Könsskillnader i utsatthet för toxiskt språk online

Lisa Kaati, Katie Cohen, Björn Pelzer, Daniel Wallgren, Jenny Yourstone, Nazar Akrami.

Ett flertal studier har visat att kvinnor och män utsätts i ungefär lika stor omfattning för kränkande kommentarer på nätet, men oftast på olika sätt. I den här studien har vi undersökt skillnader mellan kvinnors och mäns utsatthet för toxiskt språk, liksom eventuella karaktärsskillnader på toxiskt språk riktat mot kvinnor respektive män. Undersökningen baseras på ett års data från några av de största svenskspråkiga sociala medierna.

6 800 000 kommentarer

Vi har analyserat kommentarer från olika sociala medieplattformar producerade under ett år.

Mansnamn och manliga pronomen nämns nästan dubbelt så ofta som kvinnonamn och kvinnliga pronomen i de undersökta källorna.

Psykisk hälsa

Kvinnor blir i större utsträckning än män utsatta för nedvärderande kommentarer om psykisk hälsa och förmåga.

Kompetens

Män utsätts för fler nedvärderade kommentarer om bristande kompetens eller prestation inom sin yrkesgärning eller allmänt.

Utseende

Kvinnor blir i större omfattning utsatta för nedvärderande kommentarer om utseende som innefattar påståenden om att en person är ful eller oattraktiv, eller objektifierande och närgångna kommentarer om en persons kropp oavsett värdeladdning.

Hot om våld och bestraffning

Män blir i större omfattning utsatta för kommentarer som innehåller hot/vålds- och bestraffningsidéer, direkt eller indirekt formulerade avsikter eller önskningar om att en person eller grupp ska, dö, försvinna eller utsättas för våld.

FOI Memo: 7741
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

2021-12-20

1. Inledning

Hotfulla och kränkande kommentarer i digitala miljöer lyfts i allt högre grad fram som ett samhällsproblem. Riskerna med ett hårdnande samtalsklimat har bland annat identifierats som ökande polarisering och radikalisering, negativa konsekvenser för de individer som utsätts för upprepade kränkningar, samt att de mer lågmälda och resonerande samtalen dränks i högljudda hatbudskap.

I den här studien använder vi begreppet *toxiskt språk* som ett paraplybegrepp för kommunikationshandlingar som i någon mån kan sägas förgifta samtalsklimatet i digitala miljöer. Toxiskt språk innefattar dels språkliga handlingar som är förbjudna i lag, exempelvis hets mot folkgrupp, förtal eller förgripelse mot tjänsteman, men kan även innefatta fall av integritetskränkning eller respektlöshet. Begreppet är tänkt att fånga in såväl vardagsbegrepp som *näthat* eller *hat och hot* som akademiska/juridiska begrepp som *hate speech* eller *dangerous speech*.

Forskning om näthat och kön visar att könsskillnader när det gäller utsatthet för trakasserier på nätet är små, men att kvinnor och män utsätts för olika typer av toxiskt språk.[1] Forskning från USA har till exempel visat att män i högre utsträckning utsätts för kränkande tilltal och fysiska hot, medan kvinnor oftare utsätts för sexuella trakasserier och förföljelse på sociala medier. En internationell kartläggning av kvinnors utsatthet för kränkningar och trakasserier online visade att 41% av de kvinnor som utsatts för näthat hade upplevt att deras fysiska säkerhet var hotad, upp till 25% att även deras familjs säkerhet var hotad. Därutöver upplevde hälften av kvinnorna ett sämre självförtroende, oro, ångest och/eller panikattacker till följd av näthatet.[2] En svensk studie av näthat mot olika yrkesgrupper visade att de kvinnor i yrkesgrupperna som undersöktes var mer utsatta än män, därtill oftare utsatta för sexuella trakasserier och utseenderelaterade förolämpningar, medan männen oftare utsattes för förolämpningar relaterade till yrke och kompetens.[3]

Den här studien syftar till att undersöka könsskillnader i utsatthet för toxiskt språk i digitala miljöer. Då strukturella trakasserier och kränkningar av kvinnor påverkar kvinnors tillgång till makt och inflytande i samhället,[4] är det angeläget att utreda i vilken mån och i vilken skepnad sådana förekommer i digitala miljöer, där en stor del av det samtida sociala livet utspelar sig.

