For det andre må det stilles krav til *språk*. Selv om PrevBOT som konsept kan brukes i alle land, må systemet trenes opp på språket som brukes i nettprat mellom overgriper og barn i det aktuelle landet. Som verktøy må PrevBOT derfor utvikles spesifikt for landet som ønsker å bruke det. I norske straffesaker om seksuelle overgrep mot barn går nettpraten hovedsakelig på norsk. For å ha verdi for norsk politi må PrevBOT følgelig trenes opp til å 'forstå' norsk. Kravene innebærer at PrevBOT må trenes opp på logger over nettprat fra norske straffesaker.

For det tredje må treningsdataene være *tidsriktige*. Språk og sjargong i nettprat utvikler seg over tid, og PrevBOT må derfor trenes opp

18 Europeiske politiledere har i felles erklæring fra mai 2022 sterkt advart mot en 'blanket regulation' av KI-systemer brukt av politiet. Se Joint Declaration of the European Police Chiefs as approved by the European Police Chiefs during their informal meeting in Berlin on 24 May 2022, pkt. 3.

19 Se PrevBOT Part I kapittel 4.3 s. 11-12

på et materiale som er representativt for nettprat på sosiale medier i tidsrommet den skal brukes.

1.5 Avgrensning

Oppdraget gjaldt som nevnt å «undersøke muligheten for å bruke chattelogger fra politiets etterforskinger til å utvikle forebyggende teknologi basert på maskinlæring». Dette ble presisert til

> å undersøke om norske straffesaker inneholder data i form av chattelogger på norsk, som utgjør et tilstrekkelig og adekvat datagrunnlag for en maskinlæringsalgoritme til bruk for forebygging av nettovergrep mot barn.[20]

Nettprat-prosjektet er med andre ord en undersøkelse av den faktiske muligheten for å kunne trene opp en lærende modell i henhold til PrevBOT-konseptet. Undersøkelsen har betydning for muligheten til å realisere prosessoren i sammenheng med chatboten, og referansedatabasen for språklige fingeravtrykk. De øvrige komponentene som meldingsfunksjonen, loggings-, lagrings- og slettingsmekanismer omfattes ikke av prosjektet.

20 Oppdragsbrevet, fn. 1.

2. Metode og datamateriale

2.1 Identifisering og innhenting av datamateriale

Kapitlet redegjør for innhentingen av nettprat-logger i norske straffesaker. Dette ble utført i flere trinn, først ved identifisering av potensielt relevante saker gjennom uttrekk fra STRASAK, som deretter måtte innhentes og gjennomgås for å avklare hvilke av dem som inneholder relevant materiale. I tillegg ble det sendt et brev direkte til alle politidistriktene med anmodning om å få tilsendt nettprat-loggene i relevante straffesaker.

2.1.1 Identifisering av straffesaker fra STRASAK

STRASAK-uttrekket ble gjort med bistand fra rådgiver og IKT-prosess-forvalter for straffesakssystemet Asle Moe ved Kripos. Grunnlaget for uttrekket var statistikkgrupper og modusbeskrivelser som var generert og kvalitetssikret i samarbeid med Asle Moe og leder for seksjon for seksuallovbrudd, Emil H. Kofoed ved Kripos. Modusbeskrivelser er koder som gir supplerende informasjon for å få frem ulike kategorier innenfor av statistikkgruppene. I overensstemmelse med tillatelsen fra riksadvokaten ble uttrekket begrenset til straffesaker fem år tilbake i tid.

a) Følgende statistikkgrupper ble brukt:

Straffeloven 1902

- 1402 SEKSUELL OMGANG MED BARN U/14 ÅR (§ 195)
- 1402 UTUKTIG OMGANG MED BARN U/14 ÅR (§ 195)
- 1403 SEKSUELL OMGANG MED BARN U/10 ÅR (§ 195)
- 1403 UTUKTIG OMGANG MED BARN U/10 ÅR (§ 195)
- 1404 SEKSUELL OMGANG MED BARN U/16 ÅR (§ 196)
- 1404 UTUKTIG OMGANG MED BARN U/16 ÅR (§ 196)
- 1405 PORNOGRAFI, SKRIFTER/FILM/VIDEO MV (§204)
- 1405 UTUKTIGE SKRIFTER/FILM/VIDEO M.V (§ 211)

- 1421 PORNOGRAFI, SKRIFTER/BILDER/VIDEO MV VIA DATASYST. (§204)
- 1422 SEKSUELT KRENKENDE/UANSTENDIG ADFERD VIA DATASYSTEMER (§201)
- 1429 Grooming - Bygge opp tillit for å utnytte barn seksuelt

Straffeloven 2005

- 1460 Voldtekt av barn under 14 år
- 1461 Voldtekt til samleie av barn under 14 år
- 1462 Grov voldtekt av barn under 14 år
- 1463 Forsøk på voldtekt av barn under 14 år
- 1464 Seksuell omgang med barn mellom 14 og 16 år
- 1465 Grov seksuell omgang mv. med barn mellom 14 og 16 år
- 1466 Seksuell handling med barn under 16 år
- 1467 Seksuelt krenkende atferd mv. overfor barn u/16 år
- 1468 Avtale om møte for å begå seksuelt overgrep
- 1469 Kjøp av seksuelle tjenester fra mindreårige
- 1470 Fremvis/still av seksu overgr mot barn el seksualiserer barn

b) Følgende modusbeskrivelser ble brukt:

- Mobil/håndholdt datamaskin
- Programvare/skadelig programva
- Ved bruk av datamaskin/PC
- Ved bruk av datanettverk
- Ved bruk av datasystem

Fremgangsmåten for utvelgelse av data tilsvarer den som Bendiksen benyttet i sin masteravhandling.[21] Relevante saker var straffesaker med forhold som gjaldt seksuell nettprat mellom identifisert gjerningsperson og fornærmede barn. Relevante data fra straffesakene var nettsamtalene, med andre ord direktemeldinger mellom identifisert gjerningsperson og barn.

21 Jørgen Bendiksen, Automated detection of perpetrators in grooming conversations in Norwegian, NTNU, 2019, masteravhandling, kapittel 3.

Følgende egenskaper var et krav til dataene:

- Nettpraten måtte være toveissamtaler, hvor samtalepartene var fornærmet og gjerningsperson.

- Nettpraten måtte være på norsk. Dialekt var ingen begrensning.

- Fornærmede måtte være mindreårig etter norsk lov, dvs. under 18 år.[22] Dersom forhold i saken gjaldt seksuell omgang eller handling med barn (inkl. å få barn til å utføre handlinger som svarer til seksuell omgang med seg selv), måtte fornærmede være under den seksuelle lavalderen, dvs. under 16 år.[23] Forhold om seksualiserte bilder eller videoer gjelder også personer som har fylt 16 år og 17 år.[24]

- Gjerningspersonen måtte være myndig etter norsk lov. Med myndige personer menes personer som har fylt 18 år og som ikke helt eller delvis er fratatt den rettslige handleevnen.[25]

- Gjerningspersonens formål måtte være minst én av følgende:

    - Fysisk møtes for å voldta, ha seksuell omgang eller utføre seksuelle handlinger med fornærmede.

    - Få fornærmede til å utføre seksuelle handlinger med seg selv.

    - Få fornærmede til å dele (1) seksualiserte bilder eller videoer av seg selv; eller (2) bilder eller videoer hvor fornærmede utfører handlinger som svarer til seksuell omgang med seg selv.

22 Det vises til FNs barnekonvensjon artikkel 1, j. menneskerettsloven § 2 nr. 4 og vergemålsloven § 8.

23 Straffeloven §§ 302 og 303 rammer (grov) seksuell omgang med barn mellom 14 og 16 år. Straffeloven § 304 rammer seksuell handling med barn under 16 år. For seksuell omgang med barn under 14 år er det straffeloven § 299 som er relevant. Etter denne bestemmelsen anses overgrepet bestandig som voldtekt.

24 Straffeloven § 311 rammer anskaffelse og annen befatning med overgrepsbilder og seksualiserte fremstillinger av barn under 18 år.

25 Vergemålsloven § 2 tredje ledd.

Hver enkelt straffesak i uttrekket ble undersøkt i politiets saksbehandlingssystem for straffesaker (BL). Dette innebar manuell gjennomgang av mer enn 2.250 straffesaker. Anmeldelsen, normalt straffesakens dokument nr. 02, ble vanligvis brukt som utgangspunkt for relevansvurderingen, eventuelt støttet av øvrige dokumenter i straffesaken. Undersøkelsen rettet seg mot informasjon om gjerningsperson og fornærmede, herunder fødselsdatoer for å beregne alder (voksen/barn), og kjønn. Videre ble det undersøkt om straffesaken inneholdt relevante data, dvs. nettpratlogger, eksempelvis fra Snapchat eller Skype. Dersom det på noe trinn i undersøkelsen viste seg at straffesaken ikke var relevant, ble undersøkelsen avsluttet.

Omtrent 23% av straffesakene viste seg å være relevante for prosjektet, og disse ble registrert i et regneark fordelt på de enkelte politidistrikt. Da flere av de relevante straffesakene inngikk i større straffesakskompleks, ble kun en tredel undersøkt i mer detalj (dvs. litt over syv prosent av de opprinnelige 2.250 straffesakene). Omtrent 1% av straffesakene i uttrekket var utilgjengelige fordi de var skjermet for innsyn.

Gjennomgangen viste at STRASAK-uttrekket inneholdt flere hundre straffesaker som *ikke* var relevante for prosjektet. Eksempler følger:

- Minst 200 straffesaker gjaldt befatning med overgrepsmateriale (allerede tilgjengelige/kjente bilder/videoer på Internett), uten at det hadde vært noen direkte kontakt med de avbildede.

- Minst 100 av straffesakene om ulovlig deling av bilder og videoer gjaldt tilfeller hvor en mottaker av et seksualisert bilde/video som i utgangspunktet var frivillig delt, selv videredeler bildet/videoen videre med andre, uten at den opprinnelige avsenderen har kjennskap til det og samtykket. Bildet/videoen i seg selv er ikke nødvendigvis ulovlig etter straffeloven § 311, men viderespredningen er en overtredelse av privatlivets fred og retten til eget bilde, jf. straffeloven § 267 og åndsverkloven § 104.I disse straffesakene var de involverte som regel mindreårige og Snapchat ble mye brukt.

- Minst 50 straffesaker gjaldt seksuell nettprat mellom voksne, hvor samtaleemnet var barn. Det ble diskutert hva de kunne tenkt seg å gjøre med barn, hva de hadde gjort med barn tidligere og hvilke barn de hadde fysisk tilgang til, ofte basert på fantasier. I disse nettsamtalene ble det typisk delt overgrepsmateriale av barn, gjerne som et utgangspunkt for diskusjonen.

- Mange av straffesakene gjaldt forhold om seksuell nettprat mellom to voksne, hvor den ene voksne utga seg for å være et barn. Dette var enten seksuelle rollespill eller at formålet til den ene var å avsløre potensielle overgripere. Det sistnevnte utgjorde noen titalls straffesaker i uttrekket, hvor anmelderen var Barnas trygghet. Ifølge organisasjonens nettside arbeider Barnas trygghet forebyggende med å redusere vold og overgrep mot barn, hovedsakelig med å avsløre personer som søker seksuell kontakt med mindreårige på Internett.[26]

2.1.2 Innhenting av saker fra politidistriktene

Vi sendte også brev til hvert politidistrikt med anmodning om å dele informasjon om og eventuelle data fra, straffesaker som fremkom i STRASAK-uttrekket for det aktuelle politidistriktet, og eventuelt andre saker som inneholdt seksualisert nettprat mellom voksen og barn som ikke var omfattet av STRASAK-uttrekket. Straffesakene måtte ha nettprat som viste barnelokking eller *grooming*, hvor språket hovedsakelig var på norsk. Partene skulle være identifisert, fornærmede skulle være et barn (under 18 år) og gjerningspersonen en voksen. Verken nettprat hvor voksne utga seg for å være barn, eller omtalte overgrep mot barn, var av interesse. I tillegg ba vi om å få en kontaktperson for eventuelle oppfølgninger. Anmodningen gjaldt, i overensstemmelse med tillatelsen fra riksadvokaten, straffesaker inntil fem år tilbake i tid.

26 https://barnastrygghet.no.

2.1.3 Innhenting av data fra sakstilfanget i uttrekket og fra politidistriktene

Datainnhentingen skjedde med utgangspunkt i de sakene i STRASAK-uttrekket som var vurdert som relevante. I tillegg mottok vi noe data direkte fra politidistriktene som følge av den rettede henvendelsen.

Data fra relevante saker i STRASAK-uttrekket

Undersøkelsen av sakene i STRASAK-uttrekket omfattet som nevnt hvorvidt de inneholdt relevant nettprat. I den forbindelse ble det fortløpende registrert hvilke dokumenter i saken som dataene var en del av, eller beskrevet i. Det viste seg at dataene kunne være en del av en rapport (bilder av nettsamtalene, uttrekk fra en database m.m.), et eget dokument (normalt regneark) eller filer i Mediebanken (multimediafiler lagt til i saken). Dette innebar at mange av nettsamtalene som var inntatt i sakens dokumenter var utdrag fra nettprat, og ikke den fullstendige nettpratsamtalen. Dette har betydning for hvor egnet dataene er til å kunne brukes som treningsdata, noe som er nærmere omhandlet i punkt 3.3. Til slutt ble det undersøkt om straffesaken inneholdt informasjon som sannsynliggjorde at det fantes (mer) relevant data hos det lokale politidistriktet (m.a.o. utenfor BL). Dette kunne resultere i nye henvendelser til politidistriktet.

Data i sakene tilsendt fra politidistriktene

Direkteanmodningene til politidistriktene resulterte i få tilbakemeldinger, og flere av de innsendte sakene viste seg ikke å være relevante for prosjektet. De dataene som ble gjort tilgjengelige etter slike henvendelser var i hovedsak fra straffesaker med mange fornærmede i større straffesakskompleks med én gjerningsperson. Enkelte av straffesakene var en del av større prosjekter. Det er gjerne mye relevant data i de store prosjektene, men dataene er kun fra én gjerningsperson. Avhengig av hva en skal bruke dataene til, må en tilpasse mengden data en kan bruke i maskinlæringen. En stor mengde data fra en gjerningsperson vil gi et bias (skjevhet) som kan føre til dårlig treffsikkerhet dersom formålet er å finne andre gjerningspersoner. Er formålet derimot å identifisere om denne gjerningspersonen har

gjenopptatt sine kriminelle aktiviteter, det vil si seksualisert nettprat med barn, vil slike data være egnet til trening av modellen.

2.2 Datamateriale

Dette kapitlet redegjør for materialet som ble innhentet i prosjektet. Det skilles mellom data på dokumentnivå og aggregert nivå.

2.2.1 Data på dokumentnivå

Dataene (nettprat-loggene) som var ansett som relevante, ble undersøkt i mer detalj. Det er tale om undersøkelser på individuelt dokumentnivå, og på aggregert nivå hvor dataenes mengde og kvalitet samlet sett er gjenstand for vurdering. Dette kapitlet gjelder dataene undersøkt på dokumentnivå i den enkelte straffesaken, mens punkt 2.2.2 beskriver undersøkelser på aggregert nivå.

Hver straffesak i STRASAK-uttrekket som ble vurdert som relevant, ble gjenstand for nærmere undersøkelser utført i BL. Den relevante informasjonen ble trukket ut og registrert i et eget Excel-dokument. Først ble det avklart hvilken tjeneste eller tjenester som gjerningspersonen og fornærmede hadde brukt i nettpraten. Tjenestene ble normalt gjenkjent på bilder i dokumentene, men informasjonen kunne også være en del av beskrivelsene i dokumentene. Videre ble det avklart hvem som var hvem i nettpraten, dvs. gjerningsperson, fornærmet eller andre. Dette er viktig med tanke på opplæring av en modell (maskinlæring). For eksempel er det ikke relevant for PrevBOT å trenes opp med data fra uidentifiserte gjerningspersoner, siden det da er usikkert om vedkommende er en voksen. En utfordring med de ufullstendige nettsamtale var at det ikke alltid var åpenbart hvem partene var. Det viste seg blant annet at flere straffesaker inneholdt nettsamtaler med andre involverte, som fornærmedes forelder i samtale med gjerningspersonen, og fornærmede i samtale med et vitne. Undersøkelsen tok følgelig sikte på å utelukke nettprat med tredjeperson slik at man bare satt igjen med samtale mellom gjerningsperson og fornærmede.

Ytterligere ble det beregnet hvor mange meldinger nettpraten inneholdt. Antallet ble et estimat utregnet i henhold til følgende eksempel: Et dokument har 20 sider og hver side har ett bilde av en Snapchat nettsamtale. Hvert bilde viser åtte til ti meldinger. Med meldinger menes i denne sammenhengen tekst og uttrykksikon *(emojis/emoticons*) skrevet og sendt av en av partene i nettsamtalen. Filoverføringer (for eksempel bilde eller video sendt mellom fornærmet og gjerningsperson) og systemmeldinger ble ikke medregnet. Tjue sider multiplisert med ni meldinger gir et dokument med 180 meldinger. Antallet registrerte meldinger ble beregnet samlet for partene i nettpraten (gjerningsperson og fornærmede), ikke per part.

Videre ble datakvaliteten på nettpraten vurdert. Dette er nærmere beskrevet på aggregert nivå i punkt 3.3.5.

Det ble også vurdert om nettpraten representerte første kontakt mellom partene. Dette ble gjort for å undersøke i hvilket omfang disse sakene startet på åpne rom for nettprat, hvor PrevBOT er ment til å brukes. I flere dokumenter fremgikk det at nettpraten var den første kontakten mellom partene *på den aktuelle tjenesten*, men at de åpenbart på forhånd visste hvem den andre var, eller til og med kjente hverandre fra før. Det var eksempelvis tilfeller hvor partene hadde møttes på Omegle (dette er en uegnet tjeneste med tanke på nettsamtaler med samme person over tid, da tjenesten velger tilfeldig ut brukere som nettprater sammen – anonymt og uten registrering), og senere fortsatte nettpraten på en annen tjeneste. Det var også tilfeller hvor fornærmede og gjerningspersonen kjente hverandre fra en organisert fritidsaktivitet, hvor partene var henholdsvis utøver og trener. I de tilfellene hvor dokumentene inneholdt nettsamtaler fra flere tjenester, ble disse tjenestene registrert hver for seg i undersøkelsen vår.

Til slutt ble dato for første og siste melding i nettsamtalene registrert. Dette ble gjort for å få oversikt over hvor gamle de innhentede nettsamtalene var, og for å kunne utelukke materiale som anses å være utdatert for trening av PrevBOT. Dersom dokumentet ikke inneholdt noen datoer, ble dato for opprettelse av dokumentet i

saksbehandlingssystemet brukt. Dersom verken dato for første eller siste melding var i dokumentet, men en dato ble funnet på en annen melding, ble denne datoen brukt, siden dette ble vurdert å være mer presist enn å bruke datoen for opprettelse av dokumentet. Det er flere eksempler på at straffesaksdokumenter om nettsamtaler er opprettet flere år i ettertid av kontakten mellom gjerningspersonen og fornærmede. Videre har datoer som åpenbart ikke passet innholdet blitt utelatt. Et eksempel er at et dokument opprettet i de senere årene inneholder nettprat fra MSN Messenger, en tjeneste som for brukere utenfor Kina ble lagt ned i 2013.[27] Dersom den eneste datoen som er tilgjengelig gjelder opprettelsen av etterforskingsdokumentet, er opplysningen utelatt.

STRASAK-uttrekket skilte ikke på om straffesaken var en hovedsak eller en vedleggsak i saksbehandlingssystemet, noe som gjorde gjennomgangen av uttrekket unødvendig uoversiktlig. Saksnummer i STRASAK-uttrekket var enten en del av et sakskompleks (hovedsak eller vedleggsak) eller en enkeltstående straffesak. Dersom et saksnummer tilhørte et sakskompleks, ble alle saksnummer i sakskomplekset også gjennomgått. Dermed kunne en i ettertid ende opp med å treffe på flere titalls saksnummer som allerede var gjennomgått. Etter å ha gjennomgått et straffesakskompleks måtte en derfor søke opp alle saksnummer i komplekset i STRASAK-uttrekket (for å merke de som gjennomgått). Innenfor den aktuelle problematikken består straffesakskompleksene gjerne av mange vedleggsaker, fordi gjerningspersoner utnytter anonymiteten internett gir og går etter mange potensielle fornærmede samtidig. Et eksempel er en straffesak ved Øst politidistrikt hvor en mann i 40-årene var siktet for å ha seksuelt misbrukt 263 barn i alderen ni til 16 år via internett.[28] Mannen ble dømt til fengsel i 13 år og 6 måneder for nettovergrep mot 256 barn over en periode på fire år, samt inndragning av utstyr og oppreisning til de fornærmede barna. [29]

27 'MSN Messenger to end after 15 years', BBC News/Tech, 29 august 2014. https://www.bbc.com/news/technology-28987797.

28 Øyvind Gustavsen, 'Mann i 40-årene siktet for overgrep mot 263 barn', NRK, 5. mars 2019. https://www.nrk.no/osloogviken/mann-i-40-arene-siktet-for-overgrep-mot-263-barn-1.14452458.

29 TNERO 2018-182729.

2.2.2 Data på aggregert nivå

Generelt gjelder det at et datasystems evne til å produsere pålitelige resultater blant annet avhenger av kvaliteten på dataene som tilføres systemet. For at en lærende algoritme som PrevBOT skal gi stabile og pålitelige prediksjoner må den trenes opp på data av god kvalitet. Kvalitetsbegrepet i denne sammenheng er svært sammensatt, reflektert i AIA artikkel 10 nr. 3 om 'high-risk' KI-systemer. Her sies det at treningsdataene må være relevante, representative, feilfrie og komplette, samt ha egenskaper som gjør dem egnet for statistisk analyse. Analysen i punkt 3.3 går nærmere inn på hvordan dataene prosjektet har avdekket står seg i forhold til disse kravene.

I inneværende punkt er dataene brutt ned på kategorier som har betydning for representativitet og fullstendighet. I punkt 3.3 diskuteres dataenes egnethet som inndata ved opptrening av den lærende algoritmen. Egnetheten har særlig betydning for kravene til at dataene må være feilfrie og fullstendige.

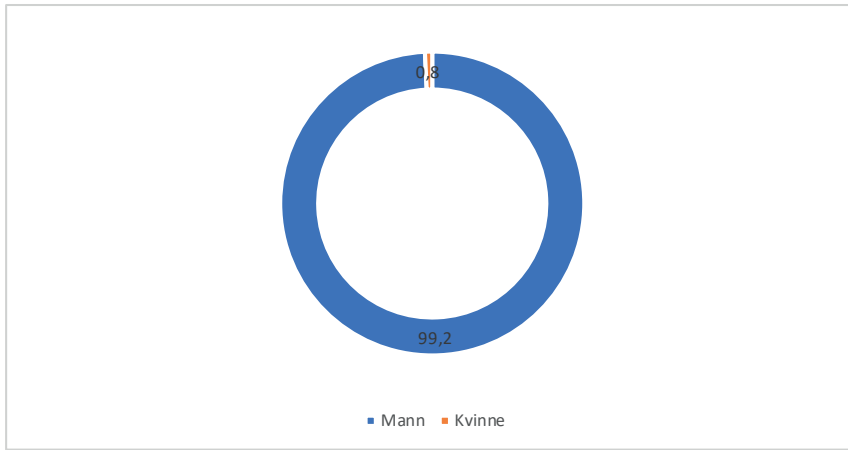
Antall meldinger

Anmodningen til politidistriktene ga som nevnt liten respons, men to av de relevante sakene (sakskompleksene) inneholdt svært mye data. Det ene straffesakskomplekset inneholdt over 500.000 meldinger, og gjaldt flere hundre fornærmede og én gjerningsperson. Det andre straffesakskomplekset inneholdt mer enn 16.000 meldinger og gjaldt et tyvetalls fornærmede og én gjerningsperson. I disse straffesakskompleksene var nettpraten primært fra 2014, 2015 og 2016.

Fra BL ble over 150 dokumenter undersøkt, noe som ga et beregnet antall på 120.341 direktemeldinger. I det videre er det kun data fra BL som er brukt i beskrivelsene.

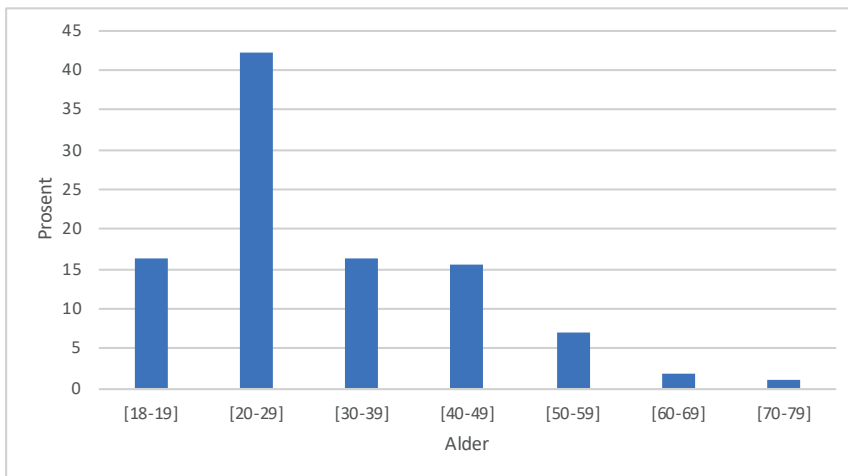
De relevante straffesakene hadde til sammen nesten 140 unike gjerningspersoner, hvorav nesten alle var menn.[30] Menn utgjør mer enn 99%, mens kvinner utgjør mindre enn 1% (se figur 1).

30 Antallet unike gjerningspersoner ble avdekket ved å sammenholde gjerningspersonens navn med navn i de øvrige straffesakene i uttrekket.



Figur 1: fordeling av kjønn for gjerningspersoner

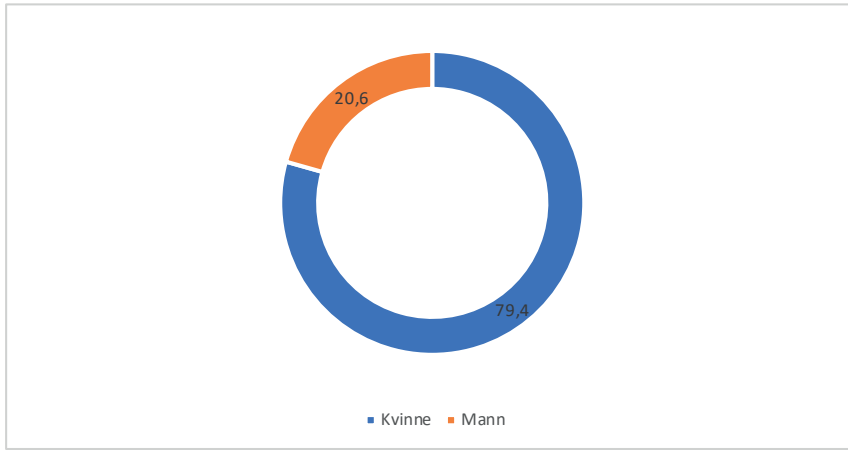
Figur 2 viser gjerningspersonenes aldersfordeling. Yngste gjerningsperson var 18 år, den eldste 73 år.[31] Gjennomsnittsalder var 31,9 år (median: 28,1 år).



Figur 2: fordeling av alder for gjerningspersoner

31 Utgangspunkt for alder på gjerningsperson var antall fylte år.

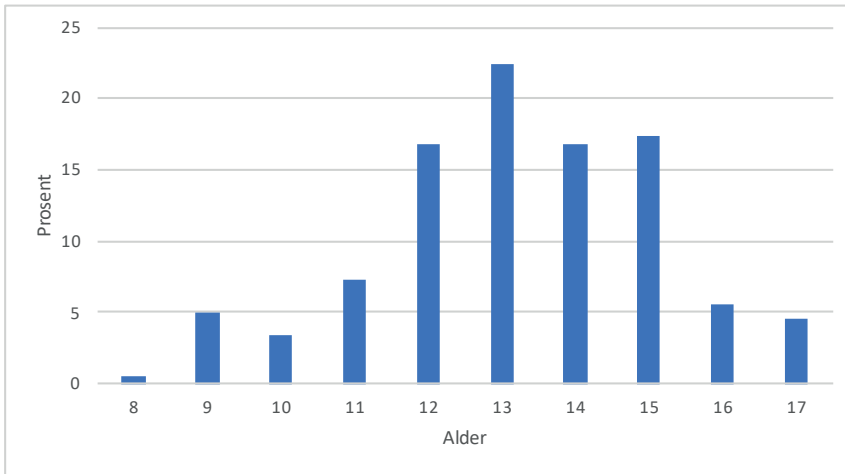
Figur 3 viser fordelingen av kjønn for de fornærmede. De fleste fornærmede (79%) var kvinner. 21% av de fornærmede var menn.



Figur 3: fordeling av kjønn for fornærmede

Figur 4 viser aldersfordelingen på de fornærmede. Yngste fornærmede var åtte år, mens den eldste var 17 år.[32] Gjennomsnittsalder: 13,7 år (median: 13,7 år).

32 Utgangspunktet for alder på fornærmede var antall fylte år.

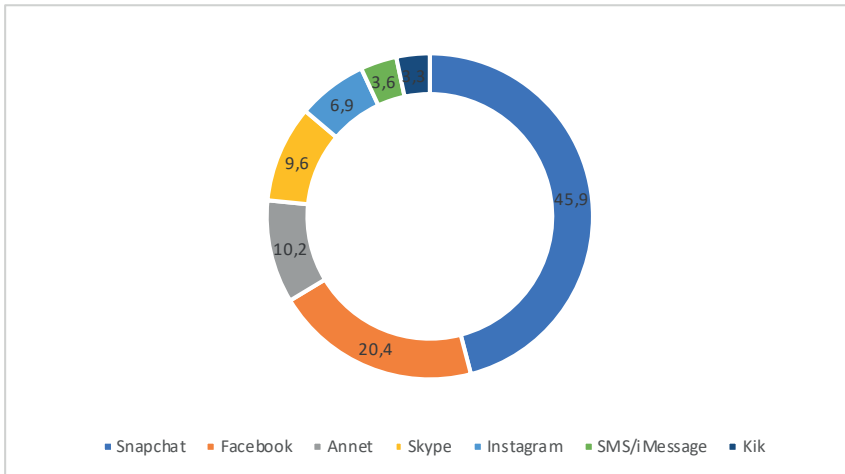


Figur 4: fordeling av alder for fornærmede

Data om nettprat-tjenestene

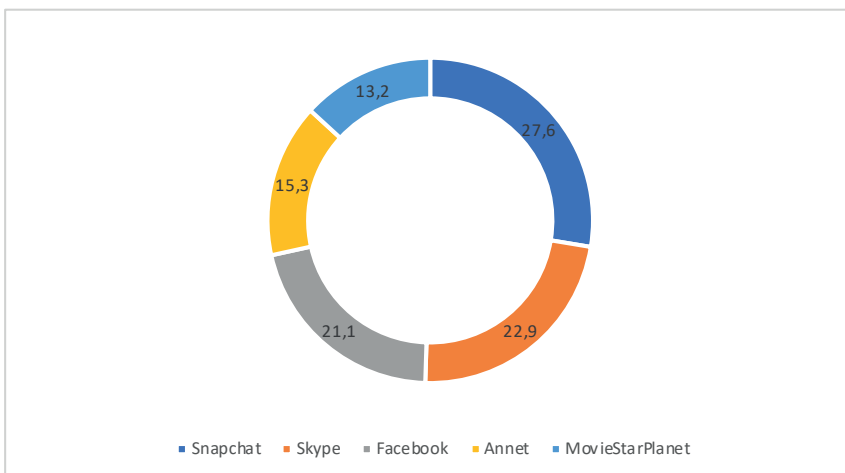
Nettpraten i dokumentene fra sakene i uttrekket stammet fra i alt 17 tjenester. Figur 5 viser dokumenter fordelt på tjenester, mens figur 6 viser nettprat-meldinger fordelt på tjenester.

Som vist i figur 5 inneholdt 46% av dokumentene nettprat fra Snapchat, og 20% nettprat fra Facebook. Kategorien 'Annet' omfatter tjenester som var i ti dokumenter eller færre. Dette gjaldt chat.no, Discord, Gaysir, Grindr, Momio, MovieStarPlanet, Omegle, Steam, sugardaters.no, TikTok/Musikal.ly og WhatsApp.



Figur 5: Dokumenter med nettprat-meldinger fordelt på tjenester

Figur 6 viser at de fleste nettprat-meldingene stammet fra tjenesten Snapchat (28%), Skype (23%) og Facebook (21%). Kategorien 'Annet' utgjør 15%, og omfatter tjenester som hadde 5.000 meldinger eller færre, dvs. chat.no, Discord, Gaysir, Grindr, Instagram, Kik, Momio, Omegle, SMS/iMessage, Steam, sugardaters.no, TikTok/Musikal.ly og WhatsApp.

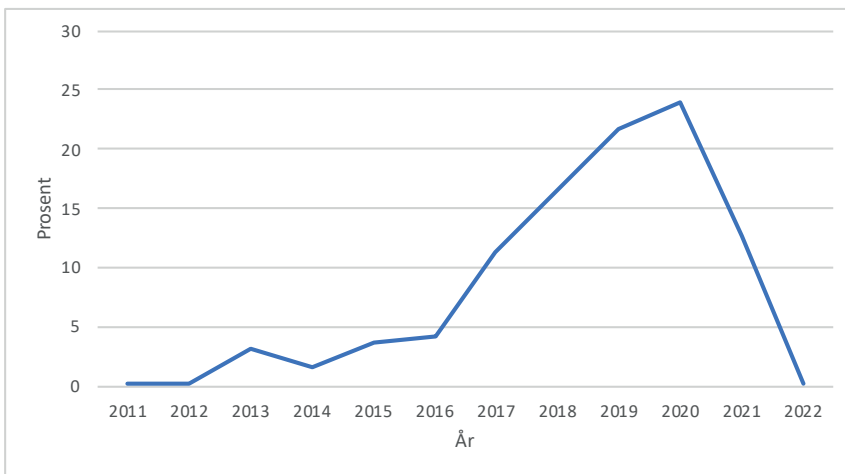


Figur 6: Nettprat-meldinger fordelt på tjenester

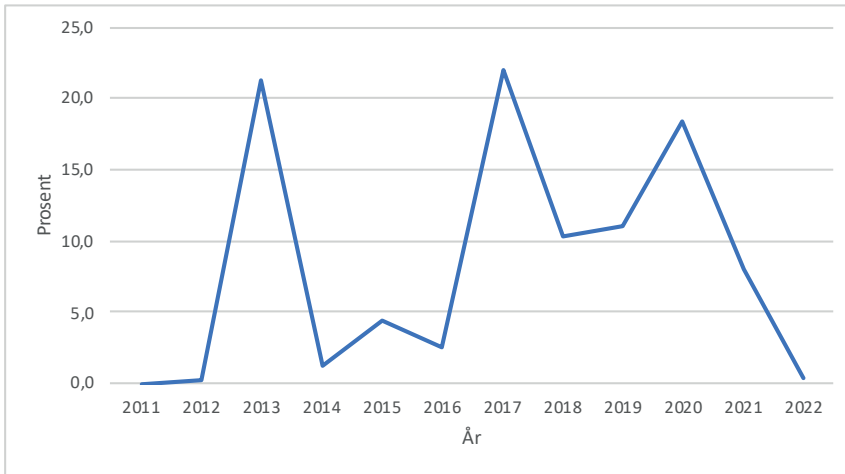
Som vist i figur 5 er Snapchat omtalt i nesten halvparten av dokumentene (46%). Meldingene fra denne tjenesten utgjør imidlertid bare litt over en fjerdedel av meldingstotalen (28%) (figur 6). Samtidig viser analysen at den mindre brukte tjenesten Skype, utgjør nesten en fjerdedel av meldingstotalen. Meldingstotalen for Skype er på 22,9% (figur 6), til tross for at tjenesten er omtalt i mindre enn 10% av dokumentene (figur 5). Disse forholdene har sammenheng med hvordan tjenestene fungerer, noe som er nærmere beskrevet under 'Prossesseringstid og kvalitet' i 3.3.5.

Nedenfor presenteres alder på det innhentede nattprat-materialet på henholdsvis dokumentnivå og aggregert nivå. Selv om de innhentede straffesakene var avgrenset til de siste 5 år, inneholdt de gjerne nettprat-meldinger fra lengre tilbake i tid. Figur 7 viser at den tidligste nettsamtalen i dokumentene daterte seg til august 2011, og den siste til januar 2022. Nesten tre firedeler (75%) av dokumentene hadde nettsamtaler med første melding fra 2018, 2019, 2020 eller 2021.

Figur 8 viser antall meldinger fordelt på datoer for nettsamtalens start. Nesten halvparten av meldingene var fra nettsamtaler med første melding fra 2018, 2019, 2020 eller 2021. Mer én av fem meldinger var fra nettsamtaler med første melding (fra) i 2013 – det samme gjelder 2017.



Figur 7: Antall dokumenter fordelt på datoer for nettsamtalenes oppstart



Figur 8: Antall meldinger fordelt på datoer for nettsamtalenes oppstart

Omtrent en tredel av dokumentene hadde nettsamtaler vurdert til å være fra første kontakt mellom partene. Dette har, som beskrevet i punkt 2.2.1, betydning for kartleggingen av om samtalene startet på åpne fora hvor PrevBOT er ment til å kunne brukes.

3. Analyse og diskusjon

3.1 Innhenting av data

3.1.1 Identifisering av aktuelle straffesaker

Den største utfordringen i prosjektet har dreid seg om å identifisere saker som inneholder nettprat-logger. Utfordringen henger sammen med hvordan saker registreres og systemenes søkefunksjonalitet. Ved søk i STRASAK med tanke på utvikling av PrevBOT, fremsto alle saker som inneholdt forhold registrert på bestemmelser i sedelighetskapitlet om krenkelser av mindreårige, som aktuelle. Dette ga totalt 21.416 saker. Bruk av modusbeskrivelsene (se punkt 2.1.1) var derfor en mulig teknikk for å avgrense tilfanget til saker som inneholdt nettprat. Det viste seg imidlertid at det er stor variasjon i hvorvidt og hvordan moduskoder benyttes. Siden vi ikke har undersøkt seksuallovbruddsaker som mangler relevant modusbeskrivelse, har prosjektet gått glipp av saker med nettprat-logger hvor modusbeskrivelse var utelatt.[33] Usikkerheten som knytter seg til avgrensningen som ble gjort, innebærer at den reelle datamengden som er relevant for PrevBOT må antas å være større enn den som ble avdekket. Hvor mye større har vi ikke grunnlag for å uttale oss om. Dette kan være en mangel ved datagrunnlaget som har betydning for alle kvalitetskriteriene i AIA artikkel 10 nr. 3.

Videre undersøkte vi muligheten for å søke etter aktuelt innhold direkte i straffesakene i BL, men slik søkefunksjonalitet finnes ikke. De aktuelle sakene fra STRASAK ble fordelt pr politidistrikt, og vi rettet en henvendelse til politidistriktene om bistand til å identifisere aktuelle saker. Det var til dels utfordrende å få pekt ut en kontaktperson som kunne bistå oss i arbeidet i det enkelte politidistrikt, og stor variasjon i hvor mye det enkelte politidistrikt bisto i identifiseringen av aktuelle saker og utlevering av relevante data.

33 Det var nødvendig å avgrense sakstilfanget fordi det innen prosjektets rammer ikke lot seg gjøre å undersøke alle sakene i totalantallet.

3.1.2 Tilgjengelighet

Den andre store utfordringen handlet om tilgang til nettprat-loggene, og har sammenheng med rutinene for lagring og sletting. Når de aktuelle sakene var identifisert, viste det seg å herske ulik praksis for om og hvordan nettprat-loggene ble oppbevart. Noen ganger ble nettprat-loggen i sin helhet lagt inn i straffesakens dokumenter i BL, mens det andre ganger kun fantes utdrag fra en lengre nettprat-logg inkludert i straffesaksdokumentet. I så fall lå resten av loggen på speilfilen som var lagret på politiets beslagsnett, utenfor sakens dokumenter. Siden fritaket fra taushetsplikt var begrenset til å gjelde nettprat-logger tilgjengelige i BL, avhang muligheten vår for å få tilgang til logger i sikringsfilen av om politidistriktet kunne avsette ressurser til å hente dem ut for overlevering til prosjektet. Når politidistrikter ga uttrykk for at de ikke hadde ressurser til å gjøre dette, var disse nettprat-loggene utenfor prosjektets rekkevidde. Videre hadde lagring i sikringsfilen direkte betydning for tilgjengeligheten av nettprat-logger i avgjorte saker, siden speilfiler i beslagsnettet da skal slettes i tråd med lokal straffesaksinstruks. Det betyr at nettprat-logger som ikke er lagt inn i dokumenter som ligger i BL, blir slettet når saken er avgjort. Følgelig går dataene tapt med tanke på muligheten for bruk i maskinlæring.

Utfordringene medførte at ressursene i prosjektet i stor grad måtte brukes til manuell gjennomgang av BL for å identifisere saker og hente ut nettprat-logger. Dersom prosjektet hadde hatt tilgang til nettprat-logger liggende i speilfil på beslagsnettet, ville datatilfanget ikke bare vært større, men dataene som ville blitt tilført prosjektet ville også vært maskinlesbare og følgelig mer egnet som inndata til den lærende algoritmen. Det vises til redegjørelsen i punkt 3.3.5.

3.1.3 Mangelfull sentral koordinering

Underveis i prosjektet ble vi gjort oppmerksom på flere tilgrensende prosjekter, og noen politidistrikt reiste spørsmål om Nettprat-prosjektet var eller burde være en del av disse prosjektene, f.eks. et AI-prosjekt i Sør-Vest PD i samarbeid med Anzyz Technologies i Grimstad; NTNU

Biometrics Laboratory's AiBA-prosjekt i samarbeid med Innlandet PD;[34] og det nevnte SOBI prosjektet hvor Trøndelag PD samarbeider med NTNUs Institutt for psykisk helse og St. Olavs Hospital.

Politidirektoratet foretar i dag ingen koordinering av politiets forskningsprosjekter og vi erfarte dermed at prosjekter med tilgrensende, tildels overlappende, forskningsformål ikke "finner hverandre." Dette er uheldig, fordi det kan lede til unødvendig belastning på politidistriktene når de må bruke ressurser på å framskaffe de samme dataene flere ganger. Politiet taper i tillegg synergier som kunne vært oppnådd ved tidlig samarbeid om prosjektinnretning, datadeling, mv.

Videre reiser vi spørsmål om håndteringen av FoU-partnerskap i prosjekter som har en kommersiell side. Når forskningsformålet gir behov for å utvikle ny teknologi, vil politiet kunne se seg tjent med å samarbeide med eksterne FoU-miljøer for å skaffe seg dette. For politiet vil den primære interessen bestå i å få teknologi som kan bidra til politiets oppgaveløsning. Dette behovet bør imidlertid vurderes i et langsiktig perspektiv som ikke bare setter verdi på den umiddelbare gevinsten ved å få tilgang på et nytt verktøy, men også setter krav til vedlikehold og videreutvikling av teknologien, og til at politiet tilføres kompetanse som sikrer noen grad av robusthet og uavhengighet av den eksterne parten.

Uten klare avtaler risikerer politiet at deltakelsen (og ressursbruken) primært gir den eksterne aktøren tilgang på problemstillinger og politidata som ellers hadde vært vanskelig tilgjengelig, uten at politiet selv tilegner seg varige verdier. Etter vårt skjønn er også dette en problemstilling som fortjener oppmerksomhet fra Politidirektoratets side. Det bør være et siktemål at spørsmål om partnerskap i prosjekter med en kommersiell side ikke kun blir behandlet på distriktsnivå for det enkelte prosjektet, men ut fra formål, hensyn og krav bestemt på sentralt nivå.

34 AiBA er/skal bli et kommersielt produkt: About us – Aiba.

3.2 Datamengde

En utfordring identifisert gjennom Nettprat-prosjektet er at formålet om forebygging av seksuelle nettovergrep på grunnlag av kunnskapsbasert politiarbeid, kommer i konflikt med straffesaksformålet. Forebygging av seksuelle overgrep mot barn basert på maskinlæring er avhengig av *store datamengder*. Straffesaksformålet derimot innebærer at dataene som innhentes må være nødvendige og forholdsmessige for å belyse tiltalespørsmålet. Innhentingen styres altså av et formål om å *begrense* datamengdene, samt *å slette* data når straffesaksformålet er oppfylt. Det er kun sakens dokumenter i BL som oppbevares i henhold til regler for arkivering.

Vi stiller derfor spørsmål om politiet er rigget for å ivareta formålet om kunnskapsbasert forebygging av seksuelle overgrep, når man er henvist til å basere seg på databehandlingen som skjer ved ivaretakelse av straffesaksformålet. For å kunne ha et egnet datagrunnlag til forebyggingsformålet må politiet tenke nytt om behandlingen av slike data.

3.3 Dataenes egnethet som treningsdata

3.3.1 Innledning

I det følgende redegjøres det for de innhentede dataenes egnethet som treningsdata for en maskinlæringsalgoritme. Først beskrives behovet for at dataene er maskinlesbare. Deretter beskrives behandlingen av databeslag i straffesak, samt de ulike kravene til datakvalitet i straffesak vs. maskinlæring. Videre diskuteres utfordringer som oppstår når formålet med datainnhentingen endres fra bevissikring i etterforsking, til å skaffe inndata for å trene opp en lærende modell. Til slutt følger en oppsummering av hvorvidt de innhentede dataene i dette prosjektet er egnet til å utvikle PrevBOT.

3.3.2 Treningsdataene må være maskinlesbare

For å kunne brukes som inndata må treningsdataene være lesbare for datamaskinen. De innsamlede dataenes egnethet for dette formålet er

vurdert ut fra hvor enkelt eller komplisert det er å mate den lærende algoritmen med dem. Som nevnt er denne forståelsen av 'egnethet' en side av et kvalitetskrav til treningsdata som er mer omfattende og sammensatt.

Det er store forskjeller på egnetheten til dataene i sakene som ble sendt til prosjektet direkte fra politidistriktene, og dataene som ble avdekket ved gjennomgang av saker i STRASAK-uttrekket. Sakene sendt fra politidistriktene inneholdt hele den relevante datamengden, i form av f.eks. sikringsfiler eller loggfiler som inneholder nettprat. Disse dataene er maskinlesbare og kan enkelt behandles for maskinlæringsformål. I tillegg er de gjerne komplette, dvs. at de omfatter alle direktesamtaler i beslaget. Sakene hvor dataene måtte hentes ut fra BL har flere begrensninger som skyldes at BL ikke inneholder de beslaglagte dataene i sin helhet, og heller ikke opplyser om og hvor i politiets systemer databeslaget eventuelt er lagret. Dokumentene i BL inneholder typisk utdrag fra nettprat-samtalene for å vise modus og ulike bevismessige forhold. For maskinlæringsformål er disse dataene ofte ufullstendige, og i tillegg vanskelige å bruke på grunn av måten de er bearbeidet på.

3.3.3 Behandling av beslaglagte data i en straffesak

I en straffesak går beslaglagte data normalt gjennom flere steg, først som ubearbeidede data (ofte omtalt som rådata), og deretter to nivåer av bearbeidede data. Data avdekket i dette prosjektet har vært fra alle stegene.

- Ubearbeidede data: Med ubearbeidede data menes i denne sammenhengen data som ennå ikke er bearbeidet av politiet på noen måte. Et eksempel kan være programdata fra Skype, som lagres lokalt på enheten programmet brukes på, typisk på gjerningspersonens eller fornærmedes smarttelefon. Deler av disse dataene er en database som brukes til å opprettholde et arkiv med informasjon om blant annet brukerkontoer og nettsamtaler.

- Bearbeidede data (1): Med bearbeidede data på det første nivået menes resultatet av uttrekk, strukturering og formatering av de

ubearbeidede dataene. Bearbeiding av data (1) inngår som en del av politiets etterforsking. Et praktisk eksempel er nettprat trukket ut fra Skypes database på gjerningspersonens eller fornærmedes smarttelefon, automatisk eller manuelt, som gjøres klart for gjennomgang ved å bli lagret i et regneark hvor data er delt opp i kolonner og rader.

- Bearbeidede data (2): Med bearbeidede data på det andre nivået menes den videre håndtering av bearbeidede data på det første nivået, for å kunne trekke logiske konklusjoner om en gitt problemstilling. Bearbeiding av data på nivå to inngår som del av politiets etterforsking eller gjennomføringen av straffesaker i retten. Bearbeidete data på nivå to vil fremgå i rapport i BL, f.eks. en rapport som viser den delen av nettpraten (sikret fra Skype på gjerningspersonens eller fornærmedes smarttelefon, jf. nivå én) som anses å ha bevisverdi.

3.3.4 Kravet til datakvalitet i straffesak vs. maskinlæring og forebygging

I utgangspunktet stiller både straffesakens behov og behov knyttet til utvikling av maskinlæring til bruk i forebygging ('ML/forebygging'), krav om god datakvalitet. Faktorene som påvirker kvaliteten er imidlertid forskjellige. I straffesaken handler kvalitet om å opplyse sakens omstendigheter og straffbarhetsvilkårene med nødvendig grad av relevant og pålitelig informasjon. For å utvikle forebyggende verktøy basert på maskinlæring handler kvalitet om at det innhentede datagrunnlaget er *relevant, representativt, feilfritt, komplett og har egenskaper som gjør det egnet for statistisk analyse*. Vi har her basert oss på vilkårene oppstilt i AIA artikkel 10 nr. 3. Siden forslaget til AIA er gjenstand for forhandlinger mellom Rådet og Europaparlamentet, kan de nevnte vilkårene komme til å bli endret. Blant annet har IMCO/LIBE foreslått å supplere kravet til representativitet med at dataene også skal være oppdaterte, og å modifisere kravet om at dataene må være komplette med at dette 'så langt som mulig' ('as complete as possible.').[35]

35 IMCO/LIBE (fn. 16), endring 96, s. 62.

Både straffesaksformålet og ML/forebyggingsformålet stiller krav til dataenes *relevans*. Hva som anses som relevant vil være forskjellig under de ulike formålene. I en straffesak vil relevansen vurderes opp mot straffbarhetsvilkårene og de øvrige etterforskingsformålene, jf. straffeprosessloven § 226. I ML/forebygging vil relevansen handle om hvorvidt datamaterialet samsvarer med problemet teknologien skal løse. Det er derfor viktig kun å inkludere samtaler som faller innenfor et klart definert formål, som her er seksualiserte samtaler mellom voksen og barn, og f.eks. utelukke seksuelle samtaler mellom to voksne personer.

Representativitet er i statistisk sammenheng et uttrykk for hvorvidt kjennetegnene ved et utvalg korresponderer med tilsvarende kjennetegn i populasjonen. Representativitet er ikke et kvalitetsmål i straffesaken, siden det ikke tas sikte på å generalisere, men snarere si noe sikkert om et konkret straffbart forhold. Representativitet er imidlertid av stor betydning for kvaliteten når det gjelder bruk som treningsdata for ML/forebyggingsformål. I lys av dette prosjektet vil det handle om hvorvidt datasettet som ML modellen trenes opp med, representerer måten nettprat foregår på. Representativiteten kan svekkes over tid, blant annet fordi sjargong og symbolbruk (f.eks. emojis) i tekstbaserte samtaler kan endre seg. Dette må kompenseres gjennom testing, kontroll og korrigering.

At et datasett er *feilfritt og komplett* (så langt som mulig) handler om at dataene som er representert er riktige og at ingen vesentlige data som skulle har vært der, ikke er inkludert eller har falt ut. Dette er viktig faktorer både for å oppnå straffesaksformålet og for ML/forebyggingsformål. I straffesaken handler det om en tilstrekkelig opplysning av saken basert på pålitelig informasjon. I ML/forebygging handler det om å unngå systematiske skjevheter i modellens output på grunn av feil eller mangler i datasettet som utgjør treningsgrunnlaget. Lav datakvalitet øker sjansen for feil i datasettet gjennom feiltolkning av data. Dette kan f.eks. skje når dataene finnes i bilder tatt med mobilkamera av tekst på en smarttelefon, som skannes i, eller skrives

ut til PDF, og legges inn i straffesaken.36 Iblant kan teksten være i så lav oppløsning at den åpner for å kunne tolkes på ulike måter (se punkt 3.3.5). Komplette data handler i dette prosjektet om hvorvidt samtaleloggene i sin helhet er inkludert eller om kun korte tekstutdrag blir representert i datasettet. Vi har avdekket at det hersker svært ulik praksis i politidistriktene omkring hvilke deler av samtaleloggene som inkluderes i straffesakens dokumenter, og som dermed ble tilgjengelig som datagrunnlag i dette prosjektet.

Vilkåret om at datasettet skal ha *egenskaper som gjør det egnet for statistisk analyse*, innebærer at datasettet har egenskaper ('features') som kan tjene som variabler i en statistisk analyse, og at variablene reflekterer eller indikerer det som søkes målt. Det kan i noen tilfeller være aktuelt å bruke statistiske metoder for å belyse straffesaksformålet, for eksempel beregne andelen seksualiserte bilder på en beslaglagt datamaskin som er i kategoriene lovlig pornografi vs. ulovlige bilder (bilder som rammes av straffeloven § 311). For ML/forebyggingsformål må egenskapene korrespondere med formålet som skal oppnås ved bruk av modellen. I dette prosjektet handler egenskapene primært om at datasettet må kunne representere seksualisert prat, og deltakernes alder og kjønn.

Oppsummert kan vi si at selv om det er høye krav til kvalitet både når formålet er etterforsking av straffbare forhold og ML/forebygging, er faktorene/indikatorene for hva som er god kvalitet noe forskjellige for disse formålene. Vi kan derfor ikke forutsette at innsamling av data til straffesaksformål uten videre ivaretar nødvendig kvalitet for ML/forebyggingsformål slik at de *direkte* kan brukes til dette.

Prosjektet har i tillegg avdekket sikringsmåter og behandling av bevis som *forringer datakvalitet* som i utgangspunktet var god. Sikring av digitale spor til etterforskingsformål skal søke å bevare og dokumentere bevisets *autentisitet* (at det er hva det framstår å være), *integritet* (bevare meningsinnhold), *fullstendighet* (ivareta alle relevante data for beviset, samt bevisets kontekst) og *ubrutt beviskjede* (chain of custody/

36 PDF er et bildeformat.

audit trail – dvs. notoritet over hvem som til enhver tid har behandlet beviset i straffesaken). Prosjektet har avdekket store ulikheter i hvordan informasjonen sikres, behandles og dokumenteres – og at manglende kunnskap trolig er årsaken til behandling som forringer datakvalitet som i utgangspunktet var god. For ML/forebyggingsformål medfører slik behandling at data blir uegnet eller av en så dårlig kvalitet at det vil kreve omfattende og tidkrevende bearbeiding for å kunne anvendes som treningsdata til en ML modell.

3.3.5 Bruk av etterforskingsdata i maskinlæring

Problemer og feilkilder forbundet med formålsendringen

Når formålet endres fra bevissikring i etterforsking til å gjelde bruk av dataene for å trene opp en lærende modell i maskinlæring, viser det seg at det kan være svært tidkrevende å arbeide med bearbeidede data. For å trene opp og teste en modell er det viktig med konsistent formatering av datasettet. Datasettet lages av de innsamlede dataene og dersom disse er samlet inn fra forskjellige kilder, må man avsette tid og ressurser for å få dataene konsekvent skrevet. Hvis kilden til og med er uegnet, dvs. gir data som ikke er maskinelt lesbare, må dataene skrives inn manuelt i datasettet. I BL har vi avdekket rapporter som dokumenterer nettprat med fotografier, dvs. fotografi av et skjermbilde som viser nettpraten, f.eks. skjermen på gjerningspersonens eller fornærmedes smarttelefon. Slik bevissikring kan gi bilder hvor teksten er av dårlig kvalitet og vanskelig å lese. Den er følgelig vanskelig og ikke minst tidkrevende, å tilføre datasettet. Dette øker risikoen for feil, en risiko som øker jo mer bearbeidet dataene er. Det er grunnleggende at bearbeidete data må gjenspeile de opprinnelige ubearbeidede dataene, men for å avklare dette må man finne ut hva som er gjort med dataene og om de er blitt behandlet på riktig måte. I dette ligger det også spørsmål om etterforskerne som har bearbeidet dataene hadde de rette forutsetningene og kunnskapen til å gjøre det.

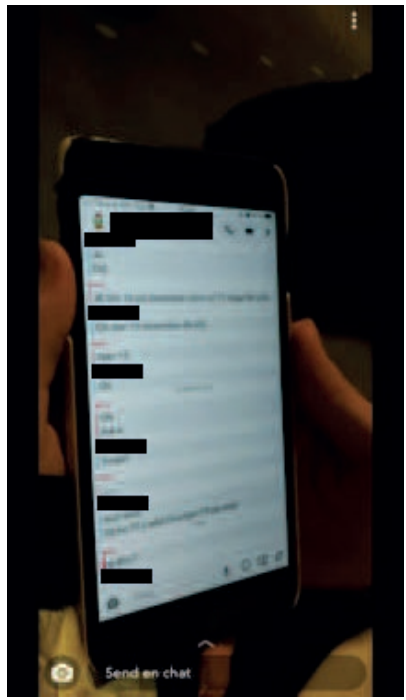
Figur 9 viser et eksempel på en feil. Figuren viser nettprat fra Snapchat, mens det i rapporten står at det er nettprat fra Skype. Det kan være en skrivefeil, men skaper usikkerhet rundt håndteringen av databeslaget.

Figur 9: Nettprat fra Snapchat beskrevet som nettprat fra Skype

Prosesseringstid og kvalitet

Under den manuelle gjennomgangen av straffesakene i uttrekket ble det funnet relevante data for prosjektet i form av nettsamtaler. Nettsamtalenes egnethet som treningsdata ble vurdert og gitt en karakter som antyder hvor lang tid en må forvente å bruke på prosesseringen, dvs. å konvertere nettsamtalene til inndata for å bygge en maskinlæringsmodell.

Karakterskalaen som ble brukt gikk fra 0 til 3, hvor 0 er den mest krevende og 3 den minst krevende kategorien. Data som var for utydelige til at det var tvil om nettsamtalen kunne fortolkes riktig ble utelukket, og ikke gitt noen karakter (se figur 10).



Figur 10: Nettprat fra Snapchat, hvor kvaliteten på dataen hvor kvaliteten på dataene ble vurdert som uleselig og ikke gitt noen karakter

Karakter 0 ble brukt for nettsamtaler hvor prosesseringen trolig krever betydelig tid. Innholdet i dokumentene kan ikke kopieres (kopier-lim-inn av tekst kan ikke brukes fordi dokumentene enten er låst eller tekst i dokumentene er av bilder), og videre fordi bilder av tekst i dokumentene er så utydelige at bruk av optisk tegnkjenning (*'Optical Character Recognition' (OCR)*) trolig ikke gir tilfredsstillende resultater. Optisk tegngjenkjenning er en teknologi som kombinerer maskinvare og programvare for å konvertere fysiske dokumenter til maskinlesbar tekst. Dermed har man ikke behov for å foreta manuell inntasting av data. Når optisk tegnkjenning blir nevnt i denne sammenhengen, gjelder det primært konvertering av bilder i PDF-filer (typisk skjermfoto av direktemeldinger på smarttelefon). Karakter 0 innebærer at teksten må tydes og skrives manuelt for å imøtekomme krav til konsistent formatering av datasettet. Antatt tidsbruk vil være fra uker til måneder

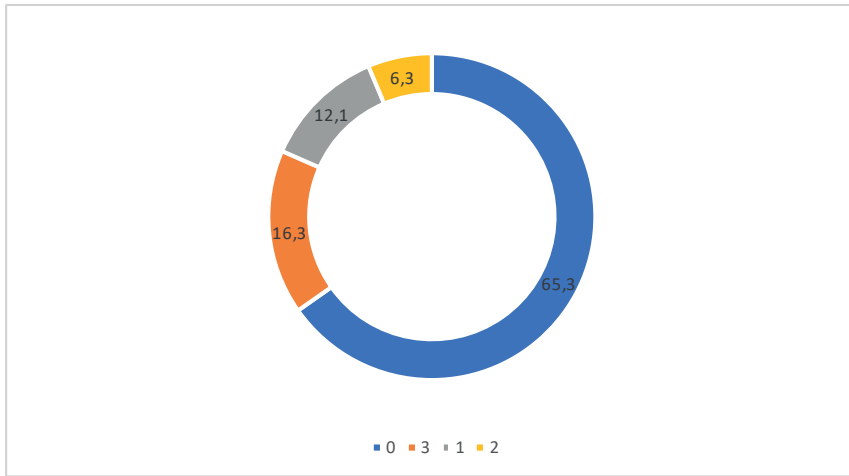
avhengig av datamengden. Et typisk eksempel på karakter 0 var straffesaksdokument på 80 sider med 400 bilder tatt av fornærmedes telefon som viser Snapchat samtale med gjerningspersonen, hvor teksten fra disse bildene må gjøres om til inndata ved manuell inntasting.

Karakter 1 innebærer at noe tekst må skrives manuelt for å imøtekomme krav til konsistent formatering av datasettet. Resultater fra optisk tegngjenkjenning må normalt gjennomgå en del rettinger på grunn av feillesing. Antatt tidsbruken vil være fra dager til uker avhengig av datamengden.

Karakter 2 ble brukt for å beskrive nettsamtaler hvor prosesseringen trolig vil ta noe tid. Innholdet i dokumentene kan kopieres (kopier-lim-inn av tekst kan brukes fordi dokumentene er åpne og tekst i dokumentene er ikke dokumentert som fotografier). Karakter 2 innebærer en del endringer på formatert tekst for å imøtekomme krav til konsistent formatering av datasettet. Antatt tidsbruk vil være fra timer til dager avhengig av datamengden.

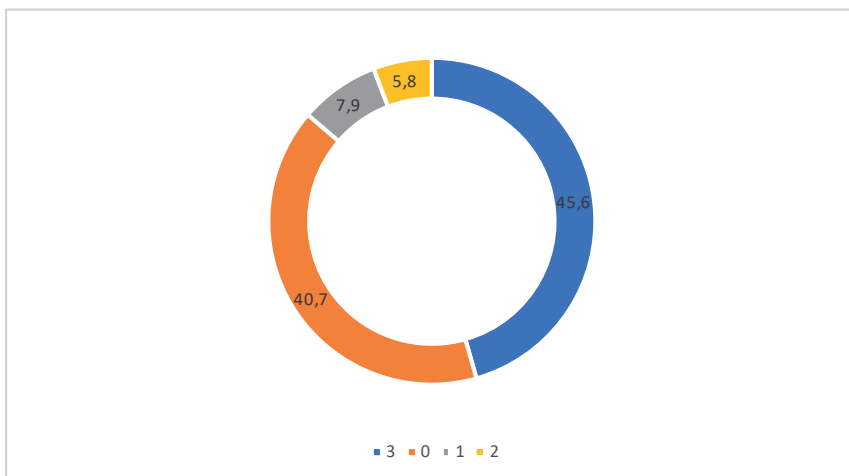
Karakter 3 ble brukt for nettsamtaler hvor prosesseringen trolig vil ta lite tid. Karakter 3 innebærer at det bare er nødvendig med mindre endringer på formatert tekst for å imøtekomme krav til konsistent formatering av datasettet. Det kan for eksempel være å fjerne eller flytte kolonner i et regneark, eller kjøre spørringer mot en database. Antatt tidsbruk vil være fra minutter til timer avhengig av datamengden.

Feil! Fant ikke referansekilden.1 viser hvordan *dokumentene* fra BL fordeler seg på karakterskalaen. Omtrent 65% inneholder nettsamtaler vurdert til karakteren 0, dvs. den mest krevende kategorien. Karakterene 1 og 2 utgjør henholdsvis 12% og 6%. Kun 16% inneholder nettsamtaler vurdert til karakteren 3, som relativt enkelt kan brukes som inndata for å trene modellen.



Figur11: Fordelingen av dokumenter fra saksbehandlingssystemet mellom karakterene 0, 1, 2 og 3

Figur 12 viser hvordan *meldingene* i dokumentene fordeler seg på karakterskalaen. Omtrent 41% av meldingene er vurdert til karakteren 0. Karakterene 1 og 2 utgjør hver for seg under 10% av meldingene, henholdsvis 6% og 8%. Karakteren 3 utgjør 46% av meldingene.

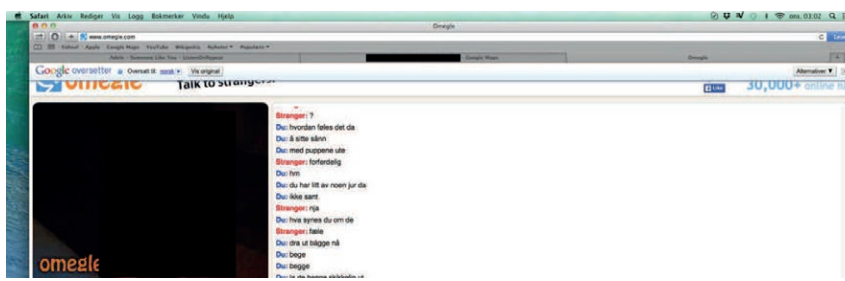


Figur 12: fordelingen av meldinger fra dokumenter mellom karakterene 0, 1, 2 og 3

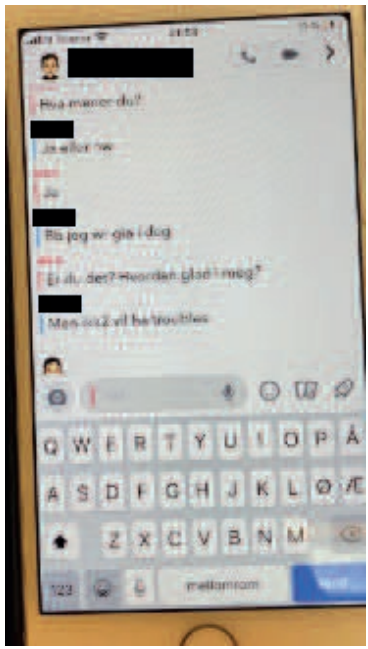
I de undersøkte sakene er Snapchat den mest brukte tjenesten. Samtidig er Snapchat blant de tjenestene som det er mest utfordrende å sikre meldinger og nettsamtaler fra i etterforskingen. Til forskjell fra andre mye brukte tjenester som Skype og Facebook, logger ikke Snapchat meldinger automatisk, fordi tjenesten har som grunnleggende funksjonalitet at meldinger skal ha begrenset levetid. For meldingstjenester som bruker logging, vil det normalt være enkelt for etterforskeren å eksportere meldinger fra brukerkontoer vedkommende har tilgang til, og over til sikringsdatamaskinen. I slike tilfeller vil de sikrede dataene inneholde tekst som er konsistent formatert og maskinelt lesbar. I motsetning til dette har meldinger fra Snapchat har ofte vært sikret i form av et foto av en mobiltelefonskjerm som viser meldinger.

Denne forskjellen mellom tjenestene forklarer hvorfor kun 16% av *dokumentene* inneholder nettsamtaler vurdert til karakteren 3, mens 45% av *meldingene* har blitt vurdert til karakteren 3. Snapchat er den tjenesten som er omtalt i flest rapporter, men gir kun en begrenset mengde meldinger og da av dårlig kvalitet. Skype og Facebook er omtalt i færre rapporter enn Snapchat, men med en større mengde meldinger av bedre kvalitet.

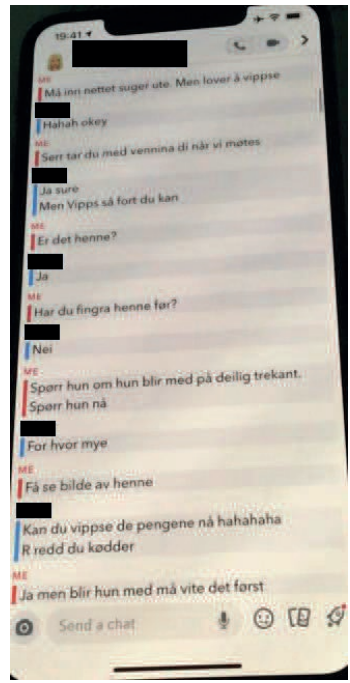
Figurene 13-16 viser eksempler på dokumentasjon av nettprat i det innhentede materialet. Mulig identifiserbar informasjon i figurene er fjernet.



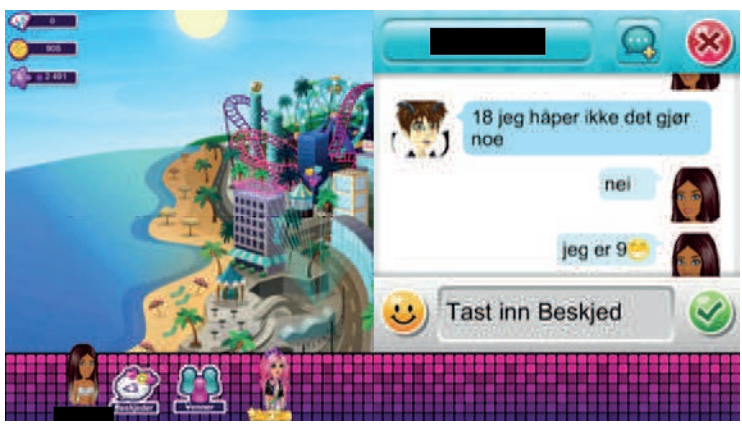
Figur 13: Nettprat fra Omegle. «Stranger» og «Du» blir automatisk satt av Omegle,



Figur 14: Eksempel på dokumentasjon av nettprat fra Snapchat. «MEG»/«ME» blir automatisk satt av Snapchat.



Figur15: Eksempel på dokumentasjon av nettprat fra Snapchat. «MEG»/«ME» blir automatisk satt av Snapchat.



16: Eksempel på dokumentasjon av nettprat fra Momio

Særlig om optisk tegngjenkjenning

Det ble undersøkt om optisk tegngjenkjenning kunne brukes for å omgjøre bilder av samtaler til tekst, men dette viste seg å tilføre mange feil i dataen. Ved bruk av optisk tegngjenkjenning avhenger resultatet av blant annet tilstanden til den opprinnelige kilden, som kvaliteten og oppløsningen på bildet. Språket, skriftstørrelsen og linjeavstanden til teksten er også viktig. Problemer med ord og tegn er å forvente. Resultatet av prosessen kan ikke umiddelbart brukes til maskinlæring, men en må manuelt sammenligne inn- og utdata og rette opp nevnte, forventede feil.

Tabellene nedenfor viser resultater etter optisk tegngjenkjenning i Adobe Acrobat Pro 2020. Utgangspunktet for resultatet i tabell 1 er figur 113 , mens utgangspunktet i **Feil! Fant ikke referansekilden.** er figur6. Merk at forsøkene på optisk tegngjenkjenning er gjort på bilder fra dokumentene, ikke selve dokumentene.[37] Da bildene i dokumentene ofte er av dårligere kvalitet i forhold til det faktiske bildet (pga. redusert størrelse), kan resultatene ved optisk tegngjenkjenning direkte på dokumentene gi dårligere resultater enn vist i tabellene.

Bortsett fra den siste meldingen i tabell 1, «Du: begge», *er det feil i all øvrig tekst.* Det andre innslaget i tabellen, «Du, hvordan!øledetda o.r.å sille sånn Du:med puppene ute lillwleef: forte,deOg», er egentlig fire meldinger – ikke én. Teksten skal egentlig være (1) «Du: hvordan føles det da»; (2) «Du: å sitte sånn»; (3) «Du: med puppene ute»; og (4) «Stranger: forferdelig». Optisk tegngjenkjenning ble også forsøkt på figur 10, men der ble ingen tegn gjenkjent.

37 Bildet er en fil integrert i det elektroniske dokumentet. Tekstgjenkjenningen er gjort på selve bildefilen, og ikke dokumentet hvor bildefilen er integrert.

Tabell 1: resultat fra optisk tegngjenkjenning av Figur 12

Avsender/melding ifølge optisk tegngjenkjenning	**Faktisk avsender/melding**
IØll!Wlir ?	Stranger: ?
Du, hvordan!øledetda o.r.å sille sånn Du:med puppene ute lillwleef: forte,de0g	Du: hvordan føles det da Du: å sitte sånn Du: med puppene ute Stranger: forferdelig
Dll, hm	Du: hm
Du: du har bH av noert jur da	Du: du har litt av noen jur da
Du:il‹ke saru	Du: ikke sant
8lrlnt,l!lrn1a	Stranger: nja
Dv: hva synes dv omd	Du: hva synes du om de
,	Stranger: fæle
Du: draut bagge nå	Du: dra ut bagge nå
llu, begi,	Du: bege
Du: begge	Du: begge

Tabell 2: resultat fra optisk tegngjenkjenning av Figur 12.

Melding ifølge optisk tegngjenkjenning	**Faktisk melding**
18 jeghaper 1kl\e det gjør Noe	18 jeg håper ikke det gjør noe
Nei	Nei
Jeg er 9	Jeg er 9

3.3.6 Hvorvidt dataene brukes til utviklingen av PrevBOT

Prosjektet resulterte i relevante data, men en mindre mengde data enn forventet. Dette skyldes som diskutert at det er tidkrevende og komplisert å identifisere saker med relevante data, og at uthenting og bearbeiding krever mye manuelt arbeid. I datagrunnlaget ligger noen få saker som skiller seg ut, med mye data som involverer én gjerningsperson og mange fornærmede. Dette gir en skjevhet i datagrunnlaget. De store enkeltsakene har imidlertid potensial for å kunne brukes til å identifisere en gjerningsperson som gjenopptar kriminelle handlinger, altså seksualisert nettprat med barn.

En stor andel av nettsamtalene viste seg å stamme fra Snapchat, noe som har betydning for fullstendigheten og omfanget i nettsamtalene. Dette er en tjeneste uten logging, og det som overleveres til politiet er gjerne fotografier av skjerm med nettprat eller lagrede enkeltmeldinger mellom gjerningsperson og barn.

Kvaliteten på nettprat i straffesakene er svært varierende, og en stor andel av sakene er på kvalitetsnivå 0, som innebærer mye arbeid før de kan brukes til maskinlæringsformål.

På grunn av utfordringene i prosjektet, ble det ikke anledning til å starte arbeidet med å utvikle og teste en maskinlæringsmodell på grunnlag av de innhentede dataene, og det er derfor ikke mulig å konkludere med hvorvidt datagrunnlaget er tilstrekkelig, eller av god nok kvalitet til at en slik modell vil kunne fungere i tråd med formålet som beskrevet i punkt 1.1. Dette må i tilfelle gjennomføres i en ny prosjektfase.

4. Oppsummering, konklusjon og tilråding

På bakgrunn av mandatet startet vi med formål om å undersøke muligheten for å bruke nettprat-logger i straffesaker om seksuelle overgrep mot barn til å utvikle en lærende PrevBOT-modell. På bakgrunn av det store antallet straffesaker som gjelder internettrelaterte seksuallovbrudd mot barn, samt straffedommer og medieoppslag fra større etterforskinger som OP Darkroom og OP Sandra, antok vi at det ville finnes en rikelig mengde data i form av nettprat-logger i norske straffesaker, og at det var gode muligheter for å kunne framskaffe et tilstrekkelig og adekvat datagrunnlag for en maskinlæringsalgoritme. Forventningen vår var dermed at prosjektet vesentlig ville dreie seg om å bearbeide og tilrettelegge dataene for innmating i algoritmen.

Prosjektet har vist at dette var altfor optimistisk, både fordi data-innhentingen var uforutsett krevende, og fordi en meget stor andel av dataene ikke er maskinlesbar. Konvertering til maskinlesbar konsistent formatert tekst krever følgelig en omfattende innsats.

Prosjektet dreide seg derfor etter hvert mer om å kartlegge og beskrive disse utfordringene enn å realisere PrevBOT-modellen. Dersom det er et politisk mål å kunne utnytte politiets data til maskinlæringsformål, er kunnskap om disse hindringene av avgjørende betydning. Det gjelder uansett for hvilket formål man ønsker å trene en lærende modell.

Datainnhenting i etterforsking ivaretar ikke behovene som reiser seg for å kunne utøve kunnskapsbasert forebygging av nettbaserte overgrep basert på kunstig intelligens. Vi understreker derfor at for å kunne ha et egnet datagrunnlag til å utvikle forebyggende KI-baserte politiverktøy, må politiet tenke nytt om behandlingen av slike data. For å kunne oppnå et anvendelig datagrunnlag, bør det *på kort sikt* utarbeides klare rutiner for hvordan data skal behandles i straffesakene for å sikre tilstrekkelig kvalitet. Videre bør sletterutinene vurderes, i og med at data som også kan ha betydning for forebygging av nye nettovergrep, pr i dag rutinemessig slettes når saken er avgjort. *På lang sikt bør* det opprettes egne databaser

for utviklingsformål med tanke på maskinlæring. Data fra straffesaker må anonymiseres, behandles og gjøres tilgjengelig for forsknings og utviklingsformål, slik at teknologi som PrevBOT og andre lignende ML baserte verktøy kan realiseres og brukes til forebyggingsformål.2. Metode og datamateriale

4.1 Identifisering og innhenting av datamateriale

Kapitlet redegjør for innhentingen av nettprat-logger i norske straffesaker. Dette ble utført i flere trinn, først ved identifisering av potensielt relevante saker gjennom uttrekk fra STRASAK, som deretter måtte innhentes og gjennomgås for å avklare hvilke av dem som inneholder relevant materiale. I tillegg ble det sendt et brev direkte til alle politidistriktene med anmodning om å få tilsendt nettprat-loggene i relevante straffesaker.

POLITIHØGSKOLEN

Politihøgskolen
Slemdalsveien 5
Postboks 2109, Vika
0125 Oslo
Tlf: 23 19 99 00
www.phs.no

PHS forskning 2022:5

ISSN 0807-1721
ISBN 978-82-7808-171-6
978-82-7808-172-3

114669 / ANDVORDGRAFISK.NO FORSIDEFOTO: SHUTTERSTOCK