Ord som sårar

Toxiskt och kränkande språk
i ungas digitala miljöer

Key insights

- Undersökningen visar att politiska diskussioner i digitala miljöer oftare innehåller toxiskt och kränkande språk jämfört med andra diskussioner.

- Mängden toxiskt och kränkande språk skiljer sig åt på olika platser på nätet. Det är mer toxiska kommentarer i svenska diskussionsforum än i influencers kommentarsfält på YouTube, Twitch och TikTok och publika inlägg på Facebook och Instagram.

- I influencers kommentarsfält på YouTube, Twitch och TikTok används ett stötande språkbruk som innehåller könsord och svordomar i högre utsträckning än i publika inlägg på Facebook och Instagram.

- I digitala miljöer förekommer toxiskt och kränkande språk både direkt genom sexistiska eller rasistiska uttryck och indirekt via sarkastiska toner med specifika emojis. En toxisk eller kränkande kommentar kan dessutom helt eller delvis bestå i bilder och GIF:ar, något som samlat gör det svårt för både plattformar och forskare att detektera det som barn och unga identifierar utgör en del av den toxiska och kränkande kommunikationen.

- Att få tillgång till anonymiserad data från plattformar är kritiskt både för forskning och för att säkerställa en trygg digital miljö. Få plattformar erbjuder öppna API:er vilket gör det svårare för forskare att förstå de digitala miljöer barn och unga befinner sig i. Genom att uppmuntra och underlätta för partnerskap mellan plattformar, akademin och andra forskningsinstitutioner kan fler forskningsprojekt få tillgång till värdefull data som ger mer insikter och en bättre förståelse för barns digitala liv.

- Undersökningar som utgår ifrån textanalys av faktiska kommentarer och innehåll som barn och unga möter i digitala miljöer är relativt ovanliga, och det särskilt i en nordisk kontext. Det finns ett behov av att möjliggöra för vidare forskning och att rikta resurser till aktörer som kan ansvara för longitudinell forskning på de plattformar och platser på nätet där barn och unga befinner sig.

- När varje ord räknas, är det ändå inte antalet som spelar roll. Ett enda ord som sårar är ett ord för mycket.

Ord som sårar är en rapport och ett projekt i samarbete mellan Prinsparets Stiftelse, Mediemyndigheten, Internetstiftelsen och Mind Intelligence Lab, med stöd från forskare från Stockholms universitet och Uppsala universitet.

Författare: Johanna Lindström, Prinsparets Stiftelse, Lisa Kaati, Institutionen för data- och systemvetenskap, Stockholms universitet och Nazar Akrami, Institutionen för psykologi, Uppsala universitet

Layout & design: Amelie Borg

Prinsparets Stiftelse 2024

Innehållsförteckning

Foto: Anna Diebitsch Antoni

Förord

När Prinsparets Stiftelse grundades var det utifrån en vision om att alla barn och unga ska ha möjlighet att vara sig själva. På många sätt är nätet en möjliggörande kraft för just det. I allt ifrån gruppchatten med de närmaste klasskompisarna till sociala plattformar där människor från hela världen kan samlas innebär nätet en möjlighet för barn och unga att uttrycka sig, lära mer om sin omvärld, och ytterst att delta i det demokratiska samtalet.

Med den här rapporten vill vi bidra till det samlade kunskapsläget om något som vi ser har en stor påverkan på barns och ungas möjligheter i det digitala: förekomsten av toxicitet och kränkande språk. Oavsett om de används i syfte att avsiktligen kränka eller ingår i en hård jargong och yttras "på skämt", har vissa ord och uttryck särskild makt och riskerar att såra när de får spridning på nätet. Det är vårt ansvar som vuxna att garantera barnets rättigheter även i det digitala, och som del av det öppna upp för dialog om hur vi skapar ett respektfullt, inkluderande samtal på nätet. Den dialogen måste börja i utökad tillgång till data och utökade möjligheter att bedriva forskning för att skapa en tydligare bild av det språk som förekommer i digitala miljöer.

Utöver att möjliggöra för tillgång till data och därmed för mer forskning är det av yttersta vikt att vi skapar förutsättningar för barns verkliga och systematiska delaktighet i arbetet för ökad trygghet på nätet. Rapporten Ord som sårar fokuserar särskilt på förekomsten av toxiskt och kränkande språk, exempelvis i form av rasististiska och sexistiska ord och uttryck, svordomar och könsord, men i samtal med barn och unga är det tydligt att vi behöver ha en bredare förståelse av vad kränkande och toxisk kommunikation kan bestå av i en digital kontext. I det expertperspektiv som åttondeklassare i Sätraskolan bidragit med till rapporten identifierar eleverna bland annat användandet av GIF:ar och emojis som helt eller delvis ersätter nedsättande ord eller uttryck som del av det som uppfattas som toxiskt.

Det ska sägas att det görs och har gjorts stora framsteg i att förhindra spridningen av toxiskt och kränkande språk i digitala miljöer, och det särskilt i fråga om de offentliga forum som rapporten avser. I händerna på kunniga moderatorer kan automatiska och inbyggda filtreringsfunktioner komma åt även omskrivningar och indirekt toxiska kommentarer, särskilt när innehållet har uppmärksammats och anmälts av användare som har identifierat det som skadligt, hatiskt eller på annat sätt kränkande. Samtidigt utgör undersökningen och förekomsten av toxiska och kränkande ord och uttryck som de redovisas i denna rapport en påminnelse om att vi behöver fortsätta arbetet för ökad trygghet på nätet.

Trots automatiserad moderering, moderatorer och möjligheter för användare att flagga och anmäla innehåll förekommer ett toxiskt och kränkande språkbruk i offentliga digitala miljöer. Vi ser att vi behöver motverka spridningen av ord som sårar på flera olika sätt, och det inte minst genom en i forskning grundad dialog om hur vi kan bidra till det respektfulla och inkluderande klimat som möjliggör för barn och unga att ta del av allt det positiva som nätet kan innebära.

Vi är glada och stolta över att ha haft möjligheten att samskapa denna rapport med en ytterst kompetent projektgrupp bestående av forskare ifrån Stockholms universitet, Uppsala universitet, Mind Intelligence Lab, Mediemyndigheten (tidigare Statens Medieråd) och Internetstiftelsen. Vi är lika glada och tacksamma för alla de bidrag med expertperspektiv som finns att läsa i denna rapport, inte minst från Sätraskolans elever. Tillsammans utgör de en ytterligare påminnelse om vikten av att vi har en fortsatt dialog om det toxiska och kränkande språket på nätet. Vi måste agera, tillsammans, för att stärka barns och ungas trygghet på nätet och ytterst deras möjlighet att vara sig själva.

Helene Öberg, generalsekreterare för Prinsparets Stiftelse

Ord som sårar

1. Det toxiska och kränkande språket i digitala miljöer

Internet och sociala medier öppnar upp fantastiska möjligheter för alla att kommunicera och delta i diskussioner när och var som helst. Tillgängligheten medför dock vissa utmaningar. Ett exempel är när de digitala miljöerna fylls av olika typer av toxiskt och kränkande språk.

I denna rapport definieras ett toxiskt och kränkande språk som kommunikation som har potential att förgifta samtalsklimatet på nätet, exempelvis genom att skada eller skapa ett obehag för den som tar plats i det demokratiska samtalet i olika digitala miljöer. Det kan röra sig om både ett mer direkt toxiskt och kränkande språk, exempelvis bestående av uttalade hot och rasistiska kommentarer, och mer indirekt toxicitet. Det senare uttrycks genom sarkasm, ironi och andra kontextbaserade uttryck som tillsammans leder till obehag eller skada. De enskilda orden eller uttrycken i kommentaren behöver i sig inte vara toxiska eller kränkande, men förstås som det i specifika sammanhang eller med en viss ton.

Något som är värt att nämna i fråga om det toxiska och kränkande språket i digitala miljöer är att det i samtal med barn och unga blir tydligt att det inte är begränsat till text. Tvärtom kan kombinationen av en komplimang och särskilda emojis göra att kommentaren i stället ska förstås som kränkande och toxisk. Det toxiska och kränkande språket kan också vara helt bildbaserat, exempelvis genom att bilder eller GIF:ar får ersätta ord. När vi i denna rapport undersöker förekomsten av toxiskt och kränkande språk i digitala fokuserar vi uteslutande på **textbaserad kommunikation**, men har också valt att inkludera resultatet av en diskussion med elever i årskurs 8 på Sätraskolan i Stockholm, där eleverna ger exempel på vad de uppfattar som toxiskt och kränkande.[1]

Alla former av toxicitet och kränkande språk kan påverka den som blir utsatt på olika sätt. Det är till exempel inte ovanligt att den som utsätts för upprepade trakasserier drabbas av psykologiska konsekvenser som kan sänka livskvaliteten och tära på det psykiska välbefinnandet. Ur ett bredare perspektiv finns det tecken på att det upplevt toxiska samtalsklimatet har en inverkan på viljan att utnyttja nätet som demokratisk mötesplats. Enligt Internetstiftelsen uppger mer än hälften av befolkningen över 16 år att de avstår från att uttrycka sina åsikter på nätet av rädsla för näthat.[2] I fråga om barn och unga, som ännu inte kan rösta men som genom nätet bereds en förhållandevis unik möjlighet att delta i det demokratiska samtalet, är detta en särskilt alarmerande konsekvens av det toxiska och kränkande språket.

» Alla former av toxicitet och kränkande språk kan påverka den som blir utsatt på olika sätt. Det är till exempel inte ovanligt att den som utsätts för upprepade trakasserier drabbas av psykologiska konsekvenser som kan sänka livskvaliteten och tära på det psykiska välbefinnandet. «

Rapportens titel, Ord som sårar, syftar till den kraft som ligger i det språk vi använder oss av och som vi tillåter förekomma i samtalet mellan människor. Förekomsten av toxiskt och kränkande språk riskerar att såra, både på individ- och samhällsnivå. Samtidigt, med synliggörandet av hur vanligt förekommande ord som sårar är, har vi en möjlighet att gemensamt diskutera och motverka de negativa konsekvenser vi ser.

1 Läs mer om detta i avsnitt 4.4.
2 Internetstiftelsen (2023), Svenskarna och internet 2023, mer om detta i avsnitt 4.3.

Syftet med rapporten är att analysera och att bidra till det nationella kunskapsläget om hur vanligt förekommande olika typer av toxiskt och kränkande språk är i några av de digitala miljöer där barn och unga i Sverige är aktiva. Genom att analysera kommentarsfält och inlägg på några av de många olika plattformar och de diskussionsforum som barn och unga använder sig av vill vi bidra med ett komplement till befintlig statistik och till en djupare förståelse av vilken typ av kommunikation unga möter på nätet.

Undersökningen utgår från följande frågeställningar:

1. I vilken omfattning förekommer toxiskt och kränkande språk i några av de digitala miljöer som barn och unga i Sverige befinner sig i?

2. I vilken omfattning förekommer ett tufft och hårt språkbruk i några av de digitala miljöer som barn och unga i Sverige befinner sig i?

1.1 Metod & datakällor

Analysmetoder

I undersökningen har vi använt oss av textanalys för att mäta omfattningen av olika typer av toxiskt och kränkande språk i olika digitala miljöer, både andelen toxiska och kränkande kommentarer och hur vanligt förekommande ett tufft och hårt språkbruk är i samma digitala miljöer.

För att mäta förekomsten av toxiskt och kränkande språk har vi använt **Hatescan**, en maskininlärningsmodell utvecklad av forskare på Stockholms universitet och **Mind Intelligence Lab.** Hatescan bygger på en språkmodell som har utvecklats av Kungliga biblioteket[3] och som med hjälp av cirka 22 000 texter som annoterats som antingen toxiska eller inte har tränats till att känna igen toxiskt och kränkande språk.[4]

En kommentar behöver inte nödvändigtvis innehålla kränkande ord eller uttryck för att klassificeras som toxisk, med Hatescan kan vi detektera både sådana direkt toxiska kommentarer och de kommentarer som är att förstå som toxiska exempelvis på grund av kontexten de förekommer i. För att säkerställa att Hatescans klassificering är korrekt har ett representativt stickprov av kommentarer annoterats manuellt och de slutgiltiga resultaten som presenteras här justerats därefter. Med hjälp av Hatescan har vi undersökt mängden toxiska och kränkande kommentarer och inlägg. Utöver det har vi valt att också undersöka hur vanligt förekommande ord som kännetecknar ett tufft och hårt språkbruk är i digitala miljöer, men i detta fall oberoende av om de förekommer i en kommentar som är avsedd att vara toxisk, eller om det snarare är del av en hårdare jargong. I de material där förekomsten av ett tufft och hårt språkbruk har analyserats, har vi inkluderat tre olika typer av ord och uttryck: stötande språk, rasistiska uttryck och könade och sexuella kränkningar.

Med **stötande språk** syftar vi på språkbruk som innehåller svordomar och könsord. **Rasistiska uttryck** omfattar uttryck som kan vara allt från negativa stereotypa uppfattningar om grupper till uttryck där minoritetsgrupper sätts samman med svordomar eller ord som anspelar på minoritetsgrupper och sexuella handlingar. **Könade och sexuella kränkningar** består av kränkande ord som antingen är sexuellt laddade, förminskande, nedsättande eller politiskt laddade. Sexuellt laddade kränkningar syftar på kränkande uttryck som anspelar på kön eller sex. Förminskande eller nedsättande ord är könade kränkningar som ger uttryck för att framförallt kvinnor är svaga, ointelligenta eller har problem att kontrollera sina känslor. Politiskt laddade kränkningar är i detta sammanhang könade kränkningar som också anspelar på politiska åsikter.

I analysen av ett tufft och hårt språkbruk har vi använt oss av en uppsättning ordlistor med ord som kännetecknar den typ av språkbruk som vi vill undersöka. För varje kategori av språkbruk som undersökts har en ordlista med representativa ord tagits fram. För att beräkna omfattningen av de olika kategorierna har vi sökt igenom respektive datakälla och därefter normaliserat resultaten. Normaliseringen innebär att resultaten anges i procent, där förekomsten av ord från respektive kategori sätts i förhållande till det totala antalet ord i respektive datakälla.[5] Normaliseringen görs för att kunna jämföra de olika källorna trots att de har olika storlek och olika mängder text.

Vi har valt att inte återge eller ge konkreta exempel från de texter vi använder i våra analyser för att undvika att reproducera det språkbruk vi undersöker och med hänsyn till den personliga integriteten för skribenter.

3 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.
4 Hatescan har en hög träffsäkerhet (94%) när det kommer till att känna igen toxiskt och kränkande språk i de tester som gjorts under utvecklingen.
5 Exempelvis, om det förekommer tre ord som är toxiska eller kränkande i en datakälla som innehåller 1000 ord, normaliseras det till 0,3 procent.

Tabell 1: De olika källor som ingår i undersökningen.

Plattform	Beskrivning	Tidsperiod	Urval
Youtube	YouTube är ett socialt nätverk där användare laddar upp videoklipp. Utöver att ladda upp klipp kan användaren titta på andras videor och diskutera dem i kommentarsfält.	2022	3 731 110 kommentarer från 24 av de mest populära svenskspråkiga YouTube-kanalerna enligt Svenskarna och internet 2022.[6]
Twitch	Twitch är en social nätverksplats som framförallt används för att livesända dataspel och chatta med andra spelare.	2023	150 767 kommentarer från streamers som är populära enligt Svenskarna och internet 2022.
TikTok	TikTok är en social medieplattform där användaren kan skapa och dela videor, ofta korta klipp. Man kan också följa, gilla och kommentera andras videor.	2023	69 897 kommentarer från ett antal svenska influencers kanaler. Valet av influencers har gjorts baserat på samtal med barn och unga i åldrarna 13-18 år samt en avvägning utifrån engagemang och följarantal.
Reddit	Reddit är diskussionsforum där användaren kan skriva inlägg och diskutera i en mängd olika delforum (subreddits).	220401-230331	100 000 slumpmässigt utvalda kommentarer från de subreddits där diskussionerna sker på svenska.
Reddit politik	Subredditen Svepol, där svensk politik diskuteras.	220401-230331	100 000 slumpmässigt utvalda kommentarer från /r/svepol/
Flashback	Flashback är ett svenskt diskussionsforum med diskussioner inom många olika ämnesområden.	220401-230331	100 000 slumpmässigt utvalda kommentarer från hela Flashback.
Flashback politik	Ett underforum på Flashback som heter Politik:inrikes, där svensk politik diskuteras.	220401-230331	100 000 slumpmässigt utvalda kommentarer från Politik:inrikes https://www.flashback.org/f77
Instagram	Instagram är en social medieplattform där användaren kan dela inlägg med bilder och korta filmklipp. Användaren kan också kommentera andras inlägg.	221111-231111	45 744 inlägg (inte kommentarer) som nämner ämnen som intresserar unga enligt Ungdomsbarometerns generationsrapport.[7]
Facebook	Facebook är en social medieplattform där användaren bl.a. kan gå med i intressegrupper och publicera inlägg. Det är också möjligt att följa, gilla och kommentera andras inlägg.	221111-231111	224 951 inlägg (inte kommentarer) som nämner ämnen som intresserar unga enligt Ungdomsbarometerns generationsrapport.[8]

6 Internetstiftelsen (2022), Svenskarna och internet 2022.
7 Ungdomsbarometern, Generationsrapporten 2023: Generation Z.
8 Ungdomsbarometern, Generationsrapporten 2023: Generation Z.

Datakällor

De datakällor som analyserats i undersökningen kommer från ett flertal olika svenskspråkiga källor från olika plattformar. Texterna kommer dels från kommentarsfälten till några av de mest populära svenska kreatörerna, dels från olika diskussionsforum. Snarare än att vara utvalda för direkta jämförelser källor emellan är datakällorna valda för att ge en bred bild av olika svenska digitala miljöer med användargenererat innehåll som barn och unga i olika åldrar möts av, och för att möjliggöra för en diskussion kring eventuella skillnader mellan dessa olika plattformar och diskussionsforum. Det ska betonas att de olika plattformarna används olika mycket och av olika åldersgrupper, där vissa åldersgrupper knappt alls använder sig av vissa av de analyserade källorna/plattformarna. Den data som har använts beskrivs i tabell 1.

Den data som kommer från svenska kreatörers kommentarsfält är inhämtad från YouTube, Twitch och TikTok. Både YouTube och TikTok är två av de sociala medier som personer födda på 2000-talet använder mest, exempelvis använder 62 respektive 65 procent av 00- och 10-talisterna YouTube dagligen. Av 00-talisterna är det 57 procent av 00-talisterna som använder TikTok dagligen. Streamingplattformen Twitch är störst bland 00-talisterna, men används av en mindre andel motsvarande 20 procent minst en gång per år. För såväl YouTube som Twitch och TikTok gäller att andelen användare som är 00- och 10-talister är överrepresenterade i förhållande till den totala användarbasen.[9]

Vi har även analyserat data från två av de största diskussionsforumen i Sverige, Flashback och Reddit, samt öppna inlägg från Facebook och Instagram som nämner ämnen som intresserar unga. Gemensamt för samtliga dessa plattformar är att de unga som besöker dem ingår i en äldre användarbas än på tidigare nämnda plattformar, TikTok, YouTube och Twitch. Av 00-talisterna och 10-talisterna är det 52 respektive fyra procent som använder Facebook minst en gång om året, och Instagram 81 respektive 22 procent. Med undantag för 00-talisternas användande av Instagram utgör detta signifikant färre användare än den totala andelen användare hos hela befolkningen. Flashback och Reddit används i en mycket liten utsträckning av de yngre barnen, men engagerar desto fler unga och unga vuxna. Av 00-talisterna använder 20 procent Flashback minst en gång om året, och för Reddit är motsvarande siffra 26 procent. Till skillnad från Facebook är 00-talisterna på Reddit fortsatt en signifikant större grupp än den totala andelen användare i alla åldrar, men delar detta med både 90- och 80-talisterna som också utgör en större andel av användarbasen. På Flashback utgör 00-talisterna en relativt jämnstor andel med användandet bland hela befolkningen. Istället är det 70- till 90-talisterna som utgör den signifikant större gruppen användare.[10]

1.2 Begränsningar

Målsättningen med undersökningarna har varit att analysera ett brett urval av data från olika plattformar och med fokus på ämnen som engagerar unga. I den data som kommer från kommentarsfälten tillhörandes svenska kreatörers kanaler är det viktigt att komma ihåg att plattformarna erbjuder verktyg för moderering. Kreatörer har därför möjligheten att filtrera bort toxiska kommentarer i sina kanaler. Vi har inte haft tillgång till de kommentarer som efter filtrering inte har publicerats, de kommentarer som skickats via direkta meddelanden till kreatörerna själva, eller till andra användare som en reaktion på inläggen. Detta återspeglas i våra mätningar som snarare ska ses som toppen på ett isberg än en heltäckande analys av förekomsten av toxiska kommentarer. I vår undersökning ingår således bara de kommentarer som inte fastnat i en automatisk eller manuell moderering av kreatörerna, och som förekommer i det offentliga kommentarsfältet till inlägget.

Med anledning av att den språkmodell som Hatescan bygger på är textbaserad omfattar inte undersökningen kommentarer som uteslutande består av s.k. emojis, och heller inte kommentarer bestående i bilder eller GIF:ar.

Inhämtningen av data från Instagram och Facebook har gjorts med Crowdtangle, Metas verktyg för journalister och forskare. Med Crowdtangle kan vi hämta in publika poster på Facebook och Instagram men däremot inte de tillhörande kommentarerna. Detta utgör en stor begränsning eftersom det enbart är texten i inläggen, inte kommentarerna som ingår i undersökningen.
Datan är från olika tidsperioder men har publicerats under 2022-2023. Vi har inte kontrollerat resultaten för eventuella händelser som kan ha påverkat innehållet i kommentarerna och diskussionerna.

9 Internetstiftelsen (2023), Svenskarna och internet 2023, tabell 9.6a och 9.6b.
10 Internetstiftelsen (2023), Svenskarna och internet 2023, tabell 9.6a.

2. Resultat

2.1 Toxiskt och kränkande språk i barn och ungas digitala miljöer

I den första undersökningen har vi analyserat mängden toxiska inlägg och kommentarer i flera olika svenskspråkiga källor: Twitch, TikTok, YouTube, Reddit, Flashback, Facebook och Instagram. Källorna skiljer sig åt på många sätt, både i hur de modereras och vilken typ av data vi har haft tillgång till. Vad gäller Flashback och Reddit har vi dessutom analyserat eventuella skillnader mellan forum som avser politiska diskussioner och andra, mer allmänna diskussioner.

Den data vi har analyserat från Twitch, TikTok och YouTube består av kommentarer som återfinns i svenska kreatörers kanaler. Såväl Twitch[11], TikTok[12] som YouTube[13] använder sig av olika automatiska filtreringsfunktioner. Det innebär att kommentarerna med stor sannolikhet är modererade både av kreatörerna och av plattformarna, exempelvis genom automatisk filtrering av utvalda ord.

I fråga om den data från Facebook och Instagram som har analyserats är den begränsad till publika inlägg som nämner ämnen som intresserar unga, men innefattar på grund av verktygets Crowdtangles utformning inga kommentarer till inläggen. Även Facebook och Instagram, som båda ägs av Meta, använder sig av automatiska filtreringsfunktioner.[14]

Reddit och Flashback är diskussionsforum där alla möjliga ämnen diskuteras. De båda forumen skiljer sig från övriga källor i undersökningen på så sätt att de i högre grad förlitar sig på moderatorer än på automatiska filterfunktioner eller en kombination av de båda. Modereringen skiljer sig åt på de olika delforumen, där varje delforum har egna regler.[15]

Vår analys visar att diskussionsforumen Flashback och Reddit har den största andelen toxiskt språk överlag. Flashback har mer än dubbelt så stor andel toxiska kommentarer som Reddit. På övriga plattformar är andelen toxiska kommentarer mellan 0,8 och 2,5 procent av antalet kommentarer. Bland de plattformar som på olika sätt erbjuder och tillämpar automatisk moderering har YouTube den största andelen toxiska kommentarer följt av Twitch och Facebook. Den minsta andelen toxiska kommentarer återfinns på TikTok och Instagram.

Toxiskt och kränkande språk i politiska diskussioner och i inlägg om ämnen som intresserar unga

För många, och särskilt den yngre generationen, är digitala miljöer en arena för politiska samtal och diskussioner. Bland förstagångsväljarna 2022, däribland de äldsta 00-talisterna tog 39 procent del av politiskt innehåll på sociala medier dagligen, då framför allt via Instagram och Facebook.[16] Nätet och digitala miljöer bereder en relativt unik möjlighet för unga, och då särskilt de unga som ännu inte är röstberättigade, att delta i olika politiska diskussioner och engagera sig i olika samhällsviktiga ämnen. Samtidigt är möjligheten att diskutera politik på sociala medier inte alltid en uteslutande positiv upplevelse för samma målgrupp. I en undersökning från valet 2018 framkom att en tredjedel av de tillfrågade 16- till 25-åringarna avstod från att uttrycka sin politiska åsikt för att undvika hård kritik, hat eller hot.[17] Toxiskt språk verkar vara vanligare i diskussioner om politik, något som även en undersökning av diskussioner på sociala medier under riksdagsvalen 2022 visade.[18]

För att få en uppfattning om hur vanligt förekommande toxiskt och kränkande språk är i den mer specifika kontexten politiska diskussioner, och i inlägg om ämnen som intresserar unga, har vi därför analyserat inlägg på Facebook och Instagram enligt vad som har redovisats ovan, och där andelen toxiska och kränkande kommentarer på plattformarna uppgick till 1,4 respektive 0,8 procent.

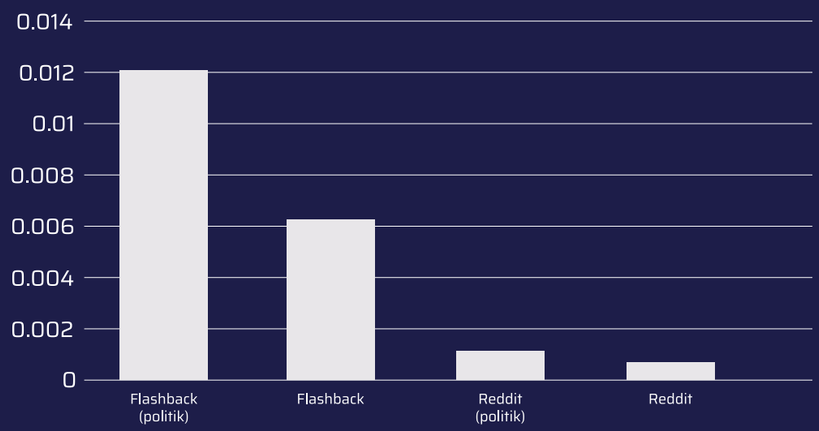
Vi har dessutom undersökt två specifika forum för diskussioner om politik: Svepol, en subreddit för diskussioner om svensk politik, samhälle och kultur samt Flashback-forumet Politik:inrikes. När innehållet i dessa specifika och politiska forum jämförs med andra forum på samma plattform, men som inte avser politiska diskussioner, bereder det tillfälle att undersöka den politiska kontextens inverkan på det toxiska och kränkande språket. Resultaten visar att 9 procent

11 Twitch, 2024. Chat Tools (twitch.tv) (Hämtad 19 juni 2024)
12 TikTok, 2024 Vår syn på innehållsmoderering | TikTok Vår syn på innehållsmoderering (Hämtad 19 juni 2024)
13 YouTube, 2024 Granska och svara på kommentarer - Dator - YouTube Hjälp (google.com) (Hämtad 19 juni 2024)
14 Meta, How technology detects violations | Transparency Center (meta.com) (Hämtad 19 juni 2024)
15 På Reddit modereras varje delforum av s.k. "mods" och i enlighet med uppsatta riktlinjer för delforumet.
16 Internetstiftelsen (2022), Svenskarna och internet: Valspecial 2022, diagram 1.7b.
17 Internetstiftelsen (2018), Valspecial 2018.
18 Kaati, Lisa, och Shrestha, Amendra (2023). Digitala diskussioner och de svenska valen 2022. Stockholm: Stockholms universitet.

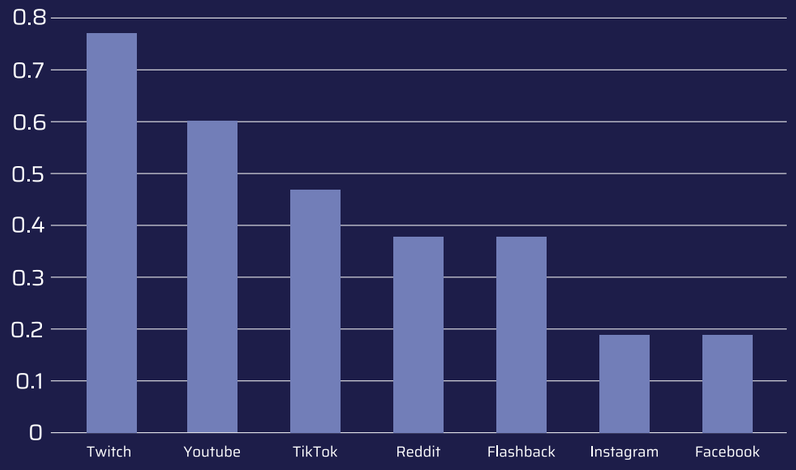
Andelen toxiska och kränkande kommentarer och inlägg (i %) i olika digitala miljöer

Andelen toxiska och kränkande inlägg (i %) underforum på Reddit och Flashback där politik diskuteras

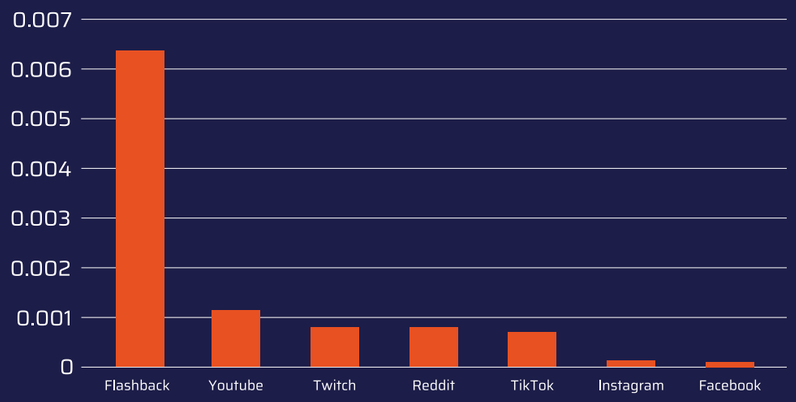
Figur 1: *Andelen rasistiska uttryck (i %) på hela Reddit och Flashback och på de underform där politik diskuteras.*



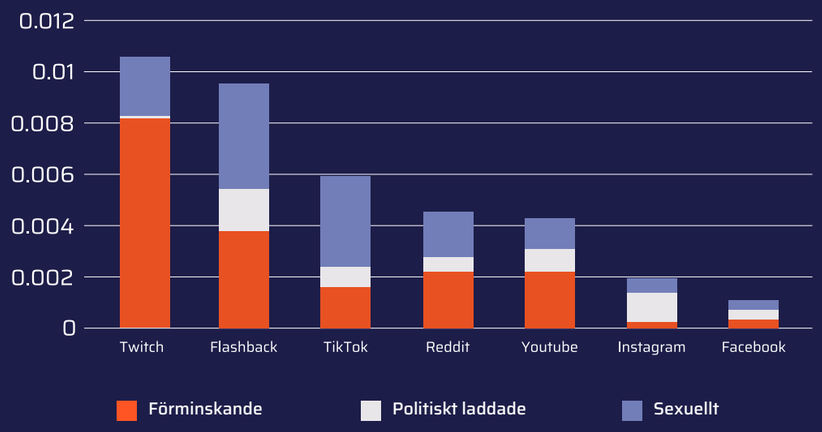
Figur 2: Andelen ord som har stötande karaktär (i %) av den totala mängden ord i de olika datakällorna.

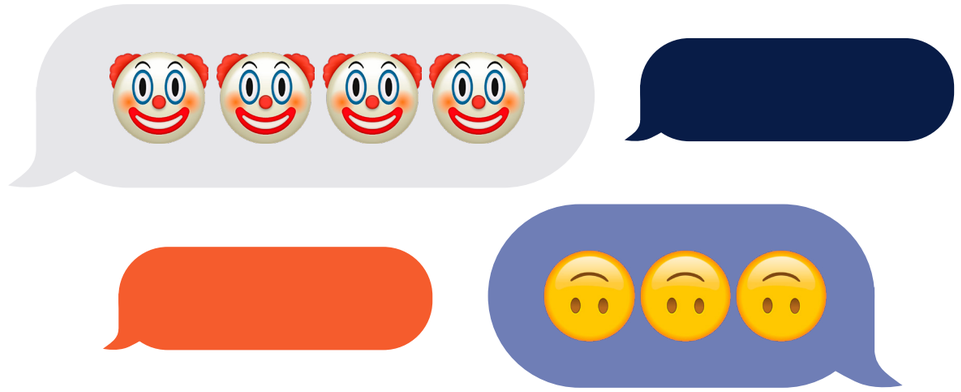


Figur 3. Andelen ord som har rasistisk karaktär (i %) av den totala mängden ord i de olika datakällorna.



Figur 4: Andelen ord (i%) som anses vara könade kränkningar av den totala mängden ord i de olika datakällorna.





respektive 10 procent av kommentarerna på Reddit Svepol respektive Flashbackforumet Politik:inrikes var toxiska. Det innebär att det är mer än dubbelt så mycket toxiskt språk på Reddit Svepol i jämförelse med resten av svenskspråkiga Reddit. På Flashback är skillnaden inte alls lika stor, 9,4 procent att jämföra med 10 procent.

För att öka förståelsen för vad i det toxiska språket består av har vi valt att särskilt mäta förekomsten av rasistiska uttryck både i de underforum som är dedikerade till politiska diskussioner samt på hela forumen. Värt att notera är att det enbart är delar av den synliga rasismen som fångas upp i våra mätningar. Resultaten visade att rasistiska uttryck förekommer oftare i politiska diskussioner – än på forumen generellt - både på Reddit och Flashback. Flashback har en högre andel rasistiska uttryck i diskussioner än vad som återfinns finns på Reddit.

2.2 Ett tufft och hårt språkbruk

Språket är i ständig utveckling och många rapporterar om ett allt hårdare språkbruk i sociala medier, serier och musik som fångas upp av unga. Den här undersökningen syftar till att synliggöra förekomsten av ett tufft och hårt språkbruk i digitala miljöer, med fokus på ett stötande språk, rasistiska uttryck och könade och sexuella kränkningar.

Stötande språk

Stötande språk är språkbruk som innehåller svordomar och könsord. Våra analyser visar att kommentarerna på Twitch har den högsta andelen stötande språk, följt av YouTube, TikTok, Reddit och Flashback. Lägst andel stötande språk återfinns i de publika inläggen på Instagram och Facebook, se Figur 2.

Rasistiska uttryck

Många av de stora sociala medieföretagen har användarregler som förbjuder användare att publicera rasistiskt innehåll men trots det publiceras det hotfulla och kränkande kommentarer som anspelar på användares hudfärg, etnicitet och religion. De rasistiska uttryck som vi har mätt är uttryck som omisskännligen är fientliga och kränkande och omfattar uttryck som kan vara allt från negativa stereotypa uppfattningar om grupper till sammanslagningar av minoritetsgrupper och svordomar eller ord som anspelar på minoritetsgrupper och sexuella handlingar.

Resultaten av undersökningen av förekomsten av rasistiska uttryck visas i Figur 3. Flashback har överlägset den största andelen rasistiska uttryck följt av YouTube och Twitch. De publika inläggen på Instagram och Facebook har den minsta andelen rasistiska uttryck.

Könade och sexuella kränkningar

Könade och sexuella kränkningar är kränkningar kopplade till kön eller könsidentitet. I vår undersökning av förekomsten av könade kränkningar har vi valt att utgå från tre olika typer av kränkande uttryck: förminskande eller nedsättande, sexuellt laddade, samt politiskt laddade. Resultaten av undersökningen visas i figur 4. Sammantaget förekommer könade och

sexuella kränkningar främst på Twitch och Flashback. Kommentarsfälten på Twitch har störst andel förminskande eller nedsättande kränkningar följt av forumen Flashback och Reddit. Instagram och Facebook har den minsta andelen könade förminskande eller nedsättande kränkningar. När det kommer till sexuellt laddade kränkningar har Flashback den största andelen följt av TikTok. Könade politiskt laddade kränkningar förekommer i störst omfattning på Flashback och Instagram och minst på Twitch.

3. Diskussion

Resultaten av undersökningen visar att mängden toxiskt språk skiljer sig mellan olika digitala miljöer. I kommentarsfälten i de kanaler som de svenska kreatörer som ingår i vår undersökning har, varierar andelen toxiska kommentarer mellan 0,8 – 2,5%. Den största andelen toxiska kommentarer återfinns på YouTube och den minsta andelen på TikTok. Publika inlägg på Facebook och Instagram har liknande nivåer av toxiskt språk. Facebook har 1,4% toxiska inlägg jämfört med Instagram som ligger på 0,8%. På svenska diskussionsforum förekommer mer toxiskt språk. Flashback har den högsta andelen av toxiskt språk i vår undersökning, följt av Reddit.

Som tidigare nämnts är det, på grund av de många skillnaderna i plattformarnas utformning och målgrupper, svårt att göra direkta jämförelser. Däremot kan resultaten ge en indikation på hur det toxiska och kränkande språket tar sig uttryck i en viss digital miljö. Tänkbara skillnader som kan ha en inverkan på de olika stora andelarna toxiskt och kränkande språk kan exempelvis vara att det finns en skillnad i hur de olika plattformarna modereras, där TikTok med den minsta andelen använder sig av automatiserad moderering och diskussionsforum som Flashback istället modereras av användare.

I den del av undersökningen som mer specifikt fokuserar på den politiska kontexten och jämför den med en allmän kontext, framgår att de politiska diskussionerna på Reddit är toxiska dubbelt så ofta som i övriga diskussioner. Även på Flashback är politiska diskussioner mer toxiska även om skillnaden inte är lika stor. Vår undersökning visar dessutom att rasistiska uttryck förekommer i en större omfattning i politiska diskussioner än i resten av forumen. Både på Flashback och Reddit förekommer rasistiska uttryck nästan dubbelt så ofta i de politiska diskussionerna.

Den andra undersökningen utgår från en frågeställning om hur vanligt förekommande ett tufft och hårt språkbruk är i de olika sociala plattformar som barn och unga använder, oberoende av kontexten orden förekommer i. Visserligen innehåller kommentarsfälten till svenska kreatörer en mindre andel toxiskt språk än vad som återfinns i diskussionsforum, men undersökningen visar att de istället innehåller en större andel tufft och hårt språkbruk, framförallt på Twitch, YouTube och TikTok. Som en möjlig förklaring till detta kan det vara att det finns en upplevd högre acceptans för en jargong som innefattar svordomar och könsord i samband med spelande som livestreamas (ett vanligt innehåll bland de största kreatörerna på både Twitch och YouTube).

Rasistiska uttryck förekommer på alla sociala medier men i störst omfattning på Flashback följt av YouTube och Twitch. Det är tydligt att rasistiska uttryck förekommer oftare i kommentarsfält och diskussionsforum än i publika inlägg som görs på Instagram och Facebook.

När det kommer till könade kränkningar ligger Twitch i topp, följt av Flashback. Twitch har den största andelen förminskande/nedsättande könade kränkningar medan Flashback har den största andelen politiskt laddade könade kränkningar och sexuella kränkningar.

Resultaten i den här undersökningen bör, liksom alltid när kvantitativa metoder tillämpas på svårdefinierade och svåravgränsade begrepp, tolkas med försiktighet. Precis som ett grovt språk kan missuppfattas som en kränkning kan också grova kränkningar gömmas med hjälp av språket. Resultaten ska därför tolkas som ungefärliga uppskattningar och inte som exakta beräkningar.

4. Expertperspektiv

I detta rapportavsnitt har vi samlat ett antal expertperspektiv på frågan om spridningen av toxiskt och kränkande språk i barn och ungas digitala miljöer. Perspektiven syftar till att ytterligare fördjupa förståelsen bland annat kring befintlig forskning på området trygghet på nätet för barn och unga, om vad barn och unga uppfattar att toxicitet och kränkningar på nätet består i, och inte minst, vikten och värdet av att upprätthålla det demokratiska samtalet på nätet.

4.1 Insikter och utmaningar i bekämpningen av toxiskt språk bland barn och unga i digitala miljöer

I snart ett decennium har Prinsparets Stiftelse arbetat för ökad trygghet på nätet och följt barn och ungas liv i det ständigt föränderliga digitala landskapet. En insikt som har blivit tydlig genom åren är att det finns en bristande kunskap och information om vad barn och unga faktiskt möter i sin vardag online. I de analyser och insatser som samhället (inklusive vi själva) utformar, förlitar vi oss i stor utsträckning på att barn och unga identifierar vad som är att se som näthat och utsatthet och inte. Samtidigt vet vi utifrån samtal med barn och unga att det finns en upplevd gråzon mellan det som å ena sidan är olagligt, skadligt eller kränkande, och det som å andra sidan inte är det. I den gråzonen hör vi barn och unga beskriva att den som väljer att vara på nätet “får räkna med lite skit” och att det “hör till” med en viss, och hård jargong. För att överbrygga detta, och för att skapa interventioner med grund i en bred bild av problematiken, behöver vi som komplement den insyn och tillgång till faktisk anonymiserad samtalsdata från de populära apparna och spelen som vi idag saknar.

När vi inledde projektet tillsammans med Internetstiftelsen och Mediemyndigheten (då Statens Medieråd) hade vi som mål att genomföra undersökningen och färdigställa en första rapport på ett halvårs tid. Nu, två år senare, är vi i mål och flertalet viktiga lärdomar rikare, men framförallt har vi kommit till en avgörande insikt för arbetet med ökad trygghet på nätet framåt: det får inte vara så här svårt att bedriva forskning och få tillgång till data om barn och ungas nätvardag. Projektet har varit betydligt mer tids- och resurskrävande än vi någonsin kunnat ana, just eftersom det har varit så svårt att få tillgång till data från plattformsbolagen. Många plattformar uttrycker att de vill öppna sina API:er för forskare och därmed bjuda in till forskning i den digitala världen av kommentarer och gruppchattar, och även den nya EU-lagstiftningen Digital Services Act ställer krav på att plattformar ska dela med sig av data, men i praktiken har detta visat sig vara mycket svårt att tillgå. Utan möjlighet att få tillgång till större mängder data genom plattformarnas försorg fick vi istället samla in den genom ett antal olika tillvägagångssätt, allt försvårat av de varierande algoritmer och andra funktioner som påverkar vad som kan sägas utgöra en helomspännande bild av “barn och ungas digitala miljöer”.

Att våra analyser visar en relativt låg andel toxiskt och kränkande språk är inte förvånande. Många plattformsföretag har satsat betydande resurser på att moderera och aktivt radera skadligt och toxiskt innehåll. Det som däremot kan tyckas förvånande är hur mycket toxicitet barn och unga vittnar om och som passerar under radarn i form av ironiska kommentarer, emojis, GIF:ar och andra subtila signaler. Detta understryker behovet av en mer nyanserad och djupare förståelse av kränkande kommunikation i digitala kontexter.
Det är också anmärkningsvärt att undersökningar där vi

genom textanalys undersöker vad barn och unga faktiskt ser och konsumerar i det digitala, inte finns in någon större utsträckning. Studier som ”Children’s Media Lives”, genomförda av brittiska Revealing Realities på uppdrag av Ofcom[19], som undersöker barns medievanor i deras hem och naturliga miljöer, är därför något vi välkomnar mer av, både i Sverige och globalt. Dessa studier ger en djupare förståelse för barns medialiv och kan bidra till att utforma effektiva strategier för att bekämpa toxiskt språk och kränkningar på nätet. För att kunna genomföra effektiva interventioner behöver vi veta vad barn och unga faktiskt upplever och möter i det digitala landskapet.

Sammanfattningsvis är det avgörande att vi fortsätter att fördjupa vår kunskap om barns och ungas digitala vardag, både genom att lyssna på dem och genom att få tillgång till den data som kan ge oss en fullständigare bild. Endast då kan vi skapa de förutsättningar som krävs för att säkra en tryggare och mer inkluderande digital miljö för framtidens generationer.

Essi Alho, verksamhetschef, Prinsparets Stiftelse

4.2 Forskning finns, men ännu mer behövs

Det digitala medielandskapet utgör i dag en viktig uppväxtmiljö för barn och unga. Sociala medier som TikTok och Snapchat, eller YouTube, erbjuder underhållning, information och inblickar i världar långt bortom den egna vardagens närmiljö. I digitala spel finns utmaningar, lek och möjligheter att interagera med kompisar från när och fjärran. Influencers, YouTubers och streamers är vanliga inslag i barns och ungdomars medievardag och utgör centrala gestalter i dagens barn- och ungdomskultur. Allt detta kan berika barns uppväxt.

Men det finns också problem förknippade med den digitala uppväxtmiljön, vilket föreliggande rapport om toxiskt och kränkande språk visar. Under senare år har det publicerats åtskilligt med forskning internationellt om medieanvändning och psykisk hälsa, studier som ibland påvisat statistiska samband mellan mer omfattande användning av till exempel mobilen eller sociala medier och olika former av psykiska besvär.

Statens medieråds egna undersökningar har visat att ungdomar med nedsatt psykiskt välbefinnande följer influencers i högre grad än andra ungdomar (rapporten Unga, medier och psykisk ohälsa 2022). Det är även väl belagt i forskningslitteraturen att utsatthet på nätet för hot, mobbning, elakheter och sexuella trakasserier korrelerar med psykisk ohälsa bland barn och unga.

Korrelationer är dock inte detsamma som orsakssamband och på ett så brett och komplext område som psykisk hälsa och medieanvändning är det möjligt att sambanden går i olika riktningar och med påverkan från åtskilliga bakgrunds- och mellanliggande faktorer. Forskning pågår för fullt och Folkhälsomyndigheten och Mediemyndigheten har arbetat på regeringens uppdrag med att ställa samman kunskap på detta område. Ofta försöker forskningen fokusera på aspekter som kan avgränsas, kontrolleras och mätas – till exempel ”skärmtiden” – eftersom det krävs för att orsakssamband ska kunna fastställas och förklaras. Men det innebär inte att andra, mer undflyende och svårundersökta aspekter av mediernas innehåll och form är mindre viktiga. Särskilt inte om vi också vill förstå den psykiska ohälsans

19 Waldie et.al. (2017). Children’s Media Lives. London: Revealing Reality.

komplexitet. Vilket språk och vilka attityder som kommer till uttryck på plattformar där barn och unga rör sig – hela den digitala uppväxtmiljön – har sannolikt betydelse för det psykiska välbefinnandet bland barnen, men vi vet inte exakt hur. Fortsatta studier av vilka språkbruk, visuella koder, normer, förhållningssätt och attityder som karaktäriserar digitala plattformar för barn och unga skulle därmed kunna bidra med värdefull kunskap på området psykisk hälsa och medieanvändning.

Den 1 januari 2024 slogs Statens medieråd och Myndigheten för press, radio och tv samman och bildade Mediemyndigheten. Mediemyndigheten koordinerar Safer Internet Center Sverige som arbetar med att på olika sätt
uppmärksamma barns och ungas liv på nätet.
Dessa frågor är högt upp på agenda hos samtliga medlemsländer i EU, liksom i andra delar av världen. Att höja kunskapen hos berörda myndigheter och organisationer, likväl som hos yrkesverksamma som möter barn, unga och föräldrar, är avgörande för att barns digitala liv ska vara trygga, roliga och givande. Denna rapport bidrar med ny kunskap som kommer att vara till nytta i centrets arbete framåt.

Yvonne Andersson, PhD, analytiker, Mediemyndigheten (tidigare Statens Medieråd)

4.3 Det toxiska språkets tystande kraft

Internetstiftelsen är en oberoende och allmännyttig organisation. Vi verkar för ett internet som bidrar positivt till människan och samhället. Vi ansvarar för internets svenska toppdomän .se och sköter drift och administration av toppdomänen .nu. Genom intäkter från vår affärsverksamhet stöder vi olika initiativ för att öka digital kompetens, skapa plattformar för kunskapsdelning och fortbildning. Dessutom driver och finansierar vi forskning samt producerar rapporter för att ge samhället och näringslivet en gedigen kunskapsgrund om digitaliseringens påverkan på individ och samhälle.

Våra årliga rapporter, Svenskarna och internet samt Barnen och internet, belyser både möjligheter och utmaningar med internet och digitalisering för både barn och vuxna. Sverige är ett land där digitaliseringen är väl etablerad, med 96 % av befolkningen som använder internet, varav 9 av 10 gör det dagligen. Nästan alla barn i åldern 8-19 år är aktiva på nätet varje dag. Vi spenderar en betydande del av vår tid i digitala miljöer där interaktion med andra är vanligt förekommande. I princip alla barn använder sociala medier dagligen, och nästan 7 av 10 barn umgås med andra genom spelplattformar. Digitala plattformar erbjuder på många sätt fantastiska möjligheter för barn att socialisera, lära sig nya saker och utvecklas. Trots detta finns det också baksidor – både barn och vuxna utsätts för hot och hat. Hård ton och ett toxiskt och kränkande språkbruk har på många sätt blivit en vardag på nätet.

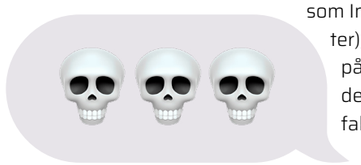
Drygt 4 av 10 har sett andra utsättas för näthat, unga har i högre grad sett andra drabbas. Dessutom är barn och unga överrepresenterade när det gäller att själva bli utsatta. Drygt var sjunde barn i åldern 12–19 år (15 %) har upplevt näthat eller negativa kommentarer riktade mot dem online det senaste året. Enligt rapporten Svenskarna och internet 2023 är näthat och kränkningar starkt kopplade till personliga eller politiska åsikter. Mer än hälften av alla som är 16 år eller äldre avstår från att kommentera inlägg, nyhetsartiklar eller dela sina åsikter på sociala medier av rädsla för näthat, kränkningar och negativa kommentarer.

Det är en påtaglig risk att användningen av toxiskt och kränkande språk påverkar vår uppfattning och möjligen riskerar att minska vår reaktion på brottsliga handlingar, såsom hot och hets mot folkgrupp. När stora grupper av människor inte vågar eller orkar delta i samtalen på nätet öppnas fältet upp för aktörer och individer som utnyttjar möjligheten att påverka för sina egna syften genom att sprida falska narrativ och desinformation.

I grunden är internet och digitaliseringen fantastiska möjliggörare för människor att mötas, föra samtal och utbilda sig. Det är viktigt att värna om alla människors rätt att dra nytta av dessa möjligheter. En gemensam ansträngning krävs där barnrättsorganisationer, vuxna i barns närhet, skolan, rättsvårdande myndigheter och inte minst företag inom spel- och teknikbranschen tar ansvar och arbetar tillsammans för att säkerställa att kunskap, lagstiftning, moderering och nya tekniska lösningar samverkar för att skapa en hållbar digital miljö för alla.

Jannike Tillå, kommunikations- och affärsområdeschef, Internetstiftelsen

4.4 Hur definierar barn och unga ett toxiskt och kränkande språk?



Följande text är resultatet av en diskussion med Sätraskolans åttondeklassare. Eleverna delade med sig av sina observationer, erfarenheter och tankar kring var och hur det toxiska och kränkande språket förekommer, hur detta påverkar individer samt föreslog lösningar för hur vi kan motverka den fortsatta spridningen.

Toxiskt och kränkande språk förekommer på många olika plattformar, och kan se väldigt olika ut beroende på kontexten. Eleverna beskriver att det kan röra sig om både den typ av direkt toxiska och kränkande kommentarer som omfattas av denna undersökning, och indirekt toxiska kommentarer, där orden i sig inte är kränkande (de kan till och med vara positiva) men där kontexten eller kombinationen med vissa emojis gör det tydligt att syftet med kommentaren är att kränka.

Eleverna ger som exempel att kommentarer som "du är så rolig" följt av ett antal dödskalleemoji, och "jag önskar jag hade samma självförtroende som du" innebär att det som objektivt kan tolkas som en komplimang egentligen är sarkastiskt menat, och ska förstås som motsatsen. Ett annat exempel som eleverna lyfter och som uppfattas som toxiskt, är olika rekommendationer på hur man kan bli "finare", exempelvis genom att använda mer smink eller att gå till gymmet. Även detta är något som eleverna vittnar om och som är att förstå som toxiskt.

I undersökningens sammanhang är det värt att nämna att eleverna lyfter att en toxisk kommentar kan bestå av endast emojis eller GIF:ar.[20] Eleverna ger som exempel att en GIF med ett troll skickas för att likna trollet vid personen som tar emot meddelandet. Både emojis och GIF:ar är exempel på kommunikationsformer som inte omfattas i denna textbaserade undersökning.

På frågan om det finns något som utmärker kontexter och situationer som tenderar att "vara" mer toxiska lyfter eleverna dels anonymitet dels kommentarer på innehåll som sticker ut som exempel. Eleverna ser att det är mer vanligt förekommande med toxicitet och kränkningar där användaren kan vara anonym, exempelvis på Reddit. Toxiskt och kränkande språk förekommer enligt eleverna både i kommentarsfält på mer offentliga plattformar, som Instagram, TikTok och X (tidigare Twitter), och i privata, direkta meddelanden på Snapchat, Discord och i direktmeddelanden på Instagram och TikTok. Det faktum att denna undersökning inte omfattar den typ av privata meddelanden som ingår i den senare gruppen kan vara del av förklaringen till diskrepansen mellan elevernas estimat på att mellan 40 och 70 procent av all kommunikation på nätet är toxisk och kränkande, och de nivåer som uppmäts i denna undersökning.

Eleverna beskriver att innehåll och inlägg "som sticker ut" eller "när man gör något annorlunda" får toxiska kommentarer till svar. Inlägg där en kille berättar om sin sminkrutin kan exempelvis mötas av nedsättande kommentarer kring sexualitet eller könsidentitet. Eleverna förklarar även att toxiska kommentarer kan hittas på s.k. exposekonton[21] på TikTok och Instagram. Eleverna var överens om att toxiska kommentarer kan påverka olika individer på olika sätt. Den som sprider hat och toxicitet kan dels ångra sig senare och må dåligt över det hen har skrivit, dels själv utsättas för toxiska kommentarer. För den som blir utsatt kan konsekvenserna vara en sänkt självkänsla, särskilt om flera personer deltar i att använda ett toxiskt språk mot den utsatte. På längre sikt lyfter

20 En GIF (Graphics Interchange Format) är ett filformat som möjliggör för att skicka animerade bilder.
21 Ett exposekonto är konto i sociala medier som syftar till att hänga ut personer i kränkande situationer, exempelvis sådana som inkluderar våld eller nakenhet. Exposekonto var del av Institutet för språk och folkminnens Nyordslista 2023, där den aktuella definitionen förekom.

eleverna att det finns en ökad risk för depression och försämrad psykisk hälsa.

När det gäller vad de önskar att vuxna ansvariga för nätet bör tänka på, gör eleverna tydligt att det är viktigt att föräldrar och andra vuxna agerar förebilder för ett respektfullt uppförande, och samtalar med sina barn om hur kränkningar påverkar. Vuxna kan inte tillåta sig själva att vänja sig vid det toxiska språkbruket, utan måste aktivt engagera sig i barnens internetvanor. Eleverna beskriver slutligen hur det är viktigt att vuxna begränsar tillgången till sociala medier för barn under en viss ålder och heller inte tillåter dem att köpa spel som de ännu inte är mogna för. Slutligen, eleverna betonar vikten av en tryggare och respektfullare digital miljö där alla kan trivas utan rädsla för negativa konsekvenser av toxiskt och kränkande språk.

4.5 Mekanismer som bidrar till aggressivitet på nätet

Avsnittet om olika mekanismer som bidrar till ett toxiskt språkbruk på internet bland barn och unga utgår från Nätmobbning. En handbok för skolan[22] och Utsatt på internet. En internationell forskningsöversikt om nätmobbning bland barn och unga.[23]

En självklar del av barn och ungas vardag idag är att spendera tid vid datorn, surfplattan eller mobilen. Möjligheterna med modern teknik och nya kommunikationskanaler ökar risken för toxiskt språkbruk bland barn och unga. Barn och unga upplever att kommunikationen kan vara mer aggressiv på internet än ansikte mot ansikte. Den ökade aggressiva kommunikationen beror på att barn och unga kan vara anonyma på vissa forum på internet och att missförstånd är vanliga. En anledning till att unga missförstår varandra beror på att kommunikationen på internet är ofullständig dvs, att man interagerar utan att se den andres kroppsspråk. Om man inte ser den andres kroppsspråk leder det till svårigheter med att sätta sig in i hur den andre upplever interaktionen speciellt avseende känslor. Ofullständig kommunikation kan ibland leda till att en individ har lättare att agera impulsivt och ohämmat, för att exempelvis hämnas en upplevd oförrätt. Ett sätt att motverka detta är att öka ungas empati, att lära barn och unga att uppmärksamma och förstå en annan människas känslor. Ökad uppmärksamhet på och förståelse för andras känslor kan hindra barn och unga från att handla aggressivt. Strävan efter att öka sin status och makt är ytterligare en förklaring till aggressivt beteende på internet. Detta verkar gälla i vissa e-spel, när man använder en rå jargong med förnedrande skämt.[24]

En annan viktig mekanism som förklarat aggressiva beteenden såväl på som utanför internet är moraliskt disengagemang (MD). Moraliskt disengagemang innebär att individer har tankemönster som rättfärdigar aggressivt beteende. Individer med hög grad av MD upplever inte ånger, skuld eller skam när de beter sig aggressivt gentemot andra. Exempel på moraliskt disengagemang är kognitiv omstrukturering som innebär skämta rått men hjärtligt och att förlägga orsaken till den utsatte. Rått men hjärtligt innebär att man förringar

22 Frisén, A. & Berne, S. (2016). Nätmobbning. En handbok för skolan. Natur och Kultur.
23 Berne, S & Frisén, A. (2021). Utsatt på internet: En internationell forskningsöversikt om nätmobbning bland barn och unga.
19 Friends (2019). Friends nätundersökning 2019: Umgängesklimatet inom e-sporten.

eller avfärdar de negativa effekter ett skämt kan få för någon. Vidare hävdar man att den utsattes beteende eller egenskaper är orsaker till de förnedrande skämten och på så vis berättigade. Man undviker att känna skuld genom att man tycker att den som utsätts för skämtet får skylla sig själv. Mekanismerna som tagits upp ger ju sin pusselbit till tavlan av vad som orsakar toxiskt språkbruk på internet.
Givetvis så interagerar dessa faktorer med varandra och det kan variera beroende på vilken plattform hur mycket toxiskt språkbruk som förekommer.

Sofia Berne, leg. psykolog/PhD och författare till forskningsöversikten Utsatt på nätet (2021) .

4.6 Polisens arbete mot näthat

Förekomsten av toxiskt språk, som används för att hota eller kränka andra människor på nätet är en oroande företeelse. Att utsättas för kränkningar av den personliga integriteten, friheten eller friden kan exempelvis påverka möjligheten att fortsätta arbeta i ett offentligt ämbete eller påverka välbefinnandet, och kan vara ett brott oavsett om det sker på nätet eller utanför. Ingen ska komma undan med att utsätta andra för brott och det är viktigt att samhället reagerar snabbt och stöttar den som har utsatts.

Det är ibland en komplex process att fastslå vad som är olagligt i fråga om toxiskt språk. Det finns en fin linje mellan yttrandefrihet och brottsliga handlingar som måste upprätthållas. Varken toxiskt språk eller näthat är juridiska begrepp, utan utgör samlingsbegrepp som kan innefatta olagliga handlingar som förtal, olaga hot eller hets mot folkgrupp. Med det sagt, att inte allt som av den utsatte upplevs som obehagligt, kränkande eller hotfullt automatiskt utgör ett lagbrott, bör situationer där hotet upplevs som verkligt och skapar rädsla, eller där allvarliga kränkningar förekommer, anmälas så polisen får avgöra om händelsen är brottslig. I de fall ärenden läggs ner betyder det inte att polisen misstror anmälaren utan det kan till exempel handla om att man inte ser det som möjligt att koppla en förövare till den brottsliga handlingen. Dock kommer möjligheten att återuppta ett ärende alltid finnas i det fall ny bevisning framkommer. Viktigt är också att veta att det inte bara är den som lagt ut den brottsliga bilden eller kommentaren som begår ett brott. Det gör också den som sprider den vidare.

Polisen strävar efter att agera snabbt och effektivt mot brott på nätet. För enskilda individer som utsätts för toxiskt språk är det rekommenderat att göra en polisanmälan om situationen upplevs som kränkande eller hotfull. Att samla bevis i form av skärmdumpar, mejl, sms eller chattkonversationer är ofta till betydande hjälp i en eventuell utredning. För dig som är ett barn rekommenderar vi att du pratar med dina föräldrar eller någon annan vuxen, och att du berättar vad som har hänt. Om du inte har en vuxen i din närhet att prata med så finns det många bra organisationer på nätet som du kan vända dig till för hjälp, råd och stöd.

» Det är viktigt att vi alla agerar ansvarsfullt på nätet och att var och en är medveten om sina egna handlingar. Genom att agera proaktivt kan vi alla bidra till att bekämpa näthatet och skapa en tryggare digital miljö för alla. «

Det är viktigt att vi alla agerar ansvarsfullt på nätet och att var och en är medveten om sina egna handlingar. Genom att agera proaktivt kan vi alla bidra till att bekämpa näthatet och skapa en tryggare digital miljö för alla. Utgå från hur du själv hade uppfattat din kommentar om det var du som läst den. Ibland är det enklare att skriva något elakt eller kränkande på nätet än att säga det direkt till en person. Ha även i åtanke att det finns en risk att det du delar på nätet kan få en oväntad och stor spridning.

Magnus Gussander
Samordnare i Polisens grupp mot demokratihotande brottslighet

Foto: Magnus Gussander

Carl Heath
Senior forskare vid RISE och tidigare utredare för Det demokratiska samtalet (SOU 2020:56)

Foto: Carl Heath

Slutligen, Polisen skulle behöva vara ännu mer aktiva på nätet för att kunna upptäcka brott alternativt förhindra brott. Ingen ska komma undan med att begå brott mot barn på internet. Gör man det ska samhället reagera snabbt och kraftfullt.

Magnus Gussander, samordnare i Polisens grupp mot demokratihotande brottslighet

4.7 Påverkan på det digitala demokratiska samtalet

I vår digitala tid genomsyrar teknologin i praktiken varje aspekt av våra liv. Detta är särskilt tydligt när det gäller unga människor, som ofta är bland de första att anamma nya teknologier. Ungas snabba anpassning till och användning av digitala tjänster och innovationer formar kultur och vardagsliv på många olika sätt. Utvecklingen av det demokratiska samtalet i en digital tid påverkar oss alla, men också demokratin i sig självt. Å ena sidan erbjuder den oöverträffade möjligheter för informationsspridning, kunskapsutbyte och politiskt deltagande. Å andra sidan innebär digitaliseringen av samhället också hot mot demokratin, som desinformation, informationspåverkan, hot och hat.

Börjar vi i att se det digitala samtalets möjligheter, syns tydligt hur unga ges en plattform för att uttrycka sina åsikter och delta i det offentliga samtalet på sätt som tidigare generationer inte haft möjlighet till. Internet och sociala medier har blivit avgörande verktyg för att engagera sig i samhällsfrågor, mobilisera sociala rörelser och till och med påverka politiska val och beslut. Ett av de mest framträdande exemplen på digitaliseringens positiva påverkan är hur unga använder sociala medieplattformar för att driva och engagera sig i samhällsfrågor. Plattformar som Instagram, Snap, TikTok och YouTube har blivit arenor där inte minst unga kan mobilisera stöd för olika samhällsfrågor, från klimatförändringar till mänskliga rättigheter. Genom kampanjer och digital aktivism har unga individer lyckats sätta viktiga frågor på dagordningen och påverka både den offentliga debatten och policybeslut.

Men detta mynt har en annan sida. Utvecklingen och användningen av sociala medier har också lett till ökad spridning av desinformation, och hot och hat på internet. Dessa fenomen utgör inte bara ett hot mot individuella användare, särskilt unga, utan också mot det bredare demokratiska samtalet. Desinformation, särskilt när den sprids via sociala medier, utgör en betydande utmaning. Detta innefattar allt från falska nyheter till vilseledande information som syftar till att påverka opinionen eller undergräva förtroendet för samhällsinstitutioner.

Desinformation på sociala medier påverkar unga i Sverige idag på flera sätt, och LVU-kampanjen är ett slående exempel på detta. Den är en desinformationskampanj som felaktigt påstår att svensk socialtjänst missbrukar LVU - Lag med särskilda bestämmelser om vård av unga för att tvångsomhänderta muslimska barn på falska eller felaktiga grunder. Sedan december 2021 har denna desinformation spridits, framför allt i sociala medier, där rykten hävdar att socialtjänsten skulle kidnappa

» När unga censurerar sig själva på grund av rädsla för hot och hat, begränsas deras förmåga att fritt uttrycka sina åsikter och delta i samhällsdebatten. «

muslimska barn.[25] Denna kampanj exponerar en betydande sårbarhet i det svenska samhället, särskilt bristen på tillit mellan svensk offentlig förvaltning och delar av landets invånare, vilket i sin tur har exploaterats av utländska aktörer. För unga användare kan exponeringen för sådan desinformation vara särskilt problematisk. Denna typ av felaktig information kan bidra till att skapa misstro och rädsla, vilket kan påverka deras syn på samhället och deras förtroende för offentliga institutioner. Desinformation som LVU-kampanjen kan även leda till att unga drar sig för att delta i samhälleliga och politiska diskussioner, på grund av den osäkerhet och förvirring som sådan information kan skapa. Detta kan i sin tur leda till en minskad känsla av samhällstillhörighet och en försämrad förmåga att kritiskt bedöma information och händelser i samhället. LVU-kampanjen illustrerar hur desinformation via sociala medier inte bara är en fråga om felaktig information, utan också hur den kan ha djupgående och negativa konsekvenser för unga individers uppfattning om och engagemang i samhället. Förekomsten av påverkanskampanjer där unga utgör en målgrupp understryker betydelsen av utbildning i medie- och informationskunnighet.

Hot, hat och toxiskt språk på nätet utgör inte bara ett problem för den enskilda unga personen, utan har också djupgående konsekvenser för demokratin. Enligt Brottsoffermyndigheten har 64% av unga i åldern 16-25 år blivit utsatta för hot och hat online, vilket negativt påverkat deras hälsa. Denna utbredda utsatthet leder till omfattande självcensur; cirka 70% av befolkningen anpassar sitt uttryck och undviker att publicera sig online för att undvika hot eller hat. Bland de som utsatts för kränkningar online är självcensuren ännu högre, särskilt bland dem vars hälsa påverkats negativt.[26]

När unga censurerar sig själva på grund av rädsla för hot och hat, begränsas deras förmåga att fritt uttrycka sina åsikter och delta i samhällsdebatten. Detta underminerar en av de grundläggande pelarna i en demokrati – yttrandefriheten. När ungdomar drar sig tillbaka från offentliga samtal, går viktiga perspektiv och röster förlorade. Det gör den demokratiska dialogen fattig och utarmad och kan leda till en försvagning av demokratiska processer.

I denna tid av digital transformation är det avgörande att samhället, inklusive myndigheter, utbildningsväsendet och civilsamhället, tar ett större ansvar för att skapa ett mer hälsosamt och inkluderande digitalt demokratiskt

25 Regeringen tar krafttag mot LVU-kampanjen, Regeringskansliet (2023).
26 Näthat leder till omfattande självcensur, Brottsoffermyndigheten (2021); Näthat och demokratiskt deltagande – en kunskapsöversikt, Brottsoffermyndigheten (2021).

samtal. Detta innebär att aktivt bekämpa desinformation, hot, hat och toxiskt språk som påverkar unga negativt och begränsar deras deltagande i samhället. Genom att öka medvetenheten om dessa frågor, stärka ungas delaktighet och inflytande i demokratin, och utveckla en digital infrastruktur som främjar öppen och respektfull kommunikation, kan vi bygga ett starkare och mer inkluderande samhälle. Dessa insatser är avgörande för att säkra en hälsosam demokratisk utveckling och garantera att alla röster, särskilt ungas, hörs och respekteras.

Carl Heath, senior forskare vid RISE och tidigare utredare för Det demokratiska samtalet (SOU 2020:56)

5. Avslutning: När varje ord räknas

Projektet Ord som sårar påbörjades under hösten 2021, och är sprunget ur en vilja att undersöka ett återkommande tema i flertalet rapporter och samtal med barn och unga angående samtalet i digitala miljöer: förekomsten av ett toxiskt och kränkande språk, bland annat bestående av rasistiska ord och svordomar. De digitala miljöerna är idag lika självklara som de fysiska för barn och unga, och det är av yttersta vikt att de är anpassade i enlighet med barnets rättigheter och särskilda behov.

Med enorma möjligheter att hantera lika enorma mängder data som skapas i vår allt mer digitala värld är de tekniska lösningarna absolut nödvändiga för att identifiera toxisk och kränkande kommunikation som har potential att skada barn och unga både på individnivå och på ett bredare, samhälleligt och demokratiskt plan. Mind Intelligence Lab (MIL), som är en av de medförfattande organisationerna till denna rapport, är ett företag som fokuserar på utveckling av teknik för att analysera och bedöma olika former skadlig kommunikation. MIL har utvecklat tekniska lösningar för att upptäcka toxiskt språk, direkta och indirekta hot samt våldsfrämjande kommunikation. Det finns dock ett flertal utmaningar med tekniken. Att utveckla modeller för att automatiskt upptäcka skadlig kommunikation med hög träffsäkerhet kräver omfattande mängder träningsdata vilket inte alltid är tillgängligt. Ytterligare en utmaning är att de flesta tekniker är språkberoende. En stor del av de tekniska lösningar som finns är på engelska och det är väldigt få som utvecklar tekniker för mindre språk som exempelvis det svenska språket.

Projektet har under de två år det fortlöpt inneburit ett konstant arbete och en vidareutveckling av en gångbar metod för att undersöka förekomsten av toxiskt och kränkande språk i barn och ungas digitala miljöer. Arbetet har inneburit återkommande avvägningar mellan metodens hållbarhet i fråga om de resurser som går åt för att samla in data, och den precision som krävs för att återspegla den diversifierade och engagemangsbaserade nätvardag som barn och unga befinner sig i. Beroende på om plattformen i fråga tillgängliggör data via öppna API:er för forskning har gruppen bitvis fått vända sig till användarna själva för att undersöka kommunikationen och förekomsten av toxiskt språk. Inte allt för sällan låg lösningen i användandet av en webbskrapningstjänst, och ett resurskrävande inklistrande av separata webbadresser för att sammanställa data. Utan en tydlig beskrivning av hur de algoritmer som bidrar till att visst innehåll vinner popularitet fungerar, har det dessutom varit svårt att skapa en helomspännande bild av vad barn och ungas nätvardag består i. Istället har projektgruppen fått utgå ifrån fokusgruppers inspel samt olika profilers större eller mindre följarantal för att indikera att ett visst innehåll och dess kommentarsfält är något som många barn och unga har interagerat med.

» Det är allt för svårt att utföra forskning på något som majoriteten av våra barn och unga använder sig av, som i den mån den innehåller toxiskt och kränkande språk påverkar barnens mående, och vidare viljan att delta i det demokratiska samtalet. «

Samlat talar detta sitt tydliga språk: Det är allt för svårt att utföra forskning på något som majoriteten av våra barn och unga använder sig av, som i den mån den innehåller toxiskt och kränkande språk påverkar barnens mående, och vidare viljan att delta i det demokratiska samtalet. I det avseendet att digitala miljöer erbjuder unika möjligheter för barn och unga att utnyttja sina demokratiska rättigheter är det särskilt alarmerande att det toxiska språket förekommer i högre grad i politiska diskussioner än i övriga.

Det finns ett tydligt behov av en systematisk och etisk tillgång till data för att möjliggöra för bredare, och extern forskning, för att med detta till grund möjliggöra för olika interventioner från olika aktörer för ökad trygghet på nätet.

Sedan projektets start har betydelsefull lagstiftning från Europeiska unionen (EU) kommit att träda i kraft. EU:s förordning om digitala tjänster, *Digital Services Act*, ställer bland annat krav på att de särskilt stora digitala tjänsterna delar med sig av data för forskning på s.k. systemrisker, däribland faktiska eller förutsebara negativa effekter på utövandet av grundläggande rättigheter. Lagstiftning likt Digital Services Act är ett viktigt steg i möjliggörandet av fortsatt forskning, och i förlängningen, möjligheten att följa eventuella förflyttningar i det toxiska och kränkande språket i digitala miljöer.

I författandet av en rapport som redovisar olika mängder toxicitet i digitala miljöer vill projektgruppen avsluta med att konstatera det uppenbara: **När varje ord räknas är det ändå inte antalet som spelar roll. Ett enda ord som sårar är ett ord för mycket.** Snarare än att lägga vikt vid hur vanligt förekommande toxiska och kränkande ord på en plattform är, och hur den förhåller sig till andra digitala miljöer bör dess närvaro ses som en tydlig indikation på att vi gemensamt behöver verka för ökad trygghet på nätet. Varje tiondels procentenhet representerar rasistiska, sexistiska eller på annat sätt kränkande ord eller uttryck som kan skada en eller flera personer.

Med vetskap om att kommunikationen i det digitala bitvis är ofullständig, och att avsaknaden av ansiktsuttryck, kroppspråk eller tonläge kan göra det svårare att sätta sig in i hur en annan person kommer att känna och reagera på det vi skrivit, bör vi därför ha större marginaler och vara än mer noggranna i hur vi samtalar i i digitala miljöer. Oavsett om det är del av en ”rå men hjärtlig” jargong, en ogenomtänkt reaktion på något som upprör eller en kommentar som lämnats i tydligt syfte att kränka, kan ord som sårar skada enskilda individer till den grad att vi behöver bemöta det med samma självklara nolltolerans som vi hade gjort i klassrummet eller på en arbetsplats. Det är i vårt gemensamma ställningstagande, vårt upprätthållande av grundläggande värderingar och mänskliga rättigheter och vårt erkännande av att varje ord räknas, som vi skapar ett demokratiskt och respektfullt samtal i digitala miljöer. ■