För att genomföra undersökningen har vi till att börja med mätt hur mycket det skrivs om män respektive kvinnor i fem svenskspråkiga digitala miljöer. Därefter har vi beräknat hur stor andel av dessa omnämnanden som innehåller toxiskt språk. I ett tredje steg har vi kategoriserat toxiska kommentarer utifrån deras innehåll och studerat huruvida kommentarernas innehåll skiljer sig mellan män och kvinnor.

2. Undersökningsmetod

Källor

Studien bygger på data från fem svenskspråkiga diskussionsforum och sociala medier. Källorna, som finns beskrivna i tabell 1, är valda för att ge en bred bild av svenska digitala miljöer med användargenererat innehåll. I studien ingår diskussionsforumen Reddit, Flashback och Familjeliv, mikrobloggen Twitter samt kommentarsfälten för fem av de största nyhetsmedierna på Facebook. Det undersökta datat har genererats mellan 2020-07-01 och 2021-06-30. Sammanlagt inhämtades 6,8 miljoner inlägg, från vilka ett slumpmässigt representativt urval[5] på 60 000 inlägg per källa användes för analys.

Enligt Internetstiftelsens undersökning om svenskars användande av internet använder män i högre utsträckning än kvinnor Reddit, Twitter och Flashback, medan kvinnor i högre utsträckning använder Facebook och Familjeliv.[6] Även om skribenternas könstillhörighet är okänd för oss, kan vi genom urvalet av källor öka sannolikheten för att båda könen finns representerade.

[1] Bladini, M. (2017). Hat och hot på nätet. En kartläggning av den rättsliga regleringen i Norden från ett jämställdhetsperspektiv. NIKK.
[2] Amnesty International commissioned Ipsos MORI to carry out an online poll of women aged 18–55 in the UK, USA, Spain, Denmark, Italy, Sweden, Poland and New Zealand. For full data set see: *Ipsos MORI survey for Amnesty International on online abuse and harassment*. https://www.ipsos.com/ipsos-mori/en-uk/online-abuse-and-harassment
[3] Fernquist, J., et al (2020) Det digitala hatets karaktär. FOI Memo 7429.
[4] Pressmeddelande från Arbetsmarknadsdepartementet, Försvarsdepartementet, 12 maj 2021. Hat och hot mot kvinnor i digitala miljöer ska analyseras.
[5] Motsvarar en konfidensnivå > 99% och en felmarginal < 1%.
[6] Internetstiftelsen. Svenskarna och internet 2021

FOI Memo: 7741
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

Tabell 1. De källor som ingår i undersökningen.

Källa	Beskrivning	Antal kommentarer
Facebook	Användargenererade kommentarer från nyhetssidorna SVT, Expressen, SvD och DN	930 000
Reddit	Svenska delen av diskussionsforumet Reddit	790 000
Twitter	1% av svenska tweets	780 000
Flashback	Diskussionsforum	3,8 miljoner
Familjeliv	Diskussionsforum	500 000

Mätning av omnämnanden av män och kvinnor

För att mäta hur ofta män respektive kvinnor nämns i kommentarerna räknade vi alla förekomster av egennamn och könade personliga pronomen i 60 000 inlägg per källa, det vill säga totalt 300 000 inlägg. Namn eller pronomen, det vill säga omnämnanden av individer, fanns i 43,2% av alla meningar i inläggen. För att automatiskt kunna identifiera kvinnliga respektive manliga namn användes offentlig statistik om förnamn från Statistiska centralbyrån (SCB). Om en persons kön inte kunde identifieras automatiskt i texten (exempelvis när bara ett efternamn eller en yrkesbeteckning nämndes) räknades omnämnandet varken som kvinnligt eller manligt.

Detektion av toxiskt språk

Identifiering och mätning av toxiskt språk sker med hjälp av en maskininlärningsmodell, närmare bestämt en språkmodell som utvecklats av Kungliga biblioteket[7] som vi tränat upp att känna igen toxiskt språk genom att träna den med 6 000 manuellt klassificerade texter. Under ideala förhållanden, det vill säga när likande data används som modellen är tränad för, identifierar modellen toxiskt språk med 83-procentig träffsäkerhet.[8] Prestandan kan dock variera beroende på bland annat skillnader i språkbruk mellan olika digitala miljöer.[9]

När flera personer nämns i samma text kan maskininlärningsmodellen inte känna igen vem det toxiska språket är riktat mot. Våra analyser görs på meningsnivå, vilket gör att det sällan nämns flera personer i samma text jämfört med om vi analyserat hela inlägg. Vi har dessutom räknat bort de meningar som nämner både kvinnor och män, eller som nämner personer där könet inte kan kännas igen automatiskt.

Maskininlärningsmodellen har tränats till hög sensitivitet på bekostnad av specificitet, vilket gör att den inte missar så många toxiska kommentarer, men att det krävs en manuell analys för att kunna sålla bort meningar som klassificeraren felaktigt pekat ut som toxiska. För att säkerställa att maskininlärnings-modellens resultat är korrekta och vid behov kunna justera resultaten har en manuell annotering genomförts på 5 000 slumpmässigt utvalda meningar från den datamängd som automatiskt klassificerats som innehållande toxiskt språk. Den manuella annoteringen syftar också till att avgöra huruvida målpersonen för det toxiska språket faktiskt är en man eller en kvinna.

Tematisk analys

För att undersöka innehållet i det toxiska språket identifierades olika kategorier genom en tematisk analys av 5 000 slumpmässigt utvalda meningar från den datamängd som automatiskt klassificerats som innehållande toxiskt språk. Totalt identifierades sju vanligt förekommande kategorier:

- **Direkta/indirekta hot:** Hot/vålds- och bestraffningsidéer, direkt eller indirekt formulerade avsikter eller önskningar om att en person eller grupp ska dö, försvinna eller utsättas för våld.
- **Utseende:** Nedvärderande kommentarer om utseende som innefattar påståenden om att en person är ful eller oattraktiv, men även objektifierande och närgångna kommentarer om en persons kropp oavsett värdeladdning.

[7] Malmsten, M., Börjeson, L., & Haffenden, C. (2020). Playing with Words at the National Library of Sweden--Making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

[8] En mer utförlig beskrivning av maskininlärnings modellen finns i Fernquist, J., et al (2020) Det digitala hatets karaktär. FOI Memo 7429.

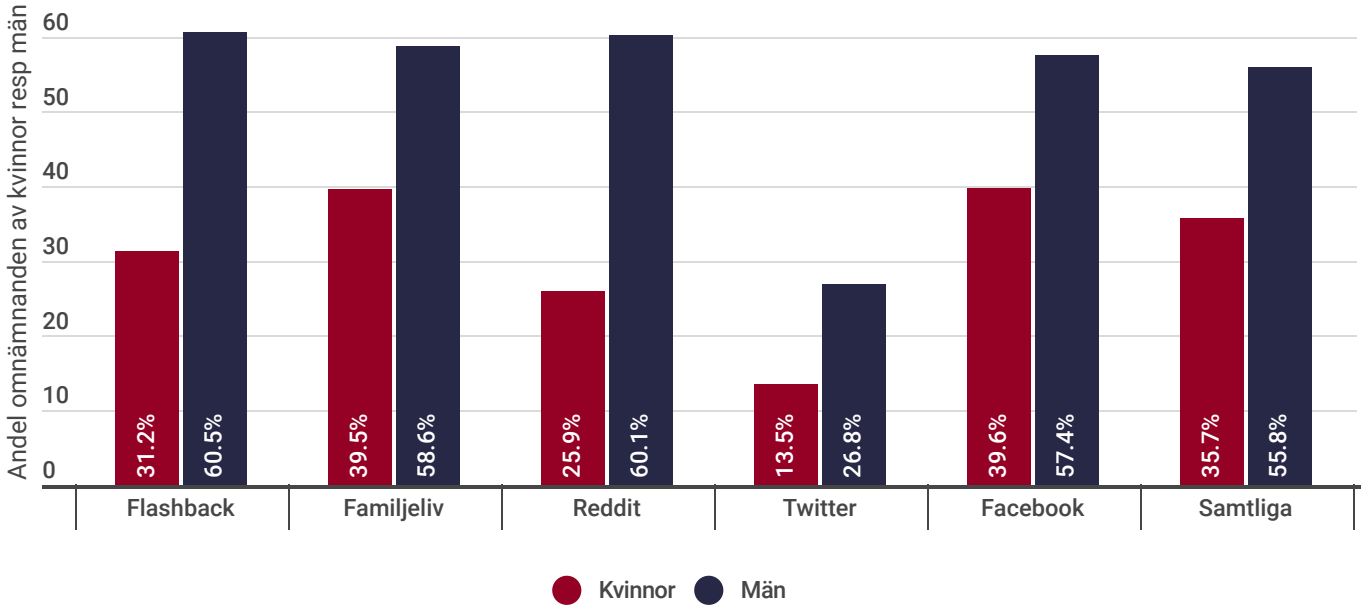
[9] Se även Kaati, L., Pelzer, B., Cohen, K., Wallgren, D., Yourstone, J & Akrami, N. (2021). Toxiskt språk i svenska digitala miljöer. Stockholm: FOI, FOI Memo 7740.

FOI Memo: 7741
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

- **Kompetens:** Nedvärderande kommentarer om en individs bristande kompetens eller prestation, inom sin yrkesgärning eller allmänt.
- **Moral:** Nedvärderande kommentarer om moralisk otillräcklighet i form av karaktärsdrag, exempelvis att en person är ond eller lögnaktig.
- **Psykisk hälsa:** Nedvärderande kommentarer om psykisk hälsa och förmåga, försök att sänka en persons trovärdighet genom att hävda att personen lider av vanföreställningar, är hysterisk eller liknande.
- **Kön och sexualitet:** Nedvärderande kommentarer om kön och sexualitet som innefattar påtaglig sexuell objektifiering, att avsiktligt felköna någon, homo- och transfobiska påståenden.
- **Förolämpningar:** Övriga nedvärderande kommentarer i form av kränkande ord eller påståenden.

3. Omnämnanden av män och kvinnor i digitala miljöer

Andelen omnämnanden av kvinnor (kvinnliga pronomen samt kvinnonamn) respektive män (manliga pronomen samt mansnamn) i de olika källorna redovisas i figur 1. Resultaten varierar mellan de olika källorna, men i alla undersökta källor finns en över-representation av omnämnanden av män. Facebook och Familjeliv är de källor som nämner kvinnor i högst utsträckning, men även där är omnämnanden av män märkbart vanligare (ca 60%) än omnämnanden av kvinnor (ca 40%). Twitter skiljer sig från de andra källorna eftersom både kvinnor och män nämns relativt sällan när personer diskuteras.

