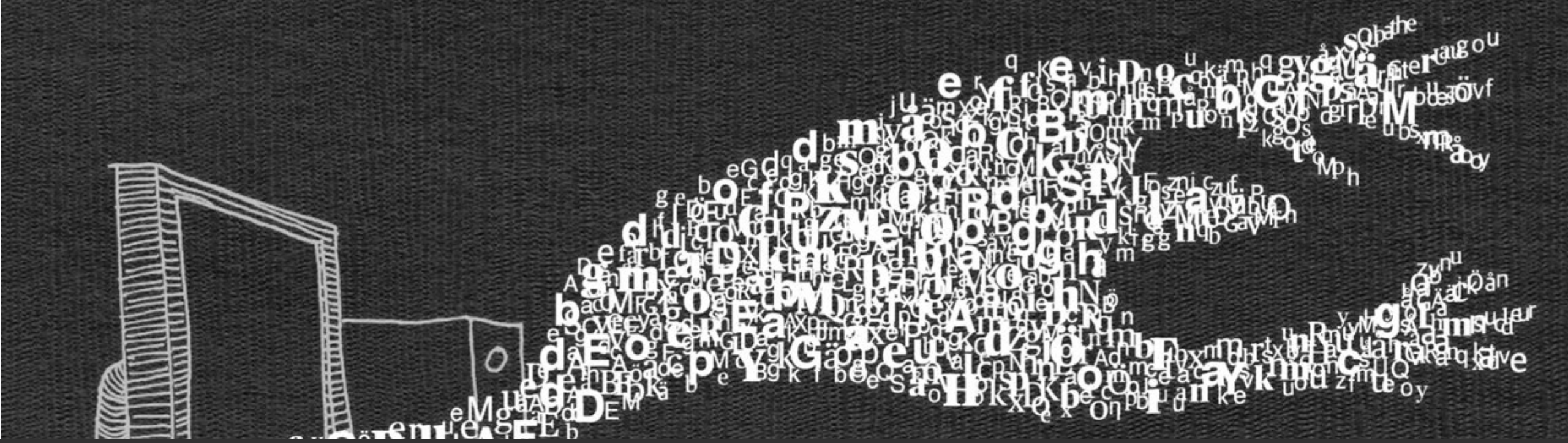


Hatbudskap och våldsbejakande extremism i digitala miljöer

Lisa Kaati
lisa.kaati@dsv.su.se

Katie Cohen
katie.cohen@foi.se

Toxiskt språk i digitala miljöer

- kommunikationshandlingar som förgiftar samtalsklimatet:
- hets mot folkgrupp, förtal, förgripelse mot tjänsteman
- integritetskränkningar, respektlöshet
- näthat, hat och hot
- hate speech, dangerous speech

Vad är toxiskt?

- Kontextberoende
- Koder
- Vi har alla olika uppfattning om vad som är toxiskt

Toxiskt språk

most government and media offices shut down or run on a skeleton crew on weekends. Financial markets don't move on weekends and thus don't drive news. Weekends are basically the downtown in which they plot the plan and agenda and how the media will promote it for the coming week. News and info always moves slower on weekends.

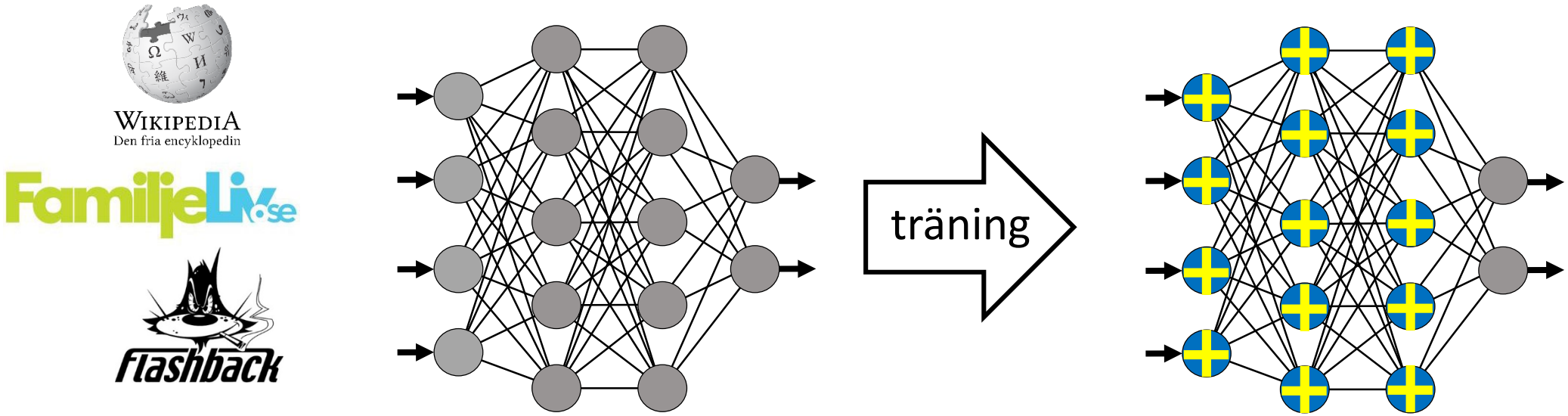
Toxiskt språk

most government and media offices shut down or run on a skeleton crew on weekends. Financial markets don't move on weekends and thus don't drive news. Weekends are basically the downtown in which (((they))) plot the plan and agenda and how the media will promote it for the coming week. News and info always moves slower on weekends.

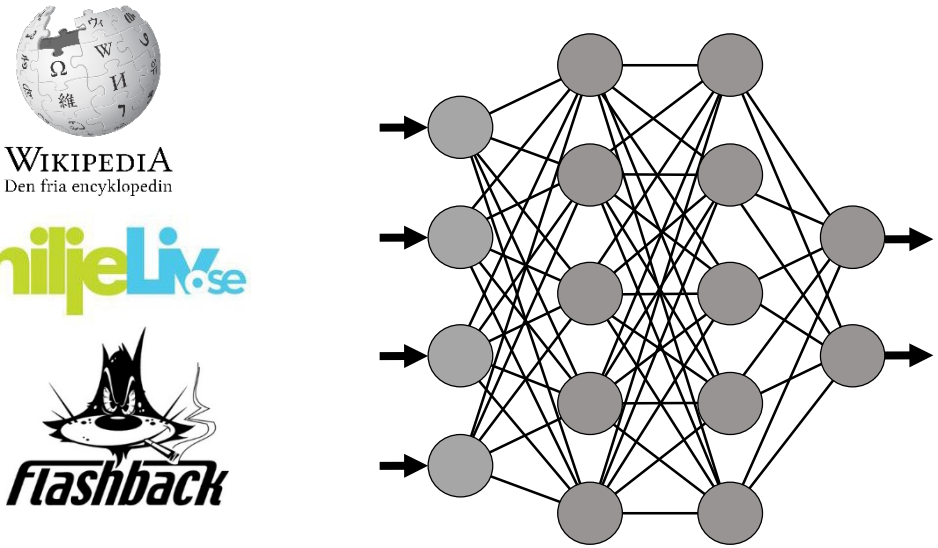
Automatiska metoder för att känna igen toxiskt språk

- Flertal olika metoder: från enkla ordlistebaserade metoder till avancerade AI-system
- Språkberoende– vissa språk är lågresurs språk
- Maskininlärningsbaserade metoder behöver massor av träningsdata

Transfer learning med språkmodell

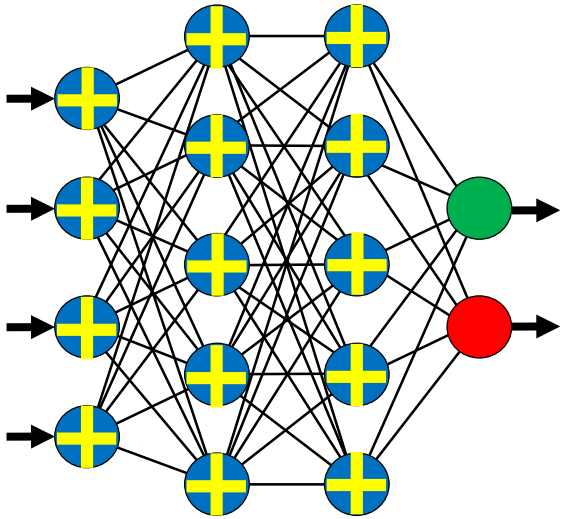


Transfer learning med språkmodell



Fine-tuning: träna vidare för speciell uppgift med annoterade exempel:

Nån borde knuffa C framför bussen!

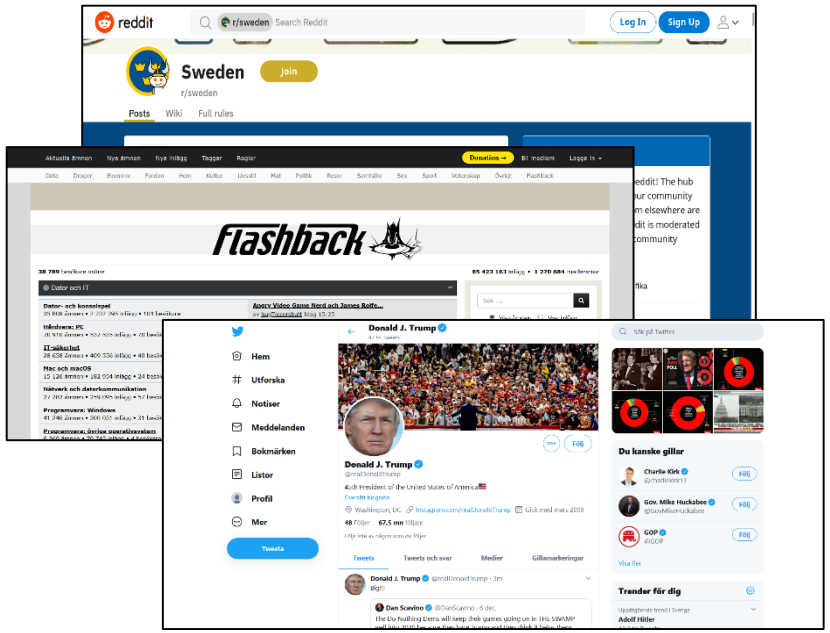


ja, toxiskt

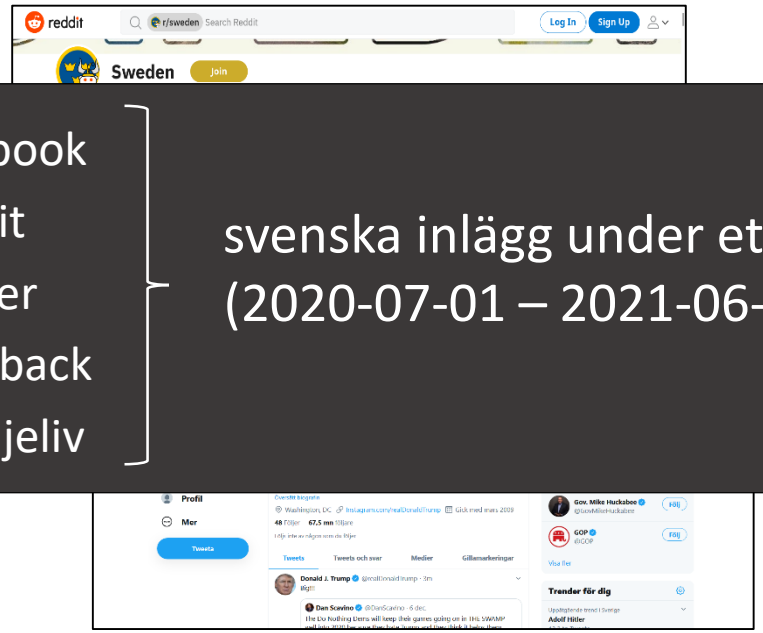
Automatisk detektion av toxiskt språk

- Alla automatiska metoder för textanalys gär fel ibland
- Inga metoder automatiska metoder kan förstå text lika bra som en människa…
- ...men dom är mycket snabbare och tröttnar inte.
- Våra automatiska metoder är lämpliga för:
  - kvantitativa analyser av stora textmängder
  - gallring: t.ex. reducera 1 miljon tweets till 3 000 som vi kan kolla för hand

Toxiskt språk i svenska digitala miljöer



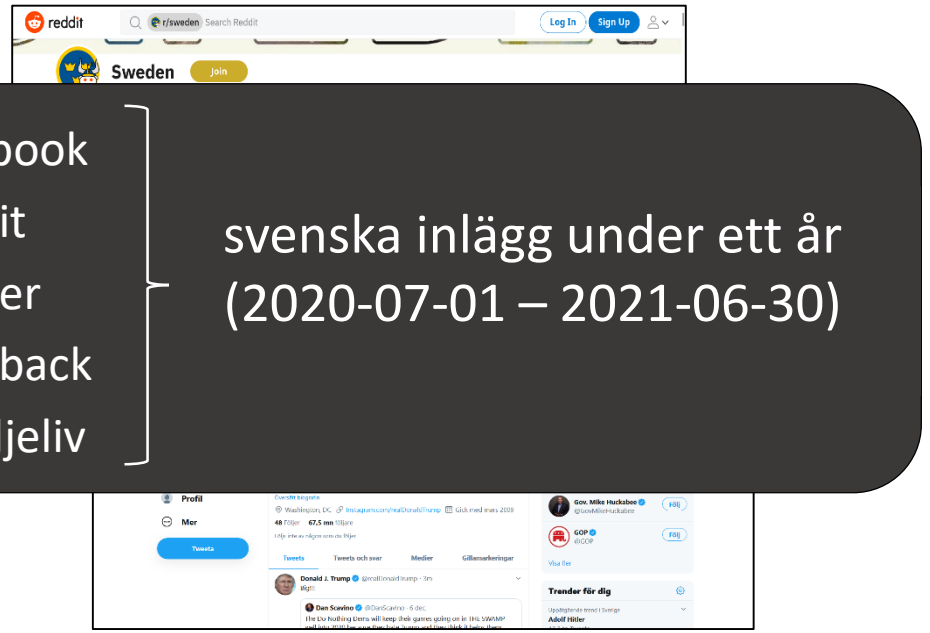
Toxiskt språk i svenska digitala miljöer



- Facebook
- Reddit
- Twitter
- Flashback
- Familjeliv

svenska inlägg under ett år
(2020-07-01 – 2021-06-30)

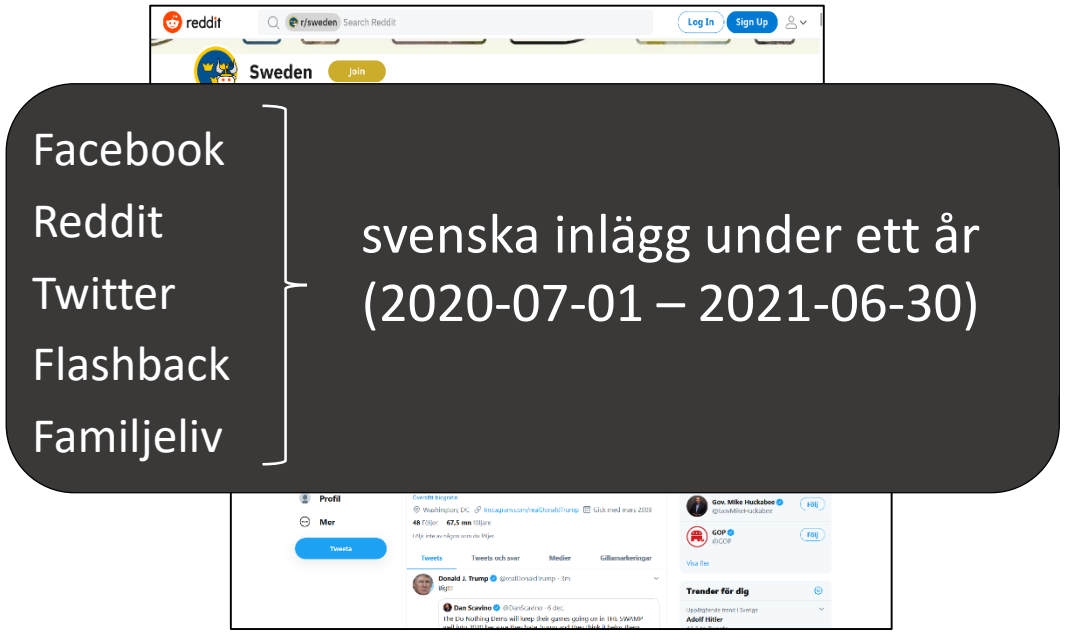
Toxiskt språk i svenska digitala miljöer

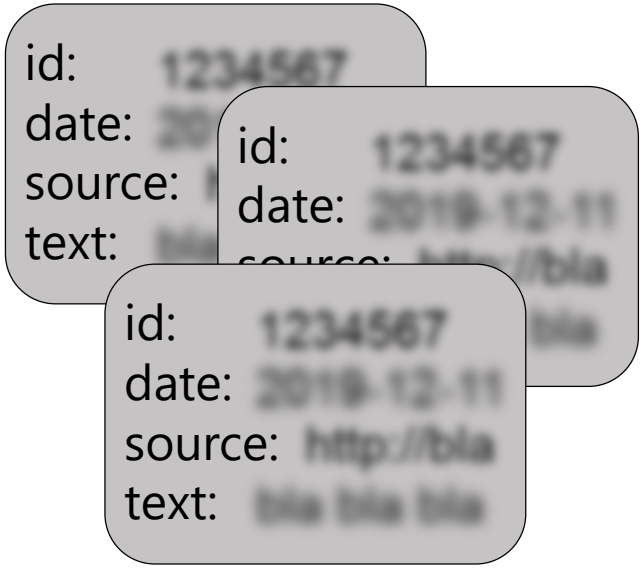


Facebook
Reddit
Twitter
Flashback
Familjeliv

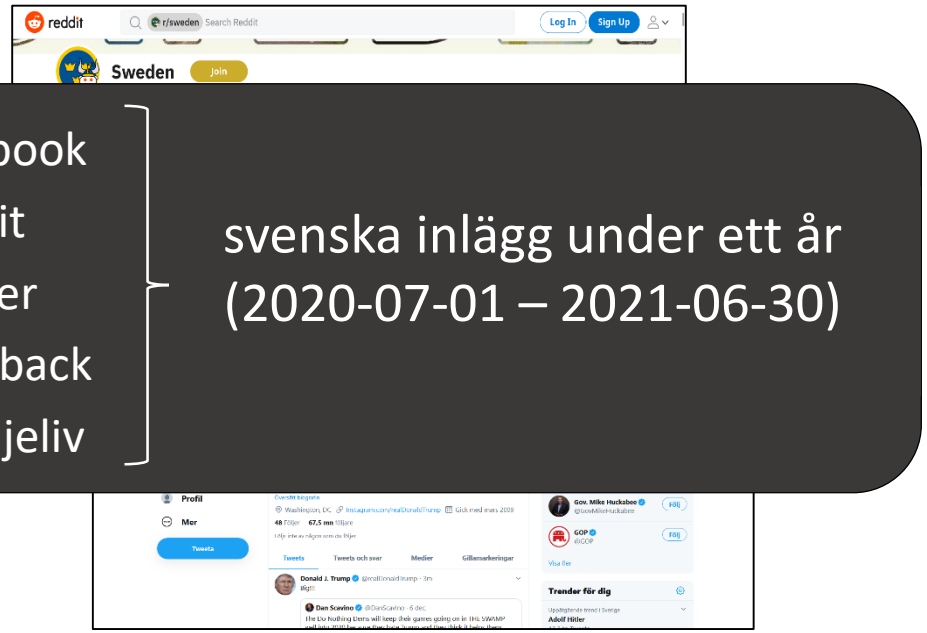
svenska inlägg under ett år
(2020-07-01 – 2021-06-30)

Toxiskt språk i svenska digitala miljöer



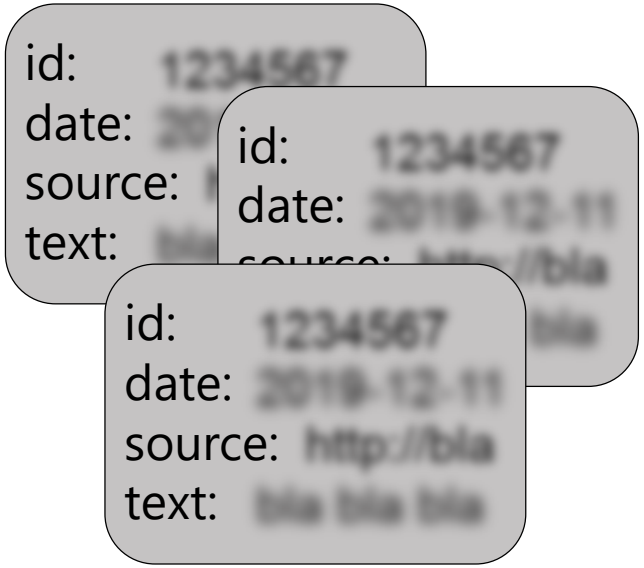


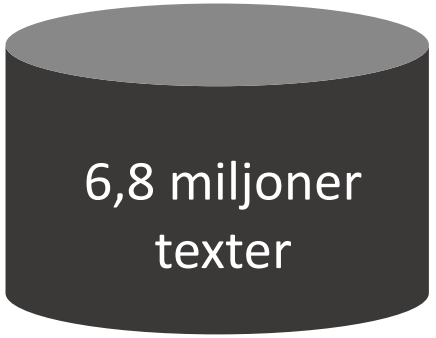
Toxiskt språk i svenska digitala miljöer



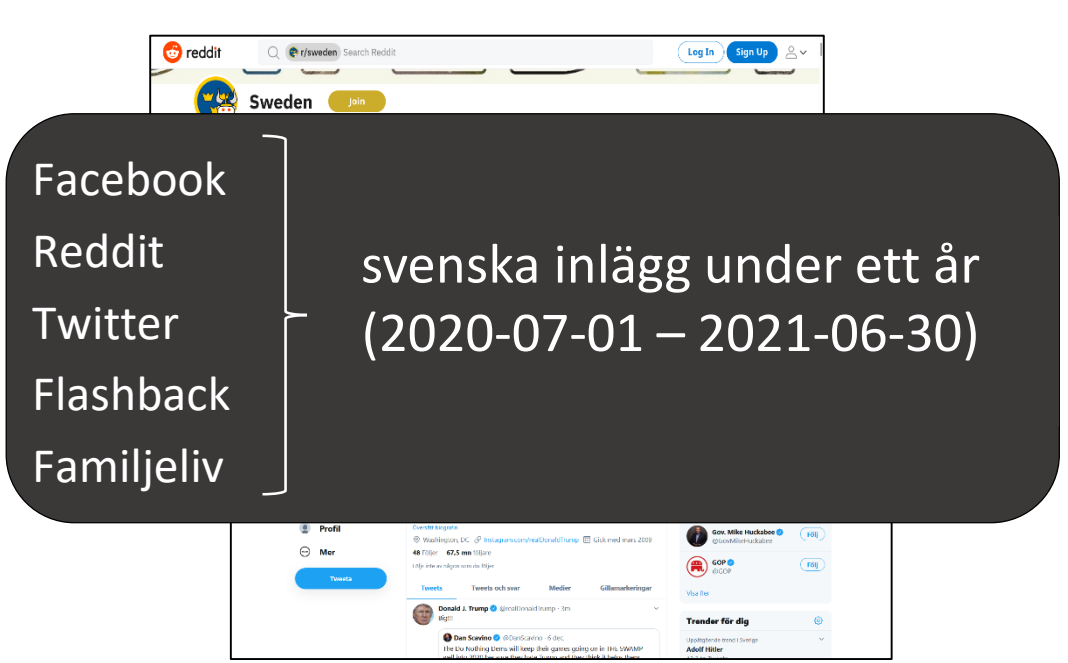
Facebook
Reddit
Twitter
Flashback
Familjeliv

svenska inlägg under ett år
(2020-07-01 – 2021-06-30)



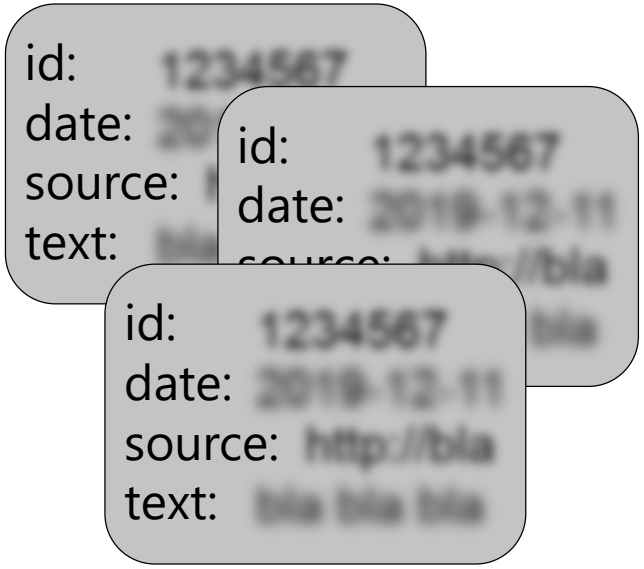


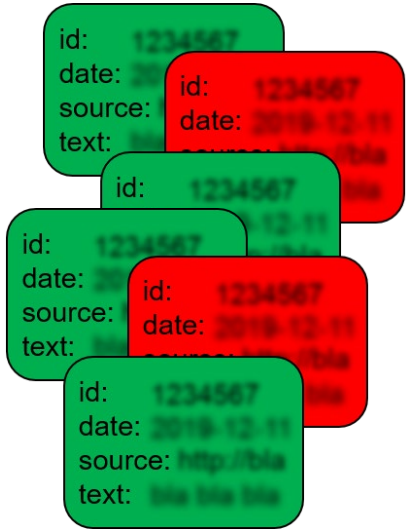
Toxiskt språk i svenska digitala miljöer

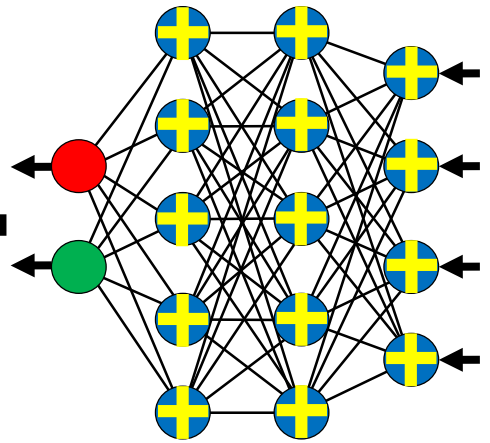


Facebook
Reddit
Twitter
Flashback
Familjeliv

svenska inlägg under ett år
(2020-07-01 – 2021-06-30)

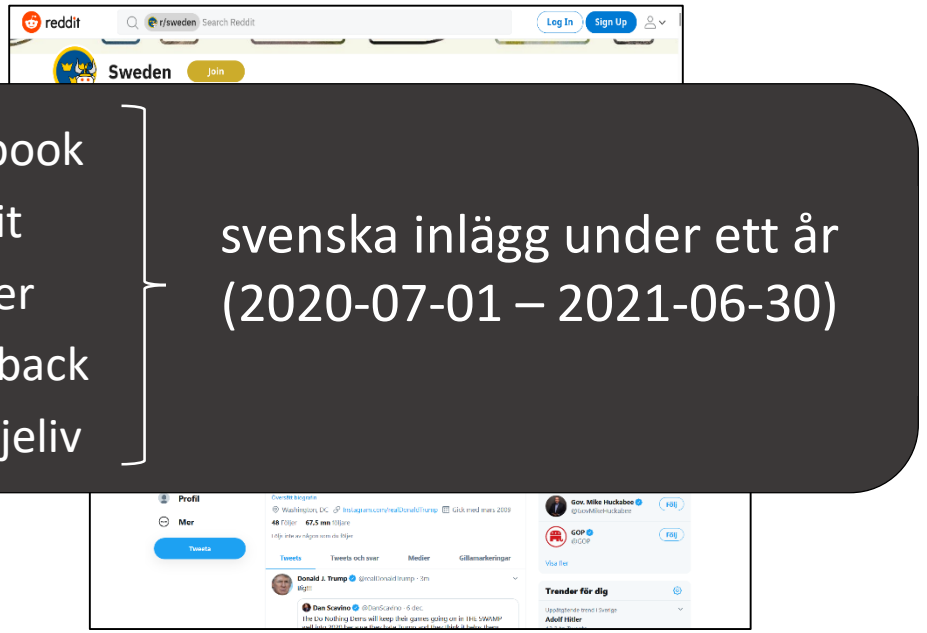






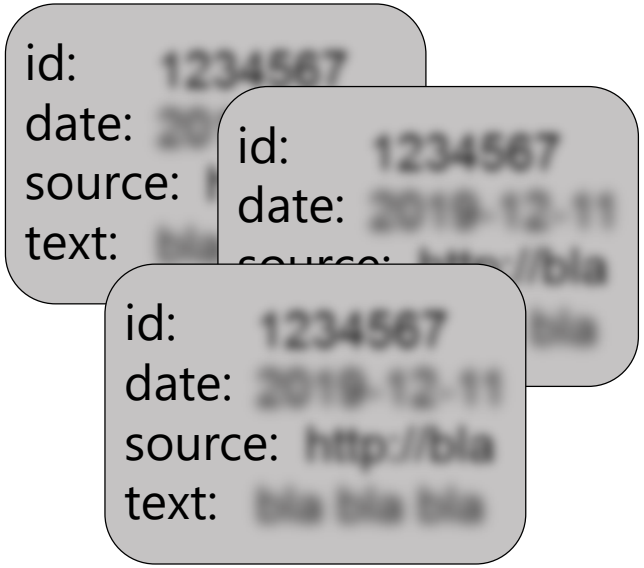
6,8 miljoner
texter

Toxiskt språk i svenska digitala miljöer

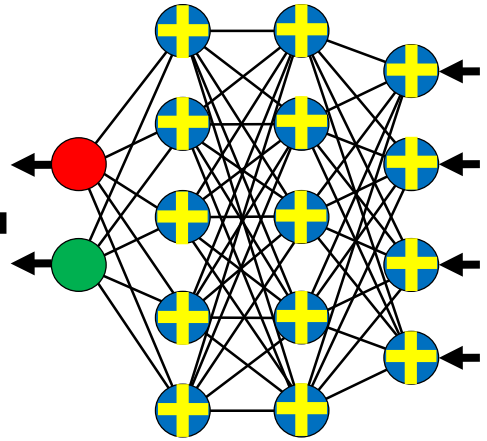


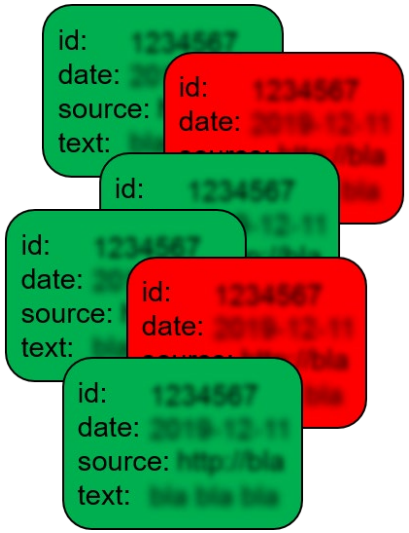
Facebook
Reddit
Twitter
Flashback
Familjeliv

svenska inlägg under ett år
(2020-07-01 – 2021-06-30)

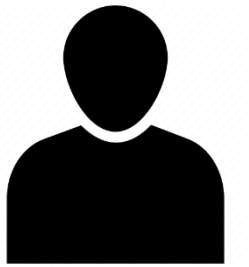


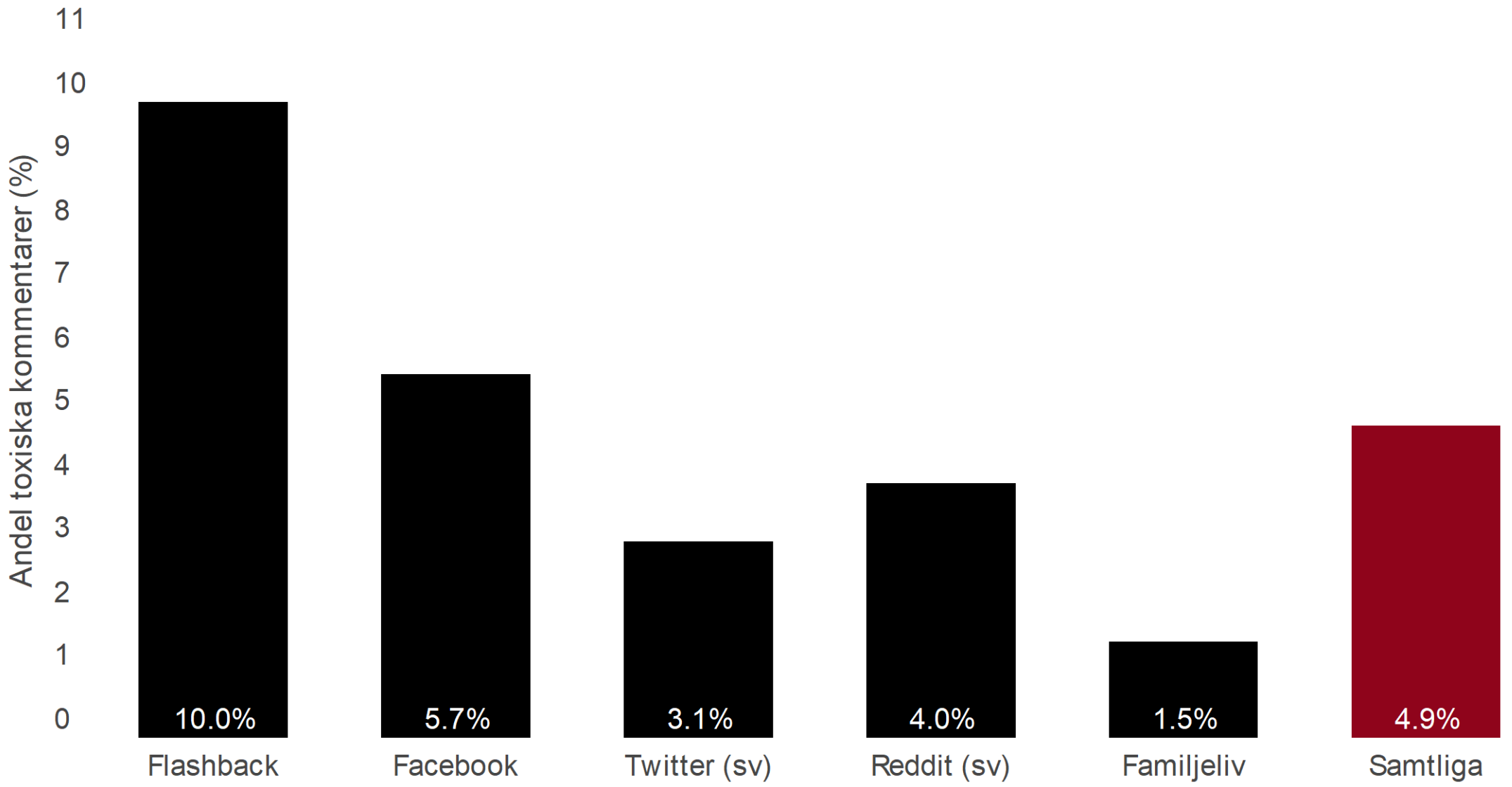
6,8 miljoner texter

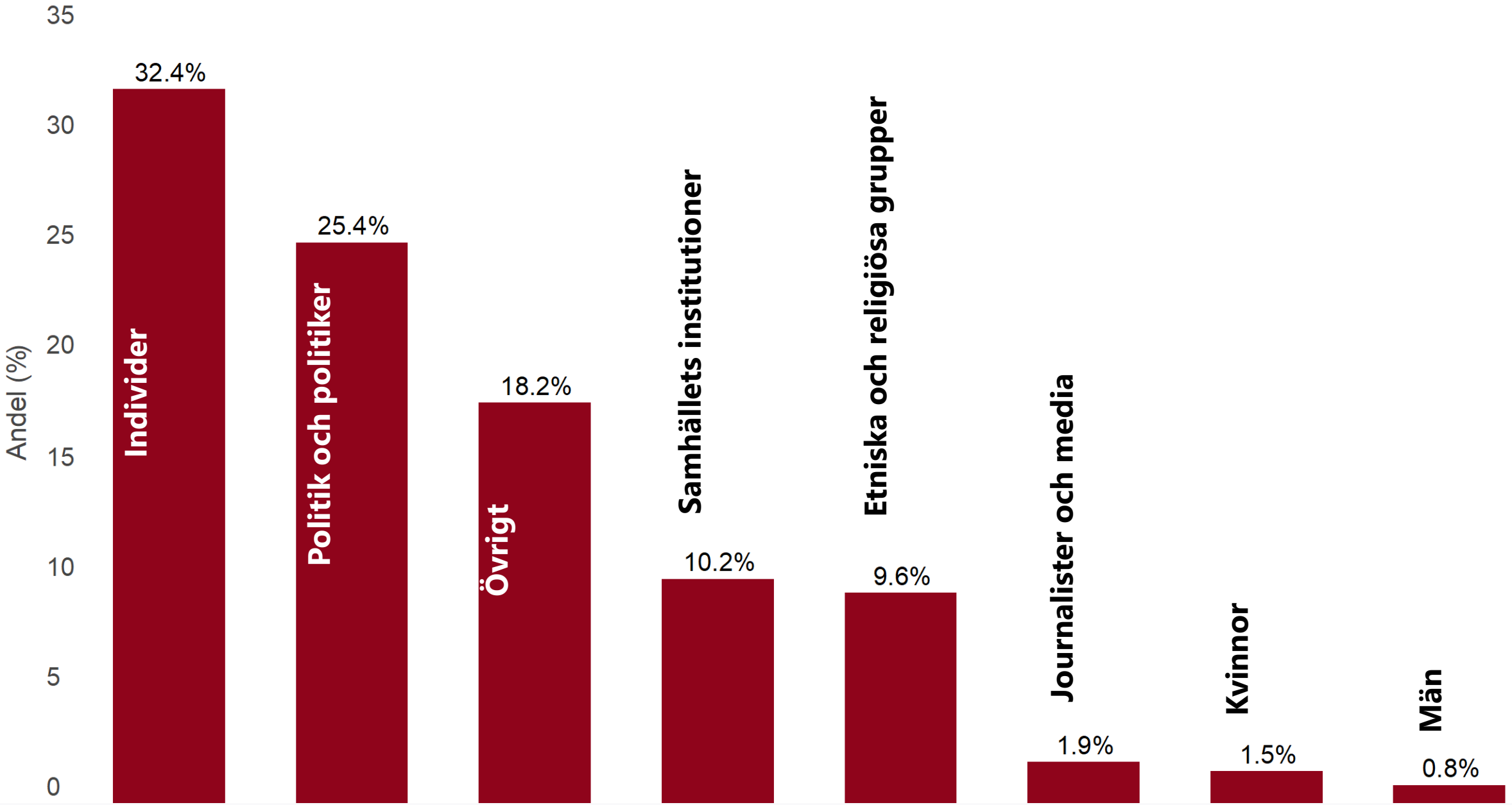


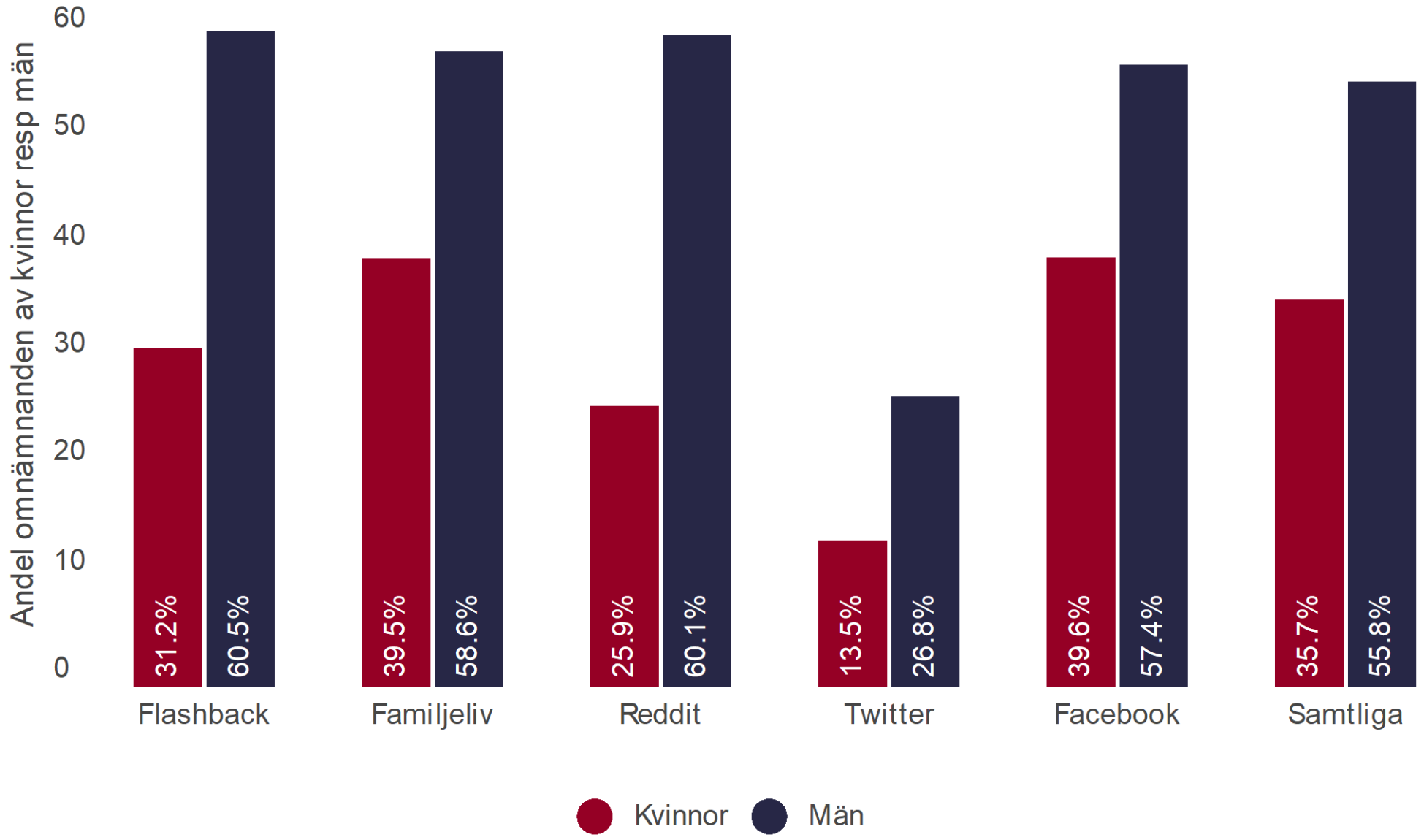


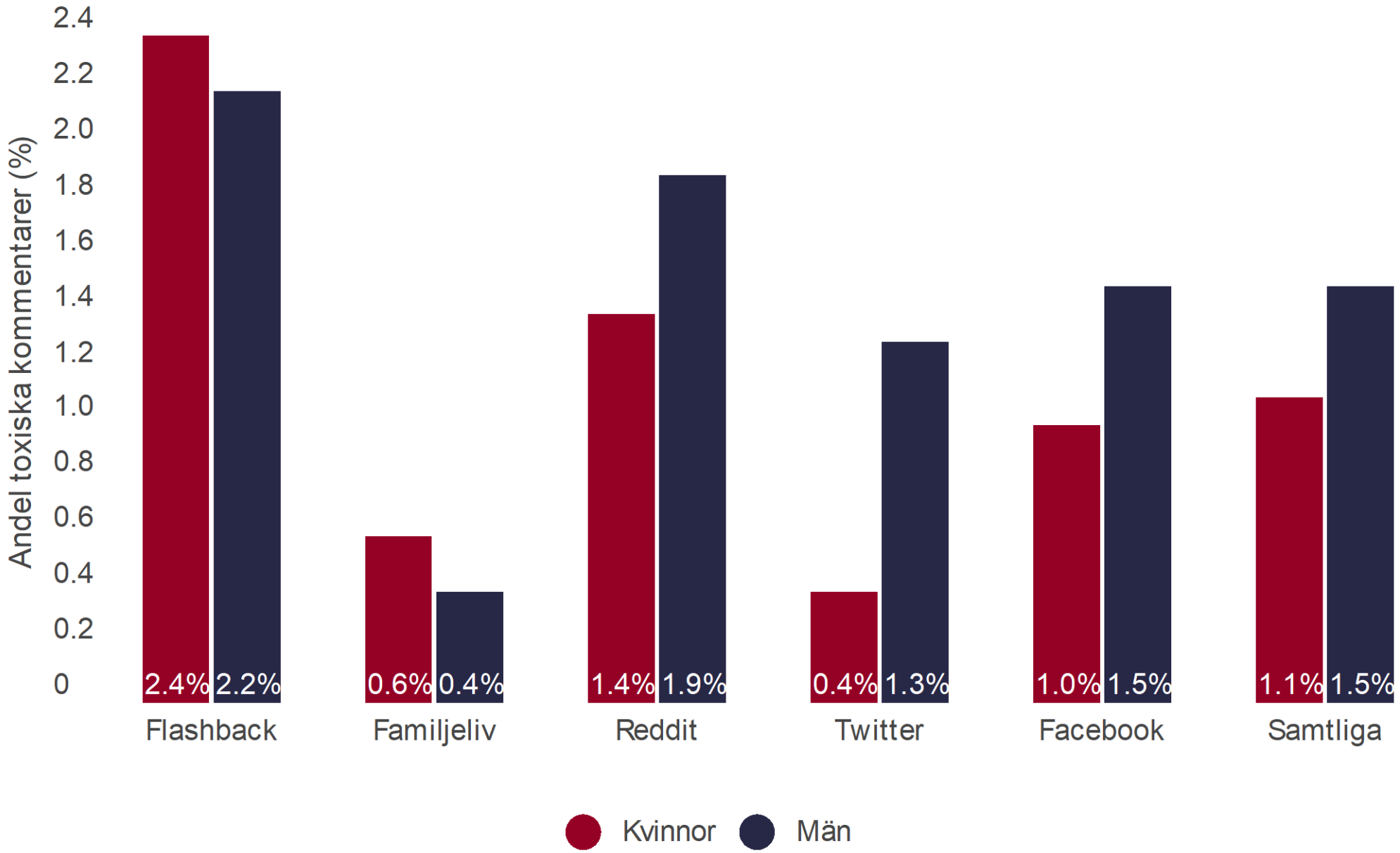
fördjupade manuella analyser av representativa urval

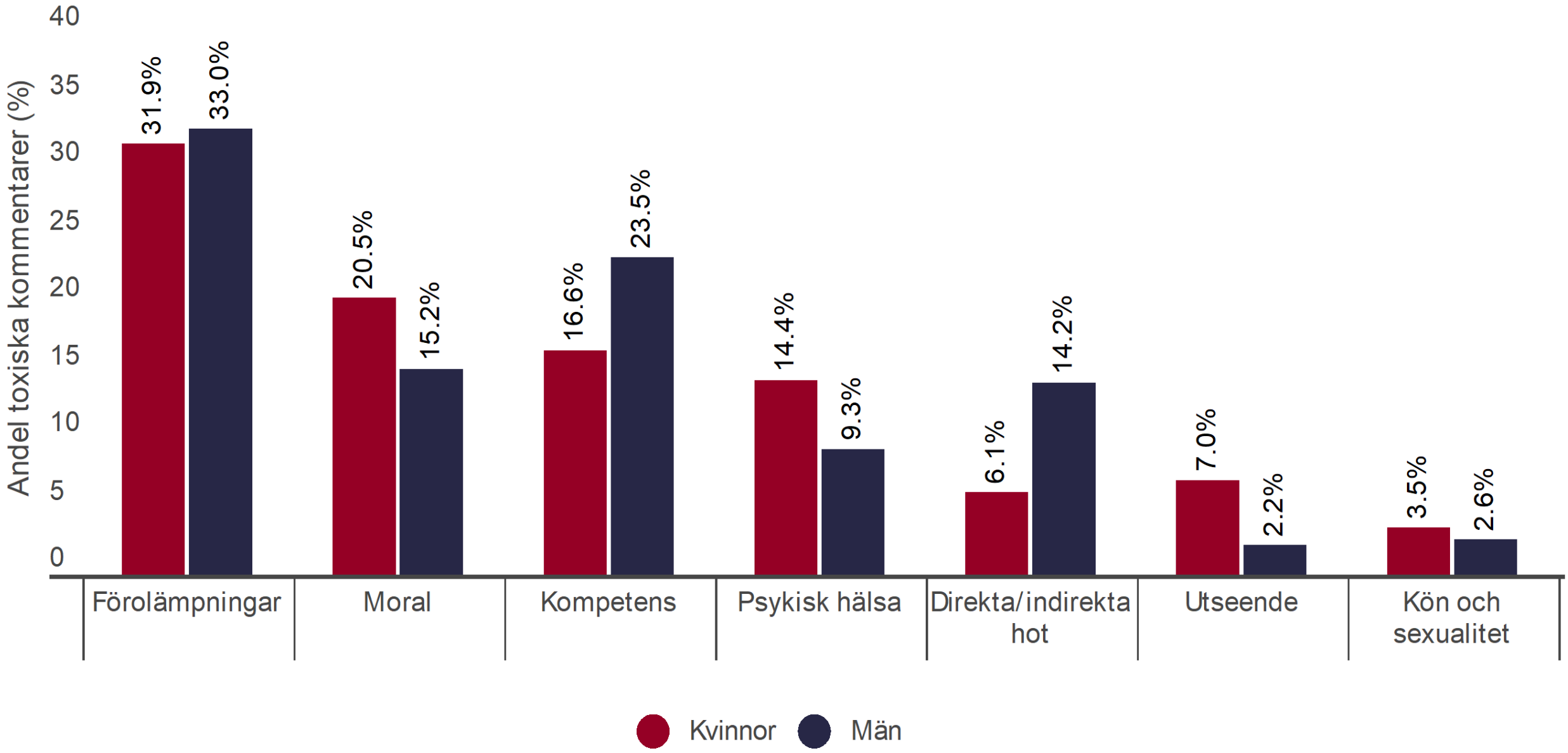



Andel toxiska kommentarer (%)
11
10
9
8
7
6
5
4
3
2
1
0
10.0%
5.7%
3.1%
4.0%
1.5%
4.9%
Flashback
Facebook
Twitter (sv)
Reddit (sv)
Familjeliv
Samtliga


35
30
25
20
15
10
5
0
Andel (%)
32.4%
25.4%
18.2%
10.2%
9.6%
1.9%
1.5%
0.8%
Individer
Politik och politiker
Övrigt
Samhällets institutioner
Etniska och religiösa grupper
Journalister och media
Kvinnor
Män


Andel omnämnanden av kvinnor resp män
60
50
40
30
20
10
0
31.2%
60.5%
39.5%
58.6%
25.9%
60.1%
13.5%
26.8%
39.6%
57.4%
35.7%
55.8%
Flashback
Familjeliv
Reddit
Twitter
Facebook
Samtliga
Kvinnor
Män


Andel toxiska kommentarer (%)
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
2.4%
2.2%
0.6%
0.4%
1.4%
1.9%
0.4%
1.3%
1.0%
1.5%
1.1%
1.5%
Flashback
Familjeliv
Reddit
Twitter
Facebook
Samtliga
Kvinnor
Män


Andel toxiska kommentarer (%)
40
35
30
25
20
15
10
5
0
31.9%
33.0%
20.5%
15.2%
16.6%
23.5%
14.4%
9.3%
6.1%
14.2%
7.0%
2.2%
3.5%
2.6%
Förolämpningar
Moral
Kompetens
Psykisk hälsa
Direkta/indirekta hot
Utseende
Kön och sexualitet
Kvinnor
Män

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

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

Hotbedömningar i digitala miljöer

Attentatspersoner använder internet

Textanalys i hotbedömningar 2012 – 2022

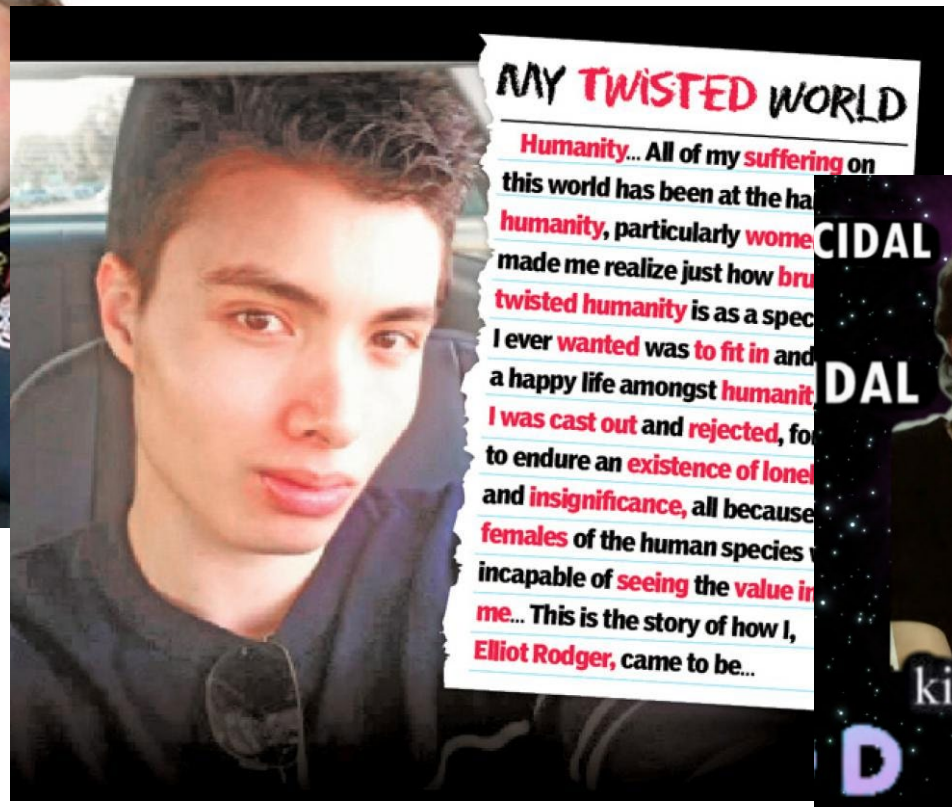
- Kan man stoppa en ensamagerande terrorist i tid?
- Finns det varningssignaler?
- Finns varningssignalerna på internet?
- Är textanalys användbart?

Textanalys i hotbedömningar 2012 – 2022

Utvidgad forskningsfråga

Utvidgad forskningsfråga

Utvidgad forskningsfråga



Varför datoriserade analyser?

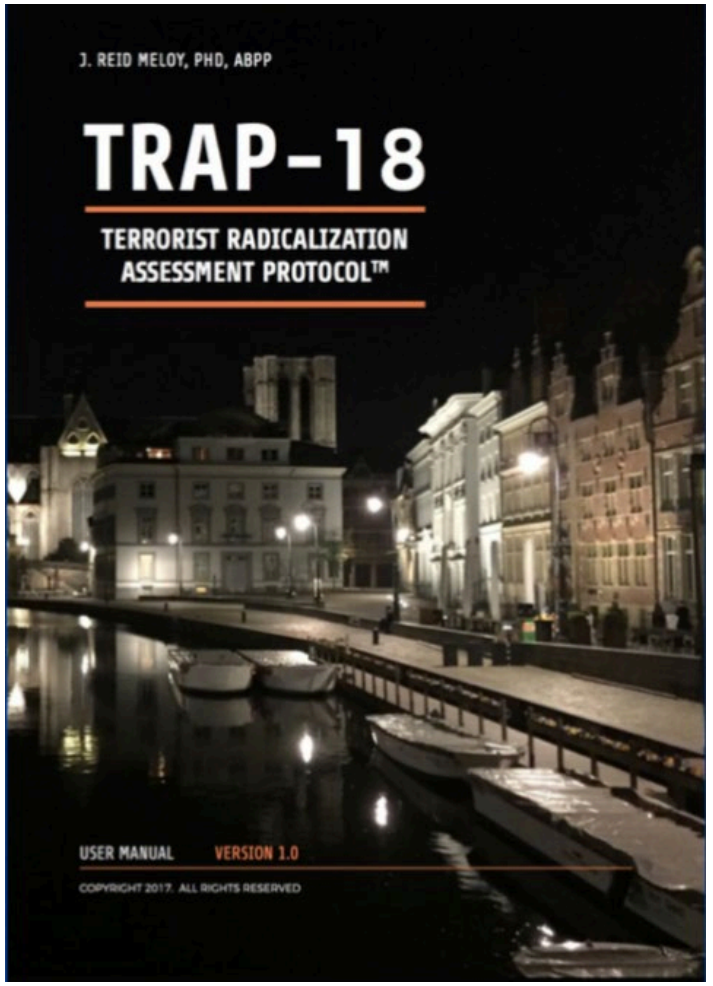
- Sparar mänskliga resurser
- Skyddar integritet
- Möjlighet att se mönster som ett mänskligt öga inte kan överblicka

Datoriserade analyser kräver människor för:

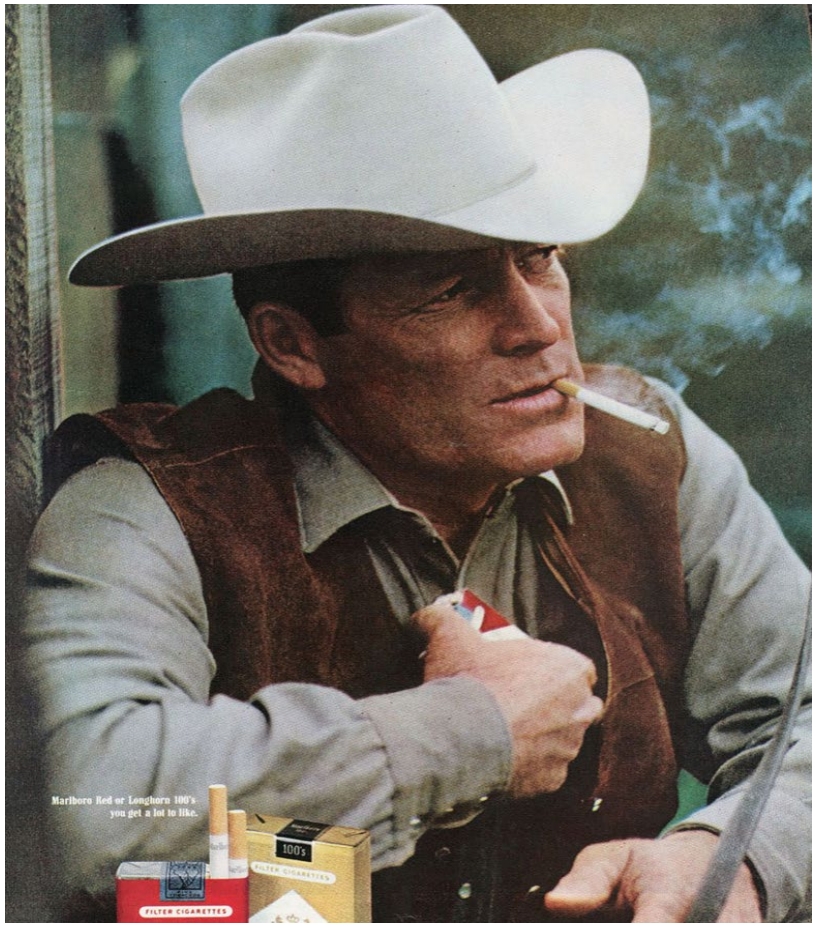
1) rätt frågeställning
2) experttolkning

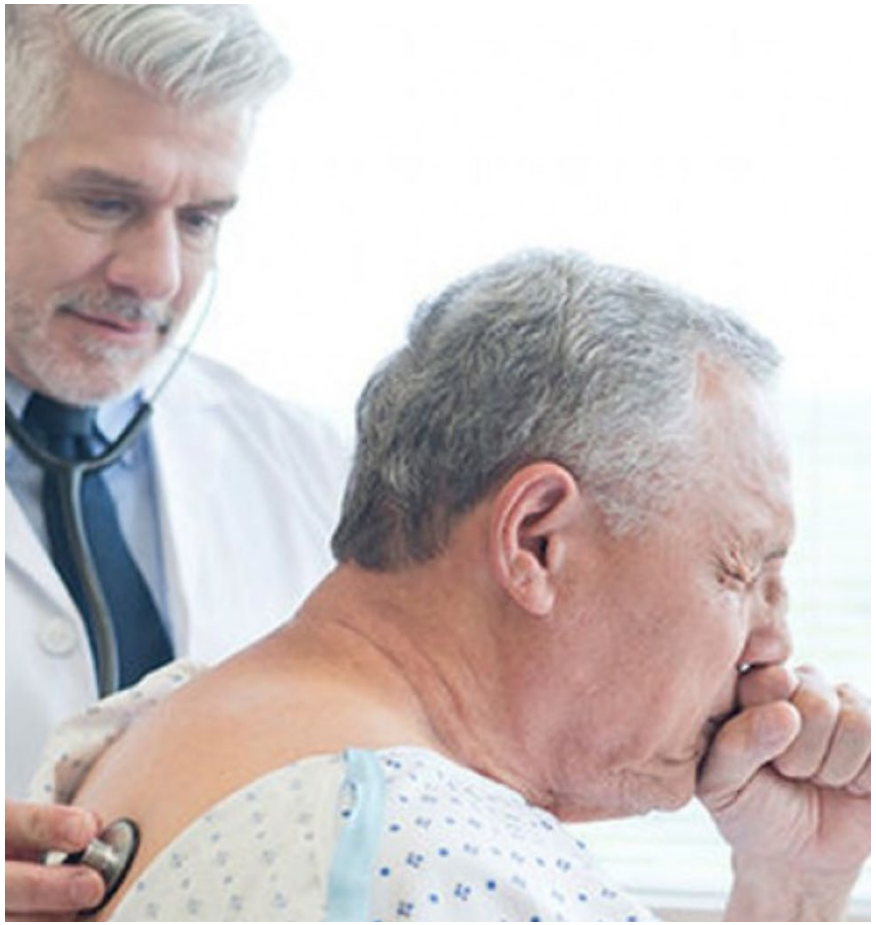
TRAP -18

- Strukturerad professionell bedömning
- Bedömer förstagångsförbrytare
- Ekologisk validitet, kan även bedöma skolskjutare och stalkers
- Baserat på 10 riskfaktorer och 8 beteendeindikatorer



Riskfaktorer och indikatorer





Beteendeindikatorer i digitala miljöer

Indikatorerna i TRAP-18 är tänkta att observeras hos individer som fysiskt är under bevakning, men…

Beteendeindikatorer i digitala miljöer

…flera av beteendeindikatorerna kan observeras där potentiella gärningspersoner ofta finns: i digitala miljöer.

Indikatorer i TRAP-18

- Förberedelse
- Testkörning
- Ökad aktivitet

} konkreta handlingar

- Identifikation
- Fixering
- Tunnelseende

} psykologiska tillstånd

- Läckage
- Explicit avsikt

} (oftast) verbala handlingar

Att observera psykologiska tillstånd…

- Identifikation
- Fixering
- Tunnelseende

} psykologiska tillstånd

…i skriven text

- Sentimentanalys – maskininlärningsmodeller
- Komplexitet i resonemang
  - Läsbarhetsmått (t ex LiX eller Flesch-Kincaid)
  - Ordlistebaserad approach (t ex LIWC)

Exempel: Identifikation

- Ser sig själv som en krigare/militär
- Ser sig själv som nyckelperson i ideologisk kamp
- Ser sig själv som efterträdare till känd våldsverkare

Exempel: Identifikation

- Ser sig själv som en krigare/militär

  Vapen, militär terminologi

- Ser sig själv som nyckelperson i ideologisk kamp

  Uttryck relaterade till handling, ex “göra något”

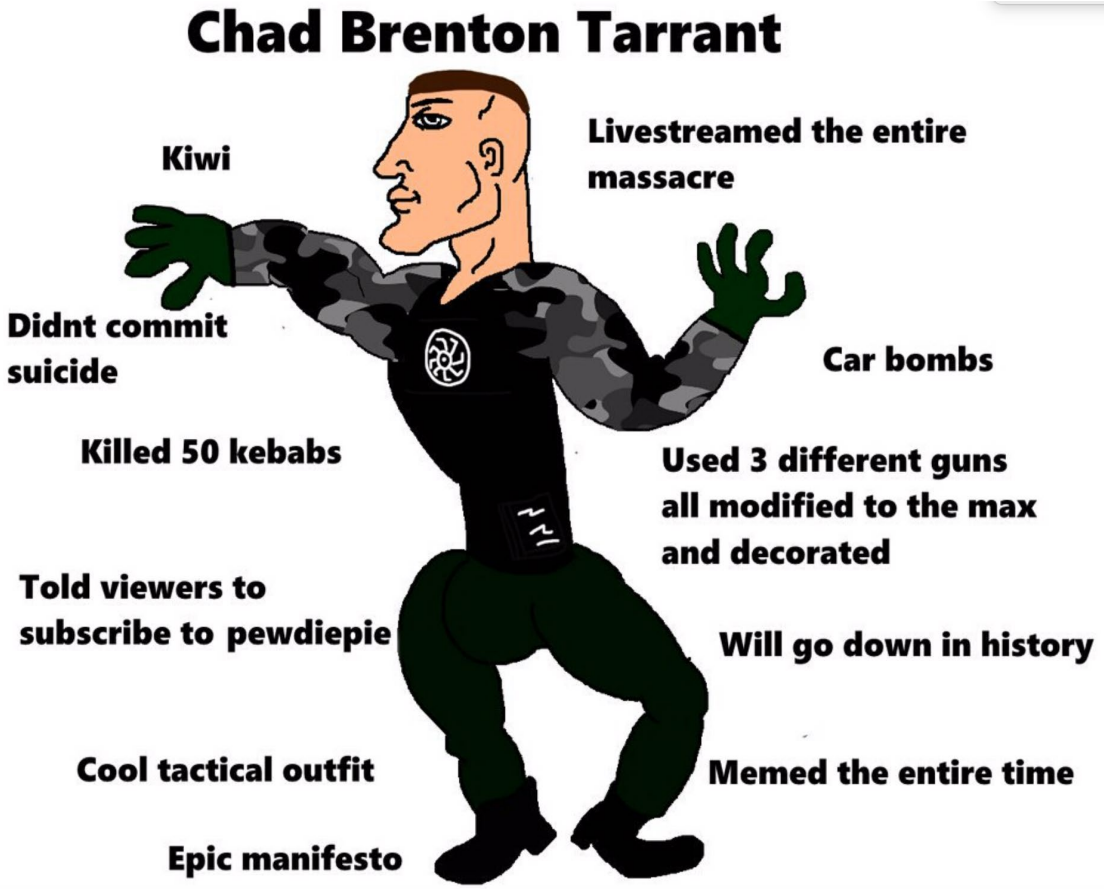
- Ser sig själv som efterträdare till känd våldsverkare

  Tar efter någon annans språkbruk

  “saint”/”chad” etc

Vadå saint och chad?

Saint



Chad

LIWC – Linguistic inquiry and word count

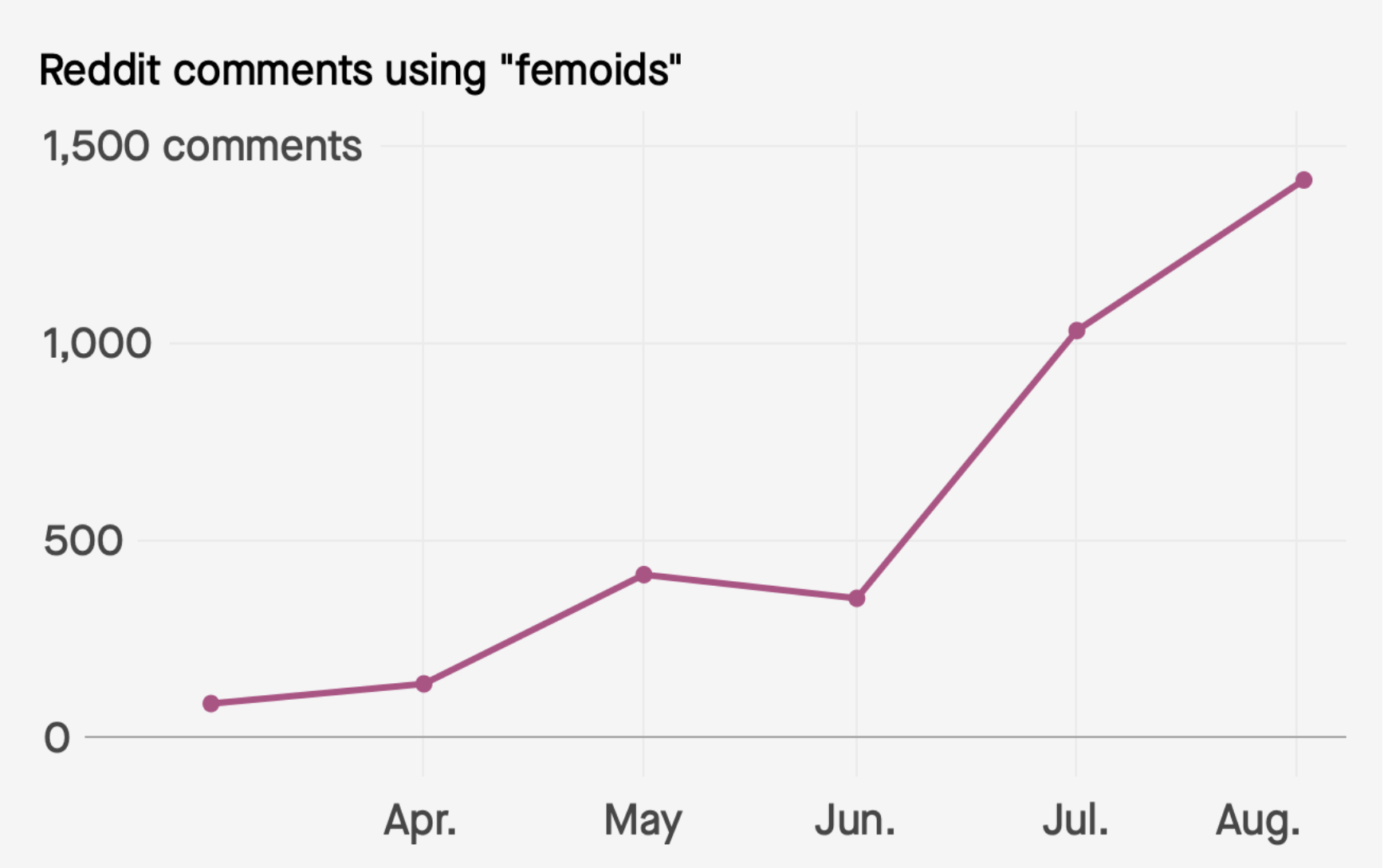
- Ordräkningsbaserat verktyg för stora textmängder
- Ursprungligen utvecklad av psykologer för att mäta effekt av terapeutiskt skrivande
- Bygger på forskning om hur ordbruk korrelerar med kön, ålder, personlighet, utbildning, känslomässiga tillstånd…
- Ordräkningsbaserade verktyg ger mycket information men har begränsningar:

Kontext och jargong

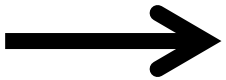
C. What the fuck did you just fucking say about me, you little bitch? I'll have you know I graduated top of my class in the Navy Seals, and I've been involved in numerous secret raids on Al-Quaeda, and I have over 300 confirmed kills. I am trained in gorilla warfare and I'm the top sniper in the entire US armed forces. You are nothing to me but just another target. I will wipe you the fuck out with precision the likes of which has never been seen before on this Earth, mark my fucking words. You think

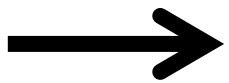
Navy seal copypasta meme

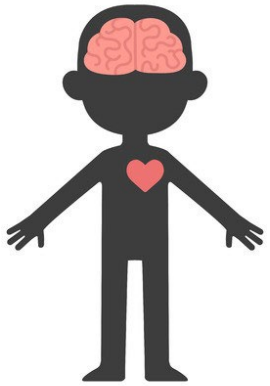
Lingvistiska normer förändras



Anpassade semi-automatiska analysverktyg







Tack!

Lisa.kaati@dsv.se
katie.cohen@foi.se