Toxiskt språk i svenska digitala miljöer

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

När samtalsklimatet i digitala miljöer blir infekterat av kränkande kommunikation äventyras även det demokratiska samtalet. I den här studien undersöker vi några svenskspråkiga digitala miljöer för att få en uppfattning om förekomsten av toxiskt språk, det vill säga kommunikationshandlingar som medför kränkningar mot mottagaren eller en tredje part. Vi undersöker även vilka grupper eller företeelser som de toxiska kommentarerna riktas mot samt förändring över tid i två av Sveriges största diskussionsforum.

6 800 000

Vi har analyserat kommentarer från olika sociala medieplattformar som producerats under ett års tid.

4,9%

Av alla kommentarerna innehåller toxiskt språk

32,4%

Av de toxiska kommentarerna riktas mot individer

25,4%

Av de toxiska kommentarerna handlar om politik eller om enskilda politiker.

10,2%

Av de toxiska kommentarerna riktas mot samhällets institutioner

9,6%

Av de toxiska kommentarerna riktas mot etniska och religösa grupper

1,9%

Av de toxiska kommentarerna riktas mot media och journalister

Kvinnor som grupp utsätts nästan dubbelt så ofta för toxiska kommentarer än män

FOI Memo: 7740
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

2021-12-20

1. Ett toxiskt samtalsklimat

Yttrandefrihet, att var och en ska ha rätt att uttrycka sina tankar och åsikter i ett fritt meningsutbyte, är en mänsklig rättighet och en förutsättning för ett demokratiskt samhälle.[1] Digitala infrastrukturer för kommunikation, så som till exempel sociala medier, förstärker individers möjlighet att utöva denna grundlagsskyddade rättighet.[2] Men när sociala medier används för att kommunicera hotfulla och kränkande kommentarer eller för att nedvärdera individer på grund av deras kön, hudfärg, etnicitet, sexuella läggning, nationalitet, religion, funktionsvariation eller politiska tillhörighet, ställs andra mänskliga rättigheter på spel. Individer som utsätts för den typen av kränkningar upplever ofta psykologiska konsekvenser som kan leda till att de drar sig tillbaka från det offentliga samtalet. I många fall kan de även uppleva att deras egen eller deras familjers säkerhet är hotad.[3] Normalisering av fördomar och intolerans i digitala miljöer kan bidra till ökad diskriminering och fientlighet mot exempelvis religiösa grupper eller etniska minoriteter.[1] Även extrema händelser som till exempel hatbrott, masskjutningar, knivattacker, rekrytering av extremister, bombdåd samt hot mot offentliga personer hämtar allt oftare inspiration från digitala miljöer.

I takt med att samtalsklimatet i digitala miljöer infekteras av kommunikation som uppviglar till eller normaliserar våld, hot och trakasserier, har det uppstått nya termer och begrepp för att beskriva fenomenet. Varken i vardagsdiskursen eller det akademiska samtalet har man enats om benämningar eller definitioner.[4] Begrepp som ”näthat” eller ”hat och hot” förekommer frekvent i nyhetsmedier utan något närmare förtydligande om vad som avses. Termen *hate speech* (ibland översatt som hatretorik eller hatbudskap) används i och utanför den akademiska världen för att benämna ett aggressivt, våldsamt eller kränkande språk som riktar sig mot en specifik grupp människor som delar en gemensam egenskap, exempelvis kön, etnisk grupp, religion eller politiska preferenser.[5] Den alltmer använda termen *dangerous speech*[6] definieras, till skillnad från *hate speech,* utifrån kommunikationens potential för våldsamma konsekvenser snarare än vilka grupper kommunikationen riktas mot. Ett ytterligare sätt att definiera fenomenet är att peka ut de beteenden man anser bör ingå: till exempel kränkande förolämpningar (glåpord), avsiktliga förödmjukelser, stalkning, fysiska hot, trakasserier som sker under en längre tid och/eller sexuella trakasserier.[6]

Vi använder begreppet *toxiskt språk* för att benämna den flora av kommunikationshandlingar som i den ovan nämnda bemärkelsen förgiftar samtalsklimatet i digitala miljöer. Toxiskt språk innefattar kommunikations-handlingar som är förbjudna i lag, så som till exempel hets mot folkgrupp, förtal eller förgripelse mot tjänsteman, men kan även i viss mån inbegripa fall av nedsättande tilltal, integritetskränkning eller respektlöshet.