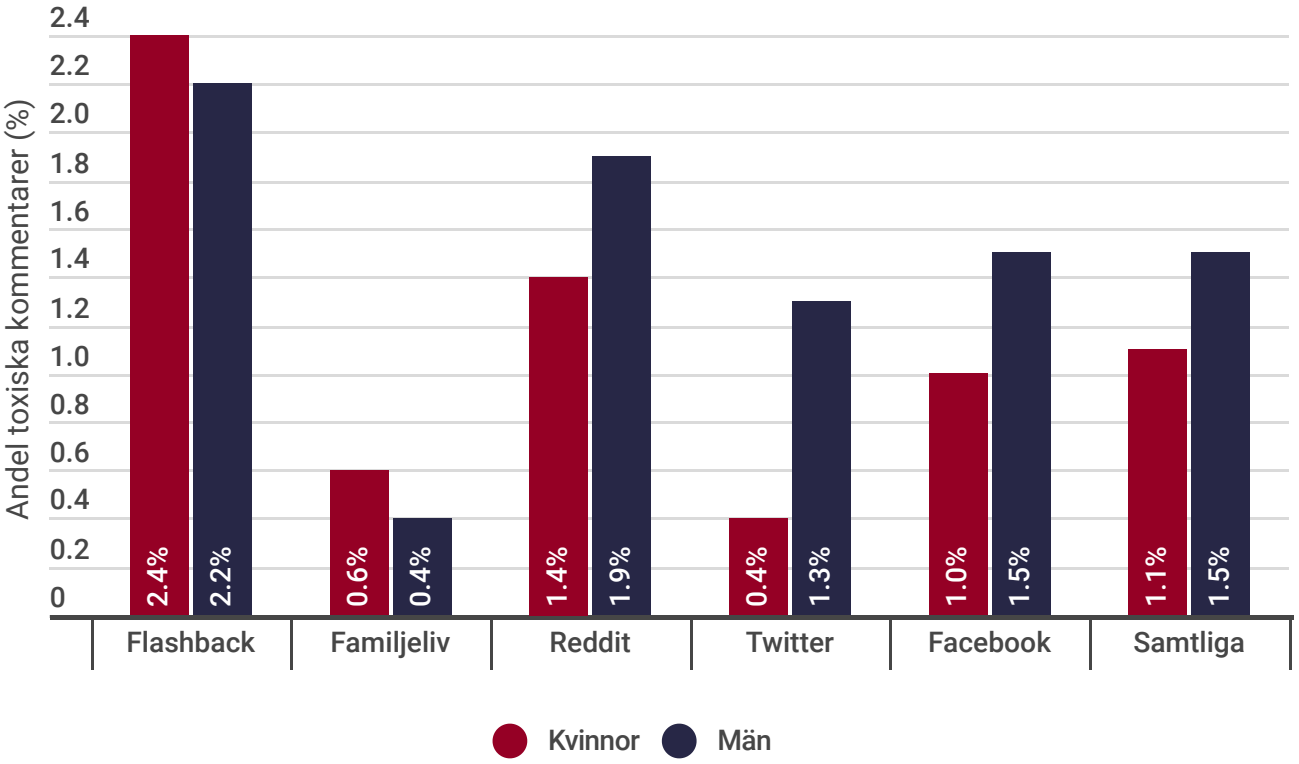


Figur 1. Andelen omnämnanden av kvinnor respektive män i de olika källorna i meningarna där personer nämns.

FOI Memo: 7741
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

4. Mängden toxiskt språk riktat mot män och kvinnor

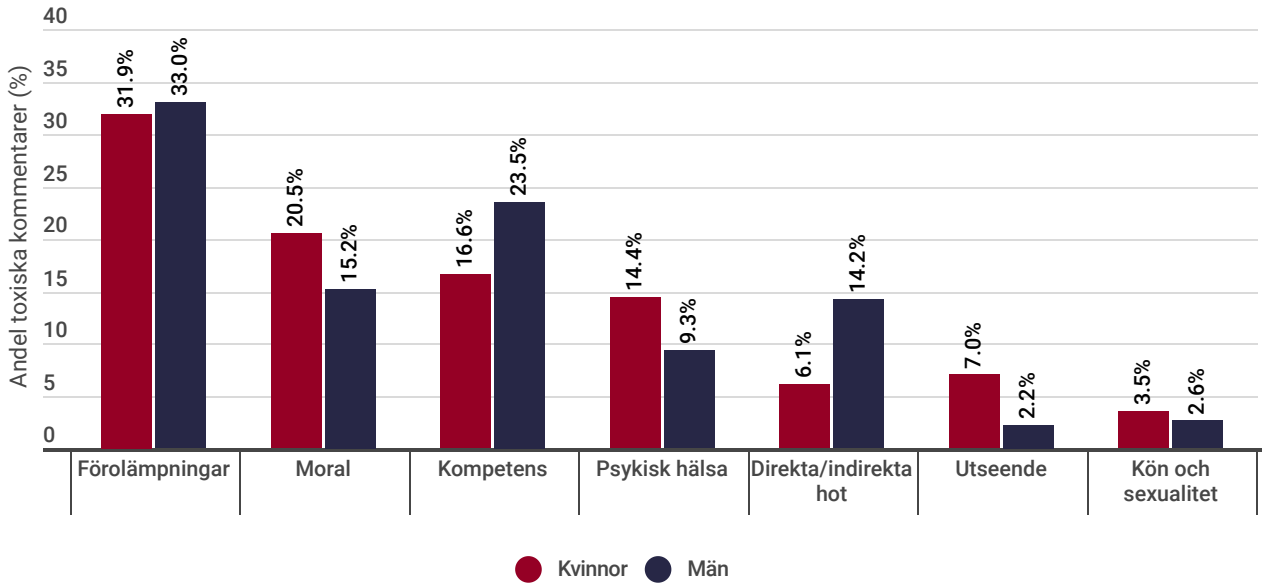
För att jämföra män och kvinnor med avseende på den mängd toxiskt språk som riktas mot respektive kön har de nästan 300 000 meningar som nämner män eller kvinnor analyserats i vår maskininlärningsmodell för detektion av toxiskt språk (se avsnitt 2). Maskininlärningsmodellen klassificerade ungefär 28 000 meningar (ca 9%) som toxiska mot antingen kvinnor eller män. Av dessa har ett slumpmässigt urval på 5 000 meningar annoterats manuellt för att säkerställa att maskininlärningsmodellens resultat är korrekta och för att vid behov kunna justera resultaten. Den manuella annoteringen syftar också till att avgöra huruvida målpersonen för det toxiska språket faktiskt är en man eller en kvinna. Maskininlärningsmodellen har tränats till hög sensitivitet på bekostnad av specificitet vilket gör att den inte missar så många toxiska kommentarer, men att det krävs en manuell analys för att kunna sålla bort meningar som klassificeraren felaktigt pekat ut som toxiska. Resultaten från den manuella annoteringen visar att män inte bara nämns oftare än kvinnor, utan också utsätts i något högre utsträckning för toxiska kommentarer än kvinnor (figur 2). Här skiljer sig dock resultaten mellan de olika källorna: på Familjeliv och Flashback utsätts kvinnor för en större andel toxiska kommentarer, medan män utsätts i högre grad på Reddit, Twitter och Facebook.



Figur 2. Andel toxiska kommentarer som riktas mot män respektive kvinnor i relation till andelen omnämnanden av kvinnor respektive män.

5. Könsskillnader i det toxiska språkets innehåll

För att undersöka det toxiska språkets karaktär annoterades 5 000 slumpvalda toxiska meningar manuellt och kodades enligt någon av de kategorier som finns beskrivna i avsnitt 2. Resultaten från undersökningen av det toxiska språkets innehåll visar att de största skillnaderna ligger i att kvinnor får fler nedvärderande kommentarer om utseende, psykisk hälsa, moral samt sexualitet och kön (se figur 3). Män får fler direkta och indirekta hot samt fler nedvärderande kommentarer om bristande kompetens. Män utsätts också i något större utsträckning för förolämpningar.



Figur 3. Innehållet i toxiska kommentarer riktade mot män respektive kvinnor.

6. Sammanfattning och diskussion

Resultaten av undersökningen av könsskillnader i antal omnämnanden visar att det generellt sett skrivs mer om män än om kvinnor i de digitala miljöer som undersökts. Även om de olika källorna skiljer sig åt i detta avseende, så nämns män nästan dubbelt så ofta som kvinnor i några av källorna. Sammantaget nämns män i strax över 55% och kvinnor i strax över 35% av de kommentarer som nämner personer. Anledningen till att män nämns oftare än kvinnor skulle kunna vara att fler män har framträdande positioner i många av de ämnen som diskuteras (exempelvis politik).

Undersökningen av andel toxiska kommentarer per omnämnande visar betydligt mindre könsskillnader. Kvinnor utsätts för något fler toxiska kommentarer på Flashback och Familjeliv, medan män utsätts för fler toxiska kommentarer på Reddit, Twitter och Facebook. Sammantaget är andelen toxiska kommentarer per omnämnande 1,1% när kommentarerna riktas mot kvinnor, medan motsvarande siffra för män är 1,5%.

Undersökningen av könsskillnader i toxiska kommentarers innehåll visar att män mer än dubbelt så ofta utsätts för kommentarer som innehåller hot samt vålds- och bestraffningsidéer. Dessa kommentarer består av direkta eller indirekta uttryck för en önskan om att personen ska dö, försvinna eller utsättas för våld. Möjligen är den större andelen direkta och indirekta hot som riktas mot män i digitala miljöer en spegling av den fysiska världen, där män också är mer utsatta för våld i form av misshandel.[10]

Kvinnor utsätts för nedvärderande kommentarer om utseende tre gånger så ofta som män. Dessa kommentarer innefattar påståenden om att en person är ful eller oattraktiv samt objektifierande och närgångna kommentarer om en persons kropp. Kvinnor utsätts också i större utsträckning än män för nedvärderande kom-mentarer om kön och sexualitet. Män blir oftare avsiktligt felkönade, medan kvinnor oftare utsätts för sexuella trakasserier.

Män utsätts i större utsträckning än kvinnor för nedvärderande kommentarer om kompetens. Kvinnor i sin tur utsätts oftare än män för kommentarer som anspelar på psykisk ohälsa i form av vanföreställningar eller oförmåga att reglera känslor. Kvinnors moral ifrågasätts också i större utsträckning än männens.

Resultaten som presenteras här är i linje med tidigare undersökningar som gjorts på svenska sociala medier. I en studie av olika yrkesgruppers utsatthet på ett svenskt

[10] Se exempelvis BRÅ:s statistik om könsfördelning och utsatthet för misshandel: https://bra.se/statistik/statistik-utifran-brottstyper/vald-och-misshandel.html#Konsfordelning

diskussionsforum visade resultaten, precis som här, att sexuella trakasserier och förolämpande kommentarer om utseende drabbade kvinnorna i större utsträckning än männen, medan männen i större utsträckning utsattes för nedvärderande kommentarer kopplat till prestationer eller kompetens. Män drabbades, enligt samma undersökning, också i större utsträckning av hot.[11]

Begränsningar och framtida forskning

Det som skulle mätas i den här studien är ett vagt, värdeladdat och kontextberoende fenomen, vilket medför att vetenskapliga värden såsom till exempel reproducerbarhet blir svåra att upprätthålla. I den här undersökningen har vi definierat toxiskt språk som kommunikation som är förbjuden i lag, men också laglig men kränkande kommunikation. Alla begrepp som ingår i begreppet toxiskt språk är svåra att definiera eftersom de betecknar företeelser som saknar unika avgränsande egenskaper.

Även om det i många fall går att identifiera kommunikation som toxisk eller icke-toxisk kommer det alltid att finnas gränsfall där bedömningen i sista hand beror på normer och värderingar hos den som bedömer. De siffror och procentsatser som återges här bör därför tolkas som ungefärliga och inte exakta.

Avslutningsvis kan konstateras att många av de skillnader som återfinns i den här undersökningen stämmer väl överens med tidigare forskning om könsskillnader i digitala miljöer, samt ytterst även normativa uppfattningar om kvinnor och män: Män är mer utsatta för fysiskt våld, män förnedras genom kommentarer som antyder omanlighet/kvinnlighet, kvinnor förnedras genom kommentarer om utseende, objektifiering och sexualisering, kvinnor utsätts för hot om sexuellt våld. Då resultaten ger en indikation på att det toxiska språket har en direkt koppling till dessa normer eller stereotyper, bör framtida forskning på området fokusera på djupare analyser av toxiskt språk i relation till könsnormer.

Den här studien är gjort inom ramen för det uppdrag som FOI tilldelats av regeringen för att analysera förekomsten av hat och hot mot kvinnor i svenska digitala miljöer. Studien är genomförd av:
Lisa Kaati, FOI
Björn Pelzer, FOI
Katie Cohen, FOI
Daniel Wallgren, FOI
Jenny Yourstone, Södertörns högskola
Nazar Akrami, Uppsala Universitet

FOI:s Data Science-grupp är tvärvetenskaplig forskningsgrupp som bedriver forskning om digitala miljöer och fenomen. För att ta del av mer forskning som vi gör, besök oss på https://www.foi.se/forskning/ledningsteknologi/data-science-gruppen.html

Kontaktinformation
Lisa Kaati, PhD, forskningsledare
lisa.kaati@foi.se

[11] Fernquist, J. et al. (2020) Det digitala hatets karaktär. En studie av hat mot kvinnor och män i utsatta yrkesgrupper. FOI Memo 7429.

FOI Memo: 7741
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria