Angrep & hat i den offentlige debatten på Facebook

Trygt Digitalt Norge

GjensidigeStiftelsen

Nordic Safe Cities

COMMON CONSULTANCY

Analyse & Tall

Forord

Mot et trygt digitalt demokrati og mer digital forebygging

Den offentlige samtalen har i økende grad flyttet seg over til sosiale medier, hvor vi alle oppholder oss. Her er hatprat blitt et fast innslag som i økende grad brer om seg. Vi opplever at samtalen og tilstedeværelsen i vårt digitale demokrati ikke er trygg for alle.

Dette er i tråd med Politiets Sikkerhetstjeneste sin trusselvurdering de siste årene, hvor digitale arenaer trekkes frem som stedene der det er størst risiko for å utsettes for ekstreme og radikaliserende budskap.

Med denne analysen ønsker vi å bidra til å finne svar på hvordan vi kan skape et tryggere digitalt demokrati i Norge. Vi vil finne metoder for å forebygge og motvirke hatprat på nettet og skape en sterk og god ytringskultur.

I denne analysen har vi tatt temperaturen på den offentlige samtalen på Facebook. Vi har undersøkt hvor mye plass de språklige angrep faktisk tar, hvem de er rettet mot og hvor de finnes.

Gjennom bruk av maskinlæringsteknologi er det nå blitt mulig å kartlegge språklige angrep i den offentlige debatten på Facebook på norsk. Angrepsalgoritmen har analysert mer enn 10 millioner kommentarer på norske Facebook-sider til politikere, medier, offentlige personer og på ulike debattsider. Dermed utgjør denne analysen den mest omfattende språklige analysen av hatprat på nettet i Norge som noensinne er utført.

Den viser med all tydelighet at måten vi debatterer og kommuniserer på sosiale medier er et problem for demokratiet. Analysen dokumenterer at over 177.000 kommentarer kan karakteriseres som språklige angrep, herunder hatprat. Den største andelen finnes i kommentarfeltene til politikere. Dette er problematisk, særlig fordi noen grupper i samfunnet er mer utsatt for angrep og hat enn andre. Risikoen er at de trekker seg fra den offentlige samtalen og at de vegrer seg for å engasjere seg politisk. Det gjør den enkelte utsatt og sårbar, og samtidig blir vårt demokrati fattigere og mindre representativt.

Hvis de som rammer andre med hatprat og udemokratiske angrep får stå uimotsagt, kan dette forsterke en situasjon hvor enda flere vegrer seg for å delta. Flere kan påvirkes og tro på hatefulle og ekstreme fortellinger. Det kan i ytterste konsekvens føre til radikalisering og ekstremistiske handlinger og angrep.

Vi må skape en bedre ytringskultur på Facebook og i våre digitale samtaler. Vi ønsker å gå i front i arbeidet med å skape nye former for digital forebygging, som skal gjøre det trygt for alle å delta i den digitale offentlige samtalen.

Jeppe Albers
Direktør
Nordic Safe Cities

Ingrid Riddervold Lorange
Administrerende direktør
Gjensidigestiftelsen

Om prosjekt Trygg Digital By

Rapporten er en del av Nordic Safe Cities' Trygg by Norge-satsing, som Gjensidige-stiftelsen har støttet med totalt 35 millioner kroner. Midlene går til prosjekter i 12 norske byer og kommuner med mål om å skape mer trygghet på nettet, god ytrings-kultur og et sterkt deltakende demokrati.

Analysens metodikk og arbeidet med å skape nye typer digital forebygging er også en del av en større satsing hos Nordic Safe Cities, Analyse & Tall og Common Consultancy, som streber etter å skape trygge digitale byer. Satsingen heter «Safe Digital City» og er i ferd med å bli implementert i en rekke av Nordic Safe Cities' medlemsbyer i hele Norden.

Om partnerne i prosjektet

GjensidigeStiftelsen

Gjensidigestiftelsen er en finansstiftelse som viderefører over 200 års historie med å skape gode liv i et trygt samfunn gjennom eierskap og utdelinger. Siden Gjensidigestiftelsen ble etablert i 2007 har over 10 000 ulike prosjekter fått til sammen over 3,2 milliarder kroner i støtte.

Nordic Safe Cities

Nordic Safe Cities er en nordisk organisasjon som setter i gang initiativer for å styrke i samholdet og skape tryggere lokale demokratier i 21 byer i Norden.

COMMON CONSULTANCY

Common Consultancy er en analyse- og rådgivningsvirksomhet, som har spesialisert seg i å anvende data fra sosiale medier, til å forstå digitale dynamikker som hatprat, politisk oppbakking og feilinformasjon.

&#

Analyse & Tall

Analyse & Tall analyserer komplekse størrelser som digitale fellesskap, bærekraft, sosiale problemer, frivillighet, netthets og spredning av feil- og desinformasjon. Ved å kombinere konvensjonelle metoder med nye digitale metoder som stordata og maskinlæring, utvikler de nyskapende analyser av fenomener i vår nye digitale virkelighet.

Om rapporten

Angrepsalgoritmen og rapporten er utarbeidet av de digitale analysebyråene Analyse & Tall og Common Consultancy på oppdrag fra Nordic Safe Cities.

Rapporten vil:

1. Redegjøre for metoden bak, og arbeidet med utviklingen av en norskspråklig algoritme til deteksjon av språklige angrep. Algoritmene gjøres tilgjengelige for forskere som open-source på plattformen Hugging Face.

2. Bruke algoritmene til en analyse av den offentlige, digitale debatten, slik den utspiller seg i nesten 10,5 millioner kommentarer på Facebook-sidene til norske politikere, medier, offentlige personer og offentlige debattsider.

God lesning!

Innhold

6 Innledning
9 Hovedresultater
14 Data & Metode

18 **Algoritmene finner angrep og anerkjennelse**
27 **Slik virker en algoritme,** og slik har vi bygd vår
36 **Det store bildet:** Språklige angrep i den brede Facebook-samtalen
38 **Tema** som genererer angrep
44 **Tendenser** innenfor hatprat
50 **På Facebook går politikk og angrep hånd i hånd**
55 **Det harde debattklimaet hos mediene**
59 **Angrep på offentlige personers sider**

63 Anerkjennelse i Facebook-debatten

Innledning

Betydningen av å kartlegge den offentlige debatten på Facebook

Facebook er vår største digitale arena for offentlig debatt
Med 3,5 millioner brukere er Facebook fortsatt det største sosiale mediet i Norge. 67 prosent av alle nordmenn logger inn på Facebook én eller flere ganger hver eneste dag[1]. Her både eksponeres vi for, og samhandler med, en strøm av kommentarer, nyheter og ytringer. Sosiale medier som Facebook har bidratt til at langt flere deltar i det offentlige ordskiftet og bruker ytringsfriheten sin. Facebook er også et sted hvor man finner meningsfeller og anerkjenner livets store øyeblikk og hverandres prestasjoner. Men vi vet også at debattene på Facebook kan virke polariserende, og at språklige angrep på nett kan ha alvorlige konsekvenser for enkeltindivider og grupper.

Med denne rapporten ønsker vi å bidra med økt kunnskap om ytringsklimaet på Facebook, både når det gjelder de opphetede og grenseoverskridende debattene og de anerkjennende samtalene. For å gjøre dette har vi for første gang i Norge brukt maskinlæring til å kvantifisere hele debatten som finner sted på et stort antall offentlige Facebook-sider, nærmere bestemt Facebook-sidene til norske politikere, offentlige personer, medier og offentlige debattsider. Dette er sider hvor det deles nyheter, holdninger og standpunkt i stor skala, og hvor Facebook-brukere kan kommentere.

Mange avstår fra å delta i debatten
Ytringsklimaet på Facebook gjør at mange avstår fra å gi uttrykk for meningene sine. En undersøkelse fra Likestillings- og diskrimineringsombudet i 2021 viste at to av tre vegrer seg for å si hva de mener i nettdebatter på grunn av den harde tonen i debatten.[2] Samtidig viser en forskningsrapport fra Politihøgskolen at hatefulle ytringer, hets og trusler mot folkevalgte politikere har vært økende de siste årene.[3] Dette indikerer at ytringsklimaet setter politisk deltakelse og demokratisk deliberasjon under press. Vi vet også at enkelte medier de siste årene har lukket kommentarfeltene sine på Facebook fordi det er for ressurskrevende å moderere.

Men hvordan er egentlig debatten? Er det noen grupper som særlig blir utsatt for språklige angrep og hat? Er det noen temaer som oftere enn andre bidrar til at debatten blir hatefull? Og som en motvekt til dette: Hva er det som genererer anerkjennende og positive kommentarer på Facebook?

Dette er noen av spørsmålene vi ønsker å besvare i rapporten.

Dette vet vi om hatprat i debatten fra før

Institutt for samfunnsforskning har gjennomført flere undersøkelser av hatefulle ytringer, blant annet en spørreundersøkelse (2019) som slår fast at sosiale medier og kommentarfelt/nettfora er de arenaene der det er vanligst å bli utsatt for hatytringer.[4] Denne og tidligere undersøkelser fra Institutt for samfunnsforskning viser at det er en markant høyere sannsynlighet for at minoritetsgrupper som LHBT-personer og personer med innvandrerbakgrunn har vært utsatt for hatytringer enn resten av befolkningen.[5]

I 2021 publiserte Likestillings- og diskrimineringsombudet en rapport basert på en spørreundersøkelse[6] som viser at syv av ti som deltar i samfunnsdebatter på Facebook, har opplevd at andre skriver noe nedsettende eller krenkende i en debattråd.

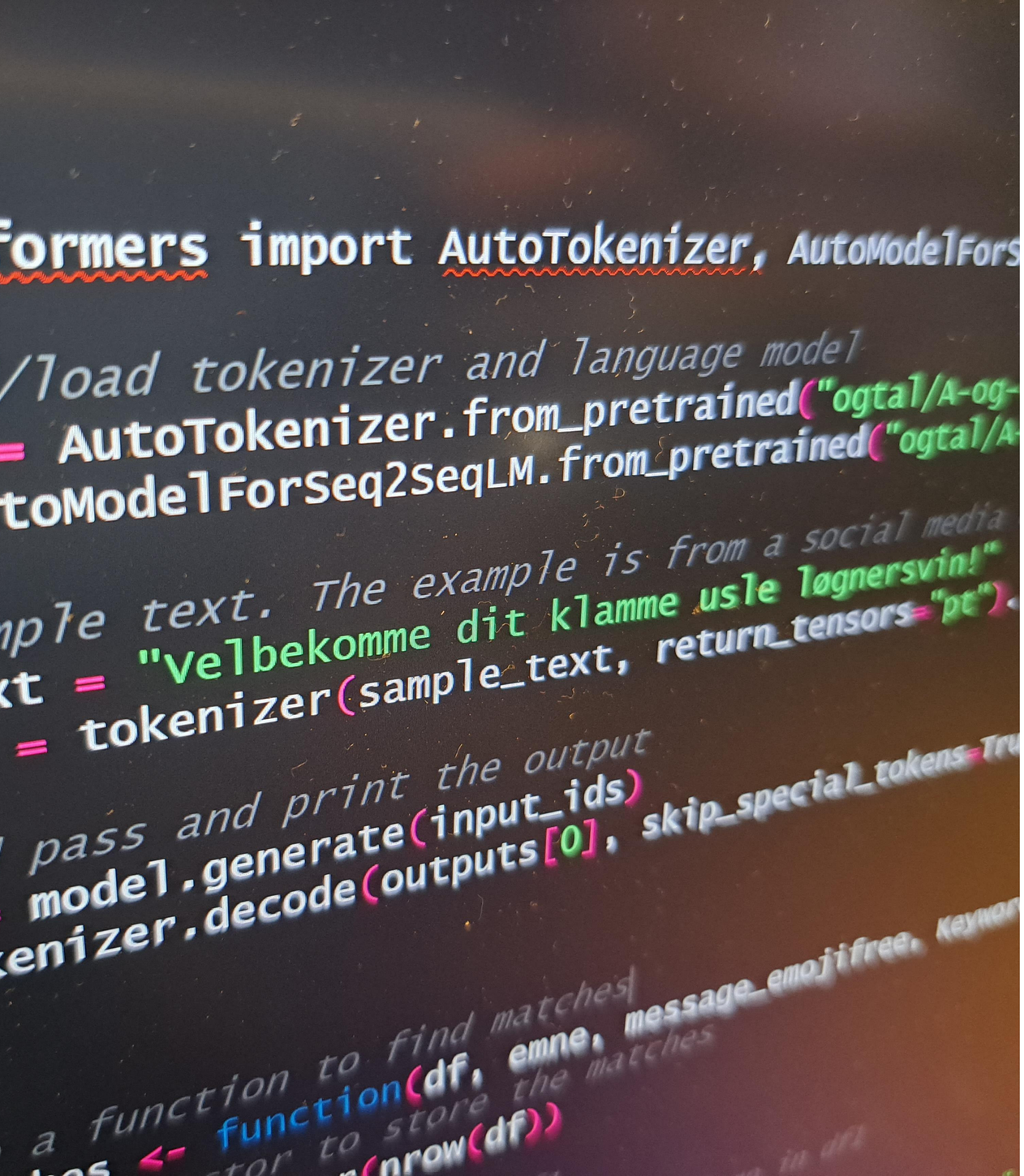
Men hva med de folkevalgte politikerne? Politikere står i en dobbeltrolle, for de har et ansvar for å legge til rette for et trygt og sunt debattklima på sine egne sider, men blir også utsatt for hets og sjikane i sosiale medier selv. En forskningsrapport fra Politihøgskolen viser at andelen folkevalgte som rapporterer indirekte og direkte trusler, har gått fra 40 prosent i 2013 til ca. 70 prosent i 2021.[7]

Det er ikke konsensus om hvor stor andelen hatprat i nettdebatten er. Både Likestillings- og diskrimineringsombudet og medieanalysebyrået Retriever har forsøkt å estimere denne andelen ved å utarbeide kvantitative innholdsanalyser av et randomisert utvalg kommentarer på Facebook-sidene til norske medier. Begge bruker en definisjon som er lik vår definisjon av hatprat, og de finner at henholdsvis 9 og 2 prosent av kommentarene er hatefulle, noe som utgjør store variasjoner innenfor et tidsrom på bare fem år.[8, 9]

Hvor mye hatefullt innhold som avdekkes, henger imidlertid tett sammen med moderering, for både mediene, politikere og andre offentlige personer kan slå ned på hatprat ved å moderere innhold. Det samme kan Facebook gjøre, og plattformen har intensivert modereringen de senere årene.

Det er likevel et iboende dilemma at innhold som grenser til forbud (for eksempel hatprat), skaper særlig mye interaksjon og aktivitet på plattformen. Facebook-gründer Mark Zuckerberg har formulert det slik: «Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average - even when they tell us afterwards they don't like the content.»[10]

Hovedresultater



Verdens første tverrskandinaviske angrepsalgoritme

Til denne undersøkelsen har vi samlet inn nesten 10,5 millioner kommentarer på Facebook-sider som tilhører norske politikere, medier, offentlige personer og offentlige debattsider. For å analysere alle disse kommentarene har vi utviklet angrepsalgoritmen **A&ttack2.**

Vi anser denne algoritmen som et av undersøkelsens viktigste resultater. Det er nemlig den første algoritmen som kan gjenkjenne angrep både på norsk og dansk. Algoritmen kan gjenkjenne om en Facebook-kommentar er et språklig angrep eller ikke, og den er bedre enn sammenlignbare, enspråklige algoritmer. A&ttack2 er dessuten fritt tilgjengelig på plattformen hugging face.

Med den nye algoritmen har vi – som de første – kunnet undersøke hele debatten i Norge, og dermed har vi også bidratt til forskningen på debattklimaet på Facebook. Samtidig har vi lagt fundamentet for at andre skal kunne ta forskningen videre.

1,7 prosent av kommentarene i den offentlige debatten på Facebook er angrep

1,7%

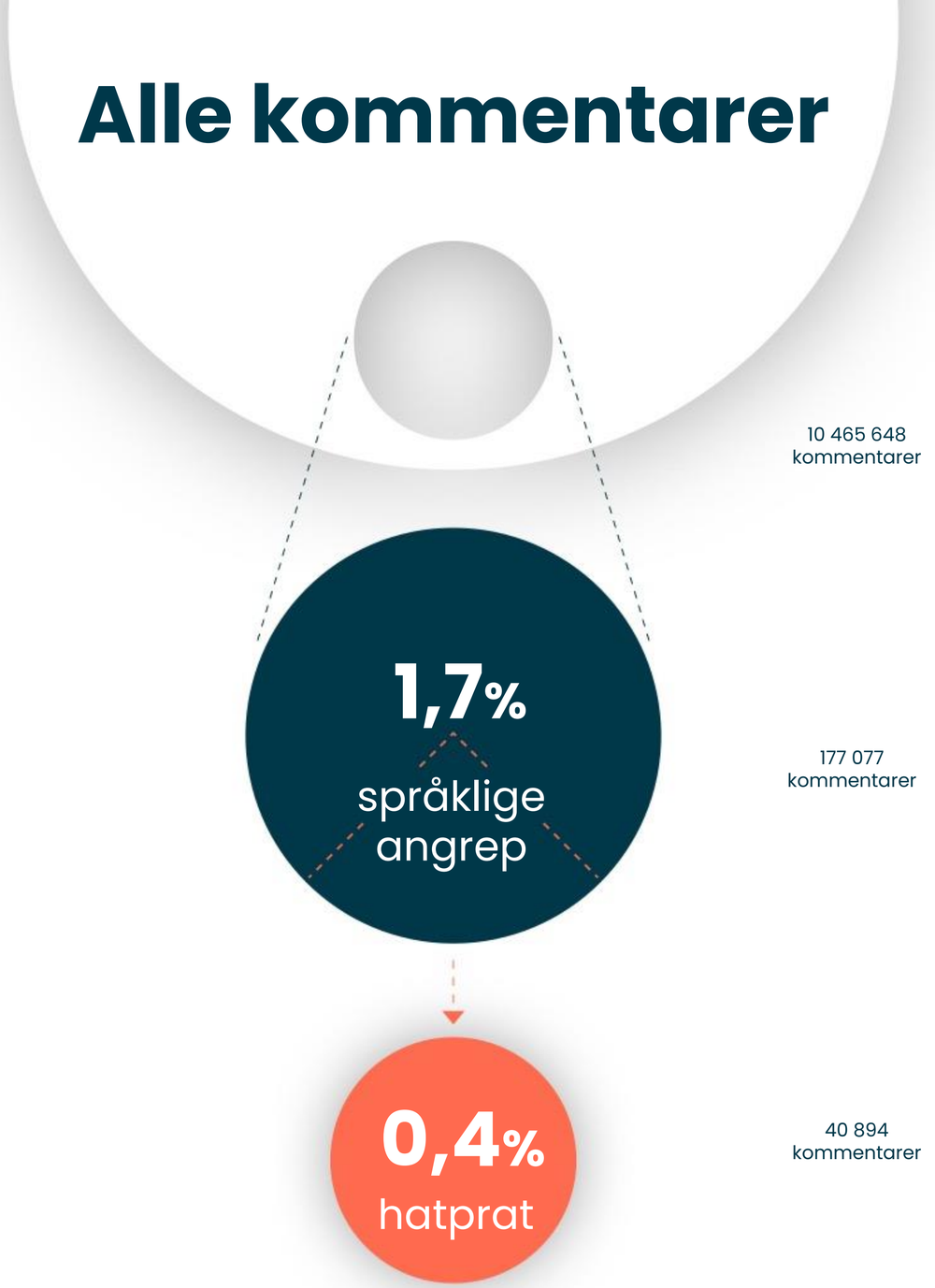
språklige angrep

0,4%

hatprat

Av de omtrent **10,5 millioner** kommentarene som utgjør den digitale, offentlige debatten på Facebook, har vi med A&ttack2 kategorisert **177 077** kommentarer som språklige angrep, det vil si kommentarer som inneholder stigmatiserende, nedsettende, krenkende, stereotype, ekskluderende, sjikanerende eller truende ytringer.

Ved hjelp av en søkenøkkel-algoritme finner vi at **40 894** av 177 077 angrepskommentarer kan defineres som hatprat. Det vil si språklige angrep som er rettet mot beskyttede karakteristika som relaterer til rase/etnisitet, hudfarge, nasjonalitet og opprinnelse, religion og tro, seksuell orientering, kjønn og kjønnsidentitet, sosial klasse og sosial status, politisk orientering, alder eller funksjonshemming og alvorlige sykdommer (både fysiske og psykiske).



Angrep og hatprat rammer ikke likt – muslimer er mest utsatt

Det er flest angrep i debatter om islam, muslimer og integrering

Islam er temaet med høyest andel angrep. Av alle kommentarer som er skrevet til innlegg om islam, er **4,9 prosent** angrep. I integreringsdebatten er andelen angrep **4,2 prosent**. Muslimer er også den befolkningsgruppen som oftest nevnes eksplisitt i de språklige angrepene. I **29,7 prosent** av alle kommentarene hvor beskyttede karakteristika nevnes, er det snakk om muslimer. Det betyr at nesten en tredjedel av all hatprat er rettet mot muslimer.

Diskusjoner om kriminalitet og rettssystemet fører til angrep

Samtaler knyttet til kriminalitet har en høy andel angrep. Kriminalomsorgen topper, da **4,9 prosent** av alle kommentarer til innlegg om dette temaet er språklige angrep, mens kriminalitet følger etter med **4,2 prosent**. Kommentarene er riktignok ikke like harde mot alle kriminelle. De går hardest utover kriminelle som er unge eller som har muslimsk bakgrunn eller innvandrerbakgrunn. Dermed blir noen grupper dobbelt stigmatisert hvis de begår kriminalitet.

Seksualitet og kjønn skaper opphetet debatt i kommentarfeltene

Kvinner er i høy grad også mottakere av angrepskommentarer. Helt nøyaktig er **16,8 prosent** av alle kommentarene vi definerer som hatprat, rettet mot kvinner. Men det er ikke bare kvinner som er gjenstand for angrep. **2,9 prosent** av alle kommentarer som er skrevet til innlegg om kjønnsidentitet og seksualkulltur, er angrep. Her er blant annet deltakerne i Pride-parader utsatt.

Debattklimaet endrer seg avhengig av typen Facebook-side

1,7%

Språklige angrep på politikersidene

Av de **2,2 millioner** kommentarene vi har hentet fra politikernes Facebook-sider, er **37 676** språklige angrep. Det betyr at vi finner den høyeste andelen angrep på politikersidene. Hos politikeren med høyest andel angrep er andelen **4,2 prosent**. Det er imidlertid store variasjoner med hensyn til om angrepene er rettet mot politikerne selv, de politiske motstanderne deres eller andre grupper i samfunnet.

1,6%

Språklige angrep på mediesidene

Av totalt **6,6 millioner** kommentarer som publiseres på medienes Facebook-sider, er **108 161** språklige angrep. Kommentarer fra mediesidene utgjør brorparten av rapportens datagrunnlag, men det er stor forskjell på debattklimaet hos de ulike mediene, og særlig alternative medier er sterkt representert. Hos mediet som scorer høyest, er **5,3 prosent** av alle kommentarene på siden et språklig angrep.

1,4%

Språklige angrep på Facebook-sidene til offentlige personer

Den laveste andelen angrep finner vi hos offentlige personer. Her inneholder **18 543** kommentarer av totalt **1,3 millioner** et angrep. Det er imidlertid blant offentlige personer at forskjellen i andelen språklige angrep er størst: Personen med høyest forekomst av angrep på Facebook-siden sin har en andel på **5,3 prosent**, mens nestemann på listen «bare» har en andel på **2,5 prosent**.

Data & metode

Med algoritmer og stordata kan vi undersøke hele debatten

Meningsmåling med hundretusenvis av respondenter

De siste årene har vi sett en fremvekst av forskning innenfor ulike samfunnsområder, blant annet demokrati og offentlighet, som utnytter mulighetene som ligger i stordata. Med en analyse som er basert på millioner av oppslag, kommentarer, reaksjoner og delinger, får vi et unikt bilde av hva vi er opptatt av i Norge, og hvordan vi samhandler med hverandre på sosiale medier.

Undersøkelser som er basert på stordata, gir oss et helt annet innblikk i den offentlige debatten enn hva spørreundersøkelser eller strukturerte intervjuer kan gi. Ved å undersøke Facebook-debatten kan vi se på langt større datamengder enn vi noensinne vil kunne gjøre med mer konvensjonelle metoder. I stedet for et tilfeldig utvalg av innlegg og kommentarer kan vi nå undersøke hele debatten. Vi ser dessuten ikke på folks oppfatning av egen adferd – vi undersøker folks faktiske digitale adferd. Vel å merke i anonymisert form.

Algoritmer tar analyser av sosiale medier et skritt videre

De danske analysene av angrep og anerkjennelse i den offentlige debatten på Facebook[11,12] markerte startskuddet for bruken av kunstig intelligens til å forstå debattklimaet på Facebook i stor målestokk. Ved å bruke maskinlæring utviklet vi algoritmer som kan finne angrep, hatprat og anerkjennelse i strømmen av danskenes Facebook-interaksjoner. Med en norsktilpasset algoritme og flere søkenøkler har vi kartlagt den norske Facebook-debatten på en måte som aldri har vært gjort tidligere.

Slik har vi analysert den offentlige Facebook-debatten

Ved hjelp av algoritmen har vi analysert alle innlegg og kommentarer på Facebook-sidene til norske offentlige personer, politikere, medier og offentlige debattsider i perioden fra 1. januar 2020 t.o.m. 27. september 2022. Datagrunnlaget består bare av innlegg og kommentarer som er tilgjengelige via Facebooks API. Det betyr også at vi bare kan hente ut så mye data som API-et tillater. I rapporten har vi anonymisert kommentarene vi fremhever ytterligere, slik at vi bare nevner navnet på offentlige personer. Det er viktig å huske at Facebook-data er dynamiske. Det betyr at Facebook selv, sideadministratorer eller Facebook-brukere kan ha slettet kommentarer uten at vi kan se at noe mangler. Dataene vi analyserer, inneholder altså bare de kommentarene som andre ikke har slettet i forveien.

Offentlige personer

Offentlige personer har mange følgere på sosiale medier som Facebook, og kommentarfeltene deres utgjør derfor en stadig større del av den offentlige samtalen. Populasjonen av offentlige personer omfatter blant annet idretts-utøvere, kulturpersoner (billedkunst, skuespill, musikk mv.), influensere, debattanter (faste og spaltister), reality-kjendiser og interesserepresentanter som har minst 10 000 følgere på Facebook-sidene sine.

Politikere

Politikerpopulasjonen består av norske politikere som har en offisiell Facebook-side, og som er eller har vært medlem av Stortinget siden 2013. Den omfatter også de kandidatene som stilte, men ikke ble valgt inn, ved siste Stortingsvalg (2021).

Medier

Denne populasjonen omfatter riks-dekkende medier, det vil si aviser, TV, nettmedier og radiokanaler. De fleste mediene i undersøkelsen følger Redaktørplakaten og Vær Varsom-plakaten. Mange av dem er leverandører av innenriks- og utenriksnyheter, men også fagmedier som Teknisk Ukeblad og Computerworld samt magasiner og ukeblader som Se og Hør er inkludert.

Offentlige debattsider

Denne populasjonen omfatter offentlige debattsider og politiske møtesteder i ytterkanten av de etablerte mediene og de politiske partiene. Fellesnevneren for de sistnevnte er at de befinner seg i periferien av den offentlige debatten, i tillegg til at de typisk utgjør et holdningsfellesskap. En slik Facebook-side må ha minst 1 500 følgere for å inngå i vår populasjon.

Vi henter data fra disse norske Facebook-sidene

Mediesider

6 556 986
Kommentarer under

▼

201 064
Innlegg fra

▼

58
Riksdekkende mediers Facebook-sider

Politikersider

2 235 959
Kommentarer under

▼

78 731
Innlegg fra

▼

336
Politikeres Facebook-sider

Offentlige personer

1 314 511
Kommentarer under

▼

39 198
Innlegg fra

▼

160
Offentlige personers Facebook-sider

Offentlige debattsider

358 192
Kommentarer under

▼

28 072
Innlegg fra

▼

54
Offentlige debattsider

Algoritmene finner angrep og anerkjennelse

Det harde debattklimaet:
Definisjon av språklige angrep

Hvor går grensen mellom opphetet debatt og kontroversielle synspunkt på den ene siden og angrep og hatprat på den andre? Det spørsmålet forsøker vi å besvare. I første omgang har vi tatt stilling til hva vi ønsker å finne. Sammen med Gjensidigestiftelsen kom vi frem til at målet var å utvikle en algoritme som kan finne **språklige angrep** i den offentlige debatten på Facebook.

Ved å legge vekt på språklige angrep kan vi kvantifisere debattklimaet i den digitale, offentlige samtalen på et mer overordnet nivå, det vil si vi kan også få med oss angrep rettet mot ikke-beskyttede karakteristika, for eksempel yrke eller utdanning.

Videre har vi laget en søkenøkkel-algoritme som identifiserer om angrepene kan karakteriseres som **hatprat**. Rapportens fire sentrale begreper er definert til høyre.

De fire begrepene har vært fundamentet for kodemanualen med regler, eksempler og unntak, som har vært førende i den menneskeassisterte treningen av algoritmen. De er også grunnlaget for størstedelen av undersøkelsen.

Slik har vi definert det:

Språklige angrep
Stigmatiserende, nedsettende, krenkende, stereotypiserende, ekskluderende, sjikanerende eller truende ytringer.

Hatprat - underkategori av språklige angrep
Et språklig angrep mot en gruppe eller et individ basert på gruppens eller individets beskyttede karakteristika. Denne definisjonen ligger tett opp til ECRIs (European Commission against Racism and Intolerance) definisjon. I undersøkelsen kaller vi det hatprat for å understreke at vår definisjon er bredere enn den juridiske definisjonen av «hatefulle ytringer».

Beskyttede karakteristika
Rase/etnisitet, hudfarge, nasjonalitet og opprinnelse, religion og tro, seksuell orientering, kjønn og kjønnsidentitet, sosial klasse og sosial status, politisk orientering, alder eller funksjonshemming og alvorlige sykdommer (både fysiske og psykiske). Her følger vi ECRIs definisjon.

Ikke-beskyttede karakteristika
Dette kan for eksempel være yrke, utdanning, lokal geografisk tilhørighet og lignende.

Eksisterende definisjoner av hatprat

Det finnes ingen klar enighet om hvordan hatprat skal defineres, verken i Norge eller internasjonalt. Det nærmeste vi kommer, er bestemmelser i straffeloven og i likestillings- og diskrimineringsloven. Slik beskriver rasismeparagrafen § 185 i straffeloven av 2005 de hatefulle ytringene som er straffbare:

«Med diskriminerende eller hatefull ytring menes det å true eller forhåne noen, eller fremme hat, forfølgelse eller ringeakt overfor noen på grunn av deres a) hudfarge eller nasjonal eller etniske opprinnelse, b) religion eller livssyn, c) homofile orientering, eller d) nedsatte funksjonsevne.»

Utgangspunktet for denne analysen er imidlertid **ikke** å kartlegge ytringer som er straffbare i juridisk forstand. Samfunnsvitenskapelig forskning viser at det er langt flere ytringer som har alvorlige skadevirkninger, og som kan avskrekke borgere fra å delta i den demokratiske debatten. Vi finner det derfor formålstjenlig å støtte oss på ECRIs og Facebooks definisjoner av hatprat, som står oppført til høyre. Våre, ECRIs og Facebooks definisjoner er videre enn den rent strafferettslige og bygger på følgende formel:

Hat = språklig angrep **basert på** beskyttede karakteristika

"Hate speech […] entails […] the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons […]– that is based on a non-exhaustive list of personal characteristics or status that includes "race", colour, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation."

– ECRI (European Commission against Racism and Intolerance)

"We define hate speech as a direct attack on people based on what we call protected characteristics — race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability"

– Facebook Community guidelines

Algoritmene: **Angrep og hatprat**

Det er egentlig litt misvisende å snakke om algoritmen i entall. Teknologien som ligger til grunn for denne rapporten, bygger på to algoritmer med hver sin funksjon. Den første oppdager angrep, og den andre identifiserer om beskyttede karakteristika nevnes i kommentaren. Algoritmene fungerer på følgende vis:

Den første algoritmen, som kalles A&ttack2, er basert på maskinlæring. Den oppdager om en kommentar er et språklig angrep eller ikke – det vil si hvorvidt det er snakk om *stigmatiserende, nedsettende, krenkende, stereotypiserende, ekskluderende, sjikanerende eller truende ytringer*.

Den andre algoritmen er en søkenøkkel som identifiserer om beskyttede karakteristika nevnes i kommentaren eller ikke. Søkenøkkel-algoritmen vurderer altså om angrepet er rettet mot en gruppes eller et individs beskyttede karakteristika – det vil si om det kan defineres som hatprat.

A&ttack2 ligger fritt tilgjengelig her, på plattformen Hugging Face.

Angrep eller ikke angrep? Det er spørsmålet

Hvor går grensen for når et utsagn kan sies å være nedsettende eller stigmatiserende? For å kunne gjøre en slik vurdering har vi utarbeidet en kodemanual som operasjonaliserer definisjonen av språklige angrep gjennom regler, eksempler og unntak. Den første delen av prosessen med å utvikle algoritmene er normativ og kvalitativ fordi det ikke finnes en verdifri, objektiv definisjon av angrep. Utsagnene nedenfor er eksempler på Facebook-kommentarer som vi har definert som henholdsvis angrep og ikke-angrep med referanse til utvalgte regler fra kodemanualen vår.

Ikke-angrep

Fordi om noen land er idioter trenger ikke vi følge etter. Har knapt lest større vrøvl en dette
Angrepet er ikke rettet mot et menneske / en gruppe mennesker. Utover det kritiseres innlegget, noe som ikke er et angrep.

Håper dette innlegget vil stå som forskrekkelse og gru ,i evig tid. På hvor virkelighets fjern det går an å bli??
Kommentaren er ikke definert som et angrep fordi det er innlegget, ikke en person, som betegnes som «forskrekkelse» og «gru».

Håper han der ryker fort
Kommentaren uttrykker et personlig håp/ønske om at mottakeren ryker ut av et program, men ikke et generelt ønske om at han skal forsvinne.

[Navn] var du der ??
For en idiotisk kommentar !!
Det er kommentaren, ikke mennesker, som blir angrepet.

Totalt vettalause handtering av energi politikk
Kritikk av håndtering og politikk er ikke et angrep.

Send dem ut av Norge.
Uttalelsen er kontroversiell, men representerer en legitim holdningsytring.

Dere fremstår som en gjeng klovne
Klovn ligger ifølge kodemanualen under «bagatellgrensen» for nedsettende språk.

Angrep

Du er bare en dritt, skyt deg selv i ansiktet, din elendige gris.
Trusler om, forherligelse av eller uttrykk for ønske om å skade mennesker betraktes som et angrep. Det er dehumaniserende å omtale mennesker med dyrebetegnelser.

Muslimene vil ikke ha eller ta til seg våres tro eller kultur,alt er urent utenom penger.
Stigmatisering, generalisering, fordommer og anklager betraktes som angrep.

Har bare en ting å si om Erna Solberg. Svakeste statsminister vi noensinne har hatt. Hun er hva jeg kaller et jæv...menneske med 0 moral....
Bruk av nedsettende beskrivelser av mennesker med uttrykk som "jæv...menneske" betraktes som et angrep.

Muslimer burde ikke ha stemmerett i noen europeiske land.
Uttrykk for at noen burde ekskluderes og fratas grunnleggende rettigheter, betraktes som et angrep.

Hun derre helvetes kjerringen altså.. måtte du få champagne korken mitt mellom øgene ditt brød naut
Bruk av nedsettende ord om beskyttede karakteristika (kjønn og alder) og uttrykk for ønske om å skade anses som et angrep.

Hjernedødt menneske.Hun bør frataes de barn hun har igjen..
Uttrykk for at noen burde ekskluderes og fratas grunnleggende rettigheter, betraktes som angrep. Å bruke kliniske diagnoser (hjernedød) som fornærmelse betraktes også som et angrep.

Forskjellen på språklige angrep og hatprat

Hvorvidt et angrep defineres som hatprat, avgjøres av om det er rettet mot ett eller flere beskyttede karakteristika, det vil si etnisitet, hudfarge, nasjonalitet og opprinnelse, religion og tro, seksuell orientering, kjønn og kjønnsidentitet, sosial klasse og/eller status, politisk orientering, alder eller funksjonsnedsettelse og sykdommer (fysiske og psykiske).

Forskjellen mellom språklige angrep og hatprat er ikke et spørsmål om hvor «hard» kommentaren er. Noen språklige angrep kan oppleves svært voldsomme fordi de for eksempel er truende, mens noen hatefulle kommentarer er hatefulle «bare» fordi de inneholder nedsettende eller stigmatiserende ord om beskyttede karakteristika (for eksempel kjerring, homse, mongo).

Språklige angrep

Hun er så ynkelig at det er bare trist.
Ordbruken er nedsettende («ynkelig»), men utgjør ikke et angrep på beskyttede karakteristika.

[Navn] kyss deg i ræva, du høres ut som en politiker.
Upassende oppfordringer er et angrep, men angrepet er ikke basert på beskyttede karakteristika.

[Navn] Stygge troll!
Kroppstype eller utseende er ikke beskyttede karakteristika, men kommentaren er nedsettende.

Ekle folk er dere..
Angrepet er ikke basert på beskyttede karakteristika.

Journalister er en gjeng med feige sauer
Arbeid (herunder titler som journalist, politiker) er ikke et av de beskyttede karakteristikaene, men å kalle folk for feige sauer er et angrep.

Dette er helt sykt og crazy , dumme, naive politikere!
Arbeid er ikke et av de beskyttede karakteristikaene.

Hatprat

Spyrrrr av hele kvinnfolket,kvalmende..
Angrepet er basert på kjønn gjennom ordet «kvinnfolket».

Alle vet at dette er p.do religion nr 1 og venstresiden elsker disse. Tror de blir seksuelt opphisset av dette samt alt det andre de driver med. Venstresiden er verre en nasistene, de er tvers igjennom ond og helt uten empati!!! ift. Politisk orientering
Angrepet er rettet mot religion, tro og politisk orientering ved å sammenligne disse med pedofili.

Og hva bidrar somalierne med til samfunnet, bortsett fra å tygge kath
Angrepet tar form av stereotypier om nasjonaliteter og opprinnelse.

Glad for at jeg alltid vil hate homofile. Misliker den Norske Kirke. Ja jeg syne alle burde forlate statskirken. Om du ikke gjør er det er du homo eller mentalt tilbakestående meld deg ut av Statskirken.
Angrepet er basert på seksuell orientering. Ordet «homo» er brukt på en nedsettende måte.

Forbanna pisspreik! Muslimene hater homofile, men elsker å voldta barn, og da bryr de seg ikke om hvilket kjønn det er!
Dette er et angrep med nedsettende og stereotypiserende innhold med hensyn til religion og tro.

[Navn],.... og du er blodrød kommunist, full av løgn! Lavere er det ikke mulig å synke!
Angrepet er basert på politisk orientering.

Den anerkjennende samtalen

Språklige angrep og hatprat er én del av Facebook-debatten. For å danne oss et samlet bilde av debattklimaet har vi også undersøkt språklig anerkjennelse i debatten på Facebook i Norge.

Den anerkjennende delen av kommentarfeltene er et mindre belyst område. Så vidt oss bekjent er det ikke gjort noen andre stordata-undersøkelser av språklig anerkjennelse på sosiale medier enn Analyse & Talls rapport fra 2021.[12] Her er det gjort et grundig stykke arbeid med å definere og operasjonalisere språklig anerkjennelse, som vi har dratt nytte av her.

På bakgrunn av arbeidet i den ovennevnte rapporten har vi definert syv former for språklig anerkjennelse, som står oppført til høyre. Et premiss for den språklige anerkjennelsen er at den skal være rettet mot et menneske. Ved å skille mellom forskjellige typer anerkjennelse får vi et nyansert bilde av den anerkjennende samtalen på Facebook. De syv definisjonene av språklig anerkjennelse har dessuten ligget til grunn for utviklingen av søkenøkkel-algoritmen vår, som oppdager anerkjennelse.

Ros og enighet
Ros for prestasjoner, kompetanse, personlighetstrekk og gjerninger. Ytringer som gir uttrykk for enighet med, eller støtte til, et synspunkt.

Empati og sympati
Språklig uttrykk for medfølelse, empati, omsorg, omtanke, toleranse og sympati.

Anerkjennelse av andre synspunkter og argumenter
Kommentarer som viser til andre mulige synspunkt enn avsenderens egne, og som erkjenner at disse kan ha mening eller vekt (uten at avsenderen nødvendigvis er enig).

Nysgjerrighet/åpenhet
Åpne spørsmål som krever andres kunnskap, holdninger og innspill. Retoriske spørsmål er imidlertid ikke uttrykk for anerkjennelse.

Uttrykk for tvil
Uttrykk for at man er påvirkelig eller tviler, for eksempel holdningsendringer og erkjennelse av feil, anger og unnskyldninger.

Ytring av ønske om dialog/forsoning
Kommentarer som uttrykker et ønske om dialog, felles grunnlag, forsoning, tilgivelse, samtale eller kompromiss.

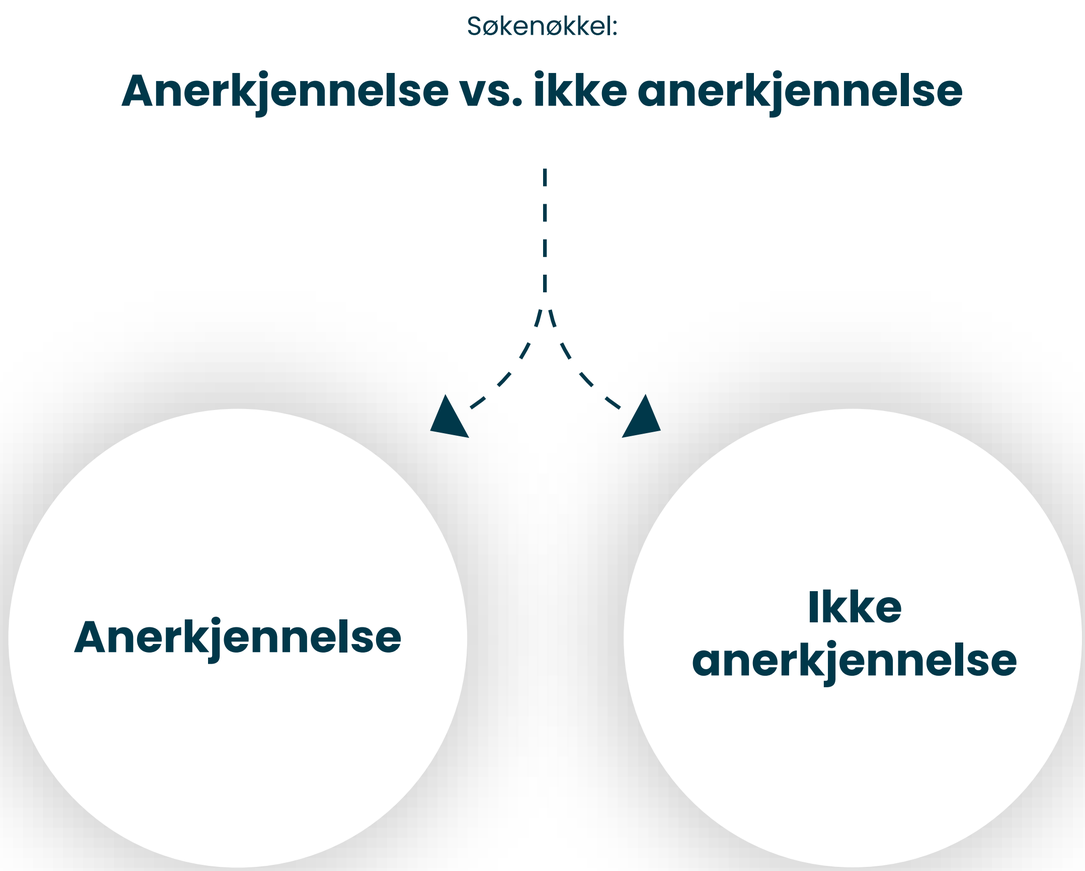
Tillit
Uttrykk for tillit eller tiltro, for eksempel til at en motpart i utgangspunktet vil oss vel. Anbefalinger av andres innspill og meninger.

Søkenøkkel-algoritmen: **Anerkjennelse**

For å undersøke den anerkjennende delen av den offentlige, digitale samtalen har vi utviklet en søkenøkkel-algoritme. Helt konkret har vi operasjonalisert de syv formene for anerkjennelse i en søkenøkkel med mer enn 250 ord og utsagn som indikerer språklig anerkjennelse. Hvis algoritmen finner én av disse i en kommentar, inneholder kommentaren språklig anerkjennelse.

Vi har testet søkenøkkel-algoritmen i flere omganger for å forsikre oss om at den er så nøyaktig som mulig. Det har vært en gjentakende prosess: Vi har revidert søkenøkkelen fortløpende frem til vi hadde ordene og begrepene som utgjør den endelige algoritmen.

Vi har valgt å operasjonalisere språklig anerkjennelse restriktivt fordi enkelte anerkjennende vendinger som «gratulerer» og «kondolerer» ofte brukes ironisk i kommentarfeltene på Facebook. Selv om mange Facebook-brukere reelt anerkjenner andre når de for eksempel gratulerer dem, har vi sett at den ironiske bruken er så utbredt at det skaper støy i datainnsamlingen.



Eksempler på anerkjennelse

Etter vår definisjon er det snakk om anerkjennelse dersom kommentarer inneholder ros, uttrykk for enighet, empati, uttrykk for at andre synspunkt kan være gyldige, åpne eller nysgjerrige spørsmål, uttrykk for påvirkelighet og selvtvil, ønske om dialog eller forsoning eller uttrykk for tillit og tiltro til andre. Anerkjennelsen skal være rettet mot et menneske. En kommentar kan godt både inneholde språklig anerkjennelse og språklige angrep. Figuren nedenfor illustrerer noen eksempler på de syv formene for anerkjennelse.

Slik virker en algoritme, og slik har vi bygget vår

Kort om maskinlæringsalgoritmer

Oppskrift, funksjon eller bruksanvisning – hypet barn har mange navn

En algoritme kan defineres som *en oppskrift på å løse et problem gjennom en regelbasert prosess*. Vi gir algoritmen en input og den gir oss en output avhengig av hvilke regler vi har angitt for prosessen. I sin enkleste form kan en algoritme sammenlignes med en oppskrift eller bruksanvisning: Den følger en trinnvis prosess og gir et ønsket resultat. De enkleste algoritmene sorterer for eksempel tall på en liste etter størrelse, mens de mest komplekse kan gjennomføre svært kompliserte operasjoner. Språkalgoritmen GPT-3 kan for eksempel produsere tekst i ulike sjangre om alle mulige temaer, nesten som et menneske. Dette kan den gjøre fordi den er trent på alle tilgjengelige tekster fra internett.

Input og output

Inputen til vår algoritme er en kommentar fra et kommentarfelt på Facebook, mens outputen er en vurdering av om kommentaren er et språklig angrep etter vår definisjon.

Nevrale nettverksalgoritmer

Moderne, avanserte algoritmer kalles nevrale nettverksalgoritmer fordi de etterligner strukturene vi kjenner fra den menneskelige hjernen. Hjernen inneholder et lag av nevroner (små prosessorer) som er forbundet i et nettverk. Når en input sendes gjennom nettverket, aktiviseres de små prosessorene etter tur. Hver nevron avkoder en liten del av inputen og sender signalet videre.

Dyp læring: mange lag gjør algoritmen «intelligent»

De enkelte «nevronene/prosessorene» i algoritmen representerer relativt enkle regler. Dermed er det dybden i det samlede nettverket, altså antall lag, som gjør algoritmen intelligent, det vil si intelligent i den forstand at den i første omgang kan komme med en kvalifisert antakelse om hvorvidt en kommentar er et språklig angrep eller ei. De ulike lagene i algoritmen inneholder tusenvis av regler som kan handle om kommentarens lengde, forekomsten av ord og emojier, store bokstaver og tegnsetting, rekkefølge og ordstilling. Alle disse reglene har en vekt og vil påvirke algoritmens vurdering av kommentaren.

A&ttack2 – verdens første tverrskandinaviske angrepsalgoritme

Første norske algoritme av sitt slag

Til denne analysen har vi utviklet en ny utgave av A&ttack-algoritmen.[13, 11] Den originale A&ttack-algoritmen er utviklet av Analyse & Tall og kan oppdage språklige angrep på dansk. Den nye algoritmen, A&ttac2, kan klassifisere språklige angrep på norsk og dansk og yter bedre enn tilsvarende enspråklige algoritmer, både på norske og danske eksempler – til og med bedre enn den originale A&ttack-algoritmen. Så vidt vi vet, er det den første (offentlig tilgjengelige) algoritmen som kan klassifisere språklige angrep blant norske Facebook-kommentarer.

Utfordringer i utviklingen av A&ttack2

Det har ikke vært en enkel oppgave å utvikle A&ttack2. Det norske språket på sosiale medier er mer komplekst enn det danske, da Norge har to offisielle skriftspråk og en rekke ulike dialekter. For å bygge A&ttack2 hadde vi behov for en teknisk infrastruktur, en språkmodell, som forstår måten det blir skrevet på i norske sosiale medier. En slik modell fantes ikke. Av den grunn eksperimenterte vi med å lage forskjellige enspråklige algoritmer på norsk, men ingen matchet den kvaliteten vi ønsket.

Gjennombruddet: den tverrskandinaviske språkmodellen

Gjennombruddet kom da vi valgte å bruke en stor, underliggende språkmodell som allerede var trent på en blanding av de skandinaviske språkene. Vi videreutviklet denne modellen ved å trene den på både danske og norske eksempler fra sosiale medier.

Forut for det store gjennombruddet måtte vi teste mange forskjellige språkmodeller. De som fungerte best tilhører familien T5: **Text-to-Text Transfer Transformer**. Disse modellene tar en tekstbit som input og gir en tekstbit som output. De egner seg spesielt godt til oppgaver som oversetting, oppsummering og besvarelse av spørsmål. De kan imidlertid også brukes til klassifisering. T5-modellene finnes i flere størrelser. De største modellene er mer komplekse og krever mer av maskinvaren som skal trene og bruke dem. Til undersøkelsen vår har vi brukt en av de større – en T5-large. De språkmodellene vi har brukt til utviklingen av A&ttack2, er trent og offentliggjort av Per E. Kummervold.

Det nødvendige og rutine-pregede forarbeidet i veiledet maskinlæring

Algoritmen vår er som nevnt basert på en språkmodell som er trent på en stor samling tekstbiter, som deretter kan fintrenes til en bestemt oppgave. I vårt tilfelle vil det være å kategorisere kommentarer som angrep eller ikke-angrep.

Når algoritmen skal lære å oppdage språklige angrep, blir den trent på kommentarer som er kategorisert av mennesker (annotører). Treningsdatasettet vårt består av ~137 000 danske og norske kommentarer fra den offentlige debatten på Facebook. Kommentarene er annotert av mennesker ut fra en vurdering av om de utgjør språklige angrep. En større mengde annoterte data gir algoritmen en bredere forståelse for hva som er et språklig angrep, selv om algoritmen nå også må lære å skille mellom flere språk.

Denne formen for trening kalles *veiledet maskinlæring*. Algoritmen settes til å kjenne igjen mønstre i kommentarene mennesker har kategorisert som språklige angrep.

Den menneskelige vurderingen av hvorvidt en kommentar inneholder et språklig angrep, og hvorvidt angrepet skal klassifiseres som hatprat, er basert på de tidligere nevnte definisjonene og den omfattende kodemanualen med regler og eksempler.

At algoritmene er trent av mennesker, gir oss mange styrker, men også forskjellige begrensninger og skjevheter. Mennesker er ikke like, og derfor er de heller ikke alltid enige, noe som betyr at det vil forekomme feilannoteringer. Heller ingen kodemanual kan sikre 100% stringens, spesielt ikke når det gjelder evalueringen av naturlig språk. Det er både positivt og negativt for algoritmen at mennesker vurderer ting på ulike måter. På den ene siden begrenser det systematiske skjevheter, på den andre siden blir algoritmens presisjonen begrenset. Til tross for dette er vi i det store og det hele begeistret for at mennesker og maskiner i samarbeid har utviklet en teknologi som med rimelig stor suksess kan kartlegge angrep i ca. 10,5 millioner kommentarer i den norske, offentlige debatten på Facebook fra 2020 til 2022.

Algoritmetrening i ti (forenklede) trinn

1 Mennesker utarbeider en definisjon av språklige angrep og hatprat.

2 Mennesker utarbeider en kodemanual til annotører med regler for, og eksempler på, hvordan man gjenkjenner språklige angrep.

3 Mennesker kategoriserer ~137 000 kommentarer etter om de inneholder språklige angrep. De kategoriserte kommentarene utgjør algoritmens trenings- og testdatasett. Ca. 5 prosent av kommentarene annoteres av mer enn én annotør. Disse brukes til å regne ut en score for annotørenes innbyrdes enighet.

4 Mennesker designer en enkel algoritme bestående av en forhåndstrent skandinavisk språkmodell og en klassifiseringsmodul. Algoritmen har en god forståelse av skandinaviske språk, men kan ennå ikke identifisere språklige angrep.

5 Algoritmen gis størstedelen av de annoterte kommentarene (vi sparer noen for å bruke dem som testdata senere) og begynner å gjenkjenne regler og mønstre for når en kommentar skal klassifiseres som angrep, og når den ikke skal det.

6 Den foreløpige algoritmen testes på en input av kommentarer. Algoritmen returnerer kommentarene med sin vurdering av om kommentarene inneholder et angrep.

7 Annotørene «etterannoterer» både kommentarene som algoritmen er sikker på, og kommentarene den er i tvil om. De forteller algoritmen i hvilke tilfeller den har rett, og i hvilke tilfeller den tar feil.

8 Algoritmen får kommentarene i retur med den menneskelige inputen og settes nok en gang til å gjenkjenne mønstre og regler for når en kommentar inneholder et angrep. På den måten korrigeres algoritmens forståelse. Denne formen for gjentakende trening kalles aktiv læring.

9 Vi eksperimenterer med best mulig inputdata til algoritmen. Algoritmen ble for eksempel ikke bedre av å kjenne konteksten til kommentarene, det vil si navnet på Facebook-siden og Facebook-innlegget der kommentaren forekom. Til gjengjeld kjenner algoritmen igjen emojier, og de inngår i algoritmens regler.

10 Algoritmen trenes og justeres frem til den oppnår best mulig resultater på et testdatasett (de sparte, annoterte kommentarene). Gjennom treningen forsøker vi å forbedre to kvaliteter hos algoritmen: presisjon (precision) og tilbakekalling (recall).

Mål for algoritmens ytelse

Presisjon (precision)

Hvis algoritmen har høy presisjon, betyr det at en stor andel av outputen algoritmen returnerer, er blitt identifisert korrekt. I vårt tilfelle vil en algoritme med en høy presisjon returnere en stor mengde kommentarer som er korrekt identifisert som språklige angrep. Dersom vi utelukkende vektlegger høy presisjon, vil algoritmen bare fange en mindre del av det samlede fenomenet vi leter etter, altså alle språklige angrep i den norske offentlige debatten.

Tilbakekalling (recall)

Hvis vi sikter på å finne så mange språklige angrep som mulig, må vi forbedre algoritmens tilbakekalling. Med en høy tilbakekallingskapasitet fanger algoritmen flere språklige angrep. Til gjengjeld vil outputen inneholde mer støy i form av kommentarer som ikke er angrep (falske positiver).

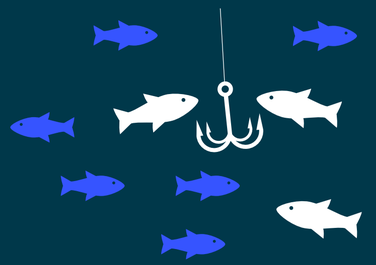
Til sammen - macro averaged f1-score

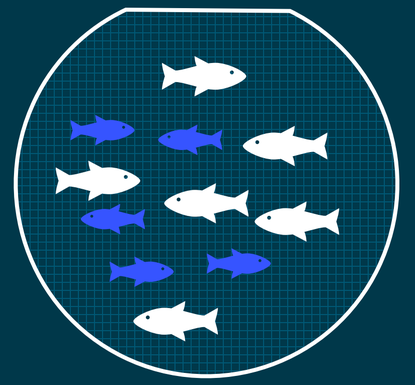
Den samlede vurderingen av algoritmens presisjon og tilbakekalling beregnes i en såkalt macro averaged f1-score. Scoren vekter algoritmens presisjon og tilbakekalling på alle parameterne.

På fisketur

Man kan sammenligne balansen mellom algoritmens presisjon og tilbakekalling med ulike fiskestrategier.

Fisker vi etter torsk, kan vi velge å kaste ut en line som vi vet at særlig torsken biter på. På den måten unngår vi å få annen fisk på kroken, men vi fanger heller ikke så mange torsk. Hvis vi kaster ut et stort garn, fanger vi mange torsk, men også en del fisk vi ikke vil ha. Kunsten er å justere garnet slik at det fanger så mange torsk som mulig og så få uønskede fisk som mulig.





A&ttack2s macro averaged F1-score: **0.76**

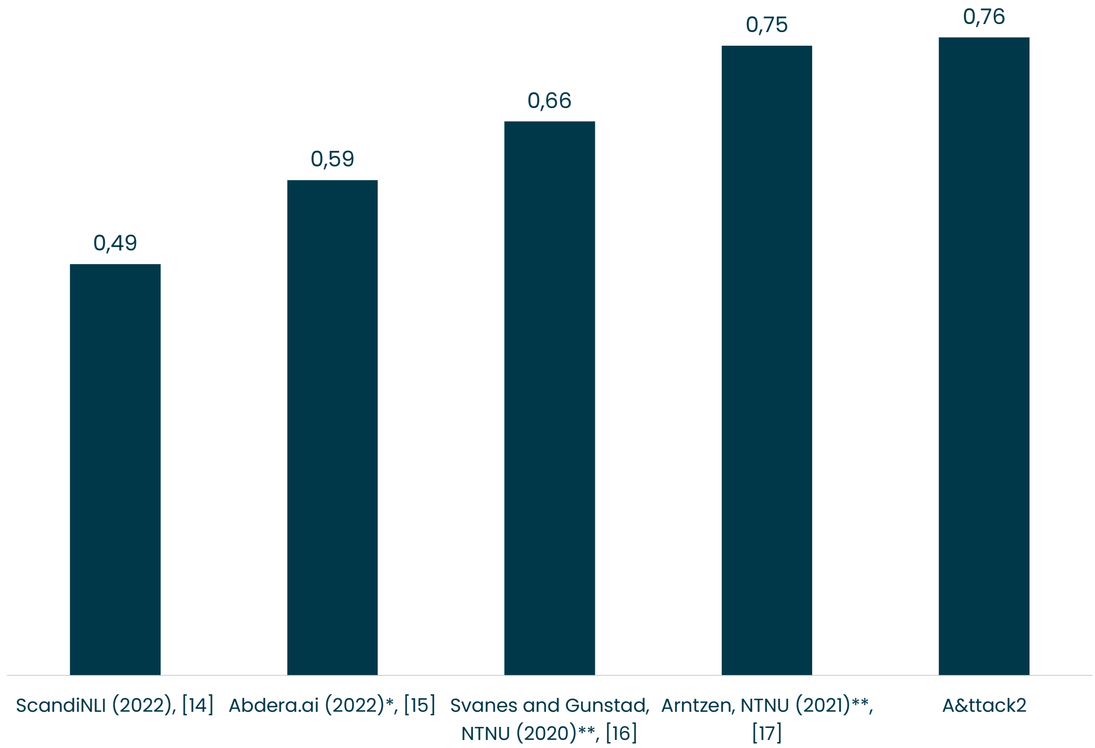
Med mindre man forsker på prosessering av naturlige språk og kunstig intelligens, betyr ikke en slik score så mye. Så la oss se på resultatene for andre algoritmer som er utviklet for å oppdage angrep.

A&ttack2 er klassens beste

Sammenlignet med lignende algoritmer, kommer A&ttack2 best ut. Vi har gått gjennom forskningslitteraturen om andre norskspråklige algoritmer som er utviklet for å oppdage og klassifisere språklige angrep.[14, 15, 16, 17] I figuren til høyre har vi sammenlignet macro averaged F1-score for de norske algoritmene vi fant i gjennomgangen.

Vi har også støtt på angrepsalgoritmer på andre språk, og inntrykket vårt er at vår algoritme, også internasjonalt, er i førersetet til tross for at det utvilsomt finnes svært gode angrepsalgoritmer på engelsk og andre store språk hvor tilgangen til treningsdata er betydelig større. Algoritmenes ytelse er spesielt avhengig av mengden og kvaliteten på treningsdataene. Jo mer og bedre treningsdata algoritmene trenes på, desto bedre blir de.

Makro F1-scorer på fem forskjellige modeller



*: Modellen er evaluert på datasettet vårt, men er trent på et annet datasett, med andre definisjoner.

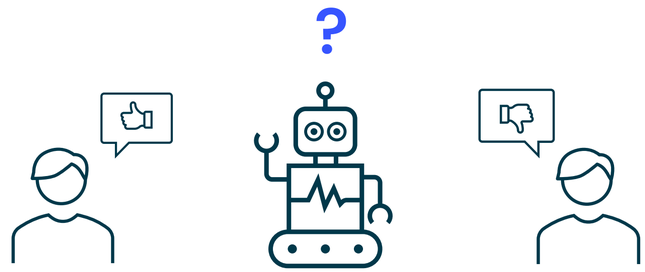
**: Modellen er trent og evaluert på et annet datasett, med andre definisjoner, og F—scoren er derfor ikke direkte sammenliknbar.

Menneskers nøyaktighet er sjelden bedre enn algoritmens

Når vi nå har sett på noen av teknologiens begrensninger, er det bare rett og rimelig at vi også tar en titt på de menneskelige feilratene når det gjelder kategorisering av innhold.

I vårt prosjekt har seks annotører kategorisert i alt ~137 000 kommentarer (~70 000 norske og ~67 000 danske) og avgjort om de inneholder angrep eller ei. De annoterte kommentarene utgjør algoritmens trenings- og testdata. 5 prosent av kommentarene er blitt kategorisert av mer enn én annotør. Disse kommentarene brukes til å regne ut en score for annotørenes innbyrdes enighet.

Beregningen av annotørenes enighet kalles interkoder-reliabilitet. Jo mer kompleks kategoriseringsoppgaven er, desto mindre enighet vil det være mellom annotørene. Til tross for at annotørene bruker en kodemanual med eksempler og regler og dermed søker å være enige i avkodingen, er de bare enige i 83 prosent av kategoriseringene i vårt prosjekt. Tallet høres kanskje ikke overbevisende ut, men med hensyn til oppgavens kompleksitet og sensitivitet og annotørenes subjektive skjevhet er resultatene overraskende gode sammenlignet med lignende undersøkelser.[18, 19]



Algoritmen lærer av menneskene som trener den

I starten av prosessen tester vi hvor godt kodemanualen funger som bruksanvisning for kategoriseringen av angrep. Her hjelper beregningen av interkoder-reliabilitet oss med å få tettet de største hullene i veiledningen til annotørene. Men vi utnytter både den innbyrdes enigheten og uenigheten til å utvikle en algoritme som er mer tro mot definisjonene i kodemanualen. På de områdene hvor annotørene er mest innbyrdes enige, vil algoritmen være mer sikker i vurderingen.

Et siste mål med å teste annotørenes innbyrdes enighet er å avgjøre hvor god vi med rimelighet kan forvente at algoritmen blir. Manglende enighet betyr at algoritmen i en del tilfeller har fått motstridende input fra annotørene, noe som legger noen naturlige begrensninger på algoritmens ytelse.

Kommentarer algoritmen feilaktig kategoriserer som angrep

Du er råtøff! Jeg hadde krøpet sammen og dødd av skrekk...
Algoritmen fanger opp «dødd», som ofte forbindes med angrep, og adjektivet «råtøff», men forstår ikke at det er sagt som et kompliment.

homser er de fineste folka, det har jeg alltid ment :D (har en sønn som er homse, ingen er så god til å skravle og flire som han)
Ordet «homse» kan både brukes som slangord for homofile og som et dehumaniserende skjellsord.

Du er en maskin!!!!!!
Algoritmen gjennomskuer ikke at noen blir omtalt med et kjæle-/kallenavn.

Herregud, du er gal! Jeg ler meg ihjel (på en bra måte, vel og merke). Digger alle dine påfunn
Religion og drap er temaer som ofte forbindes med angrep.

Fuck canser Lykke til videre.
Algoritmen fanger opp banne-/skjellsordet «fuck», men ikke at det er rettet mot en sykdom – som tegn på sympati – ikke mot et annet menneske.

Du er jo klllllllliiiin gæærn
Her brukes skjellsord som ofte er forbundet med angrep. Kommentaren er skrevet til et oppslag om isbading.

Mennesker og algoritmer gjør ulike feil

Når mennesker er uenige om kategoriseringer i språket, er det på en annen måte enn algoritmer. Mennesker diskuterer nyanser: Hvilke fornærmelser er grove nok til å telle som språklige angrep? Kjerring? Klovn? Dust? Og hva med budskap som er underforstått eller sagt som en spøk? Språkforståelsen vår bygger på erfaringer, handlingene som utføres med språket, konteksten og kulturen.

Algoritmen, derimot, har først lært språk av en språkmodell, og deretter har den lært å kjenne igjen mønstre på tvers av kommentarer som er kategorisert av mennesker. Algoritmen gjenkjenner helt andre mønstre enn oss, noe som resulterer i en annen type feil. Hvis vi leser gjennom algoritmens output, finner vi av den grunn jevnlig feil som for oss mennesker er helt åpenbare.

Til venstre kan du se eksempler på kommentarer algoritmen feilaktig kategoriserer som angrep. Vi kan gjerne gjette på hva som medvirker til feilkategoriseringene, men det er ikke slik at vi kan se at algoritmen gjør systematiske feil i de gjeldende situasjonene. De mest systematiske feilene har vi korrigert gjennom treningen.

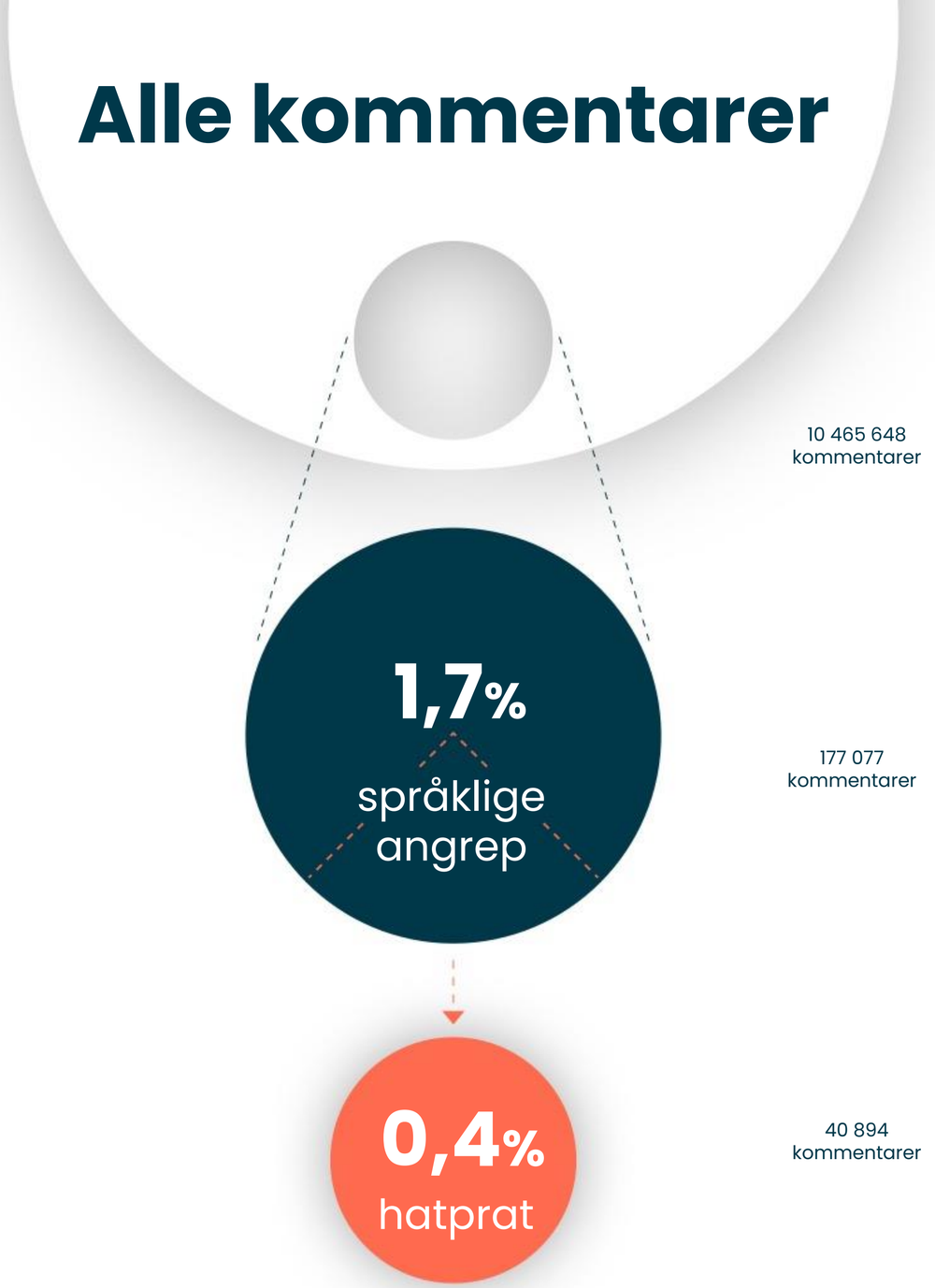
Det store bildet: **Språklige angrep i den brede Facebook-samtalen**

Angrep i den offentlige samtalen på Facebook

I hele den norske Facebook-samtalen er 1,7 prosent av de ca. 10,5 millioner kommentarene kategorisert som språklige angrep. Hele den norske Facebook-samtalen viser her til den delen av samtalen som står igjen etter eventuell moderering, fra personen som har skrevet kommentaren, Facebook selv eller fra moderatorer eller administratorer av den enkelte siden.

Ut av 1,7 prosent av kommentarene som karakteriseres som språklige angrep er 0,4 prosent hatprat. Det vil si at en knapp fjerdedel av alle angrepskommentarer er definert som hatprat.

I denne undersøkelsen er den offentlige samtalen i Norge definert som innlegg og kommentarer på Facebook-sider som tilhører norske politikere, medier, offentlige personer og offentlige debattsider. Vi har samlet inn data i perioden 1. januar 2020 til 27. september 2022.



4

Tema som genererer angrep

Tema som fører til en hard debatt

Figuren til høyre viser klart og tydelig at debatter om islam og integrering og kriminalitet dominerer når det gjelder språklige angrep i kommentarfeltene. De ti temaene med størst andel språklige angrep er i all hovedsak relatert til dette på ulike måter.

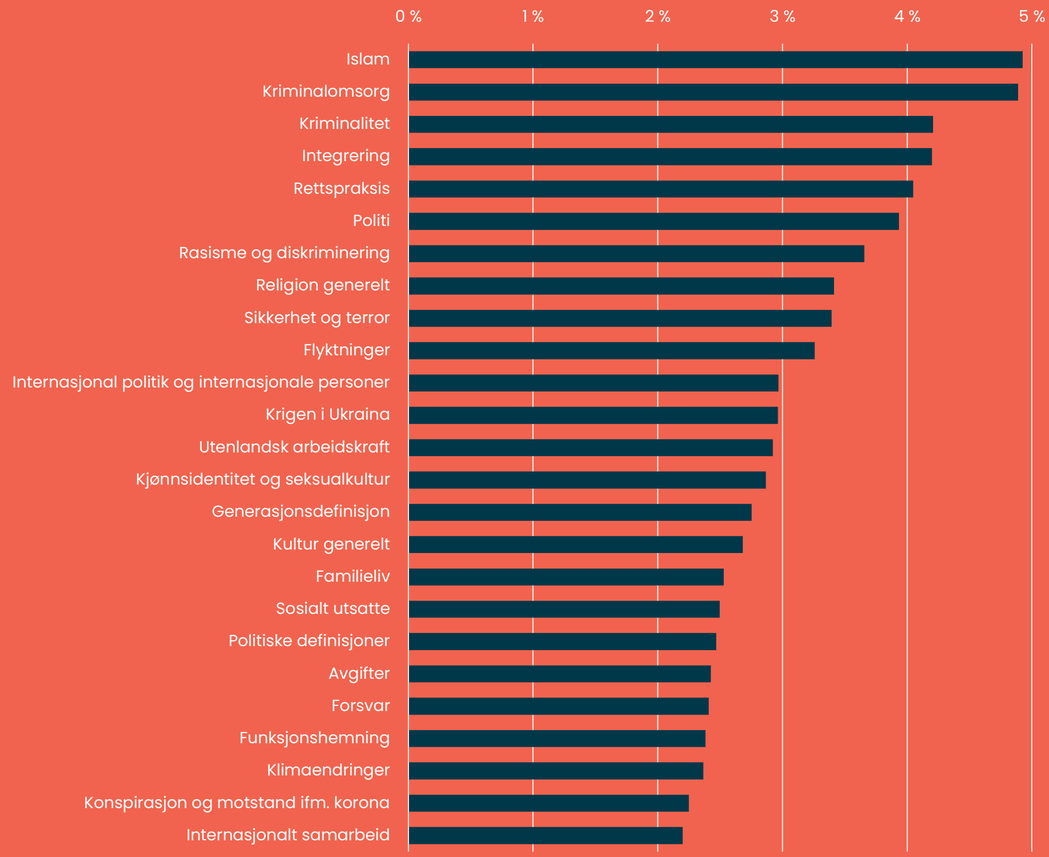
Først på ellevte plass finner vi et nytt tema: internasjonale forhold. Her er det snakk om onlinedebatter som typisk handler om den sikkerhetspolitiske situasjonen i Europa, blant annet Norges NATO-medlemskap, og om amerikansk politikk med stor vekt på Biden vs. Trump. Utover det fører også temaene kjønn og seksualitet til språklige angrep, og da dreier det seg særlig om LHBT+-personer og diskusjoner om feminisme og kjønn.

For å undersøke hvilke tema som fører til en hard debatt, har vi delt de vel 10,5 millioner kommentarene som utgjør analysens datagrunnlag, i 94 temaer. Ved hjelp av angrepsalgoritmen og en emnesøkenøkkel har vi kunnet se hvilke tema som har størst andel angrepskommentarer. Figuren til høyre viser hvilke 15 temaer som ligger på topp.

De tøffe temaene i samfunnsdebatten

For å kartlegge hvilke tema som har den høyeste andelen språklige angrep, har vi utviklet en emnebasert søkenøkkel. Figuren nedenfor viser de 15 temaene som har høyest andel språklige angrep. Bak hvert tema gjemmer det seg lange lister med søkeord.

Topp 15 temaer i den offentlige, norske debatten med størst andel språklige angrep



Diskusjoner om islam og muslimer er voldsomme

Islam er det temaet som setter de norske følelsene mest i sving, og som fører til den høyeste andelen angrep i den offentlige, digitale samtalen. Av den grunn har tema som kan knyttes til islam og Midtøstens kultur, for eksempel integrering, rasisme og diskriminering, religion generelt og flyktninger, også en høy andel angrep.

Mange av angrepskommentarene er spesifikt rettet mot muslimer. De anklager dem som befolkningsgruppe for å være «ikke-norske», «voldelige» og bærere av en «middelaldersk» religion. I disse tilfellene defineres kommentarene som hatprat nettopp fordi de går på beskyttede karakteristika som religion og etnisitet.

I denne delen av kommentarfeltene finner vi også elementer av tankegangen som sier at det foregår en bevisst utskiftning av de etnisk-vestlige befolkningene til fordel for muslimer. De kommer blant annet til uttrykk i kommentarer som anklager venstreorienterte politikere eller statsministere for å «importere» muslimer til Norge.

Det er imidlertid ikke bare muslimer som står i skuddlinjen. Mange angrepskommentarer er rettet mot de politiske fløyene. Folk med innvandrerkritiske holdninger på høyresiden kalles «rasister» og «nazister», mens den flerkulturelle venstrefløyen beskyldes for å være «hjernedøde» og «landsforrædere».

Eksempler på angrep og hatprat innenfor temaene islam, flyktninger, integrering, rasisme og diskriminering samt religion generelt

> *Islam er en diagnose,*
> *en ond jævel sådan*

> *Alle surrehuene her som kneler for Islam kan dra til Iran eller Saudi-Arabia og oppleve litt ekte shariaretferdighet som pisking, steining og kapping av noen hender. Og skulle du være så uheldig å bli voldtatt så vil du kjenne pisken flerre huden av ryggen din.....Det er Islam.*

I debatten om lov og orden settes hardt mot hardt

Andelen angrep er høy når det gjelder temaene kriminalitet, rettspraksis, kriminalomsorg og politi. Det blir en følelsesladd debatt når folk bryter loven, men ikke alle forbrytelsene straffes like hardt i kommentarfeltene.

I angrepskommentarene ønskes kriminelle med muslimsk bakgrunn eller innvandrerbakgrunn utvist når debatten handler om dem. Dessuten står det i mange kommentarer at muslimer og innvandrer i Norge er «hevet over loven» og får lov til å oppføre seg som de vil, og at de kan bryte loven uten at det får konsekvenser.

Hvis kriminalitetsdebatten derimot handler om unge som begår kriminalitet, utvikler det seg raskt til anklager om manglende oppdragelse av disse «drittungene». Selv om dette gjelder alle unge kriminelle, er det spesielt uttalt når samtalen handler om unge demonstranter.

Den gruppen kriminelle som straffes desidert hardest i den offentlige debatten på Facebook, er imidlertid de som voldtar eller skader dyr eller barn. Her kan ikke straffen bli hard nok. I debatter om denne typen kriminelle oppfordres det til voldsbruk eller til at de må fratas sine grunnleggende menneskerettigheter.

Eksempel på angrep innenfor temaene kriminalitet, rettspraksis, kriminalomsorg og politi

”

[Navn] æ kan si dæ [navn], penisen skulle bli kutta av og sydd i panna, så fengsel blant bare homoer, dette for dem som har forgrepet seg mot barn og damer!!!

Konflikter og krig i verden sprer seg til kommentarfeltene

Når konflikter, krig og storpolitiske kamper utkjempes andre steder i verden merkes dette godt også i den norske, digitale debatten. Dette gjelder spesielt diskusjoner om terrorangrep (særlig når de begås av muslimer), NATO-medlemskap og Jens Stoltenbergs evner som generalsekretær.

Mange av de språklige angrepene i kommentarfeltene, nærmere bestemt 3 prosent, handler om krigen i Ukraina. I mange kommentarer uttrykkes det sinne mot krigens hovedpersoner: den russiske presidenten Vladimir Putin og den ukrainske presidenten Volodymyr Zelenskyj. Men Facebook-brukere med ulike holdninger til krigen angriper også hverandre. I kommentarfeltene angriper de motstanderne med betegnelser som «trolls», «nazielskere» og «konspirasjonsteoretikere».

En lignende dynamikk og retorikk ser vi også når det gjelder amerikansk politikk, hvor samtalen er delt mellom de som er for Trump, og de som er for Biden. Disse kommentarene inneholder typisk ord som «idiot» og «dum» og anklager om lav IQ.

Det er altså utbredt å anklage meningsmotstanderne sine for å være uintelligente eller troll. På den måten avviser de motstandernes argumenter som ugyldige og fremmer sin egen side av saken.

Eksempel på angrep innenfor temaene internasjonal politikk og internasjonale personer, krigen i Ukraina samt sikkerhet og terror

”

[Navn] her viser du hvor infantil og uvitende naiv du er. Stakkars lille menneske. Skaff deg hjelp hos noen voksne du lille venn.....

Kjønn og seksualitetens rolle i offentligheten trekker folk til tastaturet

Det personlige er politisk – og har potensiell sprengkraft i debatter på Facebook. Dette ser vi ved at nesten 3 prosent av alle kommentarer som handler om temaet kjønnsidentitet og seksualkultur, er språklige angrep.

Pride er en av tingene som trekker folk til tastaturet. Det er et offentlig uttrykk for seksualiteter og kjønnsidentiteter som i noens øyne faller utenom de tradisjonelle normene. Pride-deltakere beskyldes for å være oppmerksomhetskrevende når de går i «homsetog», og regnbueflagget kalles politisk propaganda. Men det er også kommentarer som går til motangrep med formuleringer som «homofobisk kjerring» og «Forstår du hvor idiotisk dine meninger er?».

Kommentarfeltene inneholder ikke bare språklige angrep når det gjelder legning og Pride. Også feministiske og/eller andre skeives synspunkt utsettes for angrep. Da handler det ofte om konsekvensene av for eksempel metoo-bevegelsen eller ikke-binære kjønnsforståelser. Noen anklager «kvinnfølka» og «woke kvinder» for å mangle biologisk viten samt for å være for sensitive, mens andre gleder seg til «de gamle hvite menn» og deres «konservative cis-agenda» dør ut.

Eksempler på angrep innenfor temaene kjønnsidentitet og seksualkultur

Herregud da. Skal homsene karre til seg oppmerksomhet på dette også nå ? Var det ikke nok med eget flagg, egen parade og egen monkepox virus ??? Forbanna sutrehuer !

Men de behøver vel ikke demostrere flere ganger i året. Har ikke noe imot homofile, men de gjør så man får de i vranghalsen. Kan vel leve uten at må gå i homsetåg. Ser ut at det blir flere og flere skeive.

Tendenser innenfor hatprat

Noen grupper rammes særlig hardt

Her viser vi hva som kjennetegner de nedsettende og sjikanerende kommentarene som er rettet mot beskyttede karakteristika, og som derfor kan betegnes som hatprat.

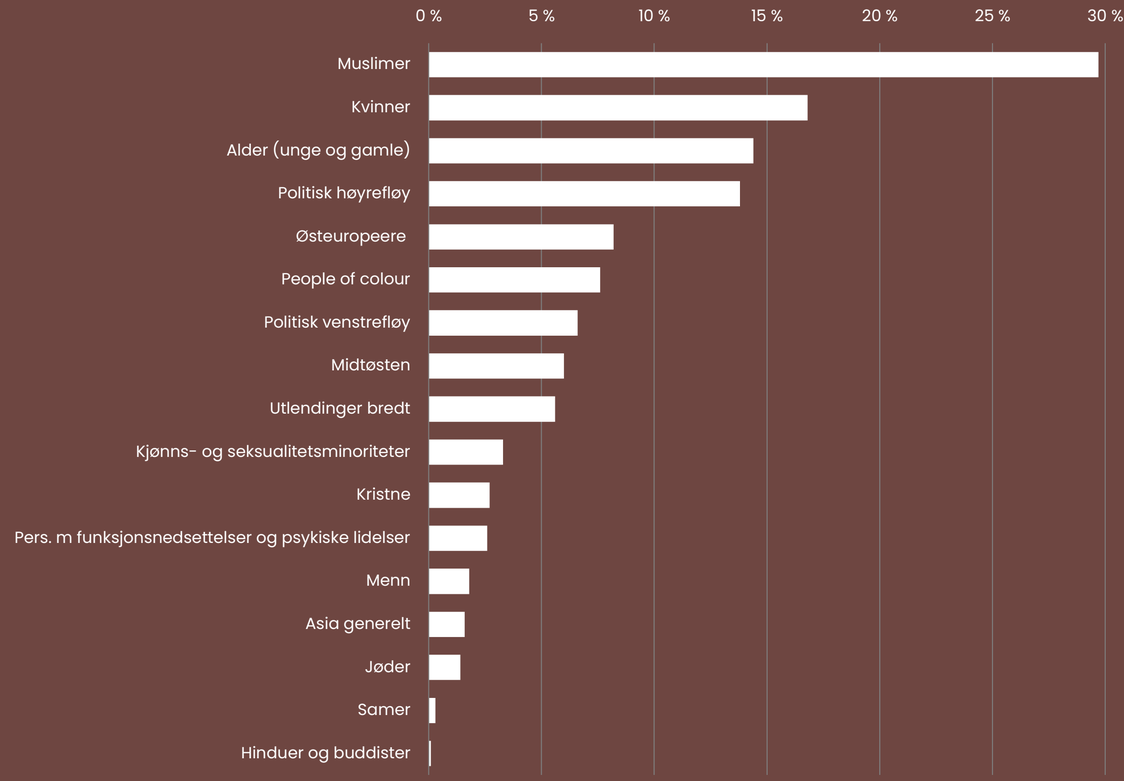
Det er nemlig ikke slik at hatet i kommentarfeltene rammer oss likt. Hvordan man oppfatter ytringsklimaet på Facebook, henger i stor grad sammen med hvilke debatter man følger, og hvilken gruppe man selv tilhører.

Ved hjelp av søkenøkkelen vår finner vi litt mer enn 40 000 av de 177 000 angrepskommentarene som nevner betegnelser for beskyttede karakteristika. Innarbeidede diskurser gjør at søkenøkkelen treffer enkelte grupper mer presist enn andre. Kommentarer om jøder avdekker for eksempel oftere muslimhat enn antisemittisme, og buddhister og hinduer blir brukt i motsetning til muslimer. Hatefulle ytringer mot de største gruppene treffer som hovedregel presist.

Hvordan vi kartlegger angrep mot beskyttede karakteristika

For å kartlegge hvilke grupper som er særlig utsatt for angrep i den offentlige debatten, har vi utviklet en egen søkenøkkel. Vi har en liste med i alt 408 ord for beskyttede karakteristika som enten er nøytrale (migrant, samisk, liberal) eller nedsettende (soper, sotrør, retard). Når søkenøkkelen får et treff blant de språklige angrepene, anser vi kommentaren som hatprat. Fordi kommentarer kan nevne mer enn ett beskyttet karakteristikon, summerer figuren til mer enn 100 prosent.

Andel hatprat fordelt på grupper



Muslimer legges oftest for hat

Så mange som ett av tre språklige angrep som rammer en gruppe basert på beskyttede karakteristika, er rettet mot muslimer. At det er et fiendtlig ytringsklima rundt muslimer, understrekes ved at dette er den gruppen som er mest utsatt i kommentarfeltene til både riksdekkende medier og norske politikere.

Det er særlig innlegg om dagsaktuell politikk, politikere med muslimsk bakgrunn og kriminalitet som fører til muslimhets i kommentarfeltene. Vi finner for eksempel følgende begrunnelse for at Abid Raja (V) ikke er egnet som statsråd: «Han er muslim. Jeg stoler ikke på muslimer.» Kommentaren illustrerer hvordan politikere med muslimsk bakgrunn angripes på bakgrunn av religion.

Det er også en tydelig sammenheng mellom politikerforakt og muslimhat. Politikere i regjeringsposisjon blir ofte beskyldt for å ødelegge Norge gjennom en for liberal flyktning- og innvandringspolitikk. Til grunn for beskyldningene er en forestilling om at muslimer skal «ta over» og fortrenge norske verdier, og om at muslimer skaper et mer voldelig samfunn.

Løsningen på voldsproblematikken blir for mange å sende dem hjem: «Send faenskapen tilbake til disse sharialandene de kommer fra !!!!!!!!!» Sharia er et begrep som dukker opp i debatten relativt ofte, og det tas blant annet til inntekt for dehumaniserende og voldelige utsagn: «Straff de etter sharia lovene de er så glad i. Koster bare 1 skudd og ingen rettsak.»

Eksempel på hatprat rettet mot muslimer

> *Du ønsker altså at Norge ditt fedreland skal bli islamisert? For et naut. Du sier du er i mot vold, men det kan du ikke være. Du støtter islamister som ønsker å dominere over deg.. håper du har burkaen klar i klesskapet. Jævla fjolle*

Opprinnelse i Midtøsten og Nord-Afrika utløser hat

I mange tilfeller er det overlappende betegnelser for etnisitet og religiøs tilhørighet i kommentarer som angriper muslimer og personer fra land i og rundt Midtøsten og Nord-Afrika (MENA-regionen). Det brukes for eksempel betegnelser som «muslimske land». Hvis vi ser på muslimer og grupper fra MENA-regionen samlet, mottar disse nesten halvparten av alt hatprat vi har identifisert. Dette understreker at ytringsklimaet rundt etniske minoriteter fra denne delen av verden og muslimer er særlig hardt. Dersom vi bryter hatpratet ned på hvilke land i MENA-regionen de er rettet mot, finner vi at personer med opprinnelse i Somalia, Pakistan og den arabiske verden («arabere») er mest utsatt.

Angrepene handler ofte om at personer fra Midøsten og Afrika utnytter det norske systemet. Her står forestillingen om personer med innvandrerbakgrunn som lykkejegere sentralt. «Syns et annenrangs sotrør burde holde kjeft, lykkejegere som trur dem er noe skal rett og slett bare sendes hjem, og ta fra dem statsborgerskapet, har ikke noe i Norge og gjøre (…)». Kommentarene inneholder også grov og rasistisk ordbruk rettet mot personer med opprinnelse i MENA-regionen.

Stereotypiske fremstillinger av personer fra Somalia og andre MENA-land er utbredt. «Det eneste en somalier kan er å lage unger, det er et lett håndverk som suger ut penger fra statskassen.»

Eksempler på hatprat rettet mot mennesker med opprinnelse i Midtøsten og Nord-Afrika

Dette er Afghaner med 11 sedelighetssaker på samvittigheten hvor 3 var mindreårige!! Alle vet hvem som står bak men media er pålagt å skjerme [navn] elskelige skjeggebarn. Når har vi fått nok av denne naïve importen av voldelige fra muslimske land

[Navn], sånne kurder som dæ er det verste type mennesker. Elsker sitt folkets fiende! Med andre ord du har Stockholm syndrom. Get help

Pakistan har skolet dem til de vanvittige voldsmænd og terrorister de er.

Kvinnehets i kommentarfeltene

Dersom vi ser på forholdet mellom hatprat rettet mot kvinnelige og mannlige karakteristika, blir det tydelig at kvinner må tåle et langt hardere ytringsklima. Rett under 17 prosent av de hatefulle kommentarene som er rettet mot beskyttede karakteristika, rammer kvinner, mens tilsvarende andel for menn er rett under 2 prosent.

En forklaring på den høye andelen er at kvinnelige karakteristika ofte blir brukt for å kneble og avfeie kvinner i debattråder eller for å undergrave kvinner som opptrer i offentligheten. Dette gjelder særlig kvinnelige politikere. Her brukes typisk nedsettende karakteristikker om alder, utseende og intelligens.

Kommentaren «Bare gamle fete kjerringer som bestemmer i Norge fra nav til regjeringen, og du ser jo åssen det går» illustrerer hvordan kropp og alder blir brukt for å avfeie kvinner i maktposisjoner.

«Kjerring» er ordet som oftest brukes nedsettende om kvinner i Facebook-debatter. Vi ser også at ordet «fitte» ofte går igjen i hatprat som rammer kvinner, sammen med «hore», «kvinnemenneske» og «heks». Selv om det går lenger mellom hver gang menn hetses i kommentarfeltene, er det noen likheter i de verbale angrepene. Også menn blir forsøkt undergravd og latterliggjort med nedsettende kommentarer om alder og seksualisert ordbruk. «Jypling», «tusseladd», «horebukk» og «pikk» er ord som går igjen.

Eksempler på hatprat rettet mot kvinner

Å være ufin mot den søppel kjerringa dær er helt innafor. Makan til ekstremist hore skall du lete lenge etter. Å tenke miljø er en ting men å legge ned Norge blir no annet.

Kan du holde kjeft di gamle fitte og innse at du burde være smigret og i god likestillingsånd sendt et bilde by request i retur! Da hadde du vunnet! Jævla taper

Hardt ordskifte om politiske standpunkt

I digitale debatter er ulike politiske meninger og holdninger ofte kime til konflikt og angrep. Når verdispørsmål og politikk diskuteres fra tastaturet, har vi i våre tidligere analyser sett at debattene ofte blir polariserte, og personer som har politiske meninger og holdninger som skiller seg fra konvensjonell politikk, er mer utsatt for språklige angrep. Til sammen er én av fem angrepskommentarer rettet mot karakteristika som er forbundet med den politiske høyre- og venstrefløyen.

Vi finner imidlertid holdepunkt for at det ikke er likegyldig hvilken fløy man tilhører. Karakteristika som er forbundet med høyreorientert politikk, går igjen i dobbelt så mange angrepskommentarer som karakteristika som er forbundet med venstreorientert politikk.

Særlig utbredt er bruken av «nazist», «fascist» og «rasist» for å stemple meningsmotstandere. Ofte brukes dette om personer som støtter høyresiden i amerikansk politikk, som argumenterer i favør av Russland og Putin, eller som ønsker en strengere innvandringspolitikk i Norge. Lignende karakteristika blir også brukt om FrP-politikere, for eksempel: «Du er en reinspikka rasist av verste sort» og «Nazikrapyl». Politikere og meningsmotstandere på venstresiden blir på sin side stemplet som «kommunister» og «jævla sosialister».

Eksempler på hatprat rettet mot politisk ståsted

… Stakkers ungene dine. Er de i det hele tatt stolte av å ha en så rasistisk far som deg? Enten så har du hatt dårlig oppvekst av dine foreldre eller så ha du blitt mobbet ganske mye i ungdomslivet ditt. Syns ganske synd i familien din, om du i det hele tatt har en…

[navn] Har du ikke evne til å lese hva han holder på med ? Om du støtter denne psykopaten er du enten en nazisympatisør eller direkte dum!!

På Facebook går politikk og angrep hånd i hånd

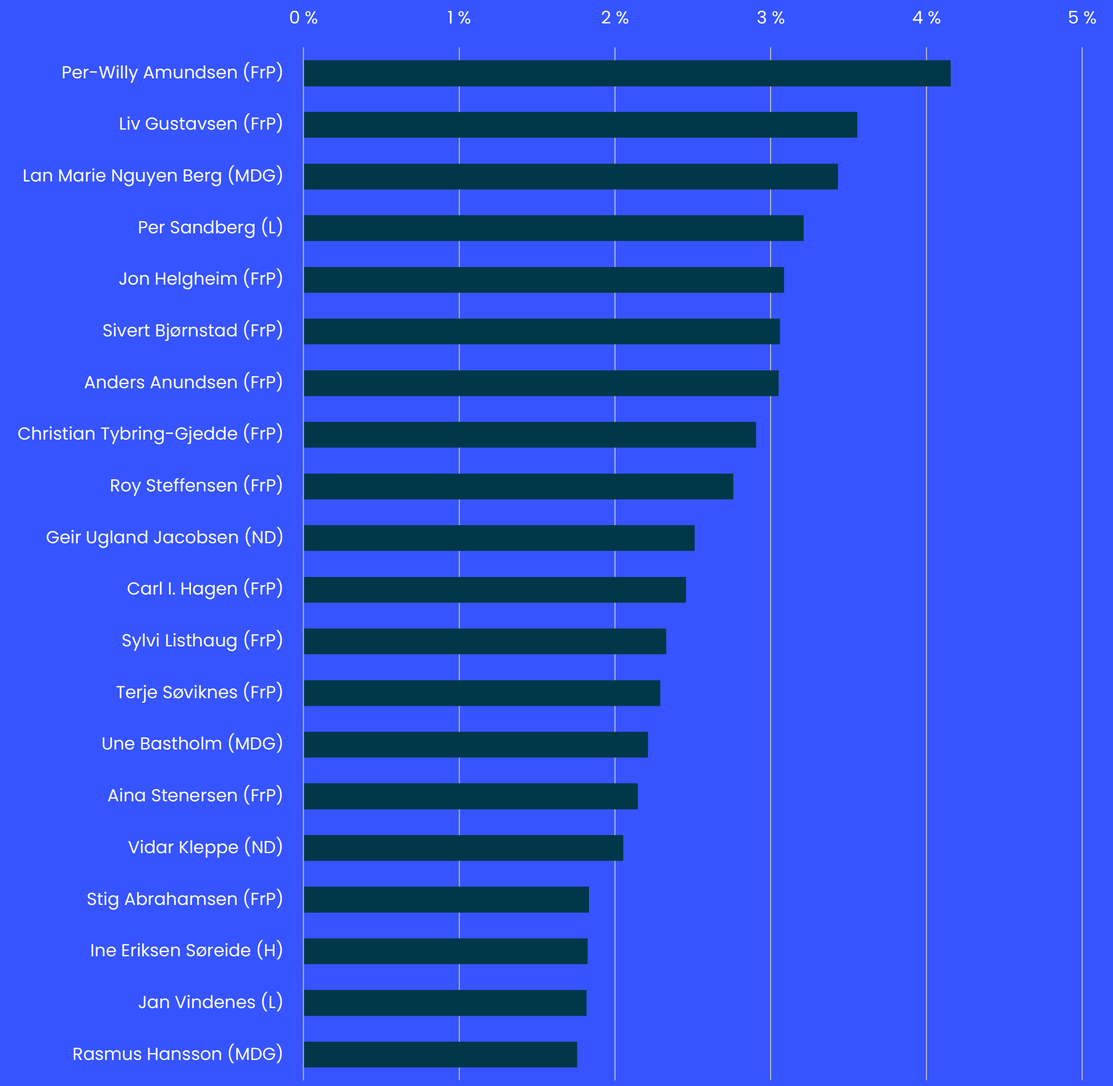
Politikerne og partiene med størst andel angrep

Figuren, eller listen, til høyre viser hvilke 20 politikere som har den største andelen språklige angrep i kommentarfeltene på de offentlige Facebook-sidene sine. Representanter fra Fremskrittspartiet (FrP) dominerer listen, med tolv politikere. Etter FrP finner vi tre politikere fra Miljøpartiet De Grønne (MDG), to fra Liberalistene (L), to fra Norgesdemokratene (ND) og én fra Høyre (H).

Til sammen har vi samlet inn over 2,2 millioner norskspråklige kommentarer på politikernes sider. Når vi legger sammen antall kommentarer på sidene etter partitilhørighet, finner vi at hele 45 prosent av kommentarene er postet på sidene til politikere fra FrP. Den høye aktiviteten i kommentarfeltene viser at FrP er svært gode til å engasjere følgerne sine. Samtidig ser vi at det er en høyere andel angrep i FrPs kommentarfelt. Dette kan indikere at postene nettopp engasjerer ved å spille på tema som vekker følelser og sinne hos følgerne.

I tillegg til at politikerne kommuniserer på ulike måter i postene de legger ut, fortoner de språklige angrepene i kommentarfeltene seg ganske ulikt. På den ene siden utsettes politikerne selv for personangrep og hets. På den andre siden blir kommentarfeltene brukt som en arena for hets og sjikane mot andre enkeltindivider og grupper. Felles for mange av politikerne som opptrer på listen til høyre, er at de ofte skriver om temaer som innvandring, integrering, kriminalitet og miljø. Og vi vet fra før at dette er konfliktfylte temaer som folk ofte har sterke meninger om.

Politikersidene med den største andelen angrep



Harde debatter hos FrP-politikerne

FrP-politikerne har både større engasjement og flere språklige angrep i kommentarfeltene sine enn de andre partiene. Vi ikke kan si noe sikkert om i hvilken grad FrP-politikerne lar flere kommentarer stå umoderert, men kombinasjonen av tema og retorikk i postene ser ut til å påvirke det som utspiller seg i kommentarfeltene.

Målt i absolutte tall finner vi flest språklige angrep i kommentarfeltet til FrP-leder Sylvi Listhaug, mens lederen av justiskomiteen på Stortinget, Per-Willy Amundsen, har den største andelen. Angrepene i FrP-politikernes kommentarfelt har flere likheter og kommer i stor grad under poster om kriminalitet, islam og muslimer, innvandring og integrering. Dette er tema som FrP-politikerne ofte tar opp, gjerne med en oppfordring til følgerne sine om å kommentere om de er enige i budskapet. Dette retoriske grepet bidrar til engasjement, men ofte også til en mer polarisert debatt. Når temaet er kriminalitet og innvandring vises det liten nåde i kommentarfeltene, og vi finner direkte oppfordringer til vold.

Pensjon og skatt er også tema som vekker mye sinne i FrP-politikernes kommentarfelt. Da er angrepene som oftest rettet mot politikere og politiske meningsmotstandere. Jan Tore Sanner (H) blir kalt «rotte», «landssviker» og «taper», mens Trygve Slagsvold Vedum (Sp) karakteriseres som «klovn» og «kukhue». Meningsmotstandere med andre politiske overbevisninger blir avfeid med formuleringer som «venstrevridde idioter» og «jævla kommunister».

Eksempler på angrep i kommentarfeltet til FrP-politikere

Elektro våpen og køller. Slå dem helseløs og gi dem 500 volt med våpen.

[Navn] de er de verste hvor enn du kommer ran, knivstikking voldtekter mm, verste folkeslag som finnes de som roper Allah akbar, svineriet Norges politikere har hentet hit til lands

Miljøpolitikk møtes med personhets

Innlegg om grønn omstilling, utslippskutt, energiforsyning og klimaendringer fyller mesteparten av Facebook-sidene til MDG-politikerne. Dette er tema som virkelig setter sinnene i kok hos enkelte meningsmotstandere, og som møtes med harde angrep i kommentarfeltene.

Mens angrepskommentarene på FrP-politikernes sider i høy grad er rettet mot «ytre fiender», som innvandrere, muslimer og politiske meningsmotstandere, er angrepene på MDG-politikernes sider som hovedregel rettet mot politikerne selv. Angrepene på MDGs sider kommer nemlig primært fra personer som ikke støtter MDGs politikk.

Kombinasjonen ung, kvinne og miljøpolitikk viser seg å være et særlig mål for mange, og vi finner mange eksempler på grov personhets og sjikane. Ord som «drittunger» og «jentunger» går igjen i kommentarfeltene til både Lan Marie Berg og Une Bastholm, sammen med spørsmål som «kan vi få stoppa disse kvinnemenneskene». Noen gir seg heller ikke med retoriske spørsmål og tar til orde for både selvmord og drap.

Istedenfor å møte MDGs politiske utspill med motargumenter avfeier mange også politikerne og støttespillerne deres som «utilregnelige», «syke» og som en «sekt med sinnsforvirrede medlemmer». Denne formen for umyndiggjørende retorikk finner vi også flere eksempler i kommentarfeltet til Rasmus Hansson. I motsetning til Berg og Bastholm er Hanssons alder og kjønn i liten grad gjenstand for angrep i seg selv, men flere bruker ulike uttrykk for Hanssons manglende intelligens og dobbeltmoral.

Eksempel på angrep i kommentarfeltet til MDG-politikere

> *Mdg er ikke et folkeparti,men en sekt for hjernevaskede idioter.Gærne jævler*

Til angrep på regjeringsmakten

Høyre og Arbeiderpartiet har begge hatt regjeringsmakt i perioden vi har undersøkt. Når vi ser nærmere på Facebook-sidene til Erna Solberg (H) og Jonas Gahr Støre (Ap), viser det seg at det følger en økt grad av språklige angrep og negative karakteristikker med statsministerjobben.

Både Solberg og Støre stilles til ansvar for regjeringens innvandringspolitikk. De blir begge anklaget for å prioritere innvandrere, muslimer og overnasjonale organer på bekostning av Norge og nordmenn. Mange av angrepene kjenner vi igjen fra kommentarfeltene på sidene til FrP og innvandringskritiske partier som Norgesdemokratene. I mange tilfeller ser det ut til at de som ytrer seg mest innvandringskritisk i kommentarfeltene, ikke skiller mellom Ap og Høyre.

Videre blir Høyres og Aps evne til krisehåndtering satt på prøve i partienes regjeringsperioder. Koronapandemien får det til å gå hardt for seg både i kommentarfeltet til Solberg og helseminister Bent Høie (H). I angrepskommentarene er Høies seksualitet også en skyteskive, med karakteristikker som «homo idiot». Flere av angrepene mot Solberg inneholder også nedsettende ord og uttrykk om hennes kjønn og utseende.

Aps håndtering av den pågående økonomiske krisen, og særlig de høye strømprisene, får hard medfart i Støres kommentarfelt. Etter kommentarene å dømme har flere mistet tillit til at venstresiden fører en politikk som er til det beste for «folk flest», og de tar frustrasjonen ut i karakteristikker av Støre som «motbydelig egosentrisk vesen», «kjeltring» og «korrupt».

Eksempel på angrep rettet mot Høyre

Ingen vits å spørre 2 mongolide apekatter. Bedre og spørre barnehageunger dem har nokk mere kunnskap en disse to fjøsnissene

Eksempel på angrep rettet mot Arbeiderpartiet

Ser du har brutt de fleste løfter, større svin enn du og politikere generelt er, må man lete lenge etter. Ser tåkefyrsten fortsatt prøver å lyve eu inn. Du er jaggu et svin jonas mot norske velgere, men du liker vel helst velgerne fra muslimske kår. Fy faen for et menneske du er, penger er alt

Det harde debattklimaet hos mediene

Medier med størst andel angrep

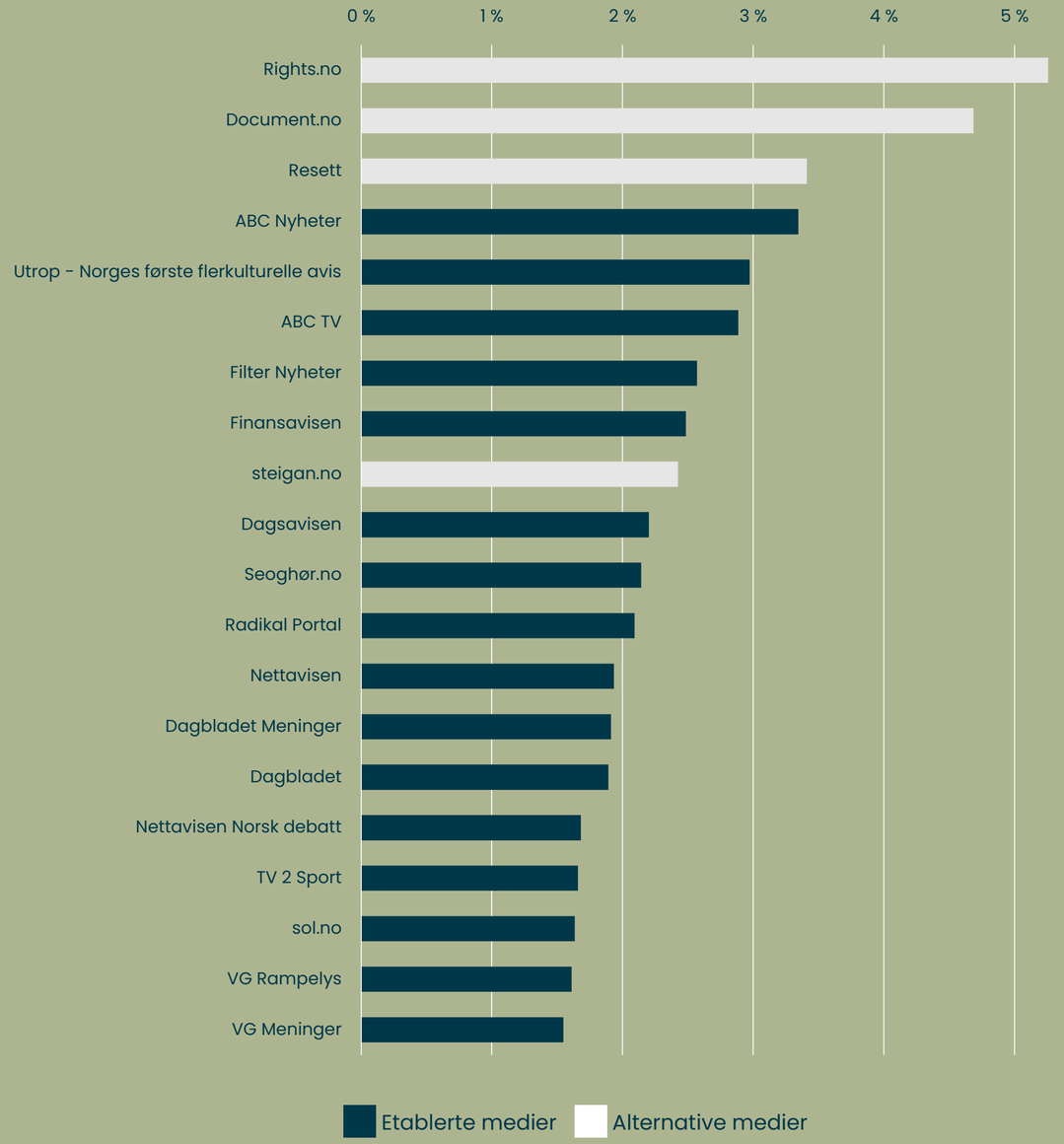
Figuren til høyre viser hvilke 20 medier som har den høyeste andelen angrep på Facebook-sidene sine. Rights.no topper listen med 5,3 prosent språklige angrep, mens Document.no ligger på andreplass med 4,9 prosent.

Både Rights.no og Document.no er blant mediene som ofte betegnes som alternative fordi de kjent for å ha en sterk politisk agenda og er i opposisjon til de etablerte mediene. Alternative medier er – til tross for størrelsen – sterkt representert på topp 20.

Flere av mediesidene som har den høyeste andelen angrep, finnes utelukkende i digital versjon, for eksempel Rights.no, Document.no, Resett og ABC Nyheter. Og nettsidene til de store, tradisjonelle (papir)mediene som Aftenposten og VG glimrer med sitt fravær. Moderering er en sannsynlig forklaring på dette.

Totalt har vi samlet inn 6 556 986 kommentarer fra mediesidene, og av dem er 108 161 kategorisert som å inneholde språklig angrep. Det vil si at 1,6 prosent av alle kommentarer på mediesidene er språklige angrep. Men majoriteten av sidene på listen til høyre har en høyere andel angrep på Facebook-sidene sine, og på de neste tre sidene ser vi nærmere på hvorfor det er slik.

Nyhets- og mediesider med den største andelen angrep



I nettmediene blir det personlig

Når temperaturen er høy, kommer debattantene ofte med personangrep i kommentarfeltene. Da går de gjerne etter motpartens intelligens eller autentisitet, med beskyldninger som «du er en falsk person» eller «du er elendig på personangrep, ikke glem hjernetrim, ditt kjøtthodet».

Personangrepene går på tvers av særlig nettmediene, og de enkelte sidene varierer med hensyn til hva som setter i gang angrepskommentarene. Hos Nettavisen Norsk Debatt finner vi for eksempel flest når islam nevnes, mens det hos ABC Nyheter dreier seg mer om krigen i Ukraina. Dette er temaer som også generelt har en høy andel språklige angrep.

Eksempel på angrep på nettmedienes Facebook-sider

”

[Navn] du lille IQ troll. Du stod tydeligvis bakerst i køen når IQ ble utlevert

Hardt debattklima hos de alternative mediene

Tre av de fem mediene som har den høyeste andelen språklige angrep på Facebook-sidene sine, kalles ofte alternative medier fordi de er klart politiske, skriver om islam og definerer seg i opposisjon til de etablerte mediene.

Her dukker de språklige angrepene særlig opp når mediene publiserer Facebook-innlegg med klare holdninger til islam og integrering. Av samme grunn er det muslimer, og deretter høyreorienterte, som oftest utsettes for språklige angrep i kommentarene. Debattene som handler om islam og integrering på disse sidene, har nemlig en tydelig grense mellom ønsket om å slakke eller stramme innvandringspolitikken. Som vi har vist gjennom denne rapporten - de ideelle betingelsene for at debatter tipper over i språklige angrep. Innvandrerkritiske Facebook-brukere kalles for eksempel for «rasist tullinger», mens venstresiden blir beskyldt for å trekke rasistkortet.

Men det er ikke bare muslimer, islam og integrering som skaper de hardeste debattene hos disse mediene. Det gjør det også når det er snakk om klimaaktivisme og Greta Thunberg eller Gunhild Stordalen. Facebook-brukerne gir i angripende formuleringer uttrykk for at «hun hylejenta der, hun er en pest og en plage». Felles for både Thunberg og Stordalen er at de angripes på bakgrunn av kjønn. I Thunbergs tilfelle brukes også alder og diagnoser aktivt i de språklige angrepene.

Eksempler på angrep og hatprat på de alternative mediesidene

En kommunist som hyler om rasisme igjen.... Stapp det jævla rasistkortet ditt opp der solen aldri skinner, det fungerer ikke lengre.

Asberger dritungen, som blir geleidet av WEF

Angrep på offentlige personers sider

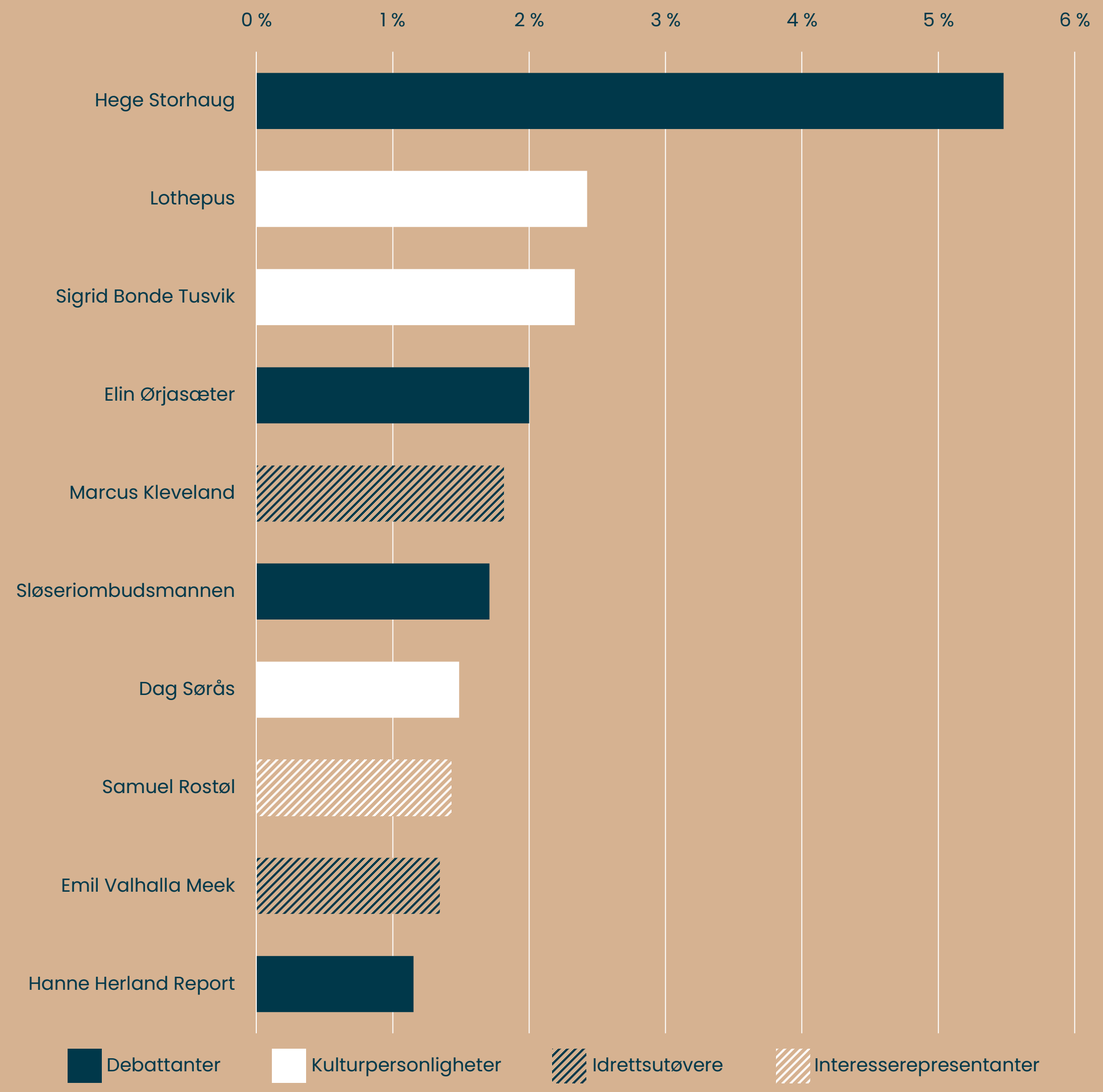
Språklige angrep hos offentlige personer

Debattanter, kulturpersonligheter, idrettsutøvere, profilerte interesserepresentanter og influensere har store følgerskarer på Facebook og er av den grunn med på å prege den offentlige samtalen. Følgelig har vi samlet inn i alt 1 314 511 kommentarer fra offentlige personer, hvorav 1,4 prosent inneholder språklige angrep. Men det er store forskjeller mellom sidene som inngår i disse 1,4 prosentene.

Andelen språklige angrep er høyest blant debattantene, da fire av ti i figuren til høyre faller inn i denne kategorien. Særlig Hege Storhaug skiller seg ut, og hele 5,5 prosent av alle kommentarene på **Facebook-siden** hennes inneholder språklige angrep. Blant de offentlige personene finner vi suverent flest angrepskommentarer på hennes side. Utover de fire debattantene finner vi tre kulturpersonligheter, to idrettsutøvere og én interesserepresentant på topp 10-listen, men ingen influensere.

Selv om vi finner et tydelig mønster når det gjelder andelen angrep i figuren til høyre, er det på de offentlige personenes Facebook-sider at vi finner de største individuelle forskjellene. Dette har sammenheng med hva de typisk skriver om. Et skarpt budskap som legger opp til diskusjon, fører til flere angrep enn en skihoppvideo. Innleggets innhold er også med å peke angrepene i kommentarfeltet i en bestemt retning. Vi ser også at privatliv og personlige historier i noen tilfeller fører til personangrep og hets i kommentarene.

Andelen språklige angrep på offentlige personers Facebook-sider



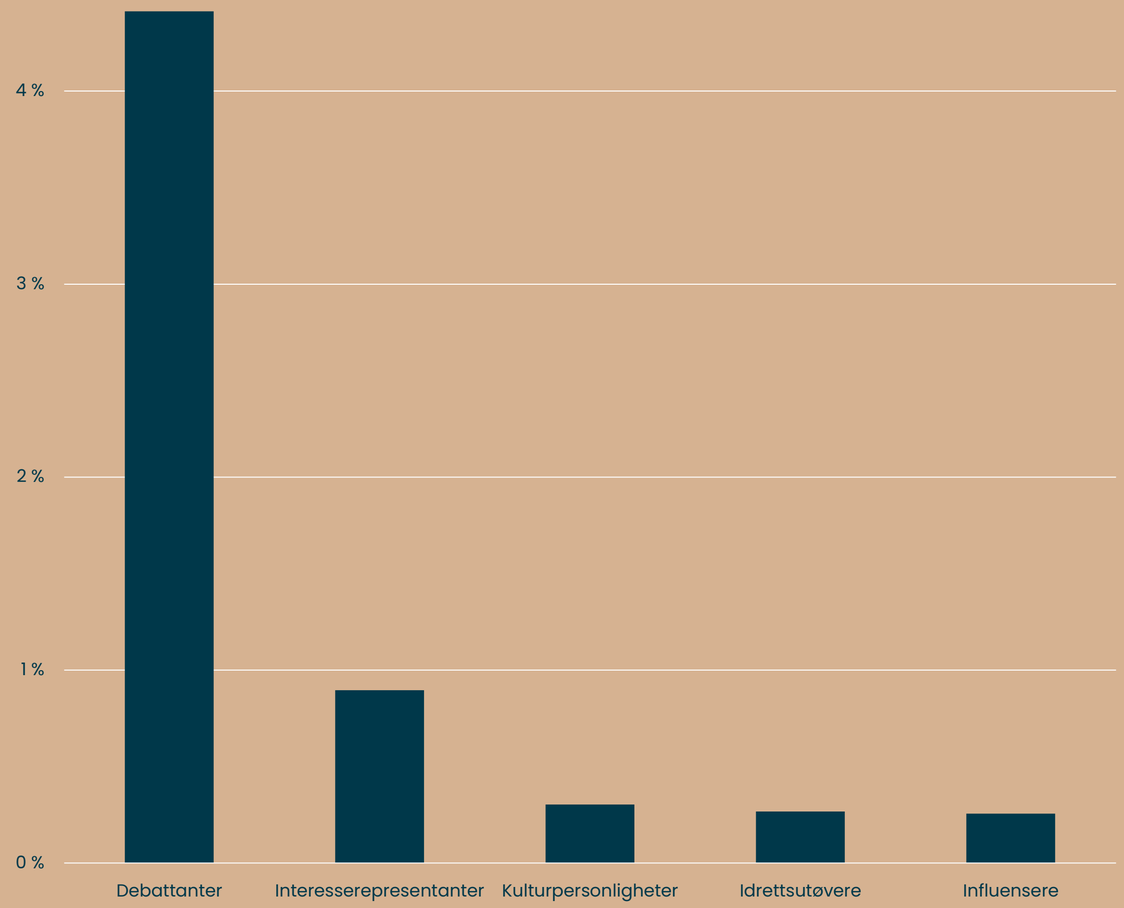
Det går hardest for seg hos debattantene

Med 4,4 prosent er debattantene de offentlige personene som har klart høyest andel språklige angrep på Facebook-sidene sine, og det er langt ned til interesserepresentantene på andreplass, som har en andel på 0,9 prosent.

Debattantene lever av å gi klart uttrykk for holdningene sine, noe som vekker sterke følelser som kan føre til angrep. Spesielt i Hege Storhaugs kommentarfelt går det hardt for seg, og 70 prosent av alle angrepskommentarene som forekommer på offentlige personers sider, stammer fra hennes side. Storhaugs innlegg handler typisk om islam, integrering og kriminalitet, og som tidligere vist er dette temaer som vekker Facebook-brukernes angrepslyst. I kommentarene til Storhaugs innlegg er angrepene derfor ofte rettet mot muslimer.

Interesserepresentantene har den nest høyeste andelen språklige angrep. Dette har en sammenheng med at de kjemper for bestemte, klart avgrensede politiske saker som det ofte kan være lett å ta standpunkt til. Dyrerettighetsaktivist Samuel Rostøl er et godt eksempel, for i sine innlegg tar han tydelig stilling til følelsesladde temaer som rovdyr og landbruk. Dette fører til at han blant mange andre ting kalles for «fuldtidsignorant og snylter», mens de som er enige med Rostøl, kaller motstanderne sine for «lettere tilbakestående».

Andelen språklige angrep fordelt på typer offentlige personer



Personlige historier vekker sterke følelser

Når offentlige personer deler fra privatlivet sitt, fører det i enkelte tilfeller til språklige angrep i kommentarene. Angrepene er rettet både mot de offentlige personene, men også mot andre Facebook-brukere som ser annerledes på sakene enn avsenderne.

Et tydelig eksempel på dette er kommentarene som langrennsløper Petter Northug fikk da han la ut innlegget om sin personlige nedtur med råkjøring og narkotika. Forståelig nok reagerer mange med kommentarer om hvor farlig det er å kjøre i påvirket tilstand, men noen går til angrep på Northug med kommentarer som «Dei hater eg uten forbehold» og «Petter Northug er ikke noe annet enn en mobber og en jævla klovn». Andre tar Northug i forsvar, men de gjør det ved å kalle andre Facebook-brukere for «idiot» og «man child».

Alle typer innhold kan føre til angrep. Alvorlige innlegg som Petter Northugs innlegg eller Hanne Kristin Rhodes innlegg om dyremishandling eller om hvordan hun har opplevd at det er å uttale seg om vold mot kvinner, setter gang i angrepskommentarene. Men også lettere innlegg, som Marna Haugens video om å tømme campingtoalettet sitt, fører til angrep.

Eksempler på angrep i kommentarene til offentlige personenes personlige innlegg

> *[Navn] Ja det er nok tungt med millioner i banken, ufortjent oppmerksomhet i media med den råtne kjeften hans, blåøyde nordmenn på enhver kant som er villig til å legge ut sin egen tunge som rød løper... Petter Northug er ikke noe annet enn en mobber og en jævla klovn som nå IGJEN har driti på draget...*

> *[Navn] og du dømmer og gjør aldri feil du eller ? Du tror du er perfekt men det er du langt ifra så klapp igjen kjeften din. Du som er idiot.*

Anerkjennelse i Facebook-debatten

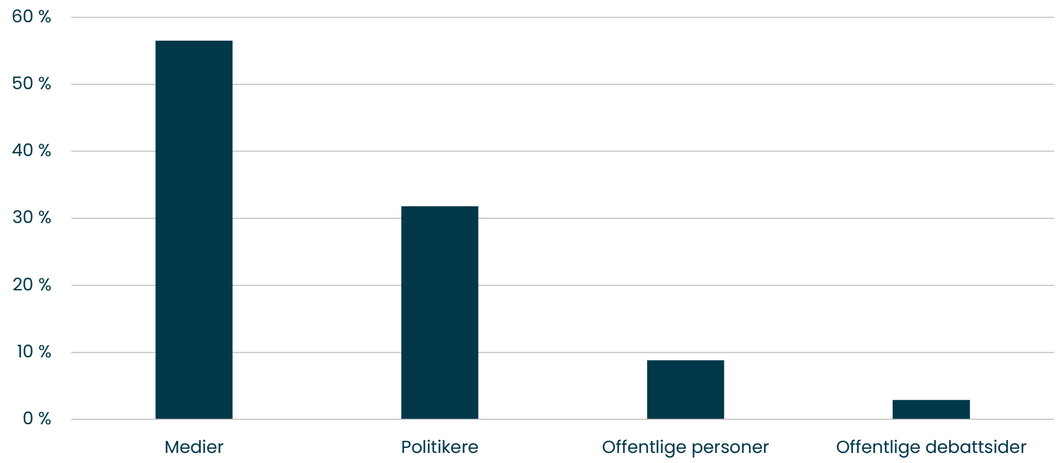
Sidene med mest anerkjennelse

For å finne anerkjennende språk i den offentlige samtalen på Facebook brukte vi en søkenøkkel med anerkjennende formuleringer og utsagn. Med den fremgangsmåten har vi funnet 129 067 anerkjennende kommentarer i den norske, digitale debatten på Facebook.

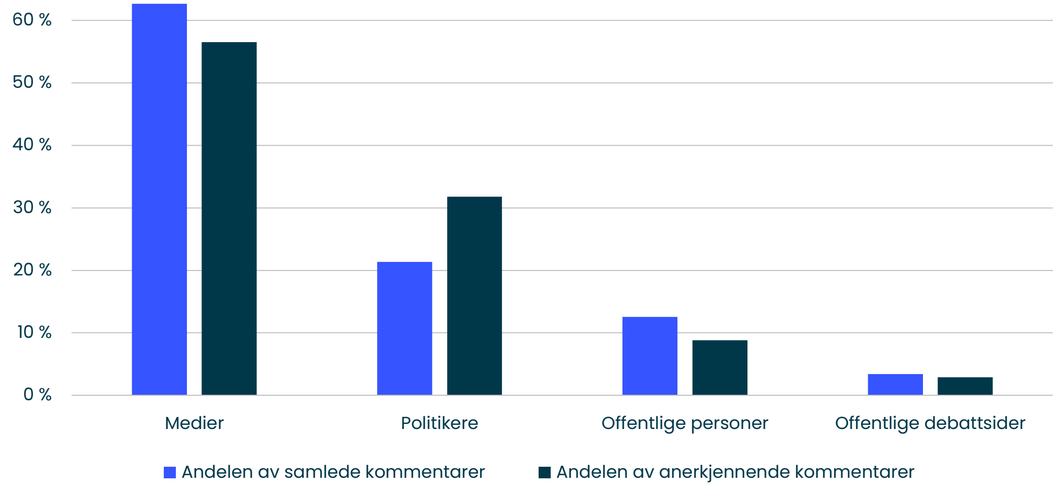
Vi finner de fleste av dem på medienes sider – 56,5 prosent av alle anerkjennende kommentarer befinner seg der. Det henger blant annet sammen med at det også er mediene som generelt får flest kommentarer til Facebook-innleggene sine.

Når ser på de relative tallene endrer bilde seg. Da har politikernes sider en høyere andel anerkjennende kommentarer enn mediene. En sannsynlig forklaring på dette er at politikere bruker Facebook til å dele politiske budskap eller informasjon om personlige bedrifter, som at de har blitt valgt inn på Stortinget. Begge typer innlegg har potensial til å få folk til tastaturet, med kommentarer som «du har helt rett» og «bra jobba». Politikerne stiller også flittig spørsmål, eller de oppfordrer folk til å stemme på partiet deres, noe som også bidrar til å øke mengden anerkjennelse i kommentarfeltene.

Fordelingen av de anerkjennende kommentarene



Andelen av kommentarer i den totale samtalen sammenlignet med andelen av anerkjennende kommentarer



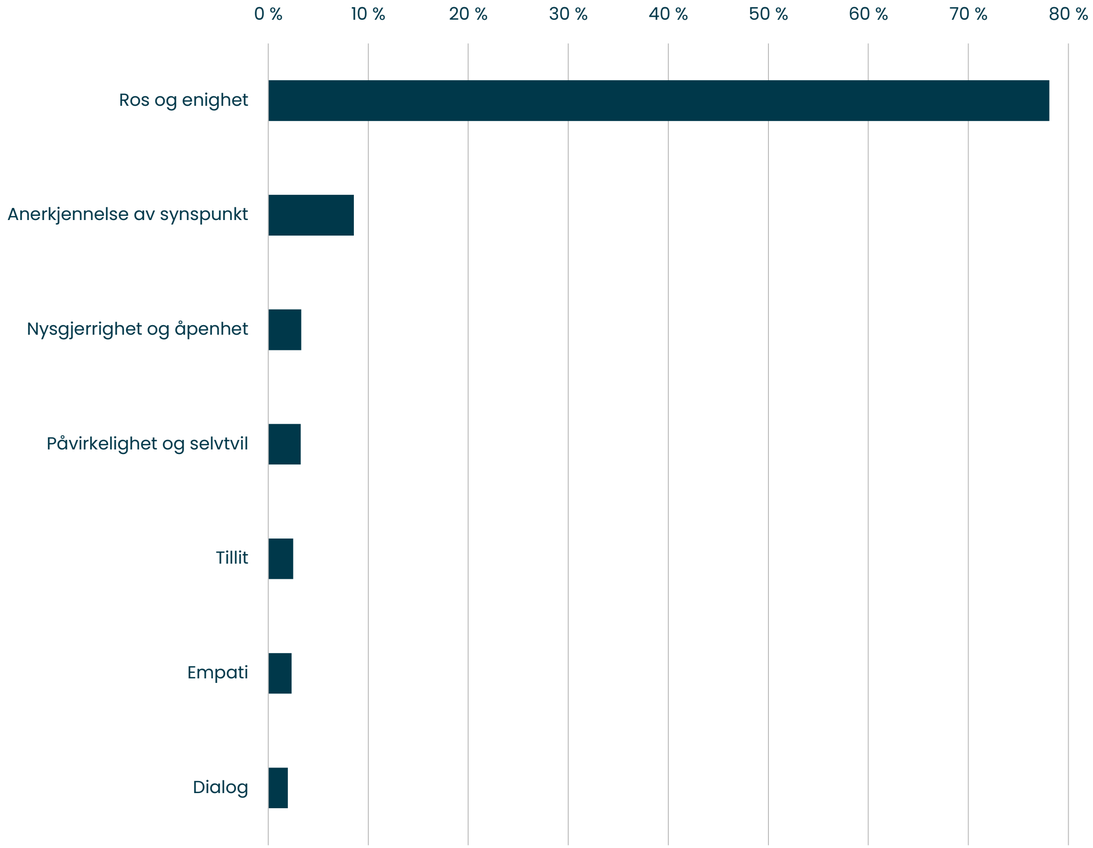
Populære synspunkt gir anerkjennelse

Det er mange måter å anerkjenne innleggsskribenter og andre Facebook-brukere på, men ros og enighet er aller mest utbredt i den digitale samfunnsdebatten i Norge. Hele 78 prosent av den anerkjennende digitale samtalen faller inn under denne kategorien.

Det er især enighet som driver anerkjennelsen i kommentarfeltene. Norske Facebook-brukere som er helt enige i andres utsagn, utgjør mer enn 60 prosent av den samlede anerkjennende samtalen. Dette peker i retning av at samfunnsdebatten inneholder flere meningsfellesskap.

Den høye graden av enighet kan imidlertid også ha to potensielle negative effekter. For det første gir enighet ikke alltid rom for motstridende argumenter eller tvil fordi det kan være vanskelig å være den første eller den eneste som ikke er helt enig. For det andre kan enighet knyttes til kommentarer som er ekskluderende eller polariserende, som når en bruker ytrer «Jepp helt enig… Øye for løye og tann for tann så får kanskje Landet respekten tilbake blant disse utskuddene». Begge deler kan føre til at bare de som er enige i innleggets innhold, tør å gi uttrykk for meningen sin, mens de uenige forblir stille eller uttrykker meningen i kommentarfeltene til andre innlegg, hvor de er enige i innholdet. Hvis det er tilfelle, begrenser det den demokratiske verdien av den offentlige debatten på Facebook.

Fordelingen av de ulike formene for anerkjennelse



Tema som gir anerkjennelse

Norskhet og sårbarhet anerkjennes

Anerkjennelse kan dukke opp mange steder i den norske, digitale samtalen. Temaet medier og debatt har flest anerkjennende kommentarer, og en av grunnene til dette er at de innleggene som har fått flest anerkjennende kommentarer på sin vei, er debatter om norsk identitet og kultur.

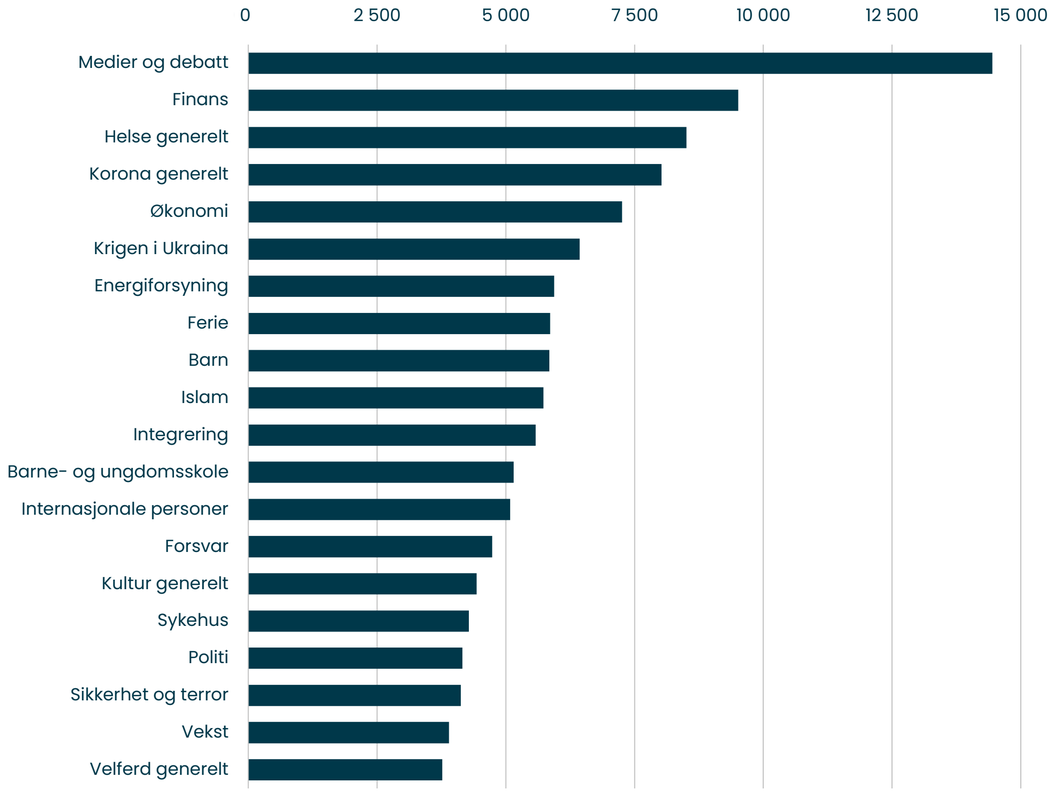
Emilie Enger Mehls (SP) innlegg om bruken av «nordmann» og Nettavisens innlegg om bevaring av kongehuset er gode eksempler på dette. Det er for det meste enighet i kommentarfeltene, men det er også eksempler på at folk som er enige eller uenige, anerkjenner hverandres synspunkt – særlig under innlegg om kongehuset.

Vi finner også mye anerkjennelse når samtalen dreier seg om sårbarhet og helse. Både helse og korona ligger nemlig høyt i figuren til høyre, og faktisk handler finans og økonomi i høy grad også om sårbare grupper og helse. Det er nemlig her vi finner samtaler om hvordan, og på hvem, skattepengene skal brukes. Finans og økonomi handler også om prisstigninger, noe som også genererer anerkjennelse.

De anerkjennende temaene i samfunnsdebatten

For å kartlegge hvilke temaer som har flest anerkjennende kommentarer, har vi utviklet en temabasert søkenøkkel. Søkenøkkelen har i underkant av 2 800 søkeord fordelt på to nivåer: hovedtema og undertema. Figuren nedenfor viser de 15 temaene som har flest anerkjennende kommentarer. Bak hvert tema gjemmer det seg lange lister med søkeord.

Topp-15 emner med flest anerkjennende kommentarer



Folk støtter dem som trenger hjelp og er utsatte

Helse er et tema som møter mye anerkjennelse. Det kommer for eksempel til uttrykk i debatter om prioriteringer i helsevesenet når det er snakk om statsbudsjettet. Også behovet for tannhelse til utsatte grupper fremheves.

”

Helt enig, helsefagarbeiderne er uunnværlige og det at hjelpepleierne ble veid og funnet for lett på sykehusene gjorde pleiegapet enda mer utfordrende.

De som sliter med **prisøkninger** på drivstoff, dagligvarer og energi, blir også anerkjent i den norske digitale debatten. Mange er enige i at de må hjelpes uansett om de økonomiske kvalene skyldes korona eller krigen i Ukraina.

”

Helt enig. Alt blir dyrere og de som har liten inntekt sliter mest.

Korona tar stor plass i den anerkjennende samtalen. Det gjelder støtte til koronatiltak som munnbind, men også støtten til de eldre, de syke og de som følte seg isolert under nedstengningene.

”

[Navn] føler med deg, vi har heller ikke sett familie og venner i England i år. Nå som jula nærmer seg er savnet stort, men det er bare å vente

Krigen i Ukraina er også et tema som går igjen i de anerkjennende kommentarene, både fordi krigen har ført til energikrise i mange land, inkludert Norge, men sympatien med ukrainerne kommer også til uttrykk.

”

Helt enig med i det du skriver her. Helt forferdelig alt det som skjer i Ukraina…

Anerkjennelse kommer til uttrykk mange steder i Facebook-debatten

Politikerne og partiene som får mest anerkjennelse

Det er de mest eksponerte politikerne som får mest anerkjennelse. Derfor er både nåværende og tidligere statsministere høyt plassert i figuren til høyre. Den som troner øverst, er imidlertid lederen av FrP, Sylvi Listhaug, som får mer enn dobbelt så mange anerkjennende kommentarer på innleggene sine som tidligere statsminister Erna Solberg, som ligger på andreplass.

Høyresiden, spesielt FrP, utgjør mer enn halvparten av politikersidene på topp 10-listen over politikersider med flest anerkjennende kommentarer, mens venstreorienterte partier ikke er like fremtredende. En forklaring på det er at politikerne fra FrP er aktive på Facebook på en slik måte som virkelig engasjerer følgerne. Ca. 45 prosent av kommentarene på samtlige politikersider stammer fra FrP-politikere. De er dessuten sterkt til stede i den offentlige debatten når diskusjonstemaene er prisstigninger og offentlig finans, to temaer som gir mye anerkjennelse.

En annen forklaring er at mye av anerkjennelsen handler om enighet, og at FrP-politikere er gode til å levere tydelige politiske budskap som legger opp til at Facebook-brukere kan erklære seg enige.

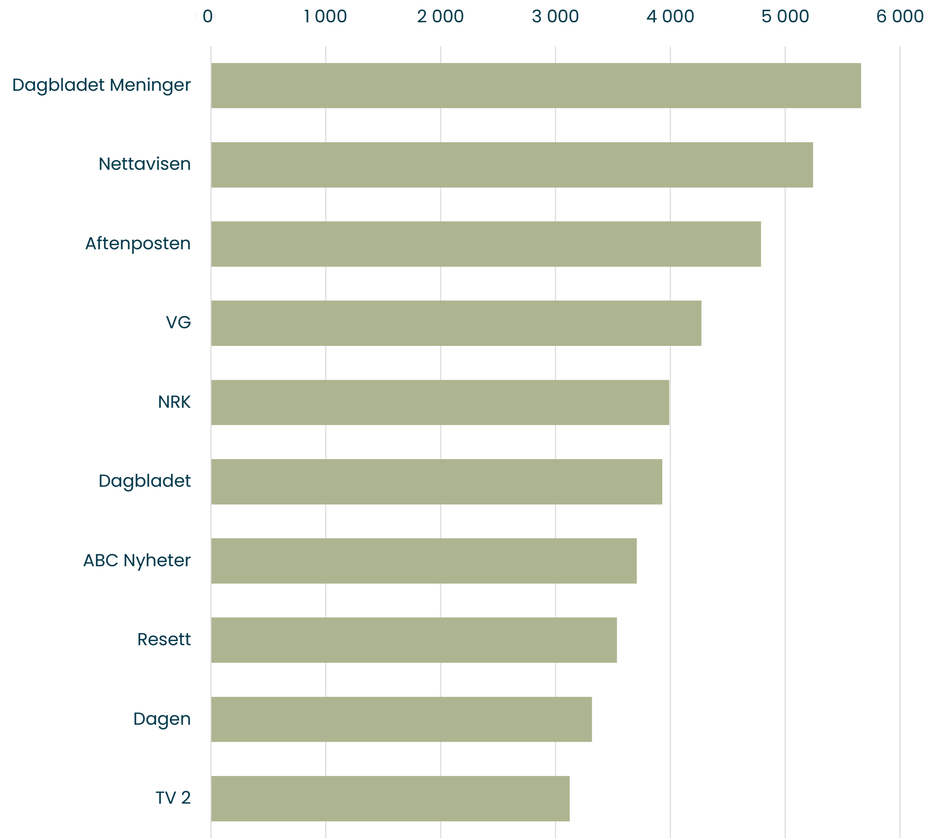
Topp 10-politikere med flest anerkjennende kommentarer på Facebook

Anerkjennelsen på mediesidene er mer variert

Det er ros og enighet som er mest utbredt på de ti mediesidene som har fått mest anerkjennelse i kommentarfeltene. Men det er også på disse medienes Facebook-sider at vi finner flest kommentarer som inneholder de andre formene for anerkjennelse. Her er brukerne altså mer tilbøyelige til å anerkjenne hverandres synspunkt eller til å være nysgjerrige.

I figuren til høyre ser vi også en del av de store, tradisjonelle mediene som Aftenposten, NRK og VG, som ikke er blant mediesidene med mange angrep. En nærliggende forklaring er at medienes modereringsinnsats i form av prinsipper og ressurser er med på å fremme den anerkjennende, digitale samtalen – også når det gjelder temaer som vekker mange følelser.

Topp-10 mediene med flest anerkjennende kommentarer



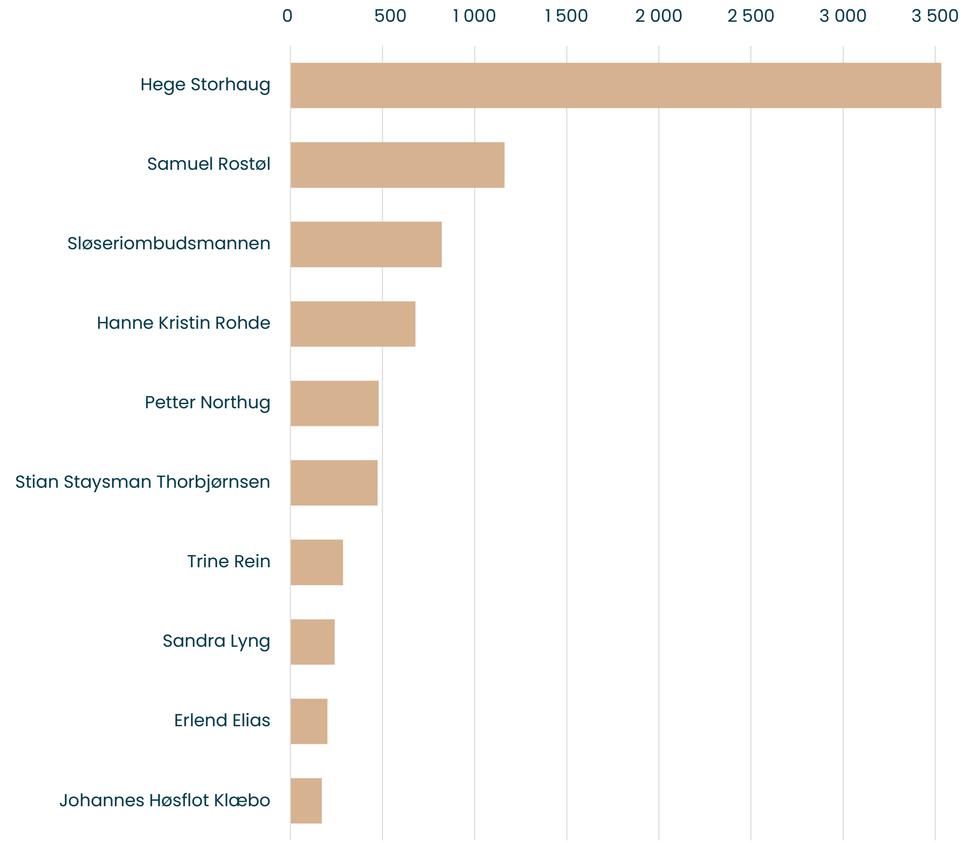
Anerkjennelse gis til offentlige personer med aktuelt innhold

Sammenlignet med politikere og medier får offentlige personer færrest anerkjennende kommentarer til innleggene sine. Ett unntak er imidlertid forfatteren, redaktøren og debattanten Hege Storhaug. Hun får mer enn tre ganger så mange anerkjennende kommentarer som dyrerettighetsaktivist Samuel Rostøl, som får nest flest.

Et fellestrekk ved både Storhaug og Rostøl er at de skriver om samfunnsaktuelt innhold, formulert på en slik måte at det oppfordrer folk til å være enige med hverandre og rose hverandre eller innlegget. Dette ser vi blant annet ved at anerkjennelsen i Storhaugs kommentarfelt er knyttet til kritikk av muslimer og folkevalgte politikere som folk erklærer seg enige i. Men både Storhaug og Rostøl er også blant de offentlige personene som har den høyeste andelen angrep på Facebook-sidene sine, noe som tyder på at Facebook-sidene deres også er hjem for en polarisert debatt.

Det er imidlertid ikke bare samfunnsorienterte offentlige personer som får anerkjennelse. Folk er for eksempel helt enige i Sandra Lyngs betraktninger om at komplimenter knyttet til vekttap er unødvendige kommentarer om kroppen. Kommentarfeltene er også fylt med folk som føler med Johannes Høsflot Klæbo i forbindelse med Klæbos innlegg om den underkjente seieren i et viktig løp under Nordic World Ski Championships 2021.

Topp 10 Facebook-sider til offentlige personer med flest anerkjennende kommentarer



En stor takk til:

Prosjektets referansegruppe

Eirik Rise, Kampanjerådgiver, Stopp Hatprat
Eliana Hercz, Koordinator, Dialogpiloterne
Elin Solberg, Fagdirektør, Justis- og beredskapsdepartementet
Elsa Skjong-Arnestad, Leder Rosa kompetanse justis, FRI
Erik Velldal, Professor ved Informatikk, UiO
Hatem Ben Mansours, Nestleder, Antirasistisk Senter
Jørgen Frydnes, Daglig Leder, Utøya
Lars M. Gudmundson, Leder, Hegnhuset - læringssenteret på Utøya
Lilja Øvrelid, Professor ved Informatikk, UiO
Linda Tinuke Strandmyr, Daglig Leder, Antirasistisk Senter
Marjan Nadim, Professor og forsker, Norsk Senter for Samfunnsforsking
Mikkel Berg-Nordlie, Forsker & Forfatter, OsloMet
Monica Lillebakken, Leder hatkrimgruppa, Oslopolitiet
Monika Kochowicz, Prosjektleder, Stopp Hatprat
Stian Lid, Forsker, By- og regionsforskningsinstituttet NIBR, OsloMet
Tore Bjørgo, Leder, Senter for ekstremismeforskning (C-REX)

Våre samarbeidspartnere hos Gjensidigestiftelsen

Ingrid Riddervold Lorange, Administrerende direktør
Ingrid Tollånes, Leder utdelinger
Annhild Mosdøl, Spesialrådgiver

Prosjektets annotører

Johanne Ringøen
Vemund Wik

Teamet bak rapporten:

Fra Analyse & Tall

Ane Kathrine Strand, Partner
Maj Baltzarsen, Partner
Niels Ørbæk Chemnitz, Partner
Ronnie Brandt Taarnborg, Partner
Ingvild Endestad, Tidligere partner

Fra Common Consultancy

Louise Madsen, Analytiker
Lucca Lund Powers Bates, Student
Eske Vinther-Jensen, Partner

Fra Nordic Safe Cities

Jeppe Albers, Administrerende direktør
Lotte Fast Carlsen, Vicedirektør
Sebastian Jørgensen, Project manager

Design

Andreas Lind Johansen, Creative director and founder

Referanser:

1. Sosiale medier tracker Q1 2023, IPSOS: https://www.ipsos.com/sites/default/files/ct/publication/documents/2023-01/Ipsos%20SoMe-tracker%20Q4%202022.pdf

2. Likestillings- og diskrimineringsombudet (2021): Hatefulle ytringer på nett. https://www.ldo.no/globalassets/_ldo_2019/03_ombudet-og-samfunnet/rapporter/hatefulle-ytringer/ldo_hatefulle_ytringer_pa_nett.pdf

3. (Bjelland, H.F., Bjørgo, T. Thomassen G. (2021). Trakassering og trusler mot politikere: En spørreundersøkelse blant medlemmer av Stortinget, Regjeringen og sentralstyrene i partiene og ungdomspartiene. PHS Forskning 2021) trakassering og trusler 2021-1.pdf (unit.no)

4. Birkvad, S. R., Fladmoe, A., Nadim, M. (2019). Erfaringer med hatytringer og hets blant LHBT-personer, andre minoritetsgrupper og den øvrige befolkningen. https://samfunnsforskning.brage.unit.no/samfunnsforskning-xmlui/bitstream/handle/11250/2584665/Erfaringer%2bmed%2bhatytringer.pdf?sequence=2&isAllowed=y

5. Fladmoe, Audun & Nadim, Marjan (2017). Silenced by hate? Hate speech as a social boundary to free speech. In Midtbøen, Arnfinn Haagensen; Steen-Johnsen, Kari & Thorbjørnsrud, Kjersti (Ed.), *Boundary Struggles : Contestations of Free Speech in the Norwegian Public Sphere.* Cappelen Damm Akademisk. p. 45–75.

6. Sletvold, B., Veledar, A. (2021): Hatefulle ytringer på nett. https://www.ldo.no/globalassets/_ldo_2019/03_ombudet-og-samfunnet/rapporter/hatefulle-ytringer/ldo_hatefulle_ytringer_pa_nett.pdf

7. Bjelland, H.F., Bjørgo, T. Thomassen G. (2021). Trakassering og trusler mot politikere: En spørreundersøkelse blant medlemmer av Stortinget, Regjeringen og sentralstyrene i partiene og ungdomspartiene. PHS Forskning 2021) trakassering og trusler 2021-1.pdf (unit.no)

8. Burkal, R., Veledar, A. (2018). Hatefulle ytringer i offentlig debatt på nett. https://www.nhri.no/wp-content/uploads/2018/12/Rapport-fra-Likestillings-og-diskrimineringsombudet-Hatefulle-ytringer-i-offentlig-debatt-p%C3%A5-nett-2018.pdf

9. Fajkovic, M,. Lindebjerg, G. (2021). Tone og hatefulle ytringer i norske kommentarfelt. https://www.bufdir.no/globalassets/global/nbbf/vold_overgrep/tone_og_hatefulle_ytringer_i_norske_kommentarfelt.pdf

10. Mark Zuckerberg (2021): A Blueprint for Content Governance and Enforcement". Facebook blogpost. https://www.Facebook.com/notes/751449002072082/

11. https://www.ogtal.dk/assets/files/Angreb-i-den-offentlige-debat-paa-Facebook.pdf

12. https://www.ogtal.dk/assets/files/Anerkendelse-i-den-offentlige-debat-paa-Facebook.pdf

13. https://github.com/ogtal/A-ttack

14. https://huggingface.co/alexandrainst/scandi-nli-large

15. https://www.abdera.ai/

16. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2777835 (model og datasæt ikke tilgængelig)

17. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2993293 (model og datasæt ikke tilgængelig)

18. Alexandra, B., & Ralf, S. (2009). Rethinking Sentiment Analysis in the News: from Theory to Practice and back.

19. Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics. Proceedings of the Fourth Workshop on Online Abuse and Harms, 184–190. https://doi.org/10.18653/v1/2020.alw-1.21