I den här studien presenteras en undersökning av toxiskt språk i några svenskspråkiga digitala miljöer. Undersökningen utgår från följande frågeställningar:

1) I vilken utsträckning förekommer toxiskt språk på några av Sveriges mest använda sociala plattformar?
2) Vilka är måltavlorna för det toxiska språket?
3) Har mängden toxiskt språk i digitala miljöer förändrats över tid?

2. Undersökningsmetod

Detektion av toxiskt språk

För att kunna identifiera och mäta toxiskt språk har vi använt oss av maskininlärning. Maskininlärning används i det här fallet för att träna en dator att lära sig känna igen toxiskt språk. Vi har utgått från en språkmodell som utvecklats av Kungliga biblioteket[7] och lärt den känna igen toxiskt språk med hjälp av ungefär 6 000 texter, vilka annoterats som antingen toxiska eller inte. Annoteringen gjordes av psykologistudenter från Uppsala universitet. Flera studenter bedömde samma mening och om mer än hälften av studenterna som bedömde en mening var överens om att meningen innehöll hat annoterades den som hat i träningsdatan.

[1] SFS 1991:1469
[2] Castaño-Pulgarín, S.A, Suárez-Betancur, N., Vega, L.M.T.V. and López, H.M.H. (2021). Internet, Social Media and Online Hate Speech. Systematic Review. *Aggression and Violent Behavior, 58*, 101608.
[3] 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
[4] MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N., Frieder, O. (2019). Hate speech detection: Challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152.
[5] Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access, 6*.
[6] https://dangerousspeech.org/
[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.*

FOI Memo: 7740
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

Maskininlärningsmodellen uppnådde en träffsäkerhet på 83% när det kommer till att känna igen toxiskt språk.[8] Modellen fungerar som ett screeningverktyg, vilket genom att sålla bort en stor mängd icke-toxiska kommentarer skapar en mer hanterbar datamängd som kan analyseras manuellt.

Källor

All data vi analyserat kommer från fem svenskspråkiga diskussionsforum och sociala medier. Källorna, som är valda för att ge en bred bild av svenska digitala miljöer med användargenererat innehåll, är: diskussionsforumen Reddit, Flashback och Familjeliv, mikrobloggen Twitter samt kommentarsfält för fem nyhetsmedier på Facebook. De källor som ingår i vår undersökning är beskrivna i tabell 1. Undersökningen baseras på data som har genererats under tidsperioden 2020-07-01 – 2021-06-30.

Undantaget är den del av undersökningen som berör det toxiska språkets förändring över tid. Flashback och Familjeliv är diskussionsforum som funnits sedan början av 2000-talet, vilket också ger oss möjligheten att studera hur mängden toxiskt språk förändrats över tid. För att analysera förändring över tid valde vi slumpmässigt ca 20 000 kommentarer per månad[9] och källa, från 2003 till 2021.

Tabell 1. Källor som ingår i undersökningen

Källa	Beskrivning	Antal kommentarer
Facebook	Användargenererade kommentarer från nyhetssidorna SVT, Aftonbladet, 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

Webbspindlar användes för att hämta in de totalt 6,8 miljoner kommentarerna och lagra dem i en databas. Därefter klassificerades ett slumpmässigt representativt urval[10] bestående av 20 000 kommentarer per källa med hjälp av maskininlärningsmodellen för bedömning av toxiskt språk.

Annotering

Av de kommentarer som maskininlärningsmodellen klassat som toxiska valdes 500 kommentarer per källa slumpmässigt ut för manuell analys, totalt 2500 kommentarer. Den manuella annoteringen har två ändamål: att justera felbedömningar av den automatiska klassificeraren och att bedöma vilka måltavlor det toxiska språket riktas mot. Resultaten presenteras i avsnitt 3 och 4.

Metodologiska begränsningar

Undersökningen är begränsad vad gäller såväl antal källor som tidsperiod för undersökningen. Trots att vi har skapat ett brett urval av de mest välbesökta digitala miljöerna i Sverige är det inte säkert att resultaten går att generalisera till andra digitala miljöer eller andra tidsperioder.

