Sammendrag: Facebook studie med lokale politikere

# Introduktion

Online politisk fjendtlighed, i form af hadefuldt sprog, er et velkendt fænomen som udgør en væsentlig trussel for demokratiet, ved at minimere det generelle politiske engagement blandt den almindelige befolkning. Denne uhensigtsmæssige tendens får potentielle debattører til at trække sig fra den politiske debat. Mens størstedelen af forskning indenfor området har figureret hovedsageligt om hvilke faktorer forsager politisk fjendtlighed samt udviklingen af adfærdsmønstre, findes der forholdsvis få studier om ’bystanders’ og pro-sociale reaktioner til had, samt hvorvidt det er muligt at øge tilstedeværelsen af denne pro-sociale adfærd på sociale medie-platforme. Med pro-social bystander adfærd hentydes der til adfærd hvor brugerne regerer på hadefuldt indhold ved sige fra og dreje kommunikationen i en mere positiv og konstruktiv retning, i kontrast til blot at ignorere indholdet eller reagere med mere had.

# Undersøgelsesformål

Dette projekt er en del af Standing by: Pro-social Bystander Reactions to Online Political Hostility (STANDBY) projektet, hvis formål er at klarlægge pro-sociale bystander reaktioner samt deres konsekvenser. Projektet er finansieret af det europæiske forskningsråd (ERC). Undersøgelsesformålet med dette specifikke studie er at undersøge hvorvidt Facebook brugere reagerer mere pro-socialt til hadefuldt indhold i kommentarsporet på opslag fra lokalpolitikere fra Århus og København, når en intervention kommentar som opfordrer brugere til at ‘Sig Fra, Anmeld, støt op’ overfor hadefuldt indhold, er en del af kommentartråden. Derudover at undersøge om mere generelle kommentar-tendenser forekommer for de to kommuner.

Projektet forsøger at besvare en række spørgsmål:

1. er der en forskel i kommentarer når interventionen er anvendt sammenholdt med når den ikke er?
2. Hvordan responderer brugere på negative kommentarer? Modtager negative kommentarer også negative underkommentarer (kommentar til kommentar) og vice versa?
3. Er der forskel kommunerne imellem? Er kommenteringen anderledes i København end i Århus?
4. Forekommer der generelle tendenser i kommenteringen på lokalpolitikkernes opslag? Kommenterer folk generelt i en hadefuld, aggressiv, positiv eller anden tone?

# Undersøgelsesdesign

Lokale politikere fra Århus og København kommune blev inviteret til at deltage i studiet. I alt valgte 11 politikere, heraf 5 fra København at deltage i interventions-forsøget. Politikerne blev instrueret i selv at kommentere en planlagt og forudbestemt tekst samt en kort animeret video omhandlende budskabet, til deres opslag. Kort fortalt, opfordrer tekst og video brugerne til at sige fra overfor hadefuldt indhold ved at gøre opmærksom på fjendtligheden i en kommentar, anvende platformens anmelderfunktion, samt udvise støtte til aktøren som fjendtligheden retter sig mod.

Et billede, der indeholder tekst, skærmbillede, pattedyr

Automatisk genereret beskrivelse

Figur 1: Intervention tekst og video

Teksten i sin helhed er vedlagt som bilag. Politikerne blev instrueret i ikke at ændre andre aspekter af opslaget, dets tekst eller emne. Treatment ugerne blev tildelt ved randomisering, således var intervention ikke en del af alle opslag fra en given politiker. Interventionsforsøget forløb over 8 uger. Derudover blev der indsamlet Facebook data fra yderligere 66 lokale politikere for at forøge dataomfanget samt muliggøre en komparativ analyse mellem treatment og kontrol opslag.

# Data

Data fra 1601 lokalpolitikers Facebook opslag er indhentet, hvoraf 51 opslag indeholder interventionen. Efter datasortering, er 13.102 kommentarer analyserbare, hvoraf 489 kommer fra intervention opslag. Opslagene stammer fra både politikernes offentlige og private profiler og er slået op mellem d. 01/05/23 og 02/07/23, svarende til 9 ugers data. Data er indsamlet som json filer mellem 02/07/23 og 25/08/23. 5 forskellige sentiment analysemodeller blev anvendt som en metode til at kvantificere date ved at bestemme en følelse eller attitude for hver kommentar repræsenteret som et tal eller et mærkat.

1. VADER: giver en negativ, positiv, neutral og sammensat score
2. ROBERTA: giver en negativ, positiv, neutral og sammensat score
3. Danish BERT: giver 1 af 9 forskellige mærkater som tillid/accept, foragt/modvilje, glæde/sindsro etc.
4. Attack: giver 1 af 2 forskellige mærkater: sprogligt angreb/ikke sprogligt angreb
5. Rec-nition: giver 1 af 2 forskellige mærkater: anerkendende tale/ ikke anerkendende tale

# Resultater

1. **Generelle tendenser for kommentering**

Gennemsnittet for ROBERTA sammensatscoren (som ligger mellem -1 og 1, hvor -1 er meget negativ og 1 er meget positiv) for alle kommentarer er 0,13 ensbetydende med en gennemsnitlig neutral tone dog med en tendens mod positiv.

Et billede, der indeholder tekst, skærmbillede, linje/række, Kurve

Automatisk genereret beskrivelse

Plot 1: Roberta sammensat score for alle kommentarer fordelt over uger

Undersøgelsens resultater af diverse mærkater viser at 35,1% af alle kommentarer indeholder anerkendende sprog, samt at meget få kommentarer indeholder et sprogligt angreb. BERT mærkaterne viser at mens 22,3% og 17,4% af kommentarerne modtager positive mærkater som tillid/accept samt forventning/interesse, markeres samtidig 19,4% af kommentarer som foragt/modvilje.

Et billede, der indeholder skærmbillede, cirkel, diagram, Font/skrifttype

Automatisk genereret beskrivelseEt billede, der indeholder tekst, skærmbillede, cirkel, diagram

Automatisk genereret beskrivelse

Et billede, der indeholder tekst, skærmbillede, diagram, cirkel

Automatisk genereret beskrivelse

Plot 2: fordeling af mærkater for alle kommentarer

1. **Sentiment af underkommentarer**

Intervention fraråder brugerne til at respondere på hadefuldt indhold med yderligere had, men modsat at gøre opmærksom på den fjendtlige tone på en neutral og konstruktiv måde. Analysen viser at underkommentarer til negative kommentarer oftest har en neutral eller negativ tone. Underkommentarer til positive kommentar er ligeledes positive eller neutrale. Dog viser et anderledes mønster sig hvis man udelukkende kigger på kommentarer og underkommentarer fra interventionsopslag, hvor positive markerede kommentarer oftest modtager negative eller neutrale underkommentarer.

Et billede, der indeholder tekst, skærmbillede, diagram, Kurve

Automatisk genereret beskrivelse Et billede, der indeholder tekst, skærmbillede, diagram, Kurve

Automatisk genereret beskrivelse

Plot 3: sentiment af underkommentarer for hele datasættet samt for treatment data

1. **Kommunale forskelle**

Der viser sig at være tydelige forskelle i kommentering mellem de to kommuner.

Et billede, der indeholder diagram, Kurve, skærmbillede, tekst

Automatisk genereret beskrivelse

Plot 4: density plots for sentiment across municipalities

Plottet demonstrerer hvordan størstedelen of Århus kommentarer er positive, færre er neutrale og mindstedelen er negativ. Selvom det samme generelle billede tegner sig for København, er spredningen mellem positive, neutral og negative kommentarer mindre fremtrædende.

Gennemsnittet for den sammensatte Roberta score er 0,3 for Århus mens den for København er 0,1, ensbetydende med kommentarer fra Århusianske lokalpolitikeres opslag gennemsnitligt har en mere positiv tone.

Et billede, der indeholder tekst, skærmbillede, diagram, Rektangel

Automatisk genereret beskrivelse

Plot 5: Gennemsnit af den sammensatte Roberta score for kommunerne

1. **Effekt af interventionen**

Hovedformålet for undersøgelsen var at fastslå hvorvidt interventionen havde en positiv indvirkning på tonen af kommentarfeltet på givne opslag. For at undersøge dette, er flere regressionsmodeller blevet afprøvet, som alle har til formål at forudsige sandsynligheden for at et opslag indeholder interventionen baseret på sentiment af kommentarerne. Efter inkorporering af den komplekse, indlejrede datastruktur (hvor sentiment af en kommentar også kan bestemmes af hvilket specifik opslag det stammer fra, kommende fra en bestemt politiker, som er en del af en bestemt kommune), afslører modellerne når at når Roberta sentiment scoren stiger med 1 (mere positiv tone), falder sandligheden for at kommentaren stammer fra et interventionsopslag, ensbetydende med at en mere positiv tone indbefatter en lavere sandsynlighed for at en intervention er tilstede. Dette tyder på en mulig ’backfire effekt’, som står i kontrast med mulige forventninger af implementeringen af interventionen.

Et billede, der indeholder tekst, skærmbillede, diagram, linje/række

Automatisk genereret beskrivelse

Plot 6: Gennemsnit af den sammensatte Roberta score for intervention

# Diskussion og fejlkilder

Flere vigtige faktorer kan potentielt skabe og forklare ’backfire effekten’, og flere overvejelser må således gøres inden der kan etableres sikkerhed omkring studiets resultater. Opslagene omhandler i flere tilfælde kontroversielle politiske emner såsom LGBT+, rasisme, Lynetteholmen, wokeness, den kommunale budgetplanlægning osv. Emner som disse indeholder oftest stridende synspunkter og indgår til tider i en debat hvor tonen blandt brugerne kan være hård og fjendtligt. Eftersom interventionen florerer på flere af sådanne opslag, er det muligt at den ønskede og ellers forventelige interventionseffekt er blevet overskygget af den ophedede debat.

Derudover er det muligt at Facebook brugeren er blevet stødt af interventionskommentaren i sig selv, og muligvis anser det som en del af en skjult politisk dagsorden mod at helgardere sig mod potentiel fjendtlighed, hvilket får brugere til at reagere negativt. Data viser flere tilfælde af sådanne situationer.

Et billede, der indeholder tekst, Font/skrifttype, Website, Webside

Automatisk genereret beskrivelse

Figur 2: screenshots af kommentarer fra intervention posts

Kommentarer som disse vil alle resultere i en negativ score.

Den endelige og formentlig den mest sandsynlige forklaring på ’backfire effekten’ hænger sammen med tilfældighed og manglende interventionsdata. Når der er taget højde for datens indlejrede struktur viser ingen modeller signifikante estimater, ergo når det er medtaget at generelle tendenser for specifikke opslag, for specifikke politikere, fra specifikke kommuner, kan påvirke resultater, findes der ikke tilstrækkeligt med evidens for at sentiment kan forklares som en udelukkende effekt af interventionen. Der er brug for mere data fra interventionsopslag for at etablere en klar effekt. Derved er det vigtigt at pointere at resultaterne omhandlende interventionseffekten er inkonklusive. Grundet randomiseringen af treatment ugerne, er det ikke muligt at fastslå hvorvidt backfire effekten skyldes interventionen alene. Dog kan resultaterne i sin helhed, baseret på over 13.000 kommentarer, give et godt fingerpeg om mere generelle tendenser og mønstre for kommenteringen af danske politiske opslag, samt den forskel der tilsyneladende eksisterer for de to kommuner, København og Århus.

# Bilag

NÅR DU OPLEVER FJENDTLIGE KOMMENTARER: SIG FRA, ANMELD, STØT OP!

Vi ønsker alle platforme, hvor vi kan diskutere åbent. Men fjendtlige kommentarer på sociale medier forekommer ofte, rammer alle og kommer i mange former. At opleve fjendtlige kommentarer giver folk lyst til at trække sig fra diskussioner. Når du oplever fjendtlighed på sociale medier, kan du vise din modstand på mange måder. Ved at vise din modstand bliver du en del af løsningen! Du kan gøre en stor forskel ved at følge tre enkle råd:

Sig fra: Gør høfligt opmærksom på indlæggets fjendtlige karakter, eller prøv at vække folks medfølelse. Reager aldrig på fjendtligt indhold med mere fjendtlighed.

Anmeld: Brug platformens anmeld-funktion til at gøre platformens moderatorer opmærksomme på den fjendtlige kommentar. Støt op: Vis din støtte til modtageren af det fjendtlige indhold i en kommentar eller i en privat besked.

Når du viser din modstand mod fjendtlige kommentarer, hjælper du offeret med at bearbejde det og få diskussionen tilbage på sporet. Målet er ikke at få afsenderen af den fjendtlige kommentar til at ændre mening, men at vise at du tager afstand fra det. Det, du synes er fjendtligt, er der sandsynligvis også andre, der synes er fjendtligt. Når du viser din modstand mod fjendtlige kommentarer er du ikke alene. Du er ikke alene, når du viser din modstand mod fjendtlige kommentarer.