FOI

Analys och detektion av vilseledning och påverkansoperationer i sociala medier

PÄR-ANDERS ALBINSSON, ULRIK FRANKE, MARIANELA GARCÍA LOZANO,
FREDRIK JOHANSSON, PETER LITSEGÅRD, BJÖRN PELZER,
MAGNUS ROSELL, TOMMY WESTMAN

FOI-R--4142--SE
ISSN 1650-1942 December 2015

Pär-Anders Albinsson, Ulrik Franke, Marianela García Lozano, Fredrik Johansson, Peter Litsegård, Björn Pelzer, Magnus Rosell, Tommy Westman

Analys och detektion av vilseledning och påverkansoperationer i sociala medier

Titel	Analys och detektion av vilseledning och påverkansoperationer i sociala medier
Title	Analysis and detection of influence operations and deception in social media
Rapportnr/Report no	FOI-R-4142-SE
Månad/Month	December
Utgivningsår/Year	2015
Antal sidor/Pages	37
ISSN	1650-1942
Kund/Customer	Försvarsmakten
Forskningsområde	1. Beslutsstödssystem och informationsfusion
FoT-område	Ledning och MSI
Projektnr/Project no	E36728
Godkänd av/Approved by	Lars Höstbeck
Ansvarig avdelning	Informations- och aerosystem
Bild/Cover	Shutterstock

Sammanfattning

Påverkansoperationer av olika slag är frekvent förekommande i politiska och såväl väpnade som icke-väpnade konflikter. Utvecklingen av sociala medier och dess utbredda användning har gjort att påverkansoperationer i ökande grad förekommer även där. Olika statsaktörer använder det för att försöka påverka såväl inhemsk som internationell opinion och terrororganisationer såsom IS har varit framgångsrika i att använda sociala medier för propagandaspridning och att rekrytera nya medlemmar.

I denna rapport presenteras tekniska lösningar för att detektera påverkansoperationer och vilseledningsförsök på mikrobloggen Twitter. De föreslagna verktygen möjliggör inhämtning av data från öppna källor och kombinerar datadrivna metoder (såsom textanalys och algoritmer för automatisk detektion av automatiserat beteende och kapade konton) med visuell datainteraktion. Detta möjliggör en arbetsfördelning där människa och maskin tillsammans kan identifiera möjliga vilseledningsförsök i de enorma datamängder som varje dag produceras i sociala medier.

Nyckelord: big data, vilseledning, påverkanskampanjer, påverkansoperationer, informationsoperationer, sociala medier, underrättelseanalys, Twitter

Summary

Influence operations of various kinds are often part of modern political and military conflicts. The use of deception and information operations in social media is also heavily increasing, due to the development and widespread use of social media platforms such as Twitter and Facebook. There are many examples of how state actors as well as non-state actors are using social media to influence the national or international opinion. Moreover, terror organizations such as Islamic State have been successful in using social media to spread propaganda and to recruit new members.

In this report we present technical solutions for detection of influence operations on the microblog Twitter. The suggested prototype tools make it possible to collect data from open sources and combine data-driven technologies and methods (such as natural language processing and algorithms for detection of automated behavior and hijacked user accounts) with interactive information visualization, enabling a workflow where human and machine can work together in order to identify potential deceptions in the large amounts of data produced every day in social media.

Keywords: big data, deception, influence campaigns, influence operations, information operations, social media, intelligence analysis, Twitter

Innehållsförteckning

1 Inledning

Påverkansoperationer i sociala medier är ett flitigt använt verktyg i såväl politiska, väpnade och icke-väpnade konflikter. FOI har inom FoT-projektet SIA (Stödverktyg för Informationsfusion och Analys) tagit fram ett tekniskt prototypsystem för att hantera, analysera och överblicka de stora datamängder som dessa operationer ger upphov till och presenterar översiktligt systemet och dess ingående komponenter i denna rapport.

I regeringens försvarspolitiska inriktningsproposition från april 2015 lyfts påverkanskampanjer och informationskrigföring fram som en typ av säkerhetshot som Sverige måste ha förmåga att hantera [24]. Regeringen konstaterar också att landet redan idag är utsatt för påverkansoperationer som syftar till att påverka vår säkerhetspolitik. Det är värt att i dess helhet citera den definition som lyfts fram i propositionen (s. 40):

> *En påverkanskampanj är centralt styrd samtidigt som ett brett spektrum av kanaler används, såväl öppna som dolda. Den kan inkludera politiska, diplomatiska, ekonomiska och militära maktmedel, både öppet och dolt, för att nå så stor effekt som möjligt. Officiella talespersoner för fram vissa budskap som är koordinerade med nyhetsförmedling i media. Det kan kompletteras med falska dokument, smutskastning och hot, inklusive med militära medel, styrke-demonstrationer samt dolda operationer såsom fiktiva personer i sociala medier, falska dokument, frontorganisationer eller planterade nyheter. På detta sätt kan exempelvis vinklad, vilseledande eller oriktig information förmedlas som påverkar mottagarens vilja, förståelse, beteenden och attityd.*

Denna rapport beskriver tekniska möjligheter att identifiera och analysera påverkansoperationer – specifika verksamheter som genomförs inom ramen för en påverkanskampanj – på sociala medier. Rapportens målsättning är att bidra till och stötta regeringens ambition att utveckla svensk förmåga att identifiera och möta påverkanskampanjer, såväl i fredstid som vid höjd beredskap. Arbetet är en del av SIA-projektet om datadriven underrättelseanalys som pågått under perioden 2013-2015.

1.1 Avgränsning och läsanvisning

Rapporten är först och främst tekniskt och metodmässigt orienterad och innehåller inte säkerhetspolitiska analyser av påverkanskampanjer. Den beskriver inte heller annat än översiktligt de aspekter av påverkanskampanjer som ligger utanför sociala medier och andra digitala plattformar.

Kapitel 2 introducerar vilseledning och påverkan främst ur ett militärt perspektiv. Det följs av ett kapitel som bredare diskuterar vilseledning och påverkan i det moderna medielandskapet. Kapitel 4 utgör rapportens kärna, med teknisk

beskrivning och diskussion av några metoder för att detektera och analysera tecken på digital vilseledning och påverkanskampanjer som utförs på mikrobloggen Twitter. I det avslutande kapitlet dras några slutsatser utifrån det som presenterats och den forskning som bedrivits inom ramen för projektet, varpå en utblick mot framtiden görs.

2 Militära perspektiv på vilseledning

Vad är vilseledning? Det finns inget självklart svar. Författarna Bell och Whaley [9], som har skrivit en av få hela böcker i ämnet, ägnar sitt förord åt att beklaga hur lite intresse ämnet har väckt i västvärlden, i synnerhet i USA. Däremot spelar vilseledning en mer framträdande roll i icke-västerländska militära kulturer. Exempelvis beskriver Lars Ulfving i FHS-antologin *Krigets traditioner* [8] hur överraskning och vilseledning är ”mycket centrala i sovjetryskt militärt tänkande och handlande” och konstaterar i sin analys av kinesiskt militärt tänkande att ”under de senaste decennierna har utbildningen i vilseledningens konst intensifierats vid de militära akademierna”.

I svensk militär doktrin definieras vilseledning enligt följande i *Handbok Informationsoperationer* [25]:

> *Medvetna åtgärder som syftar till att ge motståndaren ett felaktigt beslutsunderlag för att få aktören att disponera sina resurser på ett sätt som gynnar våra syften.*

Det är värt att notera att syftet alltså inte enbart är att uppnå att en aktör får en felaktig bild av verkligheten, utan att denna bild också ska leda till beslut som omsätts i handling, till exempel truppförflyttningar och förberedelser, som gynnar ”våra” syften.

Försvarsmaktens *Handbok Informationsoperationer* [25] delar in vilseledning i två kategorier:

- Aktiv vilseledning som syftar till att framhäva falsk information (eng. *simulation, showing the false*). I denna typ av vilseledning ”erbjuds” motståndaren falsk information som kan plockas upp av sensorer och bearbetningssystem. Handboken framhåller särskilt att *alla* motståndarens tillgängliga informationskällor kan utnyttjas, inte bara militära.
- Passiv vilseledning som syftar till att dölja sann information (eng. *dissimulation, hiding the real*). I denna typ av vilseledning döljs egen verksamhet med olika sorters kamouflage och maskering. Det enklaste exemplet är naturligtvis materiel som målats för att likna omgivningen, men det kan lika gärna handla om att uppträdandet anpassas, som när ett fartyg under minering stannar upp och låtsas lägga minor på långt fler platser än det faktiskt lägger ut.

Handboken argumenterar också för att vilseledning traditionellt har varit effektivast på de politiska och strategiska nivåerna, eftersom det där går att rikta ansträngningarna direkt mot beslutsprocessen. Därmed närmar sig alltså den militära synen på vilseledning begreppet påverkanskampanj, såsom det används i regeringens proposition. Denna konvergens är särskilt intressant i dagens

informationssamhälle, där sociala medier erbjuder en direktkanal till beslutsfattare och media, men även medborgarna i stort. Vi återkommer till detta tema i nästa kapitel.

En viktig poäng som lyfts fram i *Militärstrategisk doktrin* [27] är att motståndaren inte måste hållas fullständigt omedveten om våra åtgärder – det avgörande är att insikten kommer så sent att det inte längre går att genomföra effektiva motåtgärder. Tidsfaktorn är alltså helt avgörande och det blir missvisande att döma ut en avslöjad vilseledning bara utifrån det faktum att sanningen till slut kom fram. Det som spelar roll är om sanningen kom fram medan det ännu fanns tid att agera. Ett bra exempel på detta är Rysslands agerande och officiella uttalanden under annekteringen av Krim. Från ryskt håll förnekades det länge att de så kallade ”små gröna männen” (ry. *зелёные человечки*), – grönklädda och ofta maskerade soldater som opererade på ukrainskt territorium med ryska vapen och utrustning men utan några insignier – var just ryska soldater. I efterhand har Vladimir Putin erkänt att detta var ryska specialstyrkor och att man från ryskt håll gjort förberedelser för Krims återinträde i Ryssland redan före folkomröstningen bland invånarna på Krimhalvön. När detta erkännande kom var dock den avsedda förvirringen redan uppnådd.

Lyckad vilseledning bygger på kännedom om motståndaren. *Operativ doktrin* [29] poängterar att insikt både om hur motståndaren *hämtar in* ny kunskap (underrättelsetjänst) och sedan *använder* denna kunskap (doktrin och beslutsprocess) är förutsättningar för framgång. Omvänt gäller för underrättelsetjänsten att när så är möjligt alltid försöka nyttja flera olika källor eller inhämtningsdiscipliner för att öka säkerheten i underlaget och minska risken för vilseledning, vilket framhålls i Försvarsmaktens Underrättelsereglemente [26]. Ovan nämnda förutsättningar understryker behovet av att skydda egna inhämtnings- och bearbetningsmetoder. Ju mer motståndaren vet om dessa, desto lättare blir det att lyckas med deras vilseledning. Det är särskilt värt att beakta i det moderna informationssamhället.

I den perspektivstudie som Försvarsmakten redovisade i oktober 2013 är informationsoperationer ett av de områden som särskilt lyfts fram [28]. Bland de faktorer som analyseras är det särskilt värt att notera den snabba utvecklingen inom cybermiljön och hur det moderna informationssamhället har ökat möjligheterna att i realtid sprida budskap – budskap som ”på ett avgörande sätt kan påverka beslutsfattandet på olika nivåer” [28]. Sociala medier kan också påverkas genom olika sorters manipulation och användning av skadlig kod. Perspektivstudien noterar att offensiv förmåga i cybermiljön bland annat kan utnyttjas för vilseledning. Dessa iakttagelser ansluter nära till temat i nästa kapitel.

3 Vilseledning och påverkan i det moderna medielandskapet

Vilseledning och påverkanskampanjer är inga nya företeelser. Däremot skapar teknikens och samhällets utveckling nya och ökade möjligheter till påverkan, vilket också regeringen konstaterar i försvarspropositionen.

En del av denna utveckling handlar om hur etablerade medier alltmer existerar i samverkan med sociala medier. Det innebär att ett paradigm byggt på envägskommunikation från en redaktion till en läsekrets ersätts av ett annat, byggt på flervägskommunikation, där alla åtminstone i teorin kan vara både sändare och mottagare. Sociala medier skapar med andra ord förutsättningar för nyhetsförmedling i nära nog realtid och möjliggör rapportering även från platser dit redaktionerna inte skickat några korrespondenter. Samtidigt har det nackdelar. I slutet av sommaren 2012 dödförklarade *Skövde Nyheter* den då ännu levande Margaret Thatcher – och kallade henne till på köpet tidigare Labourledare – efter att en journalist hade lurats av ett falskt Twitterkonto som såg ut som *Sky News* [44]. När orkanen Sandy drabbade den amerikanska östkusten spred nyhetsmedier snabbt massor av spektakulära men manipulerade fotografier från sociala medier, i tron att de var äkta ögonvittnesskildringar av vad som skedde [42]. Exemplen är inte avsedda att förringa användandet av sociala medier – misstag har alltid begåtts och begreppet nyhetsanka är av betydligt äldre datum. Däremot står det klart att samspelet mellan redaktioner och sociala medier medför nya utmaningar. Tidningen *Metro* har antagit utmaningen genom satsningen Viralgranskaren[1], där tidningens journalister kritiskt granskar och nyanserar historier som har fått stor spridning i sociala medier, inte minst relaterade till dagens stora flyktingströmmar från länder som Syrien.

Forskningen bekräftar bilden av att Twitter och andra sociala medier alltmer används som journalistisk källa och att det påverkar hur källkritik betraktas och utövas [35]. En studie av de brittiska och nederländska valen 2010 visar att enstaka tweets kunde utlösa nyhetsrapportering och författarna spekulerar i att detta beteende förändrar maktbalansen mellan journalister och politiker [12]. Svensk journalistpraxis är inget undantag. Svenska journalister använder ofta Twitter som primärkälla [3] och ger ibland direkta källhänvisningar till olika sociala medier [1]. När så är fallet försöker journalister dock verifiera sanningsenligheten med andra källor före publicering [1]. Samtidigt finns det ett problem i att journalisten inte möter sina källor fysiskt – en Twitter-användares identitet går inte nödvändigtvis att verifiera, vilket riskerar att leda till misstag i stil med exemplen ovan [45] [20]. Det tycks heller inte ovanligt att svenska journalister läser på Facebook, bloggar och Twitter för att hitta unika nyheter [38]. Medieforskaren

[1] http://www.metro.se/nyheter/viralgranskaren/

Erik Lindenius vid Umeå universitet har varnat för att källkritiken från media riskerar att brista ”till förmån för att mer oreflekterat publicera det som uppmärksammats på Twitter som trender” [55].

Denna förändring medför nya möjligheter för den som vill bedriva påverkansoperationer. Facebook uppskattades 2012 ha ungefär 83 miljoner falska användare [51], och så kallade likes går enkelt att köpa av både internationella [48] och svenska [37] företag. Det är med andra ord långt ifrån säkert att det som för en journalist (eller för den delen en vanlig medborgare) ser ut att vara sant eller populärt faktiskt är det.

En stor andel av den falska eller vilseledande information som sprids på sociala nätverk är kommersiellt betingat. Falska recensioner av produkter och tjänster kan dra in pengar till ett företag om de är positiva, eller till en konkurrent om de är negativa. Konsultfirman Gartner har uppskattat att mellan 10 och 15 % av alla recensioner på nätet är köpta [39] och att automatiskt känna igen sådana recensioner är problem som har uppmärksammats inom datavetenskapen [41]. Ett annat sätt att omsätta vilseledning i pengar är otillbörlig marknadspåverkan. TT rapporterade i augusti 2015 om två läkarstudenter som köpte aktier i läkemedelsbolag, skrev positivt om dem i välbesökta elektroniska fora och sedan sålde sina innehav när kurserna hade gått upp [34]. I april 2015 såg Vattenfall sig föranlett att skicka ut ett pressmeddelande med korrekta hemsidor, Twitter-konton och Facebook-sidor, eftersom företaget utsattes för en falsk kommunikationskampanj där någon använde sig av falska elektroniska kanaler för att sprida budskap i Vattenfalls namn [56].

Det är naturligtvis långtifrån all falsk eller vilseledande information som florerar på Internet som håller för en journalistisk granskning avseende sanningsenlighet och relevans. Det som sorteras bort når aldrig traditionella medier. Större delen av diskussionen på Twitter stannar på Twitter – men den påverkar ändå användarnas och deltagarnas syn på det som diskuteras. Ett flertal studier på senare år visar hur politiska diskussioner på Twitter och andra sociala nätverk ger upphov till uppdelade användningsmönster, där interaktion främst sker med likasinnade [15] [50] [7]. Debatten har inte gått Sverige förbi. Journalisten Maria Sköld har beskrivit polariseringen i det svenska medieklimatet: ”Allt fler nöjer sig med att få sin nyhetsdos genom en förmedlare som bekräftar den egna världsbilden: vänstermänniskor väljer LO-stödda Politism, feminister Feministiskt Perspektiv och sverigedemokrater Avpixlat” [49].

Medan nyhetsagendan och urvalet tidigare sattes av redaktionerna är det alltså numera alltmer vi själva – och våra ”vänner” – som väljer vad vi läser. Det betyder att sociala medier är en nyckelarena för aktörer som vill genomföra påverkanskampanjer. Ett sätt att uppnå detta är så kallade strumpdockor (eng. *sock puppets*) – falska användare på sociala nätverk som förses med trovärdig bakgrund, inklusive en IP-adress med rätt geografisk placering. I mars 2011 avslöjades att amerikanska myndigheter upphandlat ett system där en operatör

skulle kunna styra tio sådana strumpdockor på sociala nätverk [21] [2]. I ett slående likartat avslöjande lyckades dagstidningen *Kommersant* röja en upphandling från den ryska utrikesunderrättelsetjänsten SVR av ett system som skulle möjliggöra "massiv spridning av information på utpekade sociala nätverk i syfte att påverka den allmänna opinionen" [6].

Hur stor roll kan sådan manipulation av sociala medier spela? En viktig insikt är att enskilda konton eller små grupper av exempelvis Twitter-användare kan få oproportionerligt stort genomslag. Forskaren Jonas Andersson Schwarz från Södertörns högskola har utifrån studier av den svenska Twitter-debatten konstaterat att åsiktsströmningar kan blåsas upp så att "[å]skådare förleds att tro att en opinion satts i rörelse eller att ett 'drev' pågår, trots att det i själva verket kanske bara rör sig om ett fåtal uttalanden" [5]. En annan aspekt värd att beakta är att politiska beslutsfattare i så hög grad är tillgängliga på sociala medier idag. I en intervju 2012 beskrev till exempel utrikesminister Carl Bildt Twitter som en "viktig nyhetskanal" [32]. På gott och ont är det på ett helt annat sätt möjligt för den som så önskar att påverka beslutsfattares nyhetsflöden idag, jämfört med när Carl Bildt var statsminister 20 år tidigare.

I februari 2015 kunde *Dagens Nyheter* i en stort uppslagen artikel berätta om företaget *Internet issledovanija* (eng. *Internet Research Agency*) i St. Petersburg, där cirka 250 personer dygnet runt producerar påhittade blogginlägg och nätkommentarer. Uppdraget är att hylla Putin och den ryska linjen i världspolitiken, medan västerländska ståndpunkter och politiker förringas och hånas [40]. Detta företag har också enligt *The New York Times* bland annat intressanta kopplingar till en desinformationskampanj kopplad till en påstådd (icke inträffad) olycka vid en amerikansk kemisk fabrik [13]. Genom att använda en Twitter-hashtag med namnet #ColumbianChemicals skickades koordinerat från en rad olika Twitter-konton hundratals meddelanden och påhittade ögonvittnesskildringar till politiker och media om en påstådd explosion som skulle ägt rum i Louisiana. Wikipedia-sidor om händelsen skapades och till och med falska YouTube-klipp producerades och publicerades där en islamistisk terrororganisation påstods ha tagit på sig skulden för explosionen.

Även om det inte var en islamistisk terrororganisation som låg bakom den ovan nämnda fiktiva händelsen så har också dessa framgångsrikt anammat det moderna mediesamhället och dess sårbarheter som metod för att sprida sin världsbild. *The Economist* rapporterade i augusti 2015 att videoproduktionerna från Islamiska staten (IS) håller mycket hög klass både avseende produktionskvalitet och bildspråk, vilket bidrar till deras stora spridning. Antalet IS-sympatisörer på Twitter uppskattas i artikeln till 50 000 [17]. Den stora användningen av sociala medier för att sprida propaganda och psykologisk krigsföring är väldokumenterad även för andra terrororganisationer, bland annat av den välkände israeliska terrorismforskaren Gabriel Weimann [58]. Radikalisering via webben och sociala medier tycks vara en stark bidragande faktor för många av de tusentals människor

som rest från Europa för att delta i väpnade strider i Syrien och Irak de senaste åren.

Det är alltså i detta nya komplexa landskap av etablerade och sociala medier som svensk förmåga att identifiera och möta påverkanskampanjer ska utvecklas. Nästa kapitel utgör rapportens kärna och diskuterar de tekniska förutsättningarna för att hitta och analysera vilseledning och påverkansoperationer på Twitter.

4 Detektion av vilseledning och påverkansoperationer

Begreppet påverkanskampanj är som tidigare konstaterats brett och bäst tillämpbart på strategisk nivå. Ett smalare begrepp är *påverkansoperationer*; de konkreta aktiviteter som genomförs inom ramen för en påverkanskampanj [24]. En rimlig ansats för att hitta och analysera påverkanskampanjer är därför att bryta ner problemet och istället leta efter ingående påverkansoperationer. Emellertid är även dessa till sin natur svårupptäckta och utformade för att i görligaste mån undgå upptäckt. De tekniker som beskrivs nedan syftar därför till att ge analytiker en uppsättning verktyg för att kunna leta efter *indikatorer* på vilseledning främst genom att interagera med stora mängder data inhämtad från sociala medier.

Målsättningen är alltså inte att helt maskinellt kunna detektera vilseledning på sociala medier, utan att hitta en lämplig arbetsdelning mellan människa och maskin som möjliggör denna typ av detektion.

Förekomsten av en indikator är i allmänhet vare sig en nödvändig eller en tillräcklig förutsättning för förekomst av vilseledning eller påverkansoperationer. Däremot kan ett flertal indikatorer tillsammans ge analytikern skäl att tro att det försiggår en påverkansoperation, om exempelvis ett visst budskap systematiskt förs ut (indikator 1) av användarkonton som med hög sannolikhet är automatiserade (indikator 2) och sprids vidare av konton som uppvisar statistiskt avvikande beteenden i sin interaktion med andra (indikator 3). När analytikern interagerar med data kan indikatorerna spela en viktig roll både för att ge uppslag till specifika hypoteser rörande påverkansoperationer och för möjligheten att skapa en normalbild. Metoden med indikatorer och tankarna bakom densamma finns mer ingående beskrivna i en separat vetenskaplig artikel [31].

Metoderna och teknikerna vi utvecklat och beskriver i denna rapport är tänkta att vara generella nog att kunna appliceras på olika typer av sociala medier. Än så länge har vi dock valt att begränsa oss till att göra faktisk datainhämtning och analys kopplad till mikrobloggen Twitter[2]. Från Twitter har vi under en längre tid hämtat hem tweets[3] och tillhörande metadata (såsom publiceringstid för ett tweet och ålder på kontot) som relaterar till de sökbegrepp vi satt upp. Dessa sökbegrepp är tänkta att täcka in tweets som har med den pågående konflikten mellan Ukraina och Ryssland att göra och genererar relativt stora datamängder (ett antal gigabyte per dag i medeltal).

I återstoden av detta kapitel redogörs för ett antal exempel på automatisk bearbetning som görs av inhämtade tweets och de konton från vilka dessa skickas.

[2] http://twitter.com/
[3] Inhämtningen har gjorts med det Streaming API som tillhandahålls av Twitter.

Syftet med denna bearbetning är att utvinna förädlad information som kan ligga till grund för de indikatorer som användaren av de utvecklade verktygen kan använda sig av för att detektera pågående försök till vilseledning eller påverkansoperationer. Efter att den automatiska bearbetningen gjorts är tanken att användaren (analytikern) interagerar med inhämtad data med hjälp av visuell datainteraktion, i syfte att få en bättre förståelse för data och att i samverkan med de automatiska verktygen kunna identifiera påverkansoperationer.

4.1 Automatisk textanalys

Bland tillgängliga data från sociala medier är det skrivna meddelandet ofta den rikaste källan till information. Tyvärr är texten ostrukturerad och därför svårare att behandla i befintligt skick än de ofta mer strukturerade metadata som också hör till (t.ex. tidpunkt för publicering av meddelandet, antal vänner och ålