INSTITUTT
FOR SAMFUNNS-
FORSKNING

Rapport 2018:1

Måling av omfang av hatefulle ytringer

Metodiske muligheter og utfordringer

Marjan Nadim, Audun Fladmoe og Bernard Enjolras

Rapport 2018:1

Institutt for samfunnsforskning
Munthes gate 31
Postboks 3233 Elisenberg
0208 Oslo

ISBN (trykk): 978-82-7763-581-1
ISBN (nett): 978-82-7763-582-8
ISSN (trykk): 0333-3671
ISSN (nett): 1891-4314

www.samfunnsforskning.no

Innhold

Forord

Formålet med denne rapporten er å gi en vurdering av muligheter og begrensninger ved ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer. Rapporten er skrevet på oppdrag fra Barne-, ungdoms- og familiedirektoratet (Bufdir) og danner grunnlag for fremtidige empiriske undersøkelser av hatefulle ytringer.

Marjan Nadim har ledet prosjektet og hatt hovedansvar for innlednings- og avsluttingskapitlene (kapittel 1 og 4), i tillegg til delkapittel 3.1 og 3.2 om henholdsvis analyser av hatsider og kvantitativ manuell innholdsanalyse. Audun Fladmoe har hatt ansvar for kapittel 2 om surveymetoder, og Bernard Enjolras har hatt ansvar for delkapittel 3.3 om stordataanalyse. Alle forfattere har likevel jobbet tett sammen om teksten som helhet. Karin Oline Kraglund har vært vitenskapelig assistent på prosjektet og har blant annet gjort en uunnværlig jobb med å systematisere eksisterende studier av hatefulle ytringer.

Prosjektet har hatt en ekspertgruppe som har bidratt med faglig kvalitetssikring og gode tilbakemeldinger. Ekspertgruppen har bestått av Kari Steen-Johnsen, Audun Beyer, Johannes Bergh (alle ISF), Anders Ravik Jupskås (Senter for ekstremismeforskning [C-REX]), Erik Velldal og Lilja Øvrelid (begge Institutt for informatikk, Universitetet i Oslo).

I tillegg vil vi takke referansegruppen for prosjektet, som har bestått av Rune Berglund Steen, Shoaib Sultan (begge Antirasistisk Senter), Ingjerd Hansen (Oslo politidistrikt), Eirik Aimar Engebretsen (Forening for kjønns- og seksualitetsmangfold, FRI), Hanne Marie Myrvold (Kommunal- og moderniseringsdepartementet), Amna Veledar (Likestillings- og diskrimineringsombudet) og Eirik Rise (Stopp hatprat-kampanjen).

Videre har Anine Kierulf, fagdirektør ved Norges nasjonale institusjon for menneskerettigheter, bidratt med gode tilbakemeldinger til drøftingen av begrepet hatefulle ytringer. Til slutt vil vi takke oppdragsgiver Bufdir, ved Helge Lyberg og Karen-Sofie Pettersen, for godt samarbeid.

Oslo, 10. desember 2017

Marjan Nadim, Audun Fladmoe og Bernard Enjolras

Sammendrag

Forfatter Marjan Nadim, Audun Fladmoe og Bernard Enjolras

Tittel Måling av omfang av hatefulle ytringer
Metodiske muligheter og utfordringer

Sammendrag Det finnes ingen omforent definisjon av hatefulle ytringer, verken i internasjonal lovgivning eller i forskningslitteraturen. Det er likevel vanlig å forstå hatefulle ytringer som ytringer som er av hatefull eller diskriminerende art, og som retter seg mot en gruppe eller mot et individs (antatte) gruppetilhørighet. Hatefulle ytringer kan uttrykkes muntlig eller i form av tekst, bilder, symboler og andre medieuttrykk.

Formålet med denne rapporten er å vurdere muligheter og begrensninger ved ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer. De metodiske vurderingene i rapporten er basert på en vurdering av eksisterende empiriske studier av hatefulle ytringer og tilgrensende fenomener, i tillegg til oppdatert metodelitteratur og en gjennomgang av relevante erfaringer med ulike metodiske tilnærminger.

De ulike metodiske tilnærmingene til å studere omfanget av hatefulle ytringer blir vurdert ut fra a) hvor nøyaktige de er til å fange fenomenet vi er interessert i, b) i hvilken grad de gjør det mulig å studere og sammenligne omfanget av hatefulle ytringer rettet mot ulike grupper og på ulike arenaer, c) hvor representative resultater metoden gir, d) muligheter for å innhente andre typer informasjon om hatefulle ytringer utover omfang, e) mulighetene for tidsserier og komparasjon og f) hvor kostnadskrevende de ulike metodiske oppleggene er.

Vi tar for oss to hovedtilnærminger til å studere omfanget av hatefulle ytringer: surveymetoder og analyser av medieuttrykk, nærmere bestemt kvantitativ manuell innholdsanalyse og automatiserte analyser av stordata (Big Data) ved hjelp av maskinlæringsalgoritmer.

En vanlig måte å studere hatefulle ytringer på er ved hjelp av surveymetodikk, altså å spørre folk om deres erfaringer med å motta eller observere hatefulle ytringer. Surveymetodikk er i utgangspunktet fleksibelt, og kvaliteten på resultatene man oppnår, avhenger av hvilke valg man tar, og hvilke budsjettrammer man har. På sitt beste kan en survey gi god og presis informasjon om befolkningens erfaringer med hatefulle ytringer. Men dersom spørsmålene er dårlig formulert og/eller utvalgene er lite representative, kan surveyer ha begrenset verdi. En survey fanger fenomenet hatefulle ytringer slik ytringene oppfattes av dem som svarer på undersøkelsen. Dette subjektive elementet gjør det utfordrende å måle hatefulle ytringer presist, siden ulike personer kan tolke den samme ytringen ulikt. Utfordringen forsterkes også av at begrepet hatefulle ytringer ikke er allment kjent.

Ved å kombinere generelle og konkrete spørsmål i en undersøkelse kan man redusere disse begrensningene og få et godt grunnlag for å analytisk skille hatefulle ytringer fra andre tilgrensede fenomener. Surveymetodikk gjør det også mulig å fange opp hatefulle ytringer rettet mot ulike grunnlag, både med hensyn til hva innholdet i ytringene er rettet mot, og med hensyn til hvilke (minoritets)grupper mottakerne tilhører.

Ved bruk av sannsynlighetsutvalg kan en survey gi representative estimater av omfanget av opplevde hatefulle ytringer i ulike befolkningsgrupper. Gode representative befolkningsundersøkelser kan imidlertid være svært kostbare, og med rimeligere løsninger er det ofte vanskelig å oppnå representative utvalg i alle minoritetsgrupper. Dette skyldes både at det kan være utfordrende å avgrense og definere en del minoritetsgruppepopulasjoner, og at tilbøyeligheten til å svare er lav i en del grupper. Enkelte alternative utvalgskilder, som medlemslister i organisasjoner og menigheter, kan bidra til å redusere disse problemene, men slike utvalg vil kun være representative for medlemmene i disse organisasjonene.

Kvantitativ manuell innholdsanalyse innebærer at et utvalg medieuttrykk (tekst, bilder, symboler osv.) kodes manuelt etter forhåndsdefinerte variabler. Et vidt begrep som hatefulle ytringer må få en klar, operasjonell definisjon, slik at de som koder materialet, har klare kriterier for å vurdere hvorvidt en enhet skal kodes som hatefull eller ikke, og slik at ulike kodere tolker ytringene så likt som mulig (høy «interkoderreliabilitet»). Denne metoden kan brukes til å kode medieuttrykk som for eksempel hatefulle eller ikke, hvilket grunnlag ytringene eventuelt retter seg mot, hvor grove ytringene er, hvilke temaer ytringene opptrer i forbindelse med, og hva slags respons de fremkaller. En fordel med denne tilnærmingen er at vurderingen av hvorvidt en ytring er hatefull eller ikke, gjøres på en stringent måte hvor kriteriene er klargjort på forhånd. Forskeren har dermed god kontroll over hva som måles. Metoden åpner for å måle en rekke relevante aspekter ved hatefulle ytringer, inkludert hvilket grunnlag ytringene retter seg mot, og hvilken kontekst ytringene falt i. Med denne metoden er det mulig å oppnå tilfredsstillende grad av representativitet, og det er mulig å designe analyseopplegget på en måte som åpner for komparasjon og å etablere tidsserier.

Stordataanalyse innebærer å hente inn store mengder tekstdata (fra for eksempel sosiale medier som Twitter eller Facebook) som inneholder et sett søkeord som potensielt indikerer hatefulle ytringer. Ved å først kode deler av innholdet manuelt kan man «lære» programmet å kategorisere innhold som hatefullt eller ikke. Deretter kan maskinlæringsalgoritmen automatisk klassifisere materialet og identifisere om det inneholder hatytringer. Dermed gir stordataanalysen oss et mål på omfanget av hatefulle ytringer i et stort utvalg poster fra sosiale medier. Denne tilnærmingen gir gode resultater ved at den gjenkjenner hatefulle ytringer i store mengder tekst. Kvaliteten på resultatene av denne metoden er avhengig av kvaliteten på dataene og den manuelle kodingsprosessen som er brukt til å trene algoritmen. Stordatatilnærmingen kan tilpasses ulike grunnlag og målgrupper for hatefulle ytringer, helst ved at man trener flere spesialiserte algoritmer. Dataene som danner grunnlag for stordataanalyse, er ikke nødvendigvis representative for den digitale offentligheten, i og med at mulighetene

for datainnsamling er begrenset av de tekniske mulighetene som ligger i ulike plattformer for sosiale medier. En stordatatilnærming krever en betydelig investering i startfasen. Likevel er den kostnadseffektiv på lengre sikt, fordi en trent og operasjonell algoritme kan brukes på et nytt materiale uten store ekstrakostnader. Stordataanalyse reiser forskningsetiske utfordringer med hensyn til behandling av personopplysninger. Imidlertid er det god grunn til å tro at automatisk gjenkjenning av hatefulle ytringer vil kunne bli unntatt informasjonsplikt og kravet om innhenting av samtykke.

Emneord hatefulle ytringer, metode, survey, stordata, innholdsanalyse

English Summary

Authors Marjan Nadim, Audun Fladmoe og Bernard Enjolras

Title Measuring the prevalence of hate speech:
Methodological possibilities and limitations

Summary There is no unified definition of hate speech in the research literature or in international law. Still, it is possible to understand hate speech as deliberately stigmatising, discriminatory, degrading or threatening speech directed towards an individual or a group based on specific (perceived) group characteristics. Thus, rather than individuals, hate speech primarily targets groups.

The purpose of this report is to assess the possibilities and limitations of different methodological approaches to studying the prevalence of hate speech. The methodological discussions are based on an assessment of existing empirical studies of hate speech and related phenomena, in addition to updated methods literature and a review of relevant experiences with different methodological approaches.

The following criteria are used to assess the different methodological approaches to studying the prevalence of hate speech: a) how accurately they can capture the phenomenon we are interested in, b) to what degree they enable capturing and comparing hate speech directed towards different groups in different sites, c) how representative results the approaches yield, d) the possibilities for gathering other types information beyond the prevalence of hate speech, e) the possibilities for comparisons across time and space, and f) the level of costs of different approaches.

We focus on two main approaches to studying the prevalence of hate speech: survey methods and analyses of media content, more specifically quantitative manual content analysis and automatic analysis of Big Data with machine learning algorithms.

Survey approaches are common in the study of hate speech. This entails asking people about their experiences with receiving or observing hate speech. Survey methods are flexible, and the quality of the results depends on the design and budgetary restrictions. A survey can yield good and precise information about the population's experiences with hate speech. However, if the questions are poorly articulated and/or the samples are not representative, surveys may have little value. A survey captures the phenomenon hate speech as it is perceived by the respondents. This subjective element makes it challenging to measure the prevalence of hate speech because the concept is not well established in a Norwegian context. By including both general and specific questions in a survey, it is, however, possible to obtain a good basis for analytically distinguishing hate

speech from related phenomena. In addition, survey methodology opens the possibilities for studying hate speech directed towards different grounds, both regarding the content of the speech and regarding the respondents' membership in different (minority) groups.

By utilising probability samples, a survey can yield representative estimates of the prevalence of experiences with hate speech in different segments of the population. However, large-scale population surveys representative of different minority groups can be very costly, and with less costly approaches it is often challenging to achieve representative samples of all minority groups. This is partly because it can be difficult to define and delimit some minority groups, and partly because some groups are known to have low response rates. Alternative sample sources, such as lists of members in organisations, congregations etc. can contribute to reduce these problems, but such samples will only be representative of the members of these organisations.

Quantitative manual content analysis entails manually coding a specified sample of media expressions (text, pictures, symbols, etc.) according to predefined variables. A vague concept such as hate speech must be given a clear operational definition, in order for the coders of the material to operate after clear criteria for what should be considered as hate speech, and in order to secure that individual coders interpret the units as similarly as possible (high "intercoder reliability"). This method can be used to code several dimensions related to hate speech, including: the amount of hate speech in a given sample of media expressions, who are the targets of hate speech, the degree of hate in specific utterances, what topics elicit hate speech and the responses to hate speech. An advantage with this approach is that the judgement of what is considered as hate speech is done in a stringent manner according to predefined criteria. The researcher thus has control over what is measured. Quantitative manual content analysis can give representative results and it is possible to design an analysis that is open to comparisons over time and across countries.

Big Data analysis involves retrieving large amounts of text data (e.g. from social media like Twitter or Facebook) containing a set of keywords that potentially indicate expressions of hate. By first encoding parts of the content manually, one can "learn" the program to better categorize content as hateful or not. Afterwards, the machine-learning algorithm can automatically classify the material and identify if it contains hate speech. Thus, Big Data analysis gives us a measure of the extent of hate speech in a wide range of media records. This approach provides good results for recognising hateful expressions in large amounts of text. The quality of the results of this method depends on the quality of the data and the manual coding process used to train the algorithm. The Big Data approach can be customised to capture hate speech directed at specific grounds and target groups, by training algorithms that are more specialised. The data that form the basis for Big Data analysis is not necessarily representative of the digital public, as data collection possibilities are limited by the technical capabilities of different social media platforms. A large-scale approach requires a significant investment in the start-up phase. Nevertheless, it is cost-effective because a trained and operational algorithm can be used on new

material without major additional costs. The use of Big Data raises ethical challenges regarding the processing of personal information. However, there is reason to believe that automatic recognition of hate speech could be exempted from the requirements of informing and obtaining consent from individual users of social media.

Index terms hate speech, methods, survey, Big Data, content analysis

1 Innledning

Hatefulle ytringer har fått økt oppmerksomhet de siste årene. Fremveksten av internett og sosiale medier har åpnet for at folk flest kan delta i det offentlige ordskiftet og ytre meninger på en helt annen måte enn før. Samtidig betyr denne utviklingen at hatytringer kan spres raskere og nå bredere ut enn noensinne. Likevel vet vi lite om omfanget av hatefulle ytringer, både på og utenfor nettet. Selv om det finnes noe empirisk forskning på tilgrensende fenomener (som mobbing, trakassering, trusler og trolling[1] på nett), er det få forskningsbidrag som direkte har forsøkt å måle omfanget av hatefulle ytringer.

Formålet med denne rapporten er å vurdere muligheter og begrensninger ved ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer. De metodiske vurderingene i rapporten er basert på en gjennomgang av eksisterende empiriske studier av hatefulle ytringer og tilgrensende fenomener i tillegg til oppdatert metodelitteratur og en gjennomgang av relevante erfaringer med ulike metodiske tilnærminger.

Vi tar for oss to hovedtilnærminger til å studere omfanget av hatefulle ytringer. For det første diskuterer vi *surveymetoder* for å studere omfanget av erfaringer med å motta hatefulle ytringer i ulike grupper i befolkningen. For det andre ser vi på mulighetene som ligger i ulike former for *innholdsanalyse* for å studere det faktiske meningsinnholdet i det offentlige ordskiftet. Vi diskuterer særskilt mulighetene som ligger i automatisert analyse av såkalte stordata (Big Data).

Vi vurderer de ulike metodiske tilnærmingene til å studere omfanget av hatefulle ytringer ut fra a) hvor nøyaktige de er til å fange fenomenet vi er interessert i, b) i hvilken grad de gjør det mulig å studere og sammenligne omfanget av hatefulle ytringer rettet mot ulike grupper og på ulike arenaer, c) hvor representative resultater metoden gir, d) muligheter for å innhente andre typer informasjon om hatefulle ytringer utover omfang, e) mulighetene for tidsserier og komparasjon og f) hvor kostnadskrevende de ulike metodiske oppleggene er.

Dette vil kunne danne grunnlag for fremtidige empiriske studier av hatefulle ytringer og andre tilgrensende fenomener, som netthets, mobbing og diskriminering.

1 Trolling handler om å provosere for provoseringens skyld, ofte ved å skrive noe man ikke mener, slik at diskusjonen utvikler en aggressiv tone og sporer av fra det opprinnelige temaet.

1.1 Hva er hatefulle ytringer?[2]

Det finnes ingen omforent definisjon av hatefulle ytringer (*hate speech*), verken i jussen internasjonalt eller i forskningslitteraturen.[3] Det er likevel vanlig å forstå hatefulle ytringer som ytringer som er av hatefull eller diskriminerende art, og som retter seg mot en gruppe eller mot et individs (antatte) gruppetilhørighet. For at en ytring skal defineres som hatefull, er det altså vesentlig at den er rettet mot bestemte former for gruppeidentiteter, det vi betegner som *grunnlaget* ytringene er rettet mot. Hatefulle ytringer kan ses på som et samlebegrep som omfatter mer konkrete begreper som rasistiske, homofobiske, antisemittiske, antimuslimske eller fremmedfiendtlige ytringer. Dette begrepet rommer altså ytringer basert på flere antipatier mot enkeltgrupper (Brudholm 2013: 31). Hatefulle ytringer kan uttrykkes muntlig eller i form av tekst, bilder, symboler og andre medieuttrykk.

Den norske straffeloven har et forbud mot hatefulle ytringer (§ 185, tidligere § 135a). § 185 verner ikke primært den som direkte utsettes for hat, men den *gruppen* mottakeren tilhører. Kjernen i bestemmelsen er ytringer som er egnet til å skape hat mot en gruppe – altså en ytring som fremkaller en følelse (hat) hos noen andre enn dem som vernes av bestemmelsen. Dermed står bestemmelsen om hatefulle ytringer i kontrast til for eksempel bestemmelsen om trusler (§ 183) som i stedet verner *individer* mot ytringer som direkte fremkaller følelsen alvorlig frykt hos den som rammes.

Den norske straffeloven § 185 verner mot diskriminerende og hatefulle ytringer som fremsettes offentlig eller i andres nærvær på grunn av noens

a) hudfarge eller nasjonal eller etnisk opprinnelse,
b) religion eller livssyn,
c) homofile orientering, eller
d) nedsatte funksjonsevne.

Gruppene som har juridisk vern mot hatefulle ytringer, er (historisk) utsatte minoritetsgrupper. Det er verdt å merke seg at bestemmelsen gir vern til alle grunnlagene som omfattes av diskrimineringslovverket, med unntak av kjønn, kjønnsidentitet og kjønnsuttrykk. Altså omfatter ikke § 185 ytringer motivert av hat mot for eksempel kvinner eller transpersoner som grupper. Imidlertid vil en del ytringer som er av hatefull karakter, men som ikke rettes mot gruppene som

2 En stor takk til Anine Kierulf, fagdirektør ved Norges nasjonale institusjon for menneskerettigheter for innspill til denne delen av teksten.
3 I denne rapporten bruker vi begrepene «hatefulle ytringer» og «hatytringer» om hverandre.

er vernet av straffeloven § 185, kunne fanges opp av andre bestemmelser i straffeloven (se diskusjon i Wessel-Aas, Fladmoe & Nadim 2016).

I samfunnsvitenskapelige og empiriske tilnærminger til fenomenet er det ofte brukt en videre definisjon av hatytringer enn den som følger av straffeloven. Dette skyldes dels at det er krevende å avgrense en empirisk studie til å fange nøyaktig de typene ytringer som ville falle inn under en juridisk definisjon, særlig på dette feltet, der de juridiske grensene ikke er klart trukket opp. Videre setter straffeloven terskelen for hvilke ytringer den omfatter, svært høyt. I empiriske studier som søker å forstå fenomenet hatytringer og dets årsaker og konsekvenser, er også andre typer ytringer enn dem som omfattes av straffeloven, relevante (LDO 2015; Nadim, Fladmoe & Wessel-Aas 2016).

Ulike definisjoner av hatytringer – både juridiske og ikke-juridiske – skiller seg gjerne fra hverandre langs følgende dimensjoner (se også Article 19 2015):

Hvilke grupper som innlemmes: Ulike juridiske og akademiske definisjoner av hatytringer varierer med hensyn til hvilke egenskaper ved et individ eller en gruppe som innlemmes i begrepet. Begrepet hatefulle ytringer er som regel ment å fange ytringer rettet mot (historisk) stigmatiserte og marginaliserte grupper. Dermed er gjentatte eller systematiske offererfaringer, historisk kontekst og det sosiale forholdet mellom de involverte aktørene direkte relevant for hvordan man forstår hatefulle ytringer (jf. Chakraborti 2010; Lawrence III, Matsuda, Delgado & Crenshaw 1993; Perry 2001).

Kjønn, kjønnsuttrykk og alder er eksempler på gruppeegenskaper det pågår diskusjoner om hvorvidt bør inngå i definisjonen av hatefulle ytringer (jf. LDO 2015). Mens noen definisjoner lister opp de vernede grunnlagene som inngår, opererer andre med en åpen liste hvor de gir eksempler på hvilke grunnlag som inngår, uten å presentere listen som endelig (for eksempel ECRI 2016; LDO 2015).[4]

Selv om det altså varierer hvilke gruppeegenskaper som innlemmes i begrepet, er hatefulle ytringer, slik det brukes i forskning, altså avgrenset til å gjelde ytringer som retter seg mot noens (antatte) tilhørighet til en bestemt gruppe, og utelukker dermed ytringer som er rettet mot rent individuelle egenskaper.

Alvorlighetsgraden i ytringen: Det er ikke enkelt å definere klart hvor alvorlig en ytring må være for å kunne bli definert som en hatytring. Straffeloven § 185

4 ECRI er European Commission against Racism and Intolerance.

har et krav om at ytringen skal «true eller forhåne noen, eller fremme hat, forfølgelse eller ringeakt overfor noen».

Bestemmelsen omfatter ytringer som kan forstås som trusler eller oppfordringer om å utsette enkeltpersoner eller grupper for ulovlige handlinger fordi de tilhører en av de vernede gruppene. Men den omfatter også ytringer som «bare» er egnet til å fremme hat mot, eller som grovt nedvurderer menneskeverdet til, personer eller grupper som tilhører en av de vernede gruppene (se diskusjon i Wessel-Aas mfl. 2016: 28).

Ifølge den britiske menneskerettighetsorganisasjonen Article 19 er det stor variasjon med hensyn til hvor stor vekt definisjoner av hatefulle ytringer legger på skadevirkningene av ytringen. Dette handler blant annet om hvorvidt ytringen blir sett på som skadelig i seg selv – fordi den er nedverdigende eller dehumaniserende – eller om ytringen blir ansett å ha en potensiell eller faktisk skadevirkning ved å oppfordre til handling, ved å forårsake en emosjonell respons hos mottakeren, eller om ytringen påvirker holdninger i samfunnet negativt gjennom å spre hat (Article 19 2015). Det varierer også hvor stor vekt det legges på at man skal kunne påvise en årsakssammenheng mellom ytringen og den spesifiserte skadevirkningen, og hvor stor vekt det legges på at skadevirkningene er sannsynlige eller umiddelbare.

Distinksjonene i hvor alvorlig en ytring må være for å kvalifisere som en hatytring, henger også sammen med en diskusjon om hvorvidt «hat» er et fruktbart begrep i denne sammenheng. Flere argumenterer for at bruken av «hat» indikerer at fenomenet kun inkluderer svært ekstreme ytringer, og at det da er en fare for at vi mister av syne mer subtile, men potensielt like skadelige uttrykk for fordommer (Chakraborti 2010). Argumentet er altså at hatytringer bør ses på som uttrykk for fordommer, mens hat er en mindre og mer ekstrem del av dette fenomenet. For mottakeren er det vel så mye summen av mange små, mindre alvorlige krenkelser som har konsekvenser over tid (Bowling 1999; Chakraborti 2010). I sin filosofiske diskusjon av hatbegrepet konkluderer imidlertid Brudholm (2013) med at «hat» er et passende begrep å bruke, fordi de mest alvorlige tilfellene av hatytringer inkluderer tre faktorer som er gjennomgående i hatets begrepshistorie: 1) markeringen av grupper og kategorier som gjenstand for hat, 2) vektleggingen av bakgrunnen for hatet i forestillinger om gruppens ondskap, foraktelighet, farlighet, utålelighet eller umenneskelighet, og 3) ønsket om at gjenstanden for hatet skal forsvinne, utryddes eller utstøtes (Brudholm 2013: 44).

Intensjon (om å skade): I den norske straffeloven (§ 185) er det ikke noe krav for straffbarhet at avsenderen av en hatefull *ytring* har et hatefullt motiv. Avsenderen må enten ha ønsket å fremsette en hatefull ytring, forstått eller regnet det som overveiende sannsynlig at han/hun gjorde det, men det holder at avsenderen sterkt kan klandres for at det skjedde (se Wessel-Aas mfl. 2016: 29).

European Commission against Racism and Intolerance (ECRI) presiserer i sin definisjon av hatefulle ytringer at det ikke bare gjelder ytringer som har til intensjon å fremme vold, trusler, fiendtlighet eller diskriminering, men også uttrykk som med rimelighet kan forventes å ha en slik effekt (ECRI 2016).

Mens kravet om hatefullt motiv ofte er sentralt i definisjonen av hatkriminalitet, behøver altså ikke definisjoner av hatytringer å innebære krav om en bestemt type motivasjon.

Hvorvidt ytringen må være fremsatt offentlig eller i andres nærvær: I norsk straffelovgivning (§ 185) og i mange andre juridiske definisjoner er det et krav at ytringen skal være fremsatt offentlig eller i andres nærvær (Article 19 2015). Lovbestemmelser om hatytringer verner mot ytringer som er egnet til å skape hat mot bestemte grupper, og ytringene er bare egnet til å gjøre det dersom flere enn de direkte mottakerne får dem med seg. Selv om hatefulle ytringer på tomannshånd kan ha like sterke konsekvenser for individet, er det ikke primært individet, men gruppen som vernes av juridiske bestemmelser om hatefulle ytringer.

Likevel er det ikke vanlig at mer samfunnsvitenskapelige definisjoner har som krav at ytringen skal være fremsatt offentlig. I en empirisk undersøkelse av omfanget av hatefulle ytringer kan det også være relevant å fange opp erfaringer med at ytringene sendes direkte til mottakeren.

1.1.1 Forståelsen av hatefulle ytringer i denne rapporten

I denne rapporten legger vi til grunn en utvidet definisjon av hatefulle ytringer som bevisst stigmatiserende, diskriminerende, nedverdigende (*degrading*) eller truende ytringer rettet mot et individ eller en gruppe på grunnlag av bestemte (oppfattede) gruppekarakteristikker (jf. Boeckmann & Turpin-Petrosino 2002; Gagliardone, Gal, Alves & Martinez 2015: 10; Institut for Menneskerettigheder 2017; Lawrence III mfl. 1993). Med begrepet hatefulle ytringer ønsker vi først og fremst å favne ytringer rettet mot karakteristikker som assosieres med medlemmer av historisk utsatte (minoritets)grupper (jf. Lawrence III mfl. 1993). Vi vil ikke begrense diskusjonen til grunnlagene som er vernet av § 185, men også

inkludere grunnlag som kjønn og klassebakgrunn for å favne et bredere spekter av gruppebaserte ytringer. I tillegg vil vi inkludere ytringer som er sendt direkte til mottakeren.

For å avgjøre hvilke ytringer som kan regnes som hatefulle, må ytringene *tolkes* i konteksten de falt i. For eksempel kan en påstand om at en bestemt innvandringsgruppe er overrepresentert på kriminalitetsstatistikken, være legitim når den fremføres som en del av en faglig diskusjon om kriminalitetsstatistikken, mens den kan være mer problematisk hvis den er fremsatt i den hensikt å spre hat i en sterkt innvandringsfiendtlig gruppe på Facebook. I en empirisk kartlegging av omfanget av hatefulle ytringer vil forskeren sjelden ha mulighet til å gjøre en konkret vurdering av hver ytring. Samtidig gir ulike metodiske tilnærminger ulik grad av kontroll over hva som fanges opp av en empirisk undersøkelse. Mens surveyundersøkelser i stor grad fanger opp ytringer som etter folks subjektive oppfatning kan regnes som hatefulle, vil det i en innholdsanalyse være forskergruppen som avgjør hvilke ytringer som kvalifiserer til å bli regnet med. I en samfunnsvitenskapelig tilnærming kan det være en fordel å gå bredt ut når man kartlegger omfanget av hatefulle ytringer, for så å analytisk kategorisere ytringer som mer eller mindre «hatefulle» i ettertid. Spørsmålet om hvilke muligheter ulike metodiske tilnærminger gir for å fange opp fenomenet vi er interessert i, blir diskutert videre i rapporten.

Hatefulle ytringer kan fremføres på mange ulike arenaer, både på og utenfor internett. Med den økte bruken og tilgjengeligheten av internett og sosiale medier har mulighetene for å ramme enkeltpersoner og grupper med hatefulle ytringer blitt sterkt utvidet, og det har den senere tiden vært økt oppmerksomhet om hatefulle ytringer på nett både i politikken og i akademia. Hatefulle ytringer på internett skiller seg fra lignende ytringer andre steder gjennom tilgjengeligheten, rekkevidden, anonymiteten og at de er umiddelbare (Brown 2017). Selv om hatytringer på internett ikke nødvendigvis er noe substansielt annet enn hatytringer fremført på andre arenaer, er det knyttet noen spesifikke utfordringer til ytringer på nett, og da særlig det faktum at de ikke nødvendigvis forsvinner. Ytringer på internett kan leve et langt liv i ulike formater, på ulike plattformer og kan lenkes til gjentatte ganger. I tillegg foregår en stor andel av interaksjonene på internett på sosiale nettverksplattformer som ikke er underlagt nasjonal jurisdiksjon (Gagliardone mfl. 2015). Det er dermed vesentlig å studere omfanget av og karakteristikker ved hatefulle ytringer på ulike arenaer, både på og utenfor nettet.

I tillegg er det vesentlig å studere ulike typer erfaringer med hatefulle ytringer, fra å ha vært direkte mottaker til «bare» å ha vært vitne til slike ytringer. Også

det å observere eller være vitne til hatefulle ytringer kan ha skadevirkninger. Hatefulle ytringer rettet mot for eksempel homofile bidrar til en nedvurdering av denne gruppen. Dette kan påvirke andre homofile, men også allmennheten mer generelt ved at det forsterker fordommer og kommuniserer en oppfatning om at det ikke er greit å være homofil (jf. Nilsen 2014: 8). Hatefulle ytringer rammer altså grupper, ikke bare enkeltpersoner, og kan ha negative skadevirkninger også for dem som observerer slike ytringer.

Til tross for studiene som er gjennomført de siste årene, har vi begrenset kunnskap om hatefulle ytringer i Norge. De ISF-ledede prosjektene «Status for ytringsfriheten» (Fladmoe & Nadim 2017; Midtbøen & Steen-Johnsen 2016) og «Hatefulle ytringer på internett. Omfang, forebygging og juridiske grenser» (Nadim mfl. 2016) har samlet inn noe data om personers erfaringer med å motta ubehagelige kommentarer, hatefulle ytringer og trusler. Men med unntak av personer med nedsatt funksjonsevne (Olsen, Vedeler, Eriksen & Elvegård 2016) har vi begrenset med kunnskap om erfaringene til de gruppene som er mest utsatt for hatytringer i Norge. I tillegg mangler vi kunnskap om avsendere av hatytringer og om innholdet i ytringene (Nadim mfl. 2016).

Også internasjonalt er det begrenset med forskning på omfanget av hatefulle ytringer. Flere studier har undersøkt omfanget av opplevelser med ulike typer ubehagelige kommentarer på nettet, men få studier har bevisst forsøkt å måle hatefulle ytringer i betydningen ytringer som er egnet til å spre hat mot bestemte grupper (Nadim mfl. 2016). I stedet har ulike studier undersøkt ulike aspekter ved ubehagelige opplevelser på nettet, fra ubehagelige kommentarer til alvorlige trusler. Dermed er det ikke alle fenomenene som er målt, som faller inn under definisjonen av det vi forstår som hatefulle ytringer.

1.2 Tilnærminger til å studere hatefulle ytringer

Man kan se for seg minst tre måter å forsøke å måle omfang av hatefulle ytringer på:

1. *mottaker- eller observatørperspektiv*: å studere hvor vanlig det er å motta eller observere hatefulle ytringer (eller tilgrensende fenomener)
2. *avsenderperspektiv*: å studere hvor vanlig det er å stå bak hatefulle ytringer
3. *innholdsperspektiv*: å studere meningsinnholdet i ytringer i et bestemt forum for å se hvor vanlig det er med hatefullt innhold

I den grad det har blitt gjort studier om omfang av hatefulle ytringer, tar de i all hovedsak utgangspunkt i enten et mottaker- eller et innholdsperspektiv. Det er også disse tilnærmingene vi vil konsentrere oss om i denne rapporten.

Studier av erfaringer med å motta eller observere hatefulle ytringer er stort sett basert på surveyundersøkelser, mens studier som analyserer innholdet i ytringer, bygger på ulike former for innholdsanalyse. Disse to tilnærmingene beskriver og diskuterer vi nærmere i henholdsvis kapittel 2 og 3. Vi vil hovedsakelig konsentrere oss om *kvantitative* metodiske tilnærminger i denne rapporten. Selv om det finnes noe kvalitativ forskning om hatefulle ytringer og tilgrensende fenomener, som kan gi viktig innsikt i hatefulle ytringer som fenomen og hvilke konsekvenser det kan få, er kvalitative tilnærminger lite egnet til å måle *omfanget* av hatefulle ytringer.

1.3 Grunnlaget for de metodiske vurderingene

Vår utredning av metodiske utfordringer og muligheter er basert på en rekke kilder:

- Eksisterende empirisk forskning, både norsk og internasjonal, om hatefulle ytringer og tilgrensende fenomener for å identifisere ulike metodiske tilnærminger i studier av omfang av hatefulle ytringer og utfordringer og muligheter med disse tilnærmingene

- Definisjoner og operasjonaliseringer av hatefulle ytringer i empirisk forskning, både i Norge og internasjonalt – vi vurderer hvor hensiktsmessige ulike måter å operasjonalisere begrepet på er, og diskuterer validiteten til konkrete spørsmålsformuleringer i ulike studier

- Oppdatert og relevant metodelitteratur som mer generelt diskuterer muligheter og begrensninger som ligger i ulike metodiske tilnærminger, og som kan bidra til å belyse innovative tilnærminger som kan være relevante for studier av omfanget av hatefulle ytringer

- En systematisk gjennomgang av egne erfaringer med gjennomføring av relevante empiriske undersøkelser. ISF har lang erfaring med surveymetodikk og har gjort flere undersøkelser om hatefulle ytringer og tilgrensende fenomener (blant annet gjennom prosjektene «Status for ytringsfriheten i Norge» og «Social media in the public sphere»). I disse undersøkelsene har vi studert erfaringer med å motta hatefulle eller ubehagelige ytringer ved hjelp av surveymetodikk. Utover dette har forskere ved ISF også erfaring

med å studere medieinnhold ved hjelp av innholdsanalyse, både ved hjelp av manuell koding og automatiserte analyser av stordata (jf. Enjolras, Steen-Johnsen & Karlsen 2014; Karlsen & Enjolras 2016).

– Erfaringer fra relevante miljøer som har erfaring med andre typer metodiske tilnærminger enn det vi har på ISF. Dette gjelder blant annet Likestillings- og diskrimineringsombudet (LDO), som har et prosjekt gående som kartlegger karakteren og omfanget av hatefulle ytringer på debattråder på nyhetssider på Facebook i to utvalgte perioder gjennom manuell koding av innhold. ISF er en del av referansegruppen til prosjektet og har fulgt det fra en tidlig fase. I tillegg har vi kontakt med forskningsgruppen for språkteknologi ved Institutt for informatikk ved Universitetet i Oslo. Som eksperter på datadrevet modellering og språkteknologisk bruk av maskinlæring har de bidratt med vurderinger av muligheter for maskinlæring ved bruk av de mest avanserte metodene på feltet.

1.4 Vurderingskriterier

I gjennomgangen av ulike metodiske tilnærminger til å måle omfanget av hatefulle ytringer vil vi blant annet legge vekt på følgende vurderingskriterier:

– *Fange fenomenet*: Hvor nøyaktig er den metodiske tilnærmingen i å fange fenomenet vi er interessert i? Selv om det ligger en klar definisjon av hatefulle ytringer til grunn, kan ulike måter å måle fenomenet på gi ulike resultater. Derfor er det sentralt å vurdere i hvilken grad ulike operasjonaliseringer av begrepet hatefulle ytringer måler det vi ønsker å måle. Dette er spørsmål knyttet til begrepsvaliditeten i ulike metodiske tilnærminger. I tillegg vil vi vurdere i hvilken grad ulike tilnærminger åpner for å studere ulike former for hatefulle ytringer.

– *Ulike grunnlag*: I hvilken grad gir den metodiske tilnærmingen mulighet til å fange opp hatefulle ytringer rettet mot relevante målgrupper? Hvilke muligheter har den til å sammenligne utsatte gruppers erfaringer med hatefulle ytringer med majoritetsbefolkningens erfaringer? Er det mulig å sammenligne de ulike diskrimineringsgrunnlagene, og i hvilken grad kan den metodiske tilnærmingen ivareta interseksjonelle perspektiver ved at det er mulig å se de ulike grunnlagene i sammenheng? Denne vurderingen vil ikke utelukkende være knyttet til diskrimineringsgrunnlagene, men også drøfte hvorvidt det er mulig å fange opp betydningen av for eksempel klassebakgrunn.

– *Ulike arenaer*: I hvilken grad kan ulike metodiske tilnærminger fange opp på hvilke *arenaer* hatefulle ytringer blir fremsatt? Er det for eksempel mulig å studere hatefulle ytringer både på og utenfor nettet?

– *Representativitet*: Et viktig vurderingskriterium er spørsmålet om hvor representative resultater ulike metodiske tilnærminger gir. Dette innebærer å vurdere faktorer som utvalgsstørrelse, frafallsproblematikk og muligheter for å identifisere og nå relevante målgrupper.

– *Mulighet til å innhente annen relevant informasjon*: I hvilken grad åpner den metodiske tilnærmingen for å innhente informasjon om andre aspekter ved hatefulle ytringer utover omfang? Særlig relevant er for eksempel informasjon om avsendere, konteksten hvor hatytringene falt, konsekvenser av hatytringer osv.

– *Tidsserier og komparasjon*: I hvilken grad er det mulig å gjenta det metodiske opplegget over tid, og i hvilken grad åpner tilnærmingen for internasjonal komparasjon?

– *Kostnadseffektivitet*: Vi vil også, på et overordnet nivå, vurdere hvor kostnadskrevende ulike metodiske opplegg er, og diskutere dette målt mot metodiske muligheter og utfordringer.

2 Surveymetoder

En vanlig tilnærming til å studere omfang av hatefulle ytringer er å spørre folk om deres erfaringer med å motta eller observere slike ytringer. Dette kan gjøres gjennom spørreundersøkelser, også kalt surveymetoder. Surveymetoden har noen åpenbare fordeler. For det første gjør metoden det mulig å fange opp erfaringer med ulike former for og definisjoner av hatefulle ytringer og andre tilgrensede fenomener blant de samme personene. Ved å stille ulike spørsmål som har relevans for studien, åpner man opp for analyser som ser på sammenhengen mellom kjennetegn ved individer og erfaringer med ulike former for hatefulle ytringer, konsekvenser av disse erfaringene og andre relevante aspekter. For det andre gjør metoden det mulig å måle omfanget av erfaringer med hatefulle ytringer og andre tilgrensende fenomener i ulike befolkningsgrupper. Dersom vi bruker sannsynlighetsutvalg – at alle medlemmer av en gitt populasjon har samme sannsynlighet for å delta i undersøkelsen – er det mulig å generalisere resultatene til populasjonen som helhet. Undersøker vi flere populasjoner, kan vi dermed sammenligne omfang og erfaringer mellom ulike befolkningsgrupper.

Samtidig kan det være utfordringer knyttet til formuleringen av spørsmålene og hvordan man oppnår kontakt med representative utvalg av relevante minoritetsgrupper. I tillegg er det viktig å understreke at man med surveyundersøkelser måler *subjektive* erfaringer med hatefulle ytringer. Å vurdere om en ytring er hatefull, er nødvendigvis et tolkningsspørsmål, og ulike personer kan tolke den samme ytringen ulikt.

Under vurderer vi nærmere muligheter og utfordringer knyttet til surveymetoden og vurderer metoden opp mot kriteriene som ble listet opp i det første kapitlet.

2.1 Muligheter for å måle fenomenet hatefulle ytringer

Surveymetoden gjør det mulig å studere ulike sider ved hatefulle ytringer. I tillegg til å studere *erfaringer* med å motta hatefulle ytringer gjør metoden det mulig å studere de samme personenes erfaringer med å *observere* hatefulle ytringer, og om erfaringer med å motta eller observere hatefulle ytringer har hatt *konsekvenser* for dem. Metoden gjør det også mulig å studere *avsendere* av

hatefulle ytringer. Under går vi nærmere gjennom ulike momenter knyttet til det å fange opp fenomenet hatefulle ytringer ved hjelp av surveymetoder. Hvordan kan vi definere og operasjonalisere hatefulle ytringer? Hvilken informasjon kan vi fange opp ved hjelp av surveymetoder? I hvilken grad er tematikken sensitiv – og hvordan løser vi i så fall dette?

2.1.1 Definisjoner og operasjonaliseringer

Hvordan spørsmål om et fenomen skal stilles, er alltid en utfordring i spørreundersøkelser. Små variasjoner i spørsmålsformuleringer kan produsere svært ulike svar. Dette kan særlig ha konsekvenser for en omfangsundersøkelse av hatefulle ytringer fordi omfanget vil kunne variere betydelig avhengig av hvordan man definerer og operasjonaliserer fenomenet. I en undersøkelse i USA fant for eksempel Pew Research Center at 27 prosent hadde opplevd å bli kalt noe *krenkende på internett*, mens 7 prosent hadde blitt *trakassert over tid* (Pew Research Center 2014). Tilsvarende har norske studier funnet at om lag 20 prosent av befolkningen har opplevd å motta det de opplevde som *ubehagelige eller nedlatende kommentarer*, i sosiale medier, mens 7 prosent har opplevd å motta det de opplevde som *hatefulle ytringer* (Nadim mfl. 2016).

Utfordringen med hvordan spørsmål om hatefulle ytringer skal formuleres, kompliseres av at definisjonen av fenomenet ikke er allment kjent og avgrenset. Som vi diskuterte i innledningskapitlet, har «hatefulle ytringer» en avgrenset juridisk definisjon, men en videre og uklar definisjon i samfunnet for øvrig. Mange har trolig ikke et bevisst forhold til begrepet hatefulle ytringer, og det kan lett blandes sammen med mer generelle former for hets og mobbing. Dette blir ytterligere komplisert av at ulike personer vil kunne tolke den samme ytringen ulikt. Hva en respondent i en survey oppfatter som «hatefullt», er nødvendigvis en subjektiv oppfatning. Studier har for eksempel vist at mange oppfatter ytringer som hatefulle selv når de rettes mot grunnlag som er forskjellig fra grunnlagene i konvensjonelle definisjoner av hatytringer, som innholdet i argumentet eller en persons yrke (Fladmoe & Nadim 2017).

Når et abstrakt begrep som hatefulle ytringer skal operasjonaliseres i en spørreundersøkelse, bør derfor fenomenet defineres så konkret som mulig. I litteraturen har dette vært forsøkt løst på ulike måter. Én tilnærming har vært å først stille relativt generelle spørsmål om erfaringer med å motta eller observere hatefulle ytringer og deretter gi oppfølgingsspørsmål om hvilke grunnlag ytringene har vært rettet mot (COWI 2015; Hawdon, Oksanen & Räsänen 2015; Midtbøen & Steen-Johnsen 2016; Nadim mfl. 2016). I en norsk undersøkelse fra 2016 (Nadim mfl. 2016) ble for eksempel respondentene spurt om de selv hadde

«[…] mottatt hatefulle ytringer via sosiale medier». Spørsmålet ble fulgt av en definisjon av hatefulle ytringer som «[…] nedverdigende, truende, trakasserende eller stigmatiserende» ytringer. De som svarte «Ja» på dette spørsmålet, fikk deretter et oppfølgingsspørsmål om hva disse ytringene oftest var rettet mot. Svarlisten inneholdt alle grunnlagene som er vernet av straffeloven, men også andre relevante grunnlag. På denne måten kan man i ettertid kategorisere ytringer avhengig av om de rettes mot juridisk vernede eller andre grunnlag.

En annen tilnærming har vært å stille flere spørsmål om ulike konkrete erfaringer. Et eksempel på dette er fremgangsmåten Nordlandsforskning valgte i rapporten «Hatytringer. Resultater fra en studie av funksjonshemmedes erfaringer» (Olsen mfl. 2016), som var basert på en metode utviklet av en gruppe engelske forskere (Beadle-Brown, Richardson, Guest, Malovic, Jill & Julian 2014). Undersøkelsen til Nordlandsforskning inneholdt flere detaljerte spørsmål både om innholdet i ytringene og på hvilke arenaer de ble fremsatt. Eksempler på erfaringer var «ytringer om at du er ubehagelig å se eller høre på», og «ytringer om at du ikke burde være født eller ikke har rett til å leve» (Olsen mfl. 2016: 17). Basert på listen over konkrete erfaringer kategoriserte forskerne opplevelsene som henholdsvis hatytringer, krenkende ytringer eller ubehagelige ytringer. Fordelen med denne tilnærmingen er at man ikke forutsetter at respondentene har lik oppfatning av hva som er «hatefullt», «krenkende» og så videre, men at man basert på respondentenes konkrete opplevelser selv kan definere dem på ulikt vis. Denne tilnærmingen åpner derfor opp for at ulike definisjoner kan diskuteres spesifikt opp mot empiriske forhold.

De to tilnærmingene kan også tenkes kombinert. Det er mulig å stille flere detaljerte spørsmål om erfaringer med ulike typer konkrete ytringer og deretter følge opp med spørsmål om hvilke(t) grunnlag ytringen(e) var rettet mot. På denne måten reduserer man problemet med at folk ikke har sammenfallende oppfatninger av hva som er hatefullt eller ikke, og man får mulighet til å skille mellom ytringer rettet mot ulike grunnlag.

I tillegg til å stille flere spørsmål om det samme fenomenet vil det være nyttig å også stille spørsmål om hvor alvorlige ytringene oppfattes å være. Som vi diskuterte innledningsvis, er terskelen for at en ytring skal være straffbar i henhold til § 185, høy. Dette betyr ikke at mindre alvorlige hatytringer ikke er problematiske. Men dersom en respondent svarer at han/hun har mottatt hatefulle ytringer, men samtidig oppfatter ytringen som lite alvorlig, kan det være en indikasjon på at vedkommende har en bred subjektiv forståelse av hva som er «hatefullt».

2.1.2 Annen informasjon som kan fanges opp

En åpenbar fordel med surveymetoden er at man kan stille flere relevante spørsmål om ulike forhold til de samme personene. Dermed kan man med én enkelt undersøkelse besvare flere problemstillinger.

For det første gjør metoden det mulig å ta hensyn til interseksjonelle perspektiv, altså at folk har sammensatte identiteter, og at erfaringene deres ikke nødvendigvis kan spores tilbake til kun ett aspekt ved identiteten. Dette er relevant både når det gjelder innholdet i hatefulle ytringer og egenskaper hos mottakerne. Innholdet i enkeltytringer kan være rettet mot flere grunnlag på en gang, og enkeltpersoner kan også motta ulike hatefulle ytringer som retter seg mot ulike grunnlag. Når det gjelder egenskaper hos dem som mottar hatefulle ytringer, kan personer som tilhører flere minoritetsgrupper (for eksempel etnisk og seksuell minoritet), være mer utsatt for hatefulle ytringer enn andre.

For det andre åpner metoden for å studere andre fenomener enn egenopplevde erfaringer med hatefulle ytringer. Nedenfor går vi gjennom tre forhold som kan studeres gjennom surveybaserte studier av hatefulle ytringer: 1) erfaringer med å observere hatefulle ytringer, 2) konsekvenser av å motta eller observere hatefulle ytringer og 3) egenskaper ved avsendere av hatefulle ytringer.

Den juridiske definisjonen av hatefulle ytringer (straffeloven § 185) rammer også ytringer som fremsettes offentlig uten noen direkte mottakere. En begrunnelse for dette er at hatefulle ytringer også kan ha negative konsekvenser for personer som ikke er direkte mottakere. Dette gjør det relevant å studere folks erfaringer med å observere hatefulle ytringer. Enkelte studier har gjort dette, og ikke overraskende er andelen som svarer at de har observert hatytringer, langt høyere enn andelen som svarer at de selv har mottatt slike ytringer (for eksempel Hawdon mfl. 2015; Pew Research Center 2014).

Tidligere studier har dokumentert at hatefulle ytringer kan ha alvorlige konsekvenser for dem som rammes, blant annet frykt og andre emosjonelle symptomer, lavere selvtillit, følelse av tap av verdighet, tilbaketrekking fra offentligheten og begrenset bevegelsesfrihet (Boeckmann & Liew 2002; Boeckmann & Turpin-Petrosino 2002; Eggebø, Sloan & Aarbakke 2016; Fladmoe & Nadim 2017; Gelber & McNamara 2016; Herek, Cogan & Gillis 2002; Leets 2002; Midtbøen & Steen-Johnsen 2016; Pew Research Center 2014). Ved hjelp av surveymetoden er det mulig å spørre om ulike former for opplevde konsekvenser av å motta (eller observere) hatefulle ytringer. I tillegg kan man sammenligne konsekvensene av hatefulle ytringer både på tvers av kjennetegn ved

individene som mottar eller observerer ytringene, og på tvers av hvilke grunnlag ytringene rettes mot.

Til sist vil det være høyst relevant å inkludere spørsmål som kan fange opp informasjon om *avsendere* av hatefulle ytringer: hvem de er, og hva som motiverer dem. Det kan være metodisk krevende å studere avsendere, siden det for mange vil være vanskelig å innrømme at de har ytret seg på en slik måte. I forskningslitteraturen har derfor enkelte studier undersøkt dette indirekte ved å spørre mottakere av hatefulle ytringer om de vet hvem som sto bak, og i så fall hva som kjennetegner dem (Hagen 2015; Pew Research Center 2014).

Det finnes imidlertid også surveyteknikker som er utviklet for å fange opp kontroversielle holdninger og handlinger. En relevant tilnærming for å studere kjennetegn ved avsendere av hatefulle ytringer er såkalte listeeksperimenter (Mutz 2011). Denne teknikken er ment å styrke respondentenes følelse av anonymitet når de svarer, og har blant annet blitt brukt til å studere rasistiske holdninger i USA (Kuklinski, Cobb & Gilens 1997) og synet på innvandreres velferdsrettigheter i Norge (Cappelen, Kuhnle & Midtbø 2016). Enkelt forklart innebærer listeeksperimenter at respondenter ikke svarer direkte «ja» eller «nei» på et spørsmål, men at de med utgangspunkt i en liste over ulike aktiviteter svarer *hvor mange* av disse de har utført. Utvalget blir delt i minst to grupper: Kontrollgruppen får en liste over tre–fire aktiviteter, som *ikke* inneholder noe om hatefulle ytringer. Eksperimentgruppen får den samme listen pluss en aktivitet som omhandler det å stå bak hatefulle ytringer. Forskjellen i gjennomsnittlig antall aktiviteter gir dermed et estimat over hvor mange som har stått bak hatefulle ytringer (Mutz 2011).

Når man gjennomfører listeeksperimenter, er det vanlig å også inkludere et direkte spørsmål om hvorvidt respondentene har fremsatt hatefulle ytringer. På denne måten kan man undersøke hvorvidt det faktisk er kontroversielt å innrømme at man står bak hatefulle ytringer (Mutz 2011).

2.1.3 Sensitivitet og etiske krav

Å bruke spørreundersøkelser til å studere hatefulle ytringer kan kompliseres av at tematikken kan oppleves som sensitiv – spesielt hvis en studerer avsendere av hatefulle ytringer. Såkalt «sosial ønskbarhet» (*social desirability bias*) kan føre til at respondenter underrapporterer atferd som ikke er sosialt ønskelig (Ringdal 2001: 271–272). I metodebøker diskuteres ulike generelle råd for hvordan man kan redusere dette problemet. Blant annet bør man bruke ord og begreper som «folk flest» er fortrolige med – ikke tekniske fagtermer. Før det stilles sensitive

spørsmål i en undersøkelse, bør det stilles spørsmål som ikke er sensitive. I noen tilfeller kan man vurdere å konstruere ledende spørsmål for å ufarliggjøre fenomenet (for eksempel: «Mange har opplevd å motta hatefulle og andre ubehagelige kommentarer på nettet. Har du opplevd dette selv?»). På den annen side må man alltid være oppmerksom på at ledende spørsmål også kan føre til overrapportering av det fenomenet en studerer (Ringdal 2001).

I forlengelsen av dette vil gjennomføringsmetoden ha mye å si. Surveyundersøkelser om hatefulle ytringer har i all hovedsak blitt gjennomført enten ved hjelp av nettskjema (Hawdon mfl. 2015; Nadim mfl. 2016; Pew Research Center 2014) eller ved hjelp av telefonintervju (COWI 2015; Politidirektoratet 2016). Tradisjonelt har det vært en oppfatning om at tilstedeværelsen av en intervjuer (over telefon og ansikt til ansikt) er fordelaktig for å bidra med motivasjon og oppklaringer, men at intervjueren også kan gi såkalte «intervjueffekter» – blant annet at respondenten gir mer sosialt ønskbare svar. Selvutfylling (post- og nettskjema) kan på sin side gi svakere respons, men styrker respondentenes følelse av anonymitet (Krumpal 2013; Ringdal 2001).

I tillegg til de mer praktiske sidene ved å stille sensitive spørsmål er det også nødvendig å vurdere de etiske sidene ved denne typen studier. Surveyundersøkelser med sensitive spørsmål er underlagt egne etiske krav. Dette gjelder særlig helserelaterte opplysninger, men Norsk samfunnsvitenskapelig datatjeneste (NSD) krever også at undersøkelser med spørsmål av særlig personlig karakter meldes særskilt. Her inngår spørsmål om:

> «[…] rasemessig eller etnisk bakgrunn, politisk, filosofisk eller religiøs oppfatning, at en person har vært mistenkt, siktet, tiltalt eller dømt for en straffbar handling, helseforhold, seksuelle forhold, og medlemskap i fagforeninger»[5]

Flere av disse egenskapene (rasemessig eller etnisk bakgrunn, religiøs oppfatning, seksuelle forhold) inngår i lovvernet mot hatefulle ytringer og vil derfor være relevante å spørre om i en surveyundersøkelse om dette temaet. I surveyundersøkelser om hatefulle ytringer bør man derfor melde NSD at man har med slike spørsmål. Erfaring viser at om man skisserer et opplegg som er i tråd med de forskningsetiske retningslinjene, går det greit å få tillatelse til å inkludere slike spørsmål.

5 Se http://www.nsd.uib.no/nsddata/arkivering/005_vanlige_sporsmal.html

2.2 Utvalg og representativitet

Surveymetodikk gjør det mulig å studere hatefulle ytringer i ulike befolkningsgrupper. Ved å gjennomføre undersøkelser med et såkalt sannsynlighetsutvalg er det mulig å generalisere resultatene til den aktuelle gruppen (populasjonen) som helhet. Gjennomfører man en survey blant flere minoritetsgrupper, kan man dermed sammenligne omfang og erfaringer mellom disse gruppene.

Det er på en annen side utfordrende å gjennomføre representative undersøkelser i en del minoritetsgrupper, dels fordi det i mange tilfeller er vanskelig å definere populasjonen, og dels fordi svartilbøyeligheten i en del minoritetsgrupper er lav. En alternativ metode er derfor å gjennomføre spørreundersøkelser blant medlemmer av organisasjoner, menigheter eller lignende.

2.2.1 Definere populasjonen

I en utvalgsundersøkelse som skal være representativ for hele gruppen man studerer, må det være tilfeldig hvem som blir trukket ut til å delta, og alle medlemmer av gruppen må ha samme sannsynlighet for å bli trukket ut. Dette kalles et sannsynlighetsutvalg. Med et representativt utvalg kan man enkelt beregne den statistiske sannsynligheten for at svarene på spørsmålene som blir stilt, er gyldige for hele gruppen (Ringdal 2001:143).

Det ideelle utvalget i en undersøkelse om omfang av hatefulle ytringer bør derfor ta utgangspunkt i hele befolkningen. I slike utvalgsundersøkelser er det vanlig å basere seg på uttrekk fra folkeregisteret, eventuelt på utvalgsbasene til kommersielle markedsanalyseselskaper (som regel sammensatt av telefon- og adresseregistre). «Gullstandarden» for en omfangsundersøkelse av hatefulle ytringer vil være å trekke ut et stort utvalg fra folkeregisteret, slik at alle minoritetsgrupper blir representert i et tilstrekkelig antall. Slik får man robuste sammenligningsgrunnlag på tvers av alle de vernede grunnlagene.

Denne tilnærmingen kan imidlertid være svært kostbar. I og med at minoritetsgruppene utgjør små andeler av befolkningen som helhet, må man operere med svært store befolkningsutvalg for å treffe nok personer i de aktuelle minoritetsgruppene til at resultatene i ettertid kan brukes til statistisk analyse. For å illustrere: Dersom en minoritetsgruppe utgjør anslagsvis 5 prosent av befolkningen og man ønsker å intervjue minst 500 personer i denne gruppen, må man intervjue et befolkningsrepresentativt utvalg på minimum 10 000 personer.

Alternativt kan man trekke ut et utvalg basert på bestemte kjennetegn. I folkeregisteret ligger det en del informasjon man kan trekke utvalg basert på. Blant

annet er det informasjon om befolkningens og deres foreldres fødeland, noe som gjør det mulig å trekke utvalg basert på innvandringsstatus og landbakgrunn.[6] Slik kan man redusere størrelsen på bruttoutvalget noe. Denne tilnærmingen er brukt i flere spørreundersøkelser blant personer med innvandrerbakgrunn, for eksempel i levekårsundersøkelsene til Statistisk sentralbyrå (Vrålstad & Wiggen 2017). Levekårsundersøkelsene er svært kostbare å gjennomføre, og undersøkelser med rimeligere gjennomføringsmetoder basert på folkeregisterutvalg av innvandrerbefolkningen har ofte hatt problemer med å oppnå god representativitet (for eksempel Eimhjellen 2016; Staksrud, Steen-Johnsen, Enjolras, Gustafsson, Ihlebæk, Midtbøen, Sætrang, Trygstad & Utheim 2014). Dette drøfter vi grundigere senere i rapporten.

I folkeregisteret er det ikke registrert informasjon om andre grunnlag som vernes av straffeloven § 185: hudfarge, religiøs tilhørighet[7], seksuell legning eller eventuelle funksjonsnedsettelser. Skal man trekke befolkningsrepresentative utvalg basert på disse grunnlagene, må man derfor spørre respondentene om deres tilhørighet til ulike minoritetsgrupper.

Kostnadene kan reduseres ved å bruke nettpaneler som er satt sammen ved hjelp av sannsynlighetsutvalg. Dette er paneler flere markedsanalyseselskaper har, og som består av forhåndsrekrutterte personer – personer som har sagt seg villig til å delta i undersøkelser. Det er viktig å merke seg at kvaliteten på nettpanelene kan variere stort. Et sentralt punkt er hvilken metode som er brukt for å rekruttere medlemmer. I forskning bør man som hovedregel bruke paneler som er satt sammen gjennom rekruttering ved hjelp av sannsynlighetsutvalg. Et eksempel på et slikt panel er Norsk medborgerpanel, som gjennomføres i regi av Universitetet i Bergen og UNI Research Rokkansenteret, og som rekrutterer medlemmer til panelet med utgangspunkt i sannsynlighetsutvalg trukket fra folkeregisteret. Enkelte kommersielle markedsanalysebyråer bruker lignende strategier når de setter sammen panelene sine. Som hovedregel bør man unngå paneler som er satt sammen ved hjelp av selvrekruttering, altså at hvem som helst kan «melde seg på».

En nylig publisert studie om helse blant lesbiske, homofile og biseksuelle tok utgangspunkt i et panel fra et markedsanalysebyrå, som var satt sammen ved hjelp av sannsynlighetsutvalg (Anderssen & Malterud 2017). Ved å kontakte et stort antall medlemmer av panelet filtrerte de ut respondenter basert på spørsmål om legning. På denne måten fikk de et utvalg bestående av 1134 personer som

6 For nærmere informasjon om folkeregisteret, se http://www.skatteetaten.no/no/Person/Folkeregister/Dette-er-folkeregisteret/.

7 Medlemskap i tros- og livssynsorganisasjoner er ikke koblet til folkeregisteret.

svarte at de var enten lesbiske, homofile eller biseksuelle. For enkelte andre minoritetsgrupper kan vi også bruke tilsvarende metodikk.

Det er imidlertid viktig å presisere at forhåndsrekrutterte paneler trolig ikke er fullt ut representative for befolkningen som helhet. De som går med på å delta i slike paneler, har i utgangspunktet vist interesse for å svare på undersøkelser, noe som kan tyde på at de er mer samfunnsengasjerte enn befolkningen for øvrig. Det er også sannsynlig at ressurssterke personer, for eksempel høyt utdannede, er overrepresenterte i slike paneler. Vi har ikke kjennskap til den demografiske sammensetningen i panelene til kommersielle aktører. Men blant dem som svarte i Norsk medborgerpanel våren 2017, hadde nesten 60 prosent utdanning på høgskole- eller universitetsnivå.[8] Dette er nesten dobbelt så mange som i befolkningen for øvrig.[9]

2.2.2 Svartilbøyelighet

Når svarprosenten er lav i en spørreundersøkelse, vil det oppstå usikkerhet om hvorvidt utvalget er representativt for den aktuelle populasjonen. Over tid har det vært fallende svartilbøyelighet i spørreundersøkelser. Enkelte studier har vist at dette ikke nødvendigvis behøver å ha store konsekvenser når man studerer befolkningen som helhet. Ottar Hellevik har sammenlignet svar på flere spørsmål i Norsk monitor, hvor svarprosenten har vært helt nede i 4–5 prosent, med offentlig statistikk (valgresultater) og svar på tilsvarende spørsmål i befolkningsundersøkelser gjennomført av Statistisk sentralbyrå, hvor svarprosenten har vært 50–60 prosent. Hellevik finner at svarene i Norsk monitor ikke skiller seg dramatisk fra andre mer robuste datakilder (Hellevik 2015). Dette tyder på at frafallet selv i undersøkelser med lav respons *kan* være tilfeldig og ikke-systematisk, slik at ulike grupper av respondenter har like stor sannsynlighet for å delta eller falle fra. I så fall behøver ikke fallende svartilbøyelighet være et stort problem. Utfordringen er imidlertid at vi sjelden kan vite sikkert hvor systematisk frafallet er i en utvalgsundersøkelse. Dersom frafallet er systematisk – at enkeltgrupper er overrepresentert blant dem som faller fra – og særlig hvis denne systematikken har sammenheng med fenomenet man studerer, kan frafallet føre til skjeve resultater (Ringdal 2001:155).

Systematisk frafall er relevant når man studerer enkelte minoritetsgrupper. Blant personer med nedsatt funksjonsevne kan funksjonsnedsettelsen for enkelte i seg selv være et hindrer. Dermed risikerer man at frafallet i en undersøkelse blant

8 Se dokumentasjon på http://www.uib.no/medborger/76892/data-og-dokumentasjon (besøkt 26.10.2017).
9 Se https://www.ssb.no/utniv (besøkt 26.10.2017).

denne gruppen vil være størst blant dem som har størst funksjonsnedsettelse. Hvis det samtidig er slik at de med størst funksjonsnedsettelse er de som oftest mottar hatefulle ytringer, vil en spørreundersøkelse derfor kunne *undervurdere* omfanget av hatefulle ytringer. Denne problematikken drøftes i studien av erfaringer med hatefulle ytringer blant personer med nedsatt funksjonsevne (Olsen mfl. 2016: 24–25). Forfatterne påpeker også at svartilbøyeligheten kan variere etter faktiske erfaringer med hatytringer. På den ene siden kan en undersøkelse om temaet tiltrekke seg flere personer som har erfart hatefulle ytringer, enn personer som ikke har det. Dette kan dra i retning av at omfanget fremstår som større enn det er. På den annen side kan imidlertid personer som har vært utsatt for hatytringer, også unngå å svare dersom de ikke ønsker å rippe opp i erfaringene. Dette kan dra i retning av at omfanget fremstår som mindre enn det er.

En gruppe som har fått mye oppmerksomhet i samfunnsforskningen mer generelt, og som er relevant i studier av hatefulle ytringer, er personer med innvandrerbakgrunn. Det har vist seg å være svært krevende å oppnå representative utvalg i en del innvandrergrupper, med mindre man legger ned betydelige ressurser i gjennomføringen (Djuve, Gulløy, Kavli & Berglund 2009). Det kan være flere årsaker til dette, og disse vil variere mellom ulike innvandrergrupper. Språk kan åpenbart være et hinder for mange, men også begrenset kunnskap om og tillit til norske institusjoner kan gjøre folk mindre tilbøyelige til å svare (Djuve mfl. 2009). Dersom svarprosenten er lav og det er sammenheng mellom kjennetegn ved respondentenes innvandrerbakgrunn (for eksempel landbakgrunn, språkferdigheter, hudfarge) og hvor utsatt de er for hatefulle ytringer, vil det være knyttet stor usikkerhet til studier av denne gruppen. Levekårsundersøkelsene til SSB, som er nevnt ovenfor, oppnår mer enn 50 prosent svar blant utvalgte innvandrergrupper trukket fra folkeregisteret. SSB gjennomfører undersøkelsene både på telefon og som besøksintervju, og intervjuerne har relevante språkkunnskaper og ofte selv innvandrerbakgrunn fra samme land som respondentene. Slik reduserer de språk- og kulturbarrierer (Vrålstad & Wiggen 2017). Utfordringen er imidlertid at en slik datainnsamling er *svært kostbar*.

Rimeligere undersøkelser blant innvandrerbefolkningen med utvalg trukket fra folkeregisteret har hatt langt svakere svarprosent. Et eksempel er undersøkelsen *Status for ytringsfriheten i Norge*, hvor det i 2013 ble gjennomført en nettundersøkelse blant et tilfeldig utvalg personer med innvandrerbakgrunn fra Afrika, Asia og Øst-Europa. Et utvalg på 5000 personer ble tilfeldig trukket fra folkeregisteret, basert på informasjon om eget og foreldres fødeland. Personene fikk invitasjon i posten til å svare på en undersøkelse på nettet. Samlet svarte kun 719 personer (15 prosent) på undersøkelsen, til tross for at det i utvalgstrekkingen ble satt som kriterium at innvandrerne skulle ha bodd minst fem år i

Norge, slik at de med høyest språkbarriere ikke ble en del av utvalget. Svarprosenten varierte imidlertid betydelig med landbakgrunn, med 30 prosent svar blant dem som hadde bakgrunn fra Estland og 7 prosent blant dem som hadde bakgrunn fra Somalia og Eritrea. Dette kan tyde på svært systematiske skjevheter blant respondenter fra enkelte land. I det endelige utvalget var da også personer med høy utdannelse og lang botid klart overrepresentert (Staksrud mfl. 2014: 7).

Tilsvarende metode ble brukt i en større holdningsundersøkelse blant muslimer i Norge, som fikk mye offentlig oppmerksomhet høsten 2017 (Ishaq 2017). I undersøkelsen ble personer med innvandrerbakgrunn fra Tyrkia, Pakistan, Iran, Bosnia-Hercegovina, Irak, Somalia, Marokko og Afghanistan med minimum fem års botid trukket fra folkeregisteret. Svarprosenten var på i underkant av 13 prosent.

Tilsvarende utfordringer knyttet til svartilbøyelighet kan også være til stede i større eller mindre grad i andre minoritetsgrupper. En hovedutfordring er at det er vanskelig å definere populasjonen (jf. drøftingen ovenfor) – vi har altså ikke full oversikt over hvem som tilhører gruppen vi er interessert i. Dermed er det vanskelig å beregne skjevhet i utvalgene i ettertid, og vi kan ikke vite om de som har svart, er representative også for dem som ikke har svart. Konsekvensen er at resultatene blir mer usikre enn hva som er vanlig i befolkningsrepresentative undersøkelser.

2.2.3 Organisasjonsutvalg som alternativ

Et alternativ til generelle befolkningsutvalg er å rekruttere respondenter via interesseorganisasjoner, menigheter og lignende, hvor vi vet at vi treffer på de aktuelle minoritetsgruppene. Dette ble blant annet gjort i studien Nordlandsforskning gjorde om erfaringer med hatefulle ytringer blant personer med nedsatt funksjonsevne, der medlemmer av flere interesseorganisasjoner ble kontaktet (Olsen mfl. 2016). Denne tilnærmingen har noen åpenbare fordeler. For det første er populasjonen (medlemmene) tydelig avgrenset, og organisasjonene har ofte en god del informasjon om dem, noe som gjør det mulig å beregne hvor representativt utvalget er, i ettertid. For det andre kan responsen bli god dersom organisasjonen bidrar til å legitimere undersøkelsen. Motivasjonen for å delta i en survey vil i mange tilfeller styrkes av at organisasjonen man er medlem i, stiller seg bak gjennomføringen. For det tredje vil denne tilnærmingen kunne være svært kostnadseffektiv, spesielt hvis organisasjonen har e-postadresser til medlemmene sine.

Hovedutfordringen ved å bruke organisasjonsutvalg er at medlemmene ikke nødvendigvis er representative for en minoritetsgruppepopulasjon som helhet. Vi kan ikke vite om alle med nedsatt funksjonsevne er medlem i en interesseorganisasjon (jf. Olsen mfl. 2016), og det samme vil gjelde for andre minoritetsgrupper. Organisasjonsgraden i Norge er imidlertid høy (Arnesen, Sivesind & Gulbrandsen 2016), og svært mange organisasjoner er registrert i Brønnøysundregisteret. Insentivene for å registrere seg er sterke, blant annet fordi det gjør organisasjonen til en juridisk enhet med de plikter og rettigheter det utløser, herunder muligheter til å søke offentlig pengestøtte. Samtidig varierer organisasjonslandskapet mellom grunnlagene som er relevante for undersøkelser om hatefulle ytringer. Ikke alle grunnlag er representert på en oversiktlig måte. Mens noen grunnlag har en hovedorganisasjon (som for eksempel seksuell orientering og Foreningen FRI), er det stor fragmentering og lav organisasjonsgrad knyttet til andre grunnlag (som for eksempel etnisitet). Når det gjelder religion som grunnlag, kan det være aktuelt å basere seg på menigheter, moskeer og lignende.

Selv om medlemmene i organisasjonene ikke er representative for minoritetsgruppene som helhet, er en fordel med å bruke organisasjonsutvalg at vi i alle fall vet noe om populasjonen av *medlemmer*. Det vil si at selv om det kan hefte usikkerhet ved i hvilken grad en undersøkelse basert på organisasjonsutvalg er representativ for minoritetsgruppen i samfunnet som helhet, vet vi bedre hvor representativt utvalget er for medlemmene i organisasjonen. Dermed kan vi si noe om systematikken i frafallet og om hvor vanlig det er med for eksempel erfaringer med hatefulle ytringer blant medlemmer av bestemte organisasjoner.

En annen potensiell utfordring med organisasjonsutvalg er at mulighetene for å studere endringer over tid svekkes dersom det skjer større endringer i organisasjonslandskapet innenfor et felt og oppslutningen om enkeltorganisasjoner innenfor en bestemt minoritetsgruppe minker (eller øker). Hvis for eksempel en organisasjon splittes i flere mindre organisasjoner, er det vanskelig å vite om den samlede medlemsmassen er sammenlignbar med hva den var tidligere.

Tilsvarende kan det være utfordrende å basere seg på medlemslister dersom man ønsker å sammenligne med andre land. Organisasjonslandskapet og organiseringspraksisen (sivilsamfunnet) vil variere mellom ulike land. Sivilsamfunnet i Norge har mange likhetstrekk med andre skandinaviske land, mens forskjellene er større hvis vi sammenligner med for eksempel søreuropeiske land (Salamon & Anheier 1998; Sivesind & Selle 2010).

2.2.4 Koble seg på andre undersøkelser

Et siste alternativ vi vil nevne her, er muligheten for å koble seg på andre større undersøkelser. Dersom man kan få inkludert spørsmål om hatefulle ytringer i undersøkelser med god representativitet, vil det være mulig å få robuste mål på omfanget av hatefulle ytringer i ulike deler av befolkningen. Eksempler på robuste undersøkelser er levekårsundersøkelsene til SSB og Ungdata.

I den seneste levekårsundersøkelsen blant innvandrerbefolkningen i Norge ble det gjennomført intervju med mer enn 8000 innvandrere i alderen 16 til 74 år fra tolv av de største innvandringslandene i Norge: Polen, Bosnia-Hercegovina, Kosovo, Tyrkia, Irak, Iran, Afghanistan, Pakistan, Sri Lanka, Vietnam, Eritrea og Somalia. I tillegg ble det gjennomført intervju med 2000 norskfødte barn av innvandrere i alderen 16 til 39 år fra fire land: Tyrkia, Pakistan, Sri Lanka og Vietnam. Som beskrevet tidligere i rapporten er svartilbøyeligheten i levekårsundersøkelsene gode, blant annet fordi undersøkelsene gjennomføres både på telefon og som besøksintervju og med intervjuere som har relevante språkkunnskaper og ofte selv innvandrerbakgrunn fra de samme landene (Vrålstad & Wiggen 2017).

Ungdata er et kvalitetssikret og standardisert system for lokale spørreskjemaundersøkelser blant ungdom, som kommunene kan ta i bruk. Undersøkelsene gjennomføres blant elever på ungdomsskole og videregående skole (13–19 år) ved hjelp av nettskjema. Siden undersøkelsene gjennomføres i skoletiden, har de høy svarprosent. Spørreskjemaene inneholder en lang rekke relevante spørsmål om mobbing og trakassering, og det er også mulig å komme med innspill til endringer i eksisterende spørsmål og forslag til nye (se https://www.hioa.no/ungdata/Om-undersoekelsen/Hva-er-Ungdata).

Mens de nevnte undersøkelsene kan gi gode estimater over omfanget av hatefulle ytringer i ulike deler av befolkningen, er det noen åpenbare utfordringer knyttet til det å koble seg på eksisterende undersøkelser. For det første kan man ikke regne med å få med veldig mange spørsmål. Dermed mister man mange av analysemulighetene man får ved å gjennomføre egne undersøkelser. For det andre har man ikke noen garanti for å få med tilsvarende spørsmål i fremtidige gjennomføringer av disse undersøkelsene. Dermed kan det bli vanskelig å følge opp resultatene.

Likevel kan det å få med spørsmål i eksisterende undersøkelser være svært nyttig som et *supplement* til egne undersøkelser. Ved å stille de samme spørsmålene i flere undersøkelser får man sikrere estimater over omfanget av hate-

fulle ytringer og metodisk kunnskap om hvordan omfanget varierer med undersøkelsesopplegg.

2.3 Vurdering av surveymetoder

Surveymetoden er i utgangspunktet svært fleksibel, og kvaliteten på resultatene man oppnår, avhenger av hvilke valg man tar, og hvilke budsjettrammer man har. På sitt beste kan en survey gi god og presis informasjon om befolkningens erfaringer med hatefulle ytringer. Men dersom spørsmål er dårlig formulert og utvalgene er lite representative, kan surveyer i verste fall ha begrenset verdi. Under diskuterer vi metoden i lys av vurderingskriteriene som ble skissert i innledningen.

Fange fenomenet: En survey fanger fenomenet hatefulle ytringer slik ytringene oppfattes av dem som svarer på undersøkelsen. Dette subjektive elementet gjør det utfordrende å måle hatefulle ytringer, fordi begrepet ikke er allment kjent, og fordi en ytring kan vurderes ulikt av ulike personer. Vi har diskutert to tilnærminger som har blitt brukt i tidligere forskning, som begge kan redusere målefeil knyttet til subjektive vurderinger: spørre om ulike konkrete erfaringer snarere enn abstrakte begreper og å spørre om hvilke(t) grunnlag ytringene er rettet mot. Samtidig kan det være viktig å fange nettopp subjektive erfaringer med hatefulle ytringer. For eksempel kan de opplevde konsekvensene av hatefulle ytringer være nærmere knyttet til en subjektiv forståelse av om man har vært utsatt for hatytringer, enn til en mer objektiv forståelse av begrepet definert av forskere.

Utover dette er imidlertid surveymetoden i seg selv også en velegnet tilnærming til å studere metodisk hvordan ulike spørsmålsformuleringer fungerer. Man kan stille respondentene flere spørsmål om ulike typer ytringer, og ved hjelp av surveyeksperimenter – der utvalget deles tilfeldig i flere deler som hver får ulike spørsmål – kan man eksperimentere med ulike spørsmålsformulinger. På denne måten kan man undersøke hvilken effekt ulike definisjoner og spørsmålsformuleringer har på hvilket omfang man finner av hatefulle ytringer og andre tilgrensende fenomener.

Ulike grunnlag: Surveymetoden gir gode muligheter til å fange opp hatefulle ytringer rettet mot ulike grunnlag, og derigjennom sammenligne erfaringer i ulike befolkningsgrupper. Dette kan gjøres på to måter. For det første får man informasjon om innholdet i ytringene ved å spørre om hvilke grunnlag ytringene er rettet mot. For det andre kan man gjennomføre undersøkelser i ulike utvalg

av ulike befolkningsgrupper, som dermed kan sammenlignes. I sum gjør dette at surveymetoden er godt egnet til å studere interseksjonelle perspektiver, ved at det er mulig å sammenligne ulike grunnlag både med hensyn til innholdet i ytringene og med hensyn til respondentenes tilhørighet i ulike (minoritets) grupper.

Ulike arenaer: I utgangspunktet er det i en survey enkelt å spørre respondentene hvor de har erfart å motta eller observere hatefulle ytringer. Derfor er denne metodiske tilnærmingen velegnet til å studere omfanget av hatefulle ytringer på ulike arenaer. En innvending kan være at metoden, som diskutert ovenfor, baserer seg på respondentenes subjektive erfaringer. Informasjonen om arenaer vil derfor være avhengig av hvordan de tolker og oppfatter spørsmålet. Men gitt at dette er klart definert og formulert i undersøkelsen, bør det ikke være et stort problem.

Representativitet: Store befolkningsrepresentative utvalgsundersøkelser basert på sannsynlighetsutvalg er velegnet til å få informasjon om omfanget av hatefulle ytringer. Men som diskusjonen ovenfor har vist, kan dette ofte være utfordrende i praksis.

Dette skyldes for det første at populasjonen av aktuelle minoritetsgrupper er vanskelig å definere. Det finnes få tilgjengelige registre over ulike minoritetsgrupper, noe som innebærer at man enten må gjennomføre svært store befolkningsstudier, eller at man må trekke utvalg basert på medlemskap i organisasjoner, menigheter eller annet. Med manglende informasjon om populasjonen vi studerer, er det vanskelig å vite om et utvalg i en undersøkelse er representativt for hele gruppen.

For det andre er det i økende grad et generelt problem med lavere svartilbøyelighet i spørreundersøkelser. Blant enkelte minoritetsgrupper er svartilbøyeligheten enda lavere enn i befolkningen for øvrig. Dersom frafallet er systematisk – at enkelte grupper har mindre sannsynlighet enn andre for å svare, er det utfordrende å vite om utvalget er representativt.

Mulighet til å innhente annen relevant informasjon: Surveymetoden er velegnet til å innhente annen relevant informasjon utover kun omfanget av hatefulle ytringer. I kapitlet har vi diskutert spørsmål knyttet til det å observere hatefulle ytringer, konsekvenser av hatefulle ytringer og muligheten til å studere avsendere av hatefulle ytringer. I praksis er det bare lengdebegrensningen for undersøkelsen som setter grenser for hvor mye man kan spørre om.

Tidsserier og komparasjon: Spørreundersøkelser kan gjentas over tid og på tvers av land og egner seg dermed godt som grunnlag for tidsserier og komparasjon. Men det er viktig å være oppmerksom på at begreper og fenomener endrer seg over tid, og at begreper ikke nødvendigvis er direkte overførbare på tvers av land. Mens hatefulle ytringer ikke er et allment avklart begrep i Norge, kan vi anta at det er mer avklart i for eksempel USA, hvor *hate speech* er tydelig koblet til rasisme (Bleich 2011).

Mulighetene for tidsserier og komparasjon er også påvirket av hva slags utvalg man bruker. Mens registerbaserte befolkningsundersøkelser er velegnet til disse formålene, kan organisasjonsutvalg ha begrensninger knyttet til endringer i organisasjonslandskapet over tid og ulikt organisasjonslandskap på tvers av land.

Kostnadseffektivitet: Kostnader knyttet til en survey er svært variable og avhenger særlig av gjennomføringsmetode. For eksempel vil undersøkelser gjennomført på nett være langt rimeligere enn undersøkelser gjennomført over telefon, fordi sistnevnte krever mer arbeidskraft (intervjuere) og tellerskritt. Undersøkelser gjennomført med nettpaneler eller gjennom organisasjoner er trolig de mest kostnadseffektive tilnærmingene.

3 Analyser av meningsinnhold

En annen tilnærming enn å spørre folk om deres erfaringer med hatytringer er å undersøke meningsinnholdet i ytringer i offentligheten, enten på internett, i tradisjonelle medier eller på andre offentlige arenaer. Da blir spørsmålet: Hva er omfanget av hatefulle ytringer innenfor en gitt del av offentligheten? For å kunne svare på et slikt spørsmål må man avgrense universet av innhold (populasjonen), det vil si hvilke deler av hvilke offentlige arenaer man studerer.

Det finnes tre hovedtilnærminger til å analysere meningsinnhold:

1) kvalitative tilnærminger, som studerer en avgrenset mengde ytringer for å undersøke for eksempel hvordan hatretorikk kommer til uttrykk, eller hvilken funksjon den har i en gitt sammenheng

2) kvantitativ innholdsanalyse, hvor man manuelt koder et mer omfattende innhold etter forhåndsdefinerte variabler – for eksempel ved at man koder innhold som hatefullt eller ikke

3) automatisert analyse av stordata hvor man analyserer store datamengder ved hjelp av programmer som læres opp til å gjenkjenne hatefulle ytringer

I dette kapitlet referer vi kort til forskning som har brukt kvalitative tilnærminger til innholdsanalyse, men vi legger mer vekt på de kvantitative tilnærmingene siden disse er bedre egnet til å studere omfang av hatefulle ytringer. Vi beskriver stordataanalyse som metode i mer detalj siden dette er en relativt ny og ukjent metodisk tilnærming til å studere omfanget av hatefulle ytringer.

3.1 Analyser av hatsider

Det finnes en rekke forskningsbidrag, spesielt fra USA, som har basert seg på *tekstanalyser* eller *kvalitativ innholdsanalyse* av rasistiske nettsider (for eksempel Douglas 2007; Douglas, McGarty, Bliuc & Lala 2005; Duffy 2003; Erjavec & Kovačič 2012; Gerstenfeld, Grant & Chiang 2003; McNamee, Peterson & Peña 2010; Meddaugh & Kay 2009). Disse studiene undersøker blant annet hvilken funksjon hatsidene har, hvilke strategier gruppene bruker for å kommunisere synspunktene sine, og hvordan de uttrykker «hatet» sitt. Stu-

diene finner at slike nettsider sjelden oppfordrer direkte til vold eller hat. I stedet prøver de å kommunisere budskapet sitt ved hjelp av overbevisende argumentasjon som bygger opp under hvit dominans (Douglas mfl. 2005; Gerstenfeld mfl. 2003).

Mens denne typen analyser er svært nyttige for å forstå hvordan hatgrupper opererer og kommuniserer, og for å studere innholdet i hatefull retorikk, er de mindre egnet til å si noe om *omfanget* av hatefulle ytringer. Som Rohlfing (2014: 298) påpeker, er dette «sluttprodukttilnærminger». De fokuserer på personer og nettsider som allerede er engasjert i hatefull retorikk, og gir dermed ikke et bilde av hvor vanlig slik retorikk er mer generelt. Skal vi måle omfanget av hatefulle ytringer, er det mer relevant å studere i hvilken grad vi finner slike ytringer mer generelt på internett.

3.2 Kvantitativ manuell innholdsanalyse

I tillegg til de mer kvalitative tilnærmingene som er beskrevet ovenfor, kan innholdsanalyse gjøres mer systematisk på større datamengder, i form av en kvantitativ innholdsanalyse. Metoden er utviklet og mye brukt innenfor medieforskning og brukes ofte til å analysere nettinnhold og sosiale medier for å kartlegge budskapene som kommuniseres (Krippendorff 2012). Institut for Menneskerettigheder i Danmark har nylig brukt denne metodikken for å studere omfanget og karakteren av hatefulle ytringer i den danske offentlige nettdebatten (Institut for Menneskerettigheder 2017), og det norske Likestillings- og diskrimineringsombudet (LDO) er i gang med en tilsvarende undersøkelse i Norge. Vi ser nærmere på fremgangsmåten i disse prosjektene, slik den er beskrevet i rapporten fra det danske prosjektet (Institut for Menneskerettigheder 2017) og i intervjuer med prosjektansvarlig hos LDO.

3.2.1 Datainnsamling

Hvis man er interessert i å studere forekomsten av hatefulle ytringer på nettet, må først universet (populasjonen) av nettinnhold defineres. På samme måte som i surveybaserte utvalgsundersøkelser er også innholdsanalyser basert på et utvalg. Utvalget består igjen av «enheter», som studeres enkeltvis (Krippendorff 2012). Utvalgskriteriene må blant annet definere hva slags innhold som skal analyseres, i hvilken tidsperiode og så videre. Den danske undersøkelsen av hatefulle ytringer ser på kommentarfeltene på Facebook-sidene til to nyhetskanaler i en periode på fire måneder (Institut for Menneskerettigheder 2017). LDOs undersøkelse analyserer på samme måte kommentarfeltene på Face-

book-sidene til to nyhetskanaler, men i to ulike tidsperioder, en «vanlig» periode og en periode rundt stortingsvalget 2017.

Utvalgsstrategien i kvantitativ innholdsanalyse varierer mellom studier. I studiene som er referert over, ble kommentarene samlet inn manuelt etter et tilfeldighetsprinsipp. Dette innebar at koderne ikke skulle kode alle de publiserte kommentarene eller lete seg frem til hatefulle ytringer, men at de skulle følge en på forhånd bestemt strategi for hvilke kommentarer som skal velges ut til koding (for eksempel den femte kommentaren under en gitt nyhetssak), før de så på innholdet. I dette tilfellet var det viktig å samle kommentarer fra så mange forskjellige debatter som mulig og fra ulike plasseringer i debattene, for å få et mest mulig riktig bilde av omfanget og karakteren av hatefulle ytringer (Institut for Menneskerettigheder 2017: 34).

Videre har kommentarene blitt samlet inn tidligst 12 timer etter at de ble offentliggjort, slik at nyhetsmediene skulle ha mulighet til å redigere og eventuelt slette kommentarer som er i strid med deres retningslinjer. Det er viktig å merke seg at nettsidene til nyhetsmediene skiller seg fra andre, mer uredigerte nettsider og nærmer seg et redigert medium på linje med nettaviser og så videre. LDOs erfaring er at man må samle inn kommentarer relativt raskt etter at sakene legges ut, siden nyhetsmediene sletter gamle saker. I tillegg vil Facebook ha kunnet redigere kommentarfeltene på bakgrunn av sine retningslinjer. Med tilnærmingen som er lagt til grunn i undersøkelsene til Institut for Menneskerettigheder og LDO, måler de på den ene siden omfanget og karakteren av hatefulle ytringer som brukerne møter på Facebook-sidene til nyhetskanalene etter redigering, og på den andre siden omfanget og karakteren av hatefulle ytringer som nyhetskanalene aksepterer på Facebook-sidene sine.

3.2.2 Identifisering av hatefulle ytringer

Kvantitativ innholdsanalyse innebærer at enheter av tekst, symboler, bilder eller lignende manuelt kodes etter forhåndsdefinerte variabler. Hvor store enheter som skal kodes, er opp til forskerne. Analyseenheten kan være alt fra en hel nyhetssak til et enkelt sitat. Enhver studie krever en klar definisjon og diskusjon av hva slags enheter som skal danne grunnlaget for analyse. I en studie av hatefulle ytringer i kommentarfelt vil det for eksempel være naturlig at enheten er (deler av eller hele) kommentarer.

De ulike kodene som skal brukes på materialet, og spesifikasjonen av i hvilke tilfeller de skal brukes, beskrives i en kodebok, og dette er et viktig instrument for å sikre felles forståelse og etterprøvbarhet. En sentral variabel å kode i en

studie av omfang av hatefulle ytringer er hvorvidt en kommentar kan regnes for å være hatefull eller ikke. Et stort begrep som hatefulle ytringer må få en klar operasjonell definisjon, slik at de som koder materialet, har klare kriterier for å vurdere hvorvidt en enhet skal kodes som hatefull eller ikke, og slik at ulike kodere ville kodet samme materiale på samme måte. Når alle enhetene i materialet er kodet, kan man regne ut andelen hatefulle ytringer i det gitte innholdet.

I tillegg er det mulig å registrere en lang rekke andre relevante variabler, som for eksempel hvilken form ytringen har (bruk av symboler/bilder), hvilket grunnlag ytringene retter seg mot, hvor grove ytringene er, hvilke temaer ytringene opptrer i forbindelse med, og hva slags respons de fremkaller (se Institut for Menneskerettigheder 2017). Når det gjelder analyse av kommentarfelt på Facebook, hvor folk (presumptivt) opptrer under eget navn, er det også til en viss grad mulig å registrere kjennetegn ved avsenderne. Undersøkelsen til Institut for Menneskerettigheder (2017) registrerer for eksempel kjønn og antatt etnisk bakgrunn på avsenderne av kommentarene.

Siden koding av denne typen innhold er tidkrevende, er det vanlig å bruke studenter eller assistenter til kodearbeidet. Det kan også være at kodingen får høyere validitet ved at det ikke er forskeren selv som koder. Det er svært viktig at de som skal utføre kodingen, og de som skal bruke materialet i etterkant, har en omforent forståelse av hvordan materialet skal kodes, og hva de ulike kodene betyr.

LDO beskriver at de brukte flere opplæringsdager på å diskutere og kode innhold sammen med koderne for at de skulle få en omforent forståelse av hvordan ulikt innhold skulle kodes. Det er også vanlig å teste reliabiliteten i kodingen ved at en mindre del av materialet blir kodet av flere personer uavhengig av hverandre for å se hvor stort avvik det er mellom de ulike koderne (interkoderreliabilitet) (Krippendorff 2012). I tillegg kan koderne føre logg over beslutninger de tar underveis, for å øke transparensen i hvordan materialet har blitt kodet.

En erfaring LDO formidlet fra arbeidet sitt, er at innholdet de analyserte, blant annet inneholdt mye bruk av ironi og lite bruk av direkte skjellsord. Det betyr at koderen må vurdere den antatte intensjonen i ytringen i den konteksten den har blitt fremsatt, før det er mulig å avgjøre om ytringen kan defineres som hatefull eller ikke.

3.2.3 Forskningsetiske utfordringer

Innholdsanalyse av for eksempel kommentartråder på Facebook innebærer i praksis at man lagrer en skjermdump av den aktuelle kommentaren, slik at det skal være mulig å gå tilbake og kvalitetssikre kodingen i ettertid. Ifølge personopplysningsloven krever elektronisk lagring av personlig informasjon (også når denne informasjonen har blitt offentliggjort) tillatelse fra hver enkelt person. Ifølge loven er personopplysninger en opplysning eller vurdering som kan knyttes til et individ som enkeltperson. Dette er en klar begrensing på mulighetene for å gjøre innholdsanalyse, fordi det vil være umulig å sikre samtykke fra alle som har ytret seg i debattene man ønsker å analysere.

Flere forskningsmiljøer beskriver at dette er et uavklart forskningsetisk spørsmål, hvor Datatilsynet og personvernombudet for forskning (Norsk samfunnsvitenskapelig datatjeneste) fremstår som svært restriktive.[10] LDOs erfaring var at de fikk tillatelse fra Datatilsynet til å lagre skjermdumper av kommentarer, gitt at navn og bilde på avsender ble sladdet. Skjermdumpene kunne kjøres gjennom et program før lagring, som sikret tilstrekkelig anonymisering til at de fikk tillatelse til å gjennomføre prosjektet.

Det er flere momenter som vil være relevante for å vurdere i hvilken grad informasjonsplikten gjelder for et bestemt forskningsprosjekt basert på kvantitativ innholdsanalyse, blant annet: grad av offentlighet, bruk av individdata, anonymisering, vanskeligheter med å informere og programvaren som brukes. Se for øvrig diskusjonen i punkt 3.3.3. om forskningsetiske utfordringer med stordataanalyse.

3.2.4 Vurdering av kvantitativ manuell innholdsanalyse

Nedenfor følger en oppsummering av vurderingen av kvantitativ manuell innholdsanalyse som metodisk tilnærming for å måle omfanget av hatefulle ytringer.

Fange fenomenet: I kvantitativ innholdsanalyse gjøres vurderingen av hvorvidt en ytring er hatefull eller ikke, på en stringent måte hvor vurderingskriteriene er klargjort på forhånd. I den forstand kan vi si at metoden baserer seg på et «objektivt» mål på hatefulle ytringer, til forskjell fra surveyundersøkelser hvor det i praksis er opp til den enkelte respondent å avgjøre hva som faller inn under fenomenet hatefulle ytringer. Selv om vurderingene i siste instans er basert på

10 Se for eksempel saken «Når etikken stopper forskningen» i Morgenbladet 7. august 2015: https://morgenbladet.no/2015/08/nar-etikken-stopper-forskningen

tolkning også i innholdsanalysen, kan man redusere innslaget av skjønn ved å la flere personer kode samme innhold uavhengig av hverandre.

Ulike grunnlag og annen type informasjon som kan samles inn: Kvantitativ innholdsanalyse åpner for å måle en rekke relevante aspekter ved hatefulle ytringer, inkludert hvem og hvilket grunnlag ytringene retter seg mot. Metoden gjør det mulig med et interseksjonelt perspektiv ved at det er mulig å fange opp ytringer som retter seg mot sammensatte identiteter (for eksempel homofil muslim). Det er også mulig å studere i hvilken kontekst ytringen falt (for eksempel hva slags debatter som frembringer hatefulle responser), og i den grad det er identifiserbare avsendere, er det også mulig å registrere informasjon om avsendere av hatefulle ytringer.

Ulike arenaer: Metoden er egnet til å studere hatefulle ytringer på internett eller i tradisjonelle medier (TV, radio, avis). Det er – mer eller mindre – offentlig tilgjengelige *medieuttrykk* (tekst, symboler, bilder, etc.) som er grunnlaget for analysene, så dermed vil denne metodiske tilnærmingen naturlig nok ikke fange opp hatefulle ytringer som sendes direkte til mottaker, eller som ytres ansikt til ansikt.

Representativitet, tidsserier og komparasjon: Metoden kan ikke si noe om omfanget av hatefulle ytringer på internett eller andre arenaer generelt. Men hvis tekstbitene som skal analyseres, blir valgt ut på en stringent og gjennomtenkt måte, kan denne metodiske tilnærmingen gi et representativt bilde av omfanget av hatefulle ytringer på en bestemt arena i en bestemt tidsperiode. Ved å studere debattene på de største offentlig tilgjengelige medienettstedene kan man få et godt bilde av innslaget av hatefulle ytringer i samfunnsaktuelle debatter.

Det er mulig å designe analyseopplegget på en måte som åpner for komparasjon og å etablere tidsserier. For eksempel er den nevnte undersøkelsen til LDO i stor grad bygget på en dansk undersøkelse, noe som vil gi muligheter for å sammenligne de to nasjonale kontekstene. Imidlertid kan det være en utfordring med tilgang til historiske data. Dermed må eventuelle tidsserier basere seg på å sammenligne en serie med undersøkelser som studerer omfanget av hatefulle ytringer mer eller mindre i sanntid.

Kostnadseffektivitet: Manuell kvantitativ innholdsanalyse kan være kostnadskrevende, avhengig av hvor stort materiale man ønsker å studere, og hvor kompleks analysen er. Det krever manuell koding av store mengder data, noe som er både tid- og ressurskrevende. LDO koder i sitt prosjekt 4000 kommen-

tarer og anslår at koderne i snitt bruker 3 minutter på å kode hver kommentar. Det vil si at de anslår 200 timer til kodearbeidet. Dette innebærer registrering av en rekke ulike forhold utover kun hvorvidt en kommentar er hatefull eller ikke, men det gir en pekepinn på hvilken ressursbruk som ligger til grunn.

3.3 Stordataanalyse

Manuell innholdsanalyse kan som nevnt være svært arbeids- og kostnadskrevende, og vi ønsker derfor å utrede hvilke muligheter og begrensinger som ligger i automatisert analyse av stordata (*Big Data*). Mengden av tilgjengelige digitale data har eksplodert de siste årene. Det dreier seg om hverdagslige statusoppdateringer på Facebook, videoer lagt ut på YouTube og Twitter-meldinger som er tilgjengelige for alle som vil lese dem. Det handler også om data fra kjøpstransaksjoner, søkemotorer og andre digitaliserte transaksjoner i offentlig sektor, helsevesenet, skoleverket og så videre. Begrepet stordata er en samlebetegnelse for data som er av et slikt omfang at det krever mer enn vanlig datakraft å samle inn, lagre og analysere dem. Begrepet brukes ofte ikke bare for å betegne selve dataene, men også for å beskrive de nye problemstillingene slike data reiser, både teknisk, juridisk og etisk. Felles for stordata er at de innebærer en registrering av faktiske handlinger, interaksjoner og transaksjoner koblet til individer.

Vi vil vurdere i hvilken grad en slik tilnærming egner seg til å identifisere hatefulle ytringer på internett, herunder meningsinnholdet i ytringene og hvilket grunnlag de retter seg mot. I tillegg vil vi undersøke om denne metoden også kan innhente kjennetegn ved avsenderne. Vi gir først en oversikt over metoder for å samle inn data fra sosiale medier og for å analysere slike data ved hjelp av maskinlæringsmetoder. Deretter oppsummerer vi den eksisterende litteraturen om bruk av stordata for å studere hatefulle ytringer, før vi trekker noen metodiske konklusjoner.

3.3.1 Generelt om innsamling av data fra sosiale medier og maskinlæring for tekstklassifisering

En stordatatilnærming innebærer at man samler inn for eksempel innlegg i sosiale medier som er relatert til et sett med søkeord som kan forekomme som del av hatefulle ytringer. Ved å først kode deler av innholdet manuelt kan man «lære» programmet til å bedre kunne kategorisere innhold som hatefullt eller ikke. Denne prosessen kan så gjentas, slik at maskinlæringsalgoritmene som automatisk identifiserer hatefulle ytringer, blir bedre.

Metoder for å samle inn data fra sosiale medier

I dag finnes det et mangfold av kilder til digitale data. I denne gjennomgangen begrenser vi oss til data fra sosiale medier. Sammenlignet med tradisjonelle medier, hvor brukerne er forbrukere, gjør sosiale medier det mulig for brukere å både forbruke og publisere informasjon. Man kan få tilgang til digitale data om interaksjoner i digitale nettverk (poster, delt innhold, kommentarer, og «likes») gjennom sosiale medier-plattformers *API* (*Application Programming Interface*). En API er et sett med prosedyrer og protokoller som gjør det mulig å interagere med plattformen og få tilgang til dataene ved hjelp av spesielle programmer. Både Twitter og Facebook har slike API-er som gjør det mulig å interagere med plattformen deres ved hjelp av applikasjoner (dataprogrammer) som er utviklet til dette formålet.

Twitter er en mikroblogg hvor brukere deler korte budskap, kalt «tweets». Twitter har 240 millioner brukere som publiserer 500 millioner tweets daglig. Twitter tilbyr flere API-er[11] som gir tilgang til Twitter-data gjennom programmerte interaksjoner: REST API-er og Streaming API.

REST API-er gjør det mulig å gå tilbake i tid og søke i eksisterende tweets. Dataene man kan få tilgang til gjennom REST API-er, er begrenset med hensyn til antall tweets man kan forespørre i minuttet, samt med hensyn til hvor langt man kan gå tilbake i tid (ca. en uke). I tillegg er det ingen garanti for at alle tweets er tilgjengelige for søk.

Streaming API gjør det mulig å samle tweetene som matcher et sett med søkeord, i sanntid, det vil si samtidig som de blir publisert. Ulempen med Streaming API er at den kun gir tilgang til én prosent av tweetene som publiseres på Twitter på et gitt tidspunkt. Hvis resultatet av søket i utgangspunktet utgjør mer enn én prosent av alle tweets på Twitter, får man tilgang til et utvalg av søkeresultatene (tilsvarende én prosent av alle tweetene), men det er ingen garanti for at utvalget er representativt for tweetene som passer til søket. Et alternativ er å kjøpe Twitter Firehose, som gir tilgang til 100 prosent av tweetene. Tilgangen til Twitter Firehose er imidlertid dyrt og begrenset til kommersielle aktører som selger Twitter-data.

Facebook har 1,5 milliarder aktive brukere. Facebook tilbyr Graph API[12], som gir tilgang til Facebook-data. I motsetning til Twitter, hvor alle postene er offentlige og tilgjengelige gjennom API-er, er tilgang til data knyttet til private brukere på Facebook begrenset for tredjeparter gjennom Graph API. For å få

11 https://dev.twitter.com/overview/api
12 https://developers.facebook.com/docs/graph-api

tilgang til data fra Facebook må man ha registrert en applikasjon som gir tilgang til data fra private brukere, men en slik applikasjon vil kun ha tilgang til data hvis brukeren har gitt tillatelse til den bestemte applikasjonen. Data fra offentlige sider på Facebook er imidlertid tilgjengelige gjennom Graph API. Noen individer, men særlig organisasjoner, selskaper og noen løst organiserte grupper har slike offentlige sider, som det altså er mulig å hente ut data fra.

Gjennom ulike Twitter-API-er og Facebook Graph API er det dermed mulig å samle svært store mengder ytringer publisert på disse plattformene som grunnlag for å analysere omfanget av, innholdet i og grunnlaget for hatefulle ytringer. Likevel er dette datagrunnlaget begrenset siden data fra private Facebook-kontoer for det meste ikke er tilgjengelig, og fordi data fra Twitter-API-er ikke nødvendigvis er representative.

Identifisering av hatefulle ytringer: bruk av maskinlæring og tekstklassifisering

For å analysere omfanget av, innholdet i og grunnlaget for hatefulle ytringer i sosiale medier er det mulig å bruke maskinlæringsalgoritmer som på en automatisert måte søker gjennom store mengder tekstdata og analyserer disse på grunnlag av bestemte kjennetegn. Maskinlæring gjør det mulig å bruke statistiske dataanalytiske metoder på tekst. Fordelen med maskinlæring er at når en maskinlæringsmodell først er «trent» på et sett med data, kan den brukes på et nytt sett med data. Med andre ord: Hvis man trener en modell som gjenkjenner hatefulle ytringer, på et sett med ytringer, kan modellen senere brukes til å klassifisere flere fremtidige ytringer uten videre trening.

Tekstklassifisering, et av mange bruksområder for maskinlæring, er godt tilpasset formålet om å identifisere hatytringer i data fra sosiale medier. Denne teknikken innebærer å organisere tekstdokumenter i ulike kategorier basert på bestemte kjennetegn. Det er vanlig å skille mellom maskinlæringsteknikker som benytter veiledet (*supervised*) og ikke-veiledet (*un-supervised*) læring. Med veiledet læring har algoritmen blitt trent opp i hvilke kjennetegn den skal se etter for å kategorisere materialet i ulike klasser gjennom et forhåndskodet treningssett. Det vil si at man først manuelt koder en mengde materiale ut fra om det er hatefulle ytringer eller ikke. Deretter bruker algoritmen kodingen som er gjort, til å lære hvordan den skal kjenne igjen hatefulle ytringer. Ikke-veiledet læring viser til teknikker som ikke trenger forhåndskodet treningssett for å hente ut meningsfulle mønstre fra dataene. Det er stort sett veiledet læring som brukes til å gjenkjenne hatytringer i sosiale medier.

Utvikling av et automatisert klassifikasjonssystem innebærer følgende trinn:

– Forberedelse av et trenings- og testdatasett: Trenings- og testdatasett består begge av tekster (poster fra sosiale medier) som er manuelt kodet langs kategoriene som er hensiktsmessige for formålet. For å trene algoritmen trenges det mellom 2000 og 3000 manuelt kodede eksempler på både hatefulle og ikke-hatefulle ytringer. Når det gjelder hatefulle ytringer, vil disse kategoriene for eksempel være knyttet til om en ytring er hatefull (ja eller nei), hvilke grunnlag ytringen er rettet mot, og så videre. Treningssettet brukes for å lære opp algoritmen, mens testsettet brukes for å evaluere prediksjonene algoritmen gjør (dvs. i hvilken grad den feilklassifiserer innleggene sammenlignet med resultatet av den manuelle klassifiseringen).

– Dokumentene gjøres klare til analyse gjennom såkalt tekstprosessering, hvor teksten «ryddes» gjennom standardisering og lemmatisering (erstatter den bøyde varianten av et ord med oppslagsformen – for eksempel «innvandrerne» med «innvandrer»), fjerning av spesielle tegn og symboler, fjerning av stoppeord og lignende.

– Tekstuelle data gjøres om til numeriske attributter (*features*), som gir en numerisk representasjon av teksten (ekstraksjon av attributter). Maskinlæringsalgoritmen behandler numeriske størrelser. Dermed er omgjøringen av de tekstuelle dataene til en numerisk representasjon en avgjørende fase i maskinlæringsprosessen. Tre hovedtyper av tekstrepresentasjoner blir brukt som input til maskinlæringsalgoritmer: Bag of Word, TF-IDF og vektoriell representasjon.[13] Mye av forskningen innen tekstuell maskinlæring har bestått i å utforske hvilken numerisk representasjon av tekstuelle data som gir best resultater.

– Valg av algoritme og modelltrening: Ved bruk av veiledede algoritmer er det behov for treningsdata, det vil si et utvalg tekstdata som er manuelt klassifisert (kodet) og danner grunnlaget for læring (estimering av parametere i modellen). Det finnes en rekke maskinlæringsalgoritmer, men når det gjelder

13 Med *Bag-of-Word* blir tekstdokumenter konvertert til vektorer, slik at hvert dokument er en vektor som representerer frekvensen av alle forskjellige ord i dokumentet. *TF-IDF* (Term Frequency-Inverse Document Frequency) er en kombinasjon av to elementer: termfrekvenser (tf) og omvendt dokumentfrekvens (idf). Termfrekvens er frekvensen av hvert ord i et gitt dokument, mens omvendt dokumentfrekvens (idf) regnes ut ved å dele det totale antall dokumenter med dokumentfrekvensen for hvert ord. Fordelen med sistnevnte er at den i større grad vektlegger ord som ikke forekommer ofte i alle dokumentene, men som kan være et viktig kjennetegn i et dokument. Det finnes flere tilnærminger til å skape en mer avansert *vektoriell representasjon* (eller *words embeddings*) av tekstdata. Den mest populære er word2vec-modellen, som ble utviklet av Google og gjort offentlig tilgjengelig i 2013. Word-2vec er en vektoriell ordrepresentasjon som er generert ved bruk av dype læringsalgoritmer (nevrale nettverk) som er trent for å rekonstruere ords språklige kontekster.

tekstklassifisering, er Multinomial Naive Bayes og Support Vector Machines blant de mest effektive og mest brukte. Tilbakevendende nevrale nettverk (Recurrent Neural Networks) har også blitt brukt til dette formålet.

– Modellprediksjon og evaluering: Evaluering av en algoritme består i å vurdere i hvilken grad den lykkes i å predikere de ulike kategoriene den har blitt trent til å gjenkjenne, når den brukes på nye data. Resultater fra en klassifiseringsalgoritme som klassifiserer tekster i to kategorier – positiv (P) og negativ (N) – kan representeres med følgende matrise, som viser hvordan den predikerte klassifiseringen forholder seg til den faktiske klassifiseringen:

	Predikert positiv (P')	**Predikert negativ (N')**
Faktisk positiv (P)	Sanne positive (SP)	Falske negative (FN)
Faktisk negativ (N)	Falske positive (FP)	Sanne negative (SN)

For å evaluere en maskinlæringsalgoritme er det vanlig å bruke følgende mål: nøyaktighet (*accuracy*), presisjon (*precision*), tilbakekalling (*recall*) og F1-score.[14]

3.3.2 Automatisk gjenkjenning av hatefulle ytringer: metodiske konklusjoner fra tidligere studier

I takt med den stadig økende mengden innhold på sosiale medier øker også omfanget av hatefulle ytringer. De senere årene har flere forskningsbidrag testet ulike metoder for automatisk gjenkjenning av hatefulle ytringer, hovedsakelig brukt på engelskspråklige data fra sosiale medier. Forskningen på dette feltet tar sikte på å ta i bruk og teste ulike metoder for naturlig språkbehandling (se Manning & Schütze 1999) til automatisk gjenkjenning av hatefulle ytringer. Her gir vi en kort oversikt over denne litteraturen for å trekke noen metodiske konklusjoner av den eksisterende forskningen.

14 *Nøyaktighet* måler andelen av korrekte (sanne) prediksjoner, det vil si hvor nøyaktig algoritmen gjenkjenner en tekst som faktisk tilhører (sanne positive) eller ikke tilhører (sanne negative) en bestemt kategori (det som tilsvarer *SP + SN / SP + FP + FN + SN*). *Presisjon* måler andelen av korrekte prediksjoner, det vil si algoritmens evne til å predikere hvorvidt en tekst tilhører en bestemt kategori (det som tilsvarer *SP / SP + FP*). *Tilbakekalling* (også kjent som *hit rate*) måler andelen av den positive klassen som har blitt predikert korrekt (det som tilsvarer *SP / SP + FN*). *F1-score* er det harmoniske gjennomsnittet av presisjon og tilbakekalling (2 x presisjon x tilbakekalling / presisjon + tilbakekalling).

Hatefulle ytringer: begrep og grunnlag

Begrepet «hatefulle ytringer» er mest brukt i denne typen litteratur for å betegne spredning i sosiale medier av fornærmende (*insulting)* brukergenerert innhold (Burnap & Williams 2015; Djuric, Zhou, Morris, Grbovic, Radosavljevic & Bhamidipati 2015; Gitari, Zuping, Damien & Long 2015; Kwok & Wang 2013; Silva, Mondal, Correa, Benevenuto & Weber 2016; Warner & Hirschberg 2012; Williams & Burnap 2015). Andre betegnelser som «voldelige meldinger» (Spertus 1997) eller nettmobbing (*cyberbullying*) (Dadvar, Trieschnigg, Ordelman & de Jong 2013; Dinakar, Jones, Havasi, Lieberman & Picard 2012; Hosseinmardi, Mattson, Rafiq, Han, Lv & Mishra 2015; Van Hee, Lefever, Verhoeven, Mennes, Desmet, De Pauw, Daelemans & Hoste 2015; Xu, Jun, Zhu & Bellmore 2012; Zhong, Li, Squicciarini, Rajtmajer, Griffin, Miller & Caragea 2016) har også blitt brukt.

Denne litteraturen har sett nærmere på en rekke grunnlag for hatefulle ytringer: etnisitet og kjønn (Badjatiya, Gupta, Gupta & Varma 2017), etnisitet og religion (Gitari mfl. 2015), religion, etnisitet og seksualitet (Burnap & Williams 2015, 2016), rasisme (Bartlett, Reffin, Rumball & Williamson 2014; Greevy & Smeaton 2004; Magu, Joshi & Luo 2017), etnisitet, religion, funksjonshemming, seksualitet og kjønn (Mondal, Silva & Benevenuto 2017; Silva mfl. 2016; Warner & Hirschberg 2012).

Gjenkjenning av hatefulle ytringer

Automatisk gjenkjenning av hatefulle ytringer i sosiale medier er basert på bruk av klassifikasjonsalgoritmer. Et avgjørende moment i slike klassifikasjonsoppgaver er å velge ut språklige kjennetegn som vil danne grunnlaget for klassifikasjon. Automatisk klassifikasjon av hatefulle ytringer er en vanskelig oppgave (Nobata, Tetreault, Thomas, Mehdad & Chang 2016) blant annet fordi (i) avsendere av slike ytringer ofte skriver om ord med vilje (som for eksempel ved å skrive «ni99er» i stedet for «nigger») eller bruker et kodet språk for å unngå moderering (som når «Googles» er brukt for å betegne afroamerikaner og «Bings» for asiat, jf. Magu mfl. 2017), og (ii) i sosiale medier brukes ofte sarkasme, som kan se ut som hatefulle ytringer.

I forskningen er det brukt ulike metoder og strategier innen naturlig språkbehandling (Manning & Schütze 1999), som legger vekt på ulike kjennetegn som grunnlaget for klassifisering (Schmidt & Wiegand 2017). Mange studier er basert på en Bag-of-Word-representasjon (unigram eller n-gram) av språket (Burnap & Williams 2015, 2016; Chen, Zhou, Zhu & Xu 2012; Hosseinmardi mfl. 2015; Nobata mfl. 2016; Sood, Churchill & Antin 2012; Van Hee mfl. 2015;

Warner & Hirschberg 2012; Waseem & Hovy 2016; Xu mfl. 2012). Denne typen attributter viser seg å være svært effektive for å predikere hatefulle ytringer.

Bruk av tegnbaserte n-gram (språkrepresentasjon på tegnnivå) kan dempe problemer knyttet til variasjoner i staving, som kjennetegner brukergenerert innhold (som for eksempel i følgende melding: «ki11 yrslef a$$hole»). Mehdad og Tetreault (2016) har systematisk sammenlignet tegnbaserte n-gram med ordbaserte n-gram for å gjenkjenne hatefulle ytringer og finner at tegnbaserte n-gram er mer effektive enn ordbaserte n-gram.

Nylig har distribuerte ordrepresentasjoner (som word2vec), også kjent som *word embeddings*, basert på nevrale nettverk blitt brukt til dette formålet. Imidlertid viser studier at denne metoden har begrenset evne til å oppdage hatefulle ytringer (Nobata mfl. 2016). Likevel viser en nyere studie (Badjatiya mfl. 2017) at en kombinasjon av nevrale nettverksalgoritmer og *words embeddings*-representasjon er mer effektiv enn tidligere brukte metoder.

Klassifiseringsmetoder

Klassifiseringsmetoder for automatisk gjenkjenning av hatefulle ytringer tilhører hovedsakelig overvåket maskinlæring. Den mest brukte algoritmen er «*Support Vector Machines*». Dyplæring med bruk av tilbakevendende nevrale nettverk har blitt brukt av Mehdad og Tetreault (2016) og Badjatiya mfl. (2017).

Data og koding

For å teste ulike metoder for automatisk gjenkjennelse av hatefulle ytringer har forskere samlet og kodet egne data fra sosiale medier. Datakilder som brukes i litteraturen, inkluderer: *Twitter* (Burnap, Rana, Avis, Williams, Housley, Edwards, Morgan & Sloan 2015; Burnap & Williams 2015; Burnap, Williams, Sloan, Rana, Housley, Edwards, Knight, Procter & Voss 2014; Silva mfl. 2016; Xiang, Fan, Wang, Hong & Rose 2012; Xu mfl. 2012), *Instagram* (Hosseinmardi mfl. 2015; Zhong mfl. 2016), *Yahoo!* (Djuric mfl. 2015; Nobata mfl. 2016; Warner & Hirschberg 2012), *YouTube* (Dinakar mfl. 2012), *ask.fm* (Van Hee mfl. 2015). Siden disse plattformene har ulike typer brukere og formål, er det sannsynlig at de gjenspeiler ulike typer og ulikt omfang av hatefulle ytringer.

Overvåket maskinlæring forutsetter at dataene er annotert eller kodet for å lære opp algoritmen og teste hvor presis den er. Det er knyttet to typer utfordringer til koding av data: For det første må man gjøre en avveining mellom å maksi-

mere antall hatefulle meldinger som skal kodes, og hvor mange grunnlag for hatefulle ytringer man ønsker å inkludere. For å få flest mulig meldinger som inneholder hatefulle ytringer, uten at datamengden blir uhåndterlig, er det en fordel å snevre inn søket i sosiale medier til bestemte temaer og typer hatefulle ytringer. Dette går da på bekostning av muligheten til å fange opp et bredt spekter av grunnlag for hatefulle ytringer.

For det andre er koding av hatefulle ytringer en krevende oppgave fordi det er vanskelig å oppnå enighet om hvilke ytringer som er hatefulle. Bartlett mfl. (2014) melder at selv etter grundig diskusjon var ulike (menneskelige) kodere ofte fortsatt uenige om meningen og hensikten med meldinger de skulle kode. Selv når koderne diskuterte uenigheter seg imellom, klarte de ikke å bli enige. Noen studier har tatt i bruk crowdsourcing-tjenester som *Amazon Mechanical Turk* (AMT) for å kode dataene. Nobata mfl. (2016) har sammenlignet koding gjennomført gjennom crowdsourcing med koding utført av eksperter og finner store ulikheter.

Andre stordatatilnærminger

Det finnes også andre tilnærminger til å studere stordata. Her vil vi kort diskutere bruk av *Google Trends* for å identifisere hatefulle ytringer. I boken *Everybody lies: Big Data, New Data, and what the Internet can tell us about who we really are* (2017), vier Seth Stephens-Davidowitz et kapittel til «sannheten om hat og fordommer». Analysene i boken er delvis basert på resultater fra Google Trends. Google Trends API gir en indeks over søkeaktivitet på Google per søkeord eller søkekategori – over tid eller etter geografisk område. I studien om rasisme bruker Stephens-Davidowitz begrepet «nigger» og relaterte ord som proxy for rasisme, og viser at graden av rasisme, estimert på denne måten, i ulike amerikanske delstater kan forklare forskjeller i stemmegivingen for Obama i presidentvalgene i 2008 og 2012.

Analysene som dreier seg om hat, er basert på data forfatteren har samlet ved å hente profilinformasjon om medlemmer av en høyreekstrem nettside, *Stormfront*, som er en kjent hatgruppe på nettet. Basert på denne informasjonen kan forfatteren si noe om både bakgrunnen til de som bruker nettsiden (alder og bosted), og grunnlaget for deres hatefulle ytringer.

Google Trends API kan være nyttig for å analysere søketrender på nettet og kan til en viss grad brukes som proxy for bestemte holdninger. Samtidig er det utfordrende å bruke søkeord som grunnlag for å studere holdninger eller fordommer. Det kan være ulike grunner til at noen har søkt på for eksempel ordet «nigger»,

og et søk trenger ikke å reflektere rasistiske holdninger. Man kan blant annet tenke seg at dette søkeordet også er relevant for forskere, journalister eller andre som er interessert i å følge med på hatefullt innhold på nettet. Dermed er bruken av bestemte søkeord dårlig egnet til å estimere omfang av hatefulle ytringer. Analyser basert på profilinformasjon, som er den andre metoden Stephens-Davidowitz bruker, er bedre egnet til å estimere omfanget av hatefulle ytringer. Men denne tilnærmingen reiser forskningsetiske utfordringer fordi den innebærer å samle personlige opplysninger uten samtykke fra dem det gjelder. En slik praksis vil etter all sannsynlighet ikke være tillatt i Norge.

Konklusjoner fra tidligere studier av hatefulle ytringer ved hjelp av stordataanalyse

Denne gjennomgangen viser at bruk av stordata og maskinlæringsmetoder er en farbar vei for å kunne gjenkjenne hatefulle ytringer i sosiale medier automatisk. Samtidig er disse metodene blitt utviklet på basis av engelskspråklige tekster, og så vidt vi vet, har de ennå ikke blitt forsøkt brukt på norskspråklig innhold. For å kunne bruke disse metodene på norskspråklig innhold vil man måtte gå gjennom tre faser: 1) datainnsamling, 2) forberedelse og koding av data og 3) trening og evaluering av modellen.

Hver av disse fasene har sine utfordringer og begrensninger. Datainnsamling er, av juridiske og tekniske grunner, begrenset til bestemte sosiale medier, som Twitter og offentlige sider på Facebook. Det er også mulig å samle data fra tjenester som Instagram og YouTube, men jo mer omfattende datainnsamlingen er, desto mer kostbar blir gjennomføringen. Å identifisere hatefulle ytringer er utfordrende, og det blir desto mer komplisert jo flere plattformer man tar med i denne fasen. Det samme gjelder hvis man kjører åpne søk uten å snevre inn søket til bestemte temaer, hendelser og/eller grunnlag for hatefulle ytringer.

Datakoding er en avgjørende fase som vil ha stor påvirkning på effektiviteten og presisjonen av klassifiseringsalgoritmen. Denne fasen er ressurskrevende fordi man skal kode en betydelig mengde data. Kodingsprosessen bør helst involvere flere kodere som jobber parallelt og uavhengig av hverandre, for å kunne teste interkoderreliabilitet – altså å teste om ulike kodere klassifiserer ytringene på noenlunde samme måte. I trenings- og evalueringsfasen er det ønskelig å sammenligne effektiviteten og presisjonen til ulike språkrepresentasjoner kombinert med ulike algoritmer.

Å bruke automatisk gjenkjenning av hatefulle ytringer på norske forhold innebærer en investering i utvikling av maskinlæringsmetoder til norsk språk samt

innsamling av norskspråklige data fra sosiale medier. Dette forutsetter støtte til et prosjekt av en viss skala og omfang hvis man vil oppnå meningsfulle resultater som kan bidra til bygge en permanent infrastruktur for å overvåke hatefulle ytringer i sosiale medier.

3.3.3 Forskningsetiske utfordringer

Digitalisering har ført til at skillet mellom det offentlige og det private forvitrer, noe som utfordrer personvernet. Samtidig er personvernsregelverket, særlig når det gjelder bruk av stordata og bruk av maskinlæringsteknologier til forskning, ikke tilpasset denne utviklingen. Bruk av maskinlæring til å gjenkjenne hatefulle ytringer reiser en del etiske problemstillinger som forskerne og myndighetene bør ta stilling til. Slik regelverket hittil har vært praktisert, legger det noen klare begrensninger på mulighetene til å bruke blant annet stordata i forskningsøyemed, men det kan argumenteres for at stordataanalyse av hatefulle ytringer kan gis unntak fra informasjonsplikt og kravet til innhenting av samtykke. Samtidig er regelverket i endring med innføring av *EU General Data Protection Regulation* (http://ec.europa.eu/justice/data-protection/). Vi vil her utdype noen av de forskningsetiske utfordringene bruken av stordata reiser.

Digitalisering og personvern

Stordata bidrar til å utfordre personvernet. De fleste selskapene som tilbyr nettbaserte tjenester (sosiale medieplattformer som Facebook eller Twitter, søkemotorer som Google eller Bing og dataprogramleverandører som Microsoft eller Apple), lagrer kontinuerlig data om hver enkelt brukers profil, sosiale «grafer» (nettverk av «venner» og følgere på sosiale medier) og nettrafikk. Disse databasene utgjør en enorm og rik mengde informasjon som kan analyseres ved hjelp av «datamining»-metoder. Det finnes teknikker for å lage personlige brukerprofiler, som i neste omgang kan brukes til å tilby målrettet reklame og markedsføring eller for å tilby produkter gjennom anbefalingssystemer (for eksempel Amazon.com). I kjølvannet av denne utviklingen har det oppstått en ny nisje for informasjonsmeglere som selger informasjon fra nettrafikk til en rekke private og offentlige aktører. Ved å bruke «metadata» for å koble sammen datakilder på individnivå gir stordata mulighet til å samle og analysere omfattende og detaljert informasjon om en persons liv, aktiviteter, preferanser og ytringer. Begrensede digitale spor som blir lagt enkeltvis på ulike nettjenester eller gjennom apparater som er koblet til internett, kan utfordre retten til privatliv når de blir aggregert og gjennomanalysert ved hjelp av stordatateknologier.

Datamining og maskinlæringsteknologier kombinert med stordata kan til sammen utgjøre en trussel mot ytringsfriheten og personvernet. Både regjeringer og private selskaper kan overvåke og analysere kommunikasjon som foregår digitalt. Aggregering av data på tvers av ulike brukerkontoer (for eksempel Google Gmail, YouTube, Chrome, Google+ osv.) øker muligheten for å samle omfattende mengder informasjon om en persons liv, vaner, preferanser, handlinger og meninger. Individuell kontroll over, og samfunnsregulering av, personlig informasjon er vanskelig å utøve ved hjelp av nasjonal lovgivning, fordi dataene som er tilgjengelig digitalt, i økende grad er kontrollert av globale selskaper og er i privat eie (Facebook, Google osv.) der brukerne har gitt fra seg rettighetene sine for å kunne bruke tjenestene.

Gjeldende begrensninger på bruk av stordata til forskning

Bruk av digitale data (sosiale medier og nettdata) for å identifisere hatefulle ytringer i forskningsøyemed vil måtte forholde seg til det gjeldende regelverket for lagring av digital personlig informasjon. Det er en utfordring at dette regelverket ikke er tilpasset den nye teknologiske virkeligheten, særlig når det gjelder forskning. Tilgangen til nye digitale data som kan brukes til samfunnsforskningsformål, har økt enormt, men dagens regelverk begrenser bruken av disse dataene til forskning, samtidig som dataene kan brukes fritt til kommersielle formål. Samfunnsforskningen i Norge utfordres av et regelverk som ble designet for «small data», og som ikke er tilpasset dagens teknologi, samtidig som den kommersielle nyttiggjøringen av stordata nesten er ubegrenset så lenge brukerne har samtykket til bruksvilkårene for kommersielle digitale tjenester som Facebook, Google og så videre.

I Norge, og i Europa for øvrig, er nettdata underlagt de samme retningslinjene som andre typer data. Med hjemmel i menneskerettighetskonvensjonen har personopplysningsloven til hensikt å verne privatpersoner mot krenkelser og mot bruk av bilder eller personopplysninger uten samtykke. Ifølge personopplysningsloven krever elektronisk lagring av personlig informasjon (også når denne informasjonen har blitt offentliggjort) tillatelse fra hver enkelt person. Ifølge loven er personopplysninger en opplysning eller vurdering som kan knyttes til et individ som enkeltperson. Når en virksomhet behandler personopplysninger, har forskerne informasjonsplikt overfor personene de skal innhente personopplysninger om, og hovedregelen er at dette i størst mulig grad skal være basert på et *samtykke*. Dersom en virksomhet behandler personopplysninger uten å ha innhentet samtykke, må den ha et annet *rettslig grunnlag* (som er tilfellet for eksempel når det gjelder skattlegging eller trygdeutbetaling). Imidlertid kan et forskningsprosjekt unntas fra informasjonsplikten og kravet om innhenting av

samtykke, avhengig av hvorvidt informasjonen blir hentet fra et åpent eller lukket forum/nettside, det er praktisk mulig å gi informasjon og innhente samtykke, opplysningene er sensitive, og hvorvidt opplysningene skal publiseres med eller uten personopplysninger.

Dette reiser spørsmål om hvilken status offentlig tilgjengelige data i sosiale medier, som for eksempel Twitter, skal ha i forskningssammenheng. Er det for eksempel rimelig at individene som blir forskningsobjekter (bloggere, Twitter-brukere osv.), skal måtte samtykke i at deres offentlig tilgjengelige data blir brukt i forskning? Gitt at innlegg på sosiale medier, som danner datagrunnlaget for analysene, blir innhentet fra åpne plattformer, og gitt at det vil være praktisk umulig å informere millioner av brukere om prosjektet, og at forskningen ikke vil bruke personopplysninger, men bare analysere og lagre postene fra sosiale medier, kan det tenkes at denne typen forskning vil kunne møte kriteriene for unntak fra informasjonsplikten og kravet til innhenting av samtykke.

I tillegg har brukere av sosiale medier allerede samtykket til lagring (og deling gjennom tjenestens API) av deres personlige informasjon når de har tatt i bruk de digitale tjenestene. Innsamling av data til forskning på hatefulle ytringer foregår gjennom disse digitale tjenestenes API, som er en del av tjenestenes vilkår som brukerne allerede har godtatt. Dermed kan det tenkes at det ikke er nødvendig å kreve nytt samtykke fra brukere når dataene skal brukes til forskning, fordi de allerede har samtykket i at disse dataene kan samles inn gjennom tjenestenes API.

Den nye *EU General Data Protection Regulation*[15], som vil tre i kraft i mai 2018, inneholder elementer som vil gjøre bruken av stordata i forskningsøyemed lettere. Dette regelverket vil gi hver enkelt bruker av digitale tjenester bedre kontroll over egne data. Brukeren vil måtte gi eksplisitt samtykke til lagring og prosessering av personlige data og vil kunne velge bort denne muligheten uten at det medfører å ikke kunne bruke tjenesten. Likevel åpner det nye regelverket for unntak når personopplysninger behandles til vitenskapelig eller historisk forskningsformål eller til statistiske formål, gitt at arkivering og prosessering knyttet til disse formålene av allmenn interesse innebærer tekniske og organisatoriske tiltak som sikrer anonymitet og dataminimering.

15 http://ec.europa.eu/justice/data-protection/

3.3.4 Vurdering av automatisert gjenkjenning av hatefulle ytringer som metodisk tilnærming

Vi avslutter denne delen med en oppsummerende diskusjon som vurderer stordataanalyse som metodisk tilnærming for å studere omfanget av hatefulle ytringer sett opp mot kriteriene vi har lagt til grunn i innledningen av denne rapporten.

Fange fenomenet: De tilgjengelige internasjonale studiene av hatefulle ytringer som bruker maskinlæring, viser at denne tilnærmingen gir gode resultater for å gjenkjenne hatefulle ytringer i store tekstmengder, i den forstand at det er mulig å trene en algoritme til å gjenkjenne bestemte typer ytringer med stor presisjon. Likevel er kvaliteten på resultatene av denne metoden avhengig av kvaliteten på dataene disse algoritmene behandler. Både definisjon og avgrensning av begrepet hatefulle ytringer og kvaliteten på kodingen og valgene som er gjort i kodingsprosessen, vil påvirke kvaliteten på resultatene. I tillegg er det en avveining mellom å begrense treningssettet til ett eller noen få grunnlag for hatefulle ytringer – noe som vil øke treffsikkerheten – og å fange opp et mangfold av grunnlag – noe som vil gå på bekostning av presisjonen. Ideelt sett burde man trene flere algoritmer som er spesialisert på bestemte grunnlag, for å kunne fange opp hele bredden av mulige grunnlag for hatefulle ytringer samtidig som man oppnår høy treffsikkerhet.

Ulike grunnlag: En stordatatilnærming kan tilpasses ulike grunnlag og målgrupper for hatefulle ytringer, helst ved at man trener flere spesialiserte algoritmer. På denne måten er den godt egnet til å måle omfang av hatefulle ytringer i den digitale offentligheten. Likevel, fordi spesialisering er mest effektivt, er den dårlig tilpasset til å ivareta interseksjonelle perspektiver, som det å sette de ulike grunnlagene i sammenheng. Denne tilnærmingen er også mindre godt egnet til å sammenligne utsatte gruppers erfaringer med hatefulle ytringer med majoritetsbefolkningens erfaringer, fordi den er basert på innsamling av «transaksjonsdata» (det folk faktisk legger ut i sosiale medier) og ikke på innsamling av subjektive data (hvordan folk opplever situasjonen). Denne tilnærmingen gir, som regel, begrensede muligheter til å samle inn bakgrunnsdata for å analysere betydningen av faktorer som for eksempel kjønn, etnisitet, alder og klassebakgrunn.

Ulike arenaer: En stordatatilnærming kan først og fremst brukes til å måle omfanget av hatefulle ytringer på offentlig tilgjengelige sider på sosiale medier.

Representativitet: En stordatatilnærming har minst to fordeler knyttet til representativitet: Den brukes på «transaksjonsdata» (faktiske handlinger) og ikke

rapporterte handlinger. I tillegg har den evnen til å behandle store mengder data. Likevel kan vi problematisere hvor representativt utvalget av dataene som analyseres, er. For det første har man tilgang til begrensede typer data fra sosiale medier. Private sider på Facebook er for eksempel ikke tilgjengelige for forskningsanalyse, det samme gjelder kommentarfeltene i de fleste digitale aviser. Dermed er ikke alle digitale offentlige arenaer tilgjengelige for stordatainnsamling. For det andre opererer man ofte med en seleksjon av data som skal brukes i analysen, for å begrense datagrunnlaget til relevante data, som for eksempel ved å velge ut poster i sosiale medier som inneholder bestemte søkeord. For det tredje er dataene det er mulig å samle inn fra sosiale medier, ikke nødvendigvis representative for det sosiale mediet, siden det finnes innebygde begrensninger i API-ens infrastruktur, som for eksempel i Twitter Streaming API, noe som undergraver dataenes representativitet.

Mulighet til å innhente annen relevant informasjon: Stordatatilnærmingen er også begrenset når det gjelder mulighetene for å innhente informasjon om andre aspekter ved hatefulle ytringer utover omfang, som informasjon om avsendere. De fleste API-ene gir tilgang til noen typer personlige opplysninger om avsendere, men i mindre omfang enn det som er mulig ved hjelp av surveyundersøkelser. I tillegg reiser bruken av denne typen personlige opplysninger etiske utfordringer som kan begrense bruken av slik informasjon.

Tidsserier og komparasjon: Stordatatilnærmingen er lett å bruke til både å overvåke omfanget av hatefulle ytringer over tid og til å sammenligne omfanget av hatefulle ytringer på tvers av land (språk). Når algoritmen er trent på et treningssett, kan den brukes over tid uten mye merarbeid. Den kan også trenes og brukes på data som er samlet inn i andre land, om man tar hensyn til språkforskjeller.

Kostnadseffektivitet: En stordatatilnærming krever en betydelig investering i startfasen, både i form av infrastruktur (datamaskiner og lagringskapasitet egnet til store mengder data) og i form av utvikling (koding av trenings- og testsett, testing og justering av algoritmen). Når algoritmen er trent og operasjonell, kan den imidlertid brukes uten store ekstrakostnader. Det er også knyttet en del kostnader til datainnsamlingen, avhengig av hvilke typer datakilder man bruker. Hvis man benytter mulighetene som er tilgjengelige gjennom API-ene til sosiale medier, er det kostnader knyttet til lagring og vedlikehold av innsamling over tid. Hvis man kjøper data gjennom en datameglertjeneste, kan kostnadene være betydelig høyere.

Forskningsetiske vurderinger: Stordatatilnærmingen aktualiserer to typer etiske utfordringer, som bør skilles fra hverandre. Den første utfordringen skyldes de strenge kravene knyttet til den norske tolkningen av regelverket om lagring av personlig informasjon. Hvis kravet er at man skal skaffe seg samtykke fra hver enkelt Twitter- eller Facebook-bruker for å kunne analysere omfanget av hatefulle ytringer i sosiale medier ved hjelp av maskinlæring, slik den nåværende tolkningen ser ut til å være, er stordatatilnærmingen umulig å implementere i Norge. Det er imidlertid grunn til å tro at det kan gis unntak fra informasjonsplikten og kravet om innhenting av samtykke for dette formålet. I tillegg kan det argumenteres med at brukerne av de aktuelle sosiale mediene allerede har samtykket i at dataene om dem kan lagres og gjenbrukes. Implementering av den nye *EU General Data Protection Regulation* i Norge vil åpne for et bredere unntaksregime når det gjelder forskning, enn det som er tilfellet i dag, noe som vil gjøre det mulig å bruke maskinlæring til gjenkjenning av hatefulle ytringer.

Den andre utfordringen er knyttet til bruken av resultater produsert av en maskinlæringsalgoritme, særlig når de kan kobles til personlig bakgrunnsinformasjon. Slike verktøy kan brukes til å skaffe nyttig informasjon om hatefulle ytringer som samfunnsfenomen, men de kan også brukes som et overvåkningsverktøy, som hvis det blir misbrukt, kan bidra til å begrense ytringsfriheten og andre friheter. Bruken av slike metoder bør derfor skje med formål og retningslinjer som er i tråd med enkeltmenneskers grunnleggende rettigheter og friheter.

4 Avslutning

Formålet med denne rapporten har vært å vurdere muligheter og begrensninger for å studere omfanget av hatefulle ytringer ved hjelp av ulike metodiske tilnærminger. Dette vil kunne danne grunnlag for fremtidige empiriske studier av hatefulle ytringer og andre tilgrensende fenomener. Vi har sett nærmere på to hovedtyper av tilnærminger til å studere omfanget av hatefulle ytringer, nemlig surveymetoder og tekstanalyse. Vi har gått særlig i dybden når det gjelder mulighetene for automatisert analyse av såkalte stordata, siden dette er en relativt ny og for mange ukjent metodikk.

I dette avslutningskapitlet vil vi først oppsummere hva slags forståelse, definisjon og operasjonalisering av begrepet hatefulle ytringer vi finner i empiriske studier av fenomenet, før vi oppsummerer vurderingen av muligheter og begrensninger ved ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer.

4.1 Definisjoner av hatefulle ytringer brukt i empirisk forskning

I denne rapporten har vi gått gjennom ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer og sett nærmere på hvordan fenomenet har blitt definert og målt i empiriske studier. Så vidt vi har klart å bringe på det rene, er det begrenset med forskningsbidrag som direkte undersøker omfanget av hatefulle ytringer, samtidig som det finnes en rekke tilstøtende fenomener, som mobbing, trakassering, netthat og nett-trusler og trolling, som i større og mindre grad har vært gjenstand for empirisk forskning (se også Eggebø mfl. 2016; Nadim mfl. 2016).

Selv om det varierer nøyaktig hvordan ulike studier av hatefulle ytringer definerer og operasjonaliserer begrepet, dekker definisjonen vi presenterer i innledningskapitlet – at hatefulle ytringer er bevisst stigmatiserende, diskriminerende, nedverdigende eller truende ytringer rettet mot et individ eller en gruppe på grunnlag av bestemte (oppfattede) gruppekarakteristikker – i stor grad forståelsen av hatefulle ytringer som ligger til grunn for de ulike empiriske studiene.

Det vesentlige skillet mellom studier av hatefulle ytringer og andre tilgrensende fenomener er hvorvidt de undersøker hva som er *grunnlaget* for den ubehagelige ytringen/opplevelsen. Mens empiriske studier av hatefulle ytringer gjerne har avgrenset studien til ytringer rettet mot definerte gruppeidentiteter, er dette sjelden gjort i studier av for eksempel mobbing, nett-trusler og så videre (se Eggebø & Stubberud 2016; Nadim mfl. 2016). Det varierer likevel *hvor mange* og *hvilke* gruppeidentiteter ulike studier av hatefulle ytringer har tatt med. Mens enkelte studier har tatt for seg én utvalgt gruppe (for eksempel Olsen mfl. 2016), har andre forsøkt å fange opp omfanget på tvers av ulike gruppeidentiteter (Nadim mfl. 2016).

Det er vanskelig å vurdere hva slags krav til alvorlighetsgrad ulike empiriske undersøkelser stiller for at noe skal regnes som en hatefull ytring. Én ting er hvilken definisjon studiene legger til grunn, noe annet er hvordan begrepet helt konkret er operasjonalisert. I surveyundersøkelser er det spørsmålsformuleringen som indikerer hvilken alvorlighetsgrad forskerne er ute etter å måle, mens svarene reflekterer respondentenes subjektive forståelse av spørsmålet og fenomenet. I undersøkelsen til Nordlandsforskning om hatefulle ytringer rettet mot personer med nedsatt funksjonsevne var hatefulle ytringer definert i et spenn mellom ytringer om at man er «hjelpeløs eller svak», og trusler om vold (Olsen mfl. 2016: 44). Andre studier har brukt mer abstrakte beskrivelser når de har definert hatefulle ytringer i undersøkelser, som for eksempel Hawdon og kollegaer, som definerte det som ytringer som er «hatefulle og nedverdigende (degrading)» (Hawdon mfl. 2015).

Når det gjelder innholdsanalyse, er terskelen for at noe skal defineres som hatytringer, vanskeligere å vurdere utenfra og er avhengig av tolkningen til koderne (ved manuell koding) eller algoritmene som skal gjenkjenne hatytringer (ved analyse av stordata).

Det kan være verdt å merke seg at ingen av de empiriske studiene vi har gjennomgått, har tatt mål av seg til å avgrense det de studerer, til *ulovlige* hatefulle ytringer. I den grad de forholder seg til skillet mellom lovlige og ulovlige ytringer, er det for å understreke at undersøkelsen ikke følger en streng juridisk definisjon (se for eksempel Silva mfl. 2016).

Vårt inntrykk er at få av de empiriske studiene har som krav at ytringene skal være offentlig fremsatt for at de skal regnes som hatefulle (med unntak av for eksempel Institut for Menneskerettigheder 2017). Likevel vil den metodiske tilnærmingen avgjøre hva slags type ytringer som kan fanges opp. Innholdsanalyser er i all hovedsak basert på offentlig tilgjengelig materiale og fanger

dermed opp «offentlige» ytringer. Som beskrevet i kapittel 2 kan man ved hjelp av surveyundersøkelser måle om folk har *observert* hatefulle ytringer – dette blir da per definisjon ytringer som er fremsatt offentlig eller i andres nærvær. I tillegg kan man måle om folk selv har *mottatt* slike ytringer – dette kan være både ytringer som er mottatt direkte, og ytringer som har blitt fremsatt offentlig.

4.2 Vurdering av ulike metodiske tilnærminger for å studere omfang av hatefulle ytringer

I denne rapporten har vi diskutert to ulike hovedtilnærminger til å studere omfanget av hatefulle ytringer: 1) å spørre folk om deres erfaringer med å observere eller motta slike ytringer gjennom surveymetoder, og 2) å analysere innholdet i den offentlige debatten og innslaget av hatefulle ytringer gjennom ulike former for innholdsanalyse. De metodiske tilnærmingene vi har gått gjennom, er alle egnet til å studere omfang av hatefulle ytringer, men de måler litt ulike ting og har sine ulike muligheter og begrensninger. Hvilken metode man velger, må dermed være basert på hva man primært er interessert i å fange opp.

I tabell 4.1 oppsummerer vi styrker og svakheter ved de ulike tilnærmingene sett opp mot vurderingskriteriene vi har lagt til grunn i rapporten (se kapittel 1). For å få en dypere forståelse av hvordan fenomenet hatefulle ytringer arter seg, og hvilke forestillinger og retoriske virkemidler det bygger på, er kvalitative tilnærminger og manuell innholdsanalyse de best egnede metodiske tilnærmingene. Er man derimot interessert i for eksempel minoritetsgruppers erfaringer med hatefulle ytringer, er surveymetoder et naturlig valg. Automatisert analyse av stordata har ikke den samme muligheten til å fange opp hatefulle ytringer som de andre metodene har, men det gir til gjengjeld mulighet til å følge trender over tid – også i sanntid – og til å overvåke omfanget av hatefulle ytringer i svært omfattende materialer.

Tabell 4.1 Vurdering av ulike metodiske tilnærminger for å studere omfang av hatefulle ytringer

Vurderingskriterier	Survey	Manuell innholdsanalyse	Stordataanalyse
Fange fenomenet	Avhengig av spørsmålsformulering Kan eksperimentere med ulike formuleringer Godt egnet til å fange subjektive forståelser av fenomenet	Gir forskeren god kontroll over hva som regnes som en hatefull ytring	Avhengig av kvaliteten på data og koding av data Godt egnet til å fange opp bestemte typer hatefulle ytringer i svært store utvalg
Ulike grunnlag	Godt egnet	Godt egnet	Godt egnet
Ulike arenaer	Godt egnet	Begrenset hvilke arenaer som kan fanges opp: offentlig tilgjengelig nett- eller medieinnhold	Begrenset til nettinnhold og tilgjengelige datakilder
Representativitet	Avhengig av populasjon, utvalg og svartilbøyelighet	Avhengig av utvalg, mulig å oppnå	Mulig å oppnå representativitet for bestemte datakilder
Mulighet til å innhente annen relevant informasjon	Godt egnet	Mulig å innhente noe tilleggsinformasjon, for eksempel om avsendere, kontekst for hatytringer og respons	Mulig, avhengig av datakilde, men som regel begrenset
Tidsserier og komparasjon	Avhengig av fenomenforståelse og utvalgskilde	Godt egnet	Godt egnet
Kostnadseffektivitet	Avhengig av gjennomføringsmetode	Avhengig av omfanget av innholdet som blir analysert, og hvor kompleks analysen er	Betydelig investering i startfasen Reduserte kostnader på lengre sikt
Helhetsvurdering	Svært fleksibel metode Gitt gode spørsmålsformuleringer og representative utvalg gjør metoden det mulig å gi presis informasjon om omfanget av folks erfaringer med å motta eller observere hatefulle ytringer Godt design kan være ressurskrevende	Godt egnet til å fange opp fenomenet og kan gi relevant tilleggsinformasjon Studerer faktiske ytringer heller enn rapportering av erfaringer Hovedbegrensningene ved metoden er at den ikke kan brukes på alle typer arenaer, og at den potensielt er ressurskrevende	Studerer faktiske ytringer heller enn rapportering av erfaringer Gir gode resultater for bestemte datakilder Kan anvendes til overvåkning og prevensjon

Selv om det kan være selvsagt, er det viktig å understreke at de ulike metodiske tilnærmingene ikke utelukker hverandre. Derimot kan de med hell kombineres for å belyse hatefulle ytringer fra litt ulike vinkler.

Litteratur

Anderssen, N. & Malterud, K. (2017). Oversampling as a methodological strategy for the study of self-reported health among lesbian, gay and bisexual populations. *Scandinavian Journal of Public Health*, *45*(6): 637–646.

Arnesen, D., Sivesind, K.H. & Gulbrandsen, T. (2016). *Fra medlemsbaserte organisasjoner til koordinert frivillighet? Det norske organisasjonssamfunnet fra 1980 til 2013*. Rapport 5. Bergen/Oslo: Senter for forskning på sivilsamfunn og frivillig sektor.

Article 19 (2015). "*Hate speech*" *explained: A toolkit.* London: Article 19.

Badjatiya, P., Gupta, S., Gupta, M. & Varma, V. (2017). *Deep learning for hate speech detection in tweets.* Paper presentert på Proceedings of the 26th International Conference on World Wide Web Companion.

Bartlett, J., Reffin, J., Rumball, N. & Williamson, S. (2014). Anti-social media. *Demos*, 1–51.

Beadle-Brown, J., Richardson, L., Guest, C., Malovic, A., Jill, B. & Julian, H. (2014). *Living in fear. Better outcomes for people with learning disabilities and autism. Main research report*. Canterbury: Tizard Centre, University of Kent.

Bleich, E. (2011). *The freedom to be racist? How the United States and Europe struggle to preserve freedom and combat racism*. Oxford: Oxford University Press.

Boeckmann, R.J. & Liew, J. (2002). Hate speech: Asian American students' justice judgments and psychological responses. *Journal of Social Issues, 58*(2), 363–381.

Boeckmann, R.J. & Turpin-Petrosino, C. (2002). Understanding the Harm of Hate Crime. *Journal of Social Issues, 58*(2), 207–225.

Bowling, B. (1999). *Violent racism: Victimization, policing, and social context*. Oxford: Oxford University Press.

Brown, A. (2017). What is so special about online (as compared to offline) hate speech? *Ethnicities,* doi: 1468796817709846.

Brudholm, T. (2013). Om had og hadetale. I R.E. Larsen, J. Lohmann & K. Slavensky (red.), *Hate speech: Fra hadetale til hadesyn*. København: Information.

Burnap, P., Rana, O.F., Avis, N., Williams, M., Housley, W., Edwards, A. mfl. (2015). Detecting tension in online communities with computational Twitter analysis. *Technological Forecasting and Social Change, 95*, 96–108.

Burnap, P. & Williams, M.L. (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet, 7*(2), 223–242.

Burnap, P. & Williams, M.L. (2016). Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science, 5*(1), 11.

Burnap, P., Williams, M.L., Sloan, L., Rana, O., Housley, W., Edwards, A. mfl. (2014). Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining, 4*(1), 206.

Cappelen, C., Kuhnle, S. & Midtbø, T. (2016). Velferdssjåvinisme i Norge? – Et listeeksperiment. *Norsk statsvitenskapelig tidsskrift, 31*(02), 122–141.

Chakraborti, N. (2010). Crimes Against the «Other»: Conceptual, Operational, and Empirical Challenges for Hate Studies. *Journal of Hate Studies, 8*(1), 9–28.

Chen, Y., Zhou, Y., Zhu, S. & Xu, H. (2012). *Detecting offensive language in social media to protect adolescent online safety.* Paper presentert på Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).

COWI. (2015). Kortlægning af hadforbrydelser i Danmark. En undersøgelse af befolkningens oplevde hadforbrydelser. Kongens Lyngby: COWI.

Dadvar, M., Trieschnigg, D., Ordelman, R. & de Jong, F. (2013). *Improving Cyberbullying Detection with User Context.* Paper presentert på ECIR.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H. & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(3), 18.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V. & Bhamidipati, N. (2015). *Hate speech detection with comment embeddings.* Paper presentert på Proceedings of the 24th International Conference on World Wide Web.

Djuve, A.B., Gulløy, E., Kavli, H.C. & Berglund, F. (2009). *Datafangst når minoritetsbefolkningen er målgruppe*. Fafo-rapport 2009:24. Oslo: Fafo.

Douglas, K.M. (2007). Psychology, discrimination and hate groups online. I A.N. Joinson, K.Y.A. McKenna, T. Postmes & U.-D. Reips (red.), *The Oxford Handbook of Internet Psychology* (s. 155–164). Oxford: Oxford University Press.

Douglas, K.M., McGarty, C., Bliuc, A.-M. & Lala, G. (2005). Understanding cyberhate social competition and social creativity in online white supremacist groups. *Social Science Computer Review, 23*(1), 68–76.

Duffy, M.E. (2003). Web of Hate: a Fantasy Theme Analysis of the Rhetorical Vision of Hate Groups Online. *Journal of Communication Inquiry, 27*(3), 291–312.

ECRI. (2016). *ECRI general policy recommendation No. 15 On combating hate speech*. Strasboug: Council of Europe.

Eggebø, H., Sloan, L. & Aarbakke, M.H. (2016). *Erfaringer med digitale krenkelser i Norge*. KUN-rapport 2016: 1. Steigen: Forlaget Nora.

Eggebø, H. & Stubberud, E. (2016). *Hatefulle ytringer delrapport 2: Forskning på hat og diskriminering*. Rapport 2016:15. Oslo: Institutt for samfunnsforskning.

Eimhjellen, I. (2016). *Innvandrarar si deltaking i norsk frivilligliv. Nye tal og metodiske utfordringar*. Rapport 3/2016. Bergen/Oslo: Senter for forskning på sivilsamfunn og frivillig sektor.

Enjolras, B., Steen-Johnsen, K. & Karlsen, R. (2014). *Valgkampen 2013 på Twitter: Sosiale medier som kritisk offentlighet*. Rapport 2014:3. Oslo: Institutt for samfunnsforskning.

Erjavec, K. & Kovačič, M. P. (2012). «You Don't Understand, This is a New War!» Analysis of Hate Speech in News Web Sites' Comments. *Mass Communication and Society, 15*(6), 899–920.

Fladmoe, A. & Nadim, M. (2017). The extent and consequences of hate speech in social media. I A.H. Midtbøen, K. Steen-Johnsen & K. Thorbjørnsrud (red.), *Boundary Struggles: Contestations of Free Speech in the Public Sphere*. Oslo: Cappelen Damm.

Gagliardone, I., Gal, D., Alves, T. & Martinez, G. (2015). *Countering Online Hate Speech*. Paris: UNESCO.

Gelber, K. & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities, 22*(3), 324–341.

Gerstenfeld, P.B., Grant, D.R. & Chiang, C.-P. (2003). Hate Online: A Content Analysis of Extremist Internet Sites. *Analyses of Social Issues and Public Policy, 3*(1), 29–44.

Gitari, N.D., Zuping, Z., Damien, H. & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering, 10*(4), 215–230.

Greevy, E. & Smeaton, A.F. (2004). *Classifying racist texts using a support vector machine.* Paper presentert på Proceedings of the 27th Annual International Acm Sigir Conference on Research and Development in Information Retrieval.

Hagen, A.L. (2015). *Meningers mot: netthat og ytringsfrihet i Norge*. Oslo: Cappelen Damm akademisk.

Hawdon, J., Oksanen, A. & Räsänen, P. (2015). Online Extremism and Online Hate. Exposure among Adolescents and Young Adults in Four Nations. *Nordicom Information, 34*(3-4), 29–37.

Hellevik, O. (2015). Hva betyr respondentbortfallet i intervjuundersøkelser? *Tidsskrift for samfunnsforskning, 56*(02), 211–229.

Herek, G.M., Cogan, J.C. & Gillis, J.R. (2002). Victim Experiences in Hate Crimes Based on Sexual Orientation. *Journal of Social Issues, 58*(2), 319–339.

Hosseinmardi, H., Mattson, S.A., Rafiq, R.I., Han, R., Lv, Q. & Mishra, S. (2015). Detection of Cyberbullying Incidents on the Instagram Social Network.

Institut for Menneskerettigheder. (2017). *Hadefulde ytringer i den offentlige online debat*. København: Institut for Menneskerettigheder.

Ishaq, B. (2017). *Hvem snakker for oss? Muslimer i dagens Norge – hvem er de og hva mener de?* Oslo: Cappelen Damm.

Karlsen, R. & Enjolras, B. (2016). Styles of Social Media Campaigning and Influence in a Hybrid Political Communication System. *The International Journal of Press/Politics, 21*(3), 338–357.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3. utg.). Thousand Oaks, CA: Sage.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity, 47*(4), 2025–2047.

Kuklinski, J.H., Cobb, M.D. & Gilens, M. (1997). Racial Attitudes and the "New South". *The Journal of Politics, 59*(02), 323–349.

Kwok, I. & Wang, Y. (2013). *Locate the Hate: Detecting Tweets against Blacks.* Paper presentert på AAAI.

Lawrence III, C.R., Matsuda, M.J., Delgado, R. & Crenshaw, K.W. (1993). Introduction. I M.J. Matsuda, C.R. Lawrence III, R. Delgado & K.W. Crenshaw (red.), *Words that wound: critical race theory, assaultive speech, and the First Amendment*. Boulder, Colorado: Westview Press.

LDO. (2015). *Hatytringer og hatkriminalitet*. Oslo: Likestillings- og diskrimineringsombudet.

Leets, L. (2002). Experiencing hate speech: Perceptions and responses to anti-semitism and antigay speech. *Journal of Social Issues, 58*(2), 341–361.

Magu, R., Joshi, K. & Luo, J. (2017). Detecting the Hate Code on Social Media.

Manning, C.D. & Schütze, H. (1999). *Foundations of statistical natural language processing*: MIT Press.

McNamee, L.G., Peterson, B.L. & Peña, J. (2010). A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs, 77*(2), 257–280.

Meddaugh, P.M. & Kay, J. (2009). Hate Speech or «Reasonable Racism?» The Other in Stormfront. *Journal of Mass Media Ethics, 24*(4), 251–268.

Mehdad, Y. & Tetreault, J.R. (2016). *Do Characters Abuse More Than Words?* Paper presentert på SIGDIAL Conference.

Midtbøen, A.H. & Steen-Johnsen, K. (2016). Ytringsfrihetens grenser i det flerkulturelle Norge. *Nytt Norsk Tidsskrift, 32*(01-02), 21–33.

Mondal, M., Silva, L. A. & Benevenuto, F. (2017). A Measurement Study of Hate Speech in Social Media.

Mutz, D. C. (2011). *Population-Based Survey Experiments*. Princeton and Oxford: Princeton University Press.

Nadim, M., Fladmoe, A. & Wessel-Aas, J. (2016). *Hatefulle ytringer på internett: Omfang, forebygging og juridiske grenser*. Rapport 2016:17. Oslo: Institutt for samfunnsforskning.

Nilsen, A.B. (2014). *Hatprat*. Oslo: Cappelen Damm akademisk.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. & Chang, Y. (2016). *Abusive language detection in online user content.* Paper presentert på Proceedings of the 25th International Conference on World Wide Web.

Olsen, T., Vedeler, J., Eriksen, J. & Elvegård, K. (2016). *Hatytringer. Resultater fra en studie av funksjonshemmedes erfaringer*. Rapport nr. 6/2016. Bodø: Nordlandsforskning.

Perry, B. (2001). *In the name of hate: understanding hate crimes*. New York: Routledge.

Pew Research Center. (2014). *Online Harassment*.

Politidirektoratet. (2016). *Politiets innbyggerundersøkelse*. Ipsos MMI.

Ringdal, K. (2001). *Enhet og mangfold: samfunnsvitenskapelig forskning og kvantitativ metode*. Bergen: Fagbokforlaget.

Rohlfing, S. (2014). Hate on the Internet. I N. Hall, A. Corb, P. Giannasi & J. Grieve (red.), *The Routledge international handbook on hate crime* (s. 293–305). New York: Routledge.

Salamon, L.M. & Anheier, H.K. (1998). Social Origins of Civil Society: Explaining the Nonprofit Sector Cross-Nationally. *Voluntas: International Journal of Voluntary and Nonprofit Organizations, 9*(3), 213–248.

Schmidt, A. & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing.* Paper presentert på Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Association for Computational Linguistics, Valencia, Spain.

Silva, L.A., Mondal, M., Correa, D., Benevenuto, F. & Weber, I. (2016). *Analyzing the Targets of Hate in Online Social Media.* Paper presentert på ICWSM.

Sivesind, K.H. & Selle, P. (2010). Civil society in the Nordic countries: Between displacement and vitality. I R. Alapuro & H. Stenius (red.), *Nordic Associations in a European Perspective*. Baden-Baden: Nomos Verlagsgesellschaft.

Sood, S.O., Churchill, E.F. & Antin, J. (2012). Automatic identification of personal insults on social news sites. *Journal of the Association for Information Science and Technology, 63*(2), 270–285.

Spertus, E. (1997). *Smokey: Automatic recognition of hostile messages.* Paper presentert på AAAI/IAAI.

Staksrud, E., Steen-Johnsen, K., Enjolras, B., Gustafsson, M.H., Ihlebæk, K.A., Midtbøen, A.H. mfl. (2014). *Ytringsfrihet i Norge: Holdninger og erfaringer i befolkningen. Resultater fra befolkningsundersøkelsen*. Oslo: Fritt Ord, ISF, IMK, FAFO.

Stephens-Davidowitz, S. (2017). *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are*. New York: Harper Collins Publishers.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G. mfl. (2015). *Detection and fine-grained classification of cyberbullying events.* Paper presentert på International Conference Recent Advances in Natural Language Processing (RANLP).

Vrålstad, S. & Wiggen, K.S. (red.). (2017). *Levekår blant innvandrere i Norge 2016*. SSB-rapport 2017/13. Oslo: Statistisk sentralbyrå.

Warner, W. & Hirschberg, J. (2012). *Detecting hate speech on the world wide web.* Paper presentert på Proceedings of the Second Workshop on Language in Social Media.

Waseem, Z. & Hovy, D. (2016). *Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.* Paper presentert på SRW@ HLT-NAACL.

Wessel-Aas, J., Fladmoe, A. & Nadim, M. (2016). *Hatefulle ytringer delrapport 3: Grenseoppgangen mellom ytringsfrihet og strafferettslig vern mot hatefulle ytringer*. Rapport 2016:16. Oslo: Institutt for samfunnsforskning.

Williams, M. L. & Burnap, P. (2015). Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data. *British Journal of Criminology, 56*(2), 211–238.

Xiang, G., Fan, B., Wang, L., Hong, J. & Rose, C. (2012). *Detecting offensive tweets via topical feature discovery over a large scale twitter corpus.* Paper presentert på Proceedings of the 21st ACM international conference on Information and knowledge management.

Xu, J.-M., Jun, K.-S., Zhu, X. & Bellmore, A. (2012). *Learning from bullying traces in social media.* Paper presentert på Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies.

Zhong, H., Li, H., Squicciarini, A.C., Rajtmajer, S.M., Griffin, C., Miller, D.J. mfl. (2016). *Content-Driven Detection of Cyberbullying on the Instagram Social Network.* Paper presentert på IJCAI.

Måling av omfang av hatefulle ytringer

Metodiske muligheter og utfordringer

Formålet med denne rapporten er å vurdere muligheter og begrensninger ved ulike metodiske tilnærminger til å studere omfanget av hatefulle ytringer. De metodiske vurderingene i rapporten er basert på en vurdering av eksisterende empiriske studier av hatefulle ytringer og tilgrensende fenomener, i tillegg til oppdatert metodelitteratur og en gjennomgang av relevante erfaringer med ulike metodiske tilnærminger.

De ulike metodiske tilnærmingene til å studere omfanget av hatefulle ytringer blir vurdert ut fra a) hvor nøyaktige de er til å fange fenomenet vi er interessert i, b) i hvilken grad de gjør det mulig å studere og sammenligne omfanget av hatefulle ytringer rettet mot ulike grupper, på ulike arenaer, c) hvor representative resultater metoden gir, d) muligheter for å innhente annen type informasjon om hatefulle ytringer utover omfang, e) mulighetene for tidsserier og komparasjon og f) hvor kostnadskrevende de ulike metodiske oppleggene er.

Vi tar for oss to hovedtilnærminger til å studere omfanget av hatefulle ytringer: surveymetoder og analyser av tekstinnhold, nærmere bestemt kvantitativ manuell innholdsanalyse og automatiserte analyser av stordata (Big Data) ved hjelp av maskinlæringsalgoritmer.

Institutt for samfunnsforskning

Institute for Social Research

Munthes gate 31
Postboks 3233 Elisenberg
0208 Oslo
Tlf: +47 23 08 61 00
samfunnsforskning.no

ISBN (trykk): 978-82-7763-581-1
ISBN (nett): 978-82-7763-582-8
ISSN (trykk): 0333-3671
ISSN (nett): 1891-4314