Att känna igen toxiskt språk med automatiserade metoder är en utmanande uppgift som ingen klassificerare kan förväntas behärska till fullo. Därtill är klassificerarens prestanda känslig för variationer i språkbruk mellan olika digitala miljöer. Klassificeraren bör därför endast användas som ett filter inför en manuell granskning. För att optimera resultatet av kombinerad automatisk och manuell screening har klassificeraren tränats till hög sensitivitet på bekostnad av specificitet.[11] Med andra ord är klassificeraren tillförlitlig vad gäller att inte missa toxiska kommentarer. Däremot behövs en manuell analys för att sålla bort kommentarer som klassificeraren felaktigt pekat ut som toxiska.

3. Nivåer av toxiskt språk i digitala miljöer

Efter att inlägg klassificerats i två steg (automatisk extraktion samt manuell annotering, se ovan) kunde nivåerna av toxiskt språk i de olika källorna uppskattas. Skillnaderna mellan de olika datakällorna är stora (se figur 1), med ett medelvärde på strax under fem procent. Delvis kan skillnaderna förklaras i termer av språklig

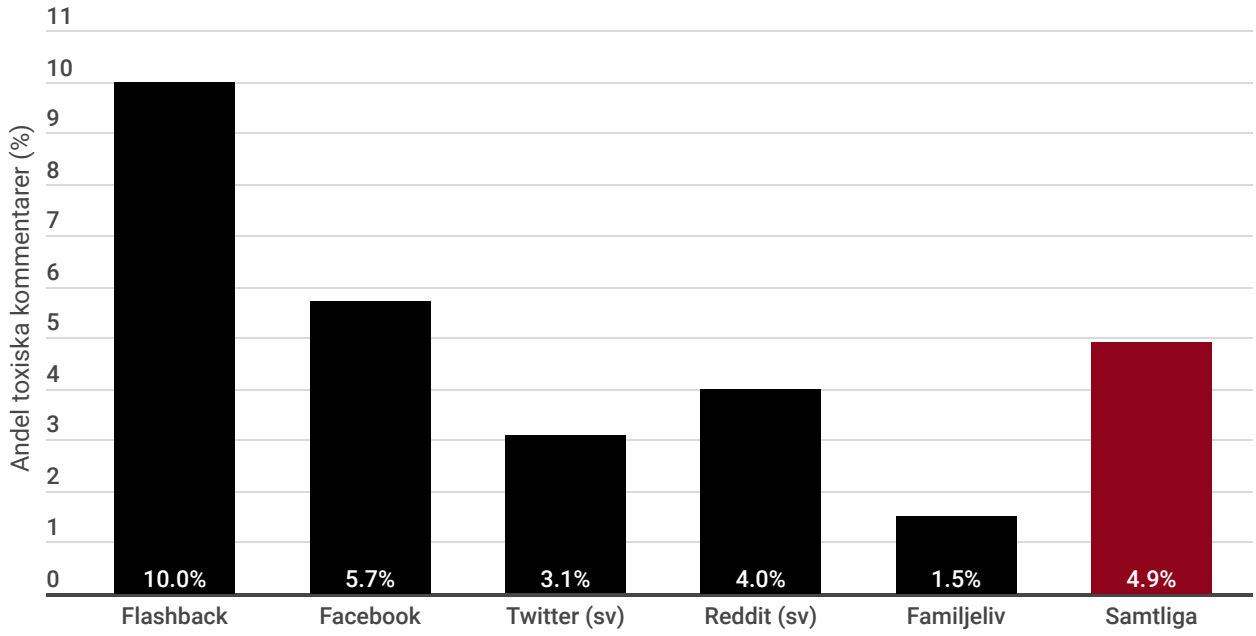
[8] En mer utförlig beskrivning av modellen finns i Fernquist, J., Kaati, L., Cohen, K., Akrami, N., Pelzer, B., Lindberg, S., Sarnecki, P. H. Det digitala hatets karaktär. En studie av hat mot kvinnor och män i utsatta yrkesgrupper Stockholm: FOI, FOI Memo 7429, 2020.

[9] Vanligtvis valdes 20 000 kommentarer per månad. Om färre än 20 000 inlägg hade publicerats under en månad så valdes samtliga inlägg den månaden.

[10] Ett representativt urval av inläggen under den tidsperioden med en konfidensnivå på minst 99% och en felmarginal på högst 1%.

[11] Bara 2,4% av klassificerarens icke-toxiska inlägg visade sig vara toxiska i Fernquist, J. Et al (2020) Det digitala hatets karaktär, FOI Memo 7429, 2020.

jargong. Flashback forum, som i undersökningen har högsta nivån av toxiskt språk, är till exempel känt för att ha vad som skulle kunna beskrivas som en ”rå samtalston”, med bland annat mycket svordomar och hårt tilltal, något som snarare återspeglar en lingvistisk norm än faktisk aggressivitet. Klassificeringsmetoden har dock inte tagit hänsyn till språkliga skillnader mellan olika digitala miljöer. Ytterligare något som kan påverka resultatet är hur och i vilken utsträckning datakällan modereras.



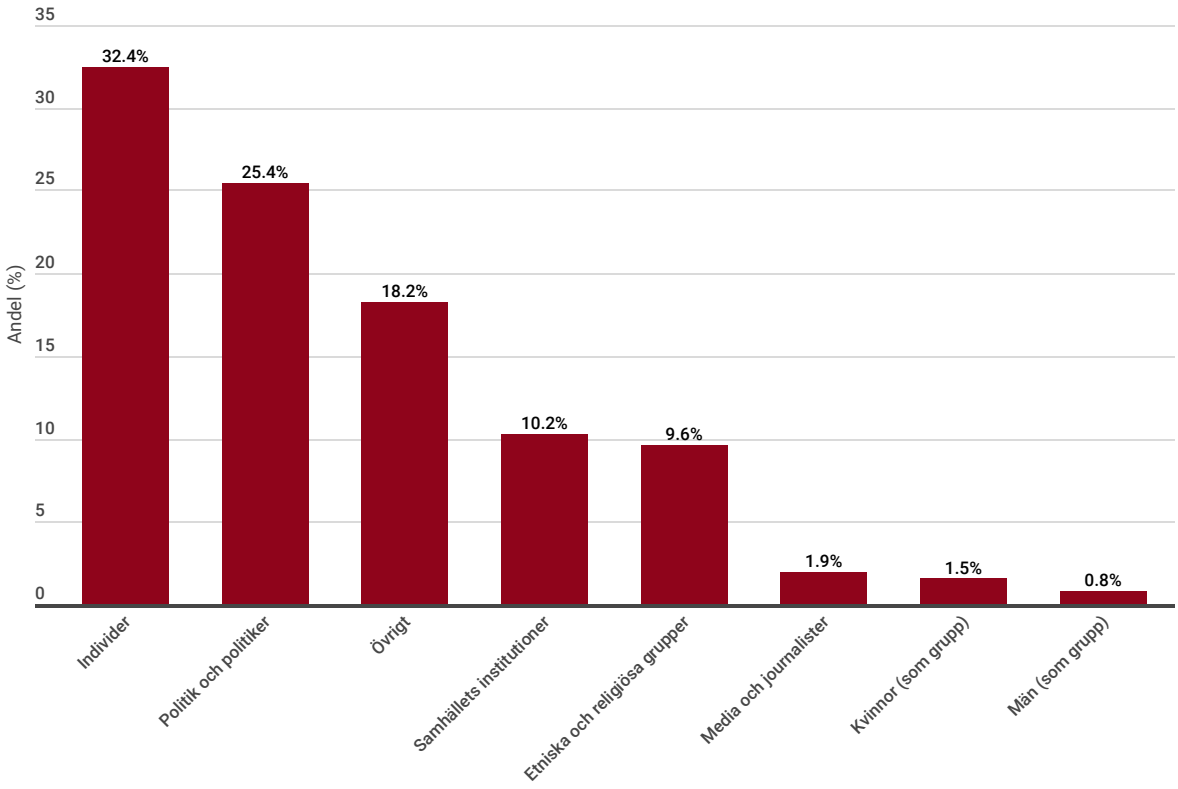
Figur 1. Mängden av toxiska kommentarer i de olika källorna

4. Måltavlor

De 2 500 kommentarerna som valdes ut för manuell analys kategoriserades utifrån *i vilken egenskap* en individ eller grupp blev föremål för de toxiska kommentarerna. Om exempelvis en grupp muslimska kvinnor blir utsatta i egenskap av muslimer kategoriseras det som toxiskt språk mot religiösa grupper, om de blir utsatta i egenskap av kvinnor kategoriseras det som toxiskt språk mot kvinnor. Analysen ledde fram till åtta olika kategorier:

- **Politik och politiker**: toxiska kommentarer som rör politik, politiker, politiska partier och anhängare till politiska partier
- **Samhällets institutioner:** toxiska kommentarer riktade mot företrädare för olika samhällsfunktioner, exempelvis sjukvård, utbildning, rättssystem och regeringen
- **Journalister och media:** toxiska kommentarer riktade mot journalister eller medier
- **Etniska och religiösa grupper**: toxiska kommentarer riktat mot etniska och religiösa grupper exempelvis judar, muslimer, invandrare
- **Kvinnor:** toxiska kommentarer riktade mot kvinnor som grupp
- **Män:** toxiska kommentarer riktade mot män som grupp
- **Individer:** toxiska kommentarer riktade mot enskilda i egenskap av privatpersoner, men även mot offentliga personer som inte tillhör någon av ovanstående kategorier
- **Övrigt:** toxiska kommentarer som inte passar in i någon av ovanstående kategorier

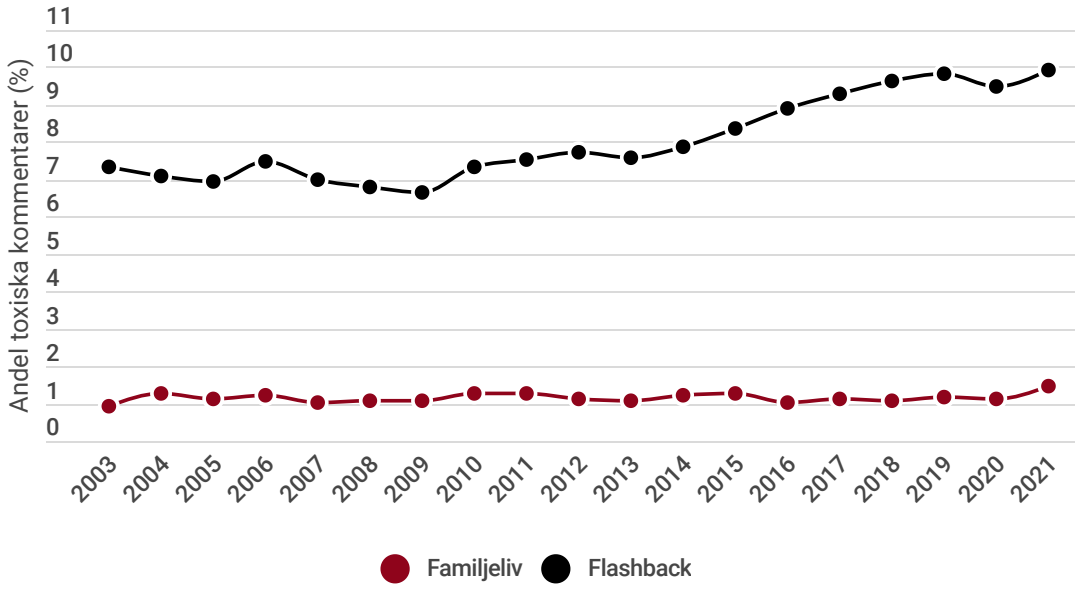
Kategoriseringen (se figur 2) visar att det främst är enskilda individer som blir måltavlor för toxiska kommentarer – dessa utgörs av allt från publika individer, forummedlemmar eller andra personer som diskuteras. I den näst största kategorin, politik och politiker, är 40% av kommentarerna riktade mot namngivna politiker. Även om toxiska kommentarer om kvinnor respektive män på gruppnivå utgör relativt små delar av den totala mängden visar resultaten att kvinnor som grupp utsätts i nästan dubbelt så stor utsträckning som män för toxiska kommentarer.



Figur 2. Måltavlor för det toxiska språket

5. Förändringar över tid

Resultaten från undersökningen av förändringar av nivåer av toxiskt språk över tid visas i figur 3. Resultaten visar att nivån av toxiskt språk har ökat på Flashback men inte på Familjeliv. På Flashback har nivån av toxiskt språk ökat med tre procentenheter: från ca 7% till 10%. På Familjeliv ligger nivån av det toxiska språket runt 1% med mycket små förändringar över tid. Förändringen över tid var inte statistiskt signifikant för något av forumen.



Figur 3. Nivåer av toxiskt språk på Flashback och Familjeliv mellan åren 2002 och 2021.

FOI Memo: 7740
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria

6. Sammanfattning och diskussion

Resultaten av denna undersökning visar att det mellan den första juli 2020 och den 30 juni 2021 har förekommit toxiskt språk i ungefär 5% av alla inlägg i ett urval av svenskspråkiga digitala miljöer. Undersökningen visar också att nivåerna av toxiskt språk varierar mellan 1,5% och 10% beroende på källa.

Måltavlorna för toxiskt språk är till största delen enskilda individer, men kritiken mot politiker, politiska system och samhällsinstitutioner tycks också ha en tendens att bli toxisk. Därtill visar resultaten att det finns nästan dubbelt så många toxiska generaliserande kommentarer om kvinnor än toxiska generaliserande kommentarer om män. Över tid (2003-2021) ligger nivån av toxiskt språk konstant på Familjeliv, medan en statistiskt icke-signifikant ökning kan observeras på Flashback.

Det finns flera faktorer som kan påverka nivån av toxiskt språk i en digital miljö, exempelvis vilka ämnen som diskuteras och i vilken omfattning det sker moderering på forumen. Flashback forum, som i undersökningen visat sig ha den högsta nivån av toxiskt språk, är till exempel känt för att ha vad som skulle kunna beskrivas som en ”rå samtalston”, med bland annat mycket svordomar och hårt tilltal, något som snarare återspeglar en lingvistisk norm än faktisk aggressivitet. Klassificeringsmetoden har dock inte tagit hänsyn till språkliga skillnader mellan olika digitala miljöer.

Flera av undersökningens begränsningar skulle kunna hanteras med hjälp av framtida forskning. Exempelvis skulle tekniker för automatisk analys kunna specialiseras till att mäta toxiskt språk i enskilda digitala miljöer för att underlätta en åtskillnad mellan faktiskt toxiskt språk och toxicitetsliknande lingvistiska normer. Vi skulle också kunna bilda oss en tydligare uppfattning om utbredningen av toxiskt språk genom att mäta hur många individer det är som använder toxiskt språk. Tidigare forskning har till exempel visat att när det gäller toxiska kommentarer riktade mot journalister är det ett fåtal individer som står för en stor andel av kommentarerna.[12] Detta skulle kunna vara fallet även i andra sammanhang.

Den största utmaningen är dock att göra klart vad det egentligen är vi mäter. Som tidigare nämnts betraktar vi toxiskt språk som innefattande kommunikationshandlingar som är förbjudna i lag, så som till exempel hets mot folkgrupp, förtal eller förgripelse mot tjänsteman, men även i viss mån nedsättande tilltal, integritetskränkning eller respektlöshet. Samtliga dessa begrepp saknar unika avgränsande egenskaper som möjliggör en tydlig definition. Trots att vissa kommunikationshandlingar är tydligt identifierbara som toxiska eller icke-toxiska kommer de flesta att vara gränsfall vilkas klassificering i slutänden är beroende av bedömarens normer och värderingar.

Eftersom människor generellt förstår sociala normer på ett intuitivt och omedvetet plan, är de flesta som lever tillsammans i ett samhälle i stora drag överens om hur hårda ord som enligt samhällsnormen är acceptabla i ett offentligt rum (som till exempel ett diskussionsforum), men få kan formulera exakt var gränsen går. Därtill varierar, vilket framgår i jämförelsen mellan Flashback och Familjeliv, normer mellan olika kontexter. På samma sätt som ett grovt språk kan misstolkas som hat och hot, kan grova kränkningar förkläs i en sympatisk “språkdräkt”. De siffror och procentsatser som angetts i denna undersökning bör, liksom alltid när kvantitativa metoder tillämpas på vaga begrepp, tolkas som ungefärliga uppskattningar och inte exakta beräkningar.

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

[12] Kaati, L. Och Olsson, F. (2018) Hatets anatomi. Rapport om hat mot journalister i digitala miljöer. TU Medier i Sverige

FOI Memo: 7740
Forskningsområde: Ledningsteknologi
Godkänd av: Daniel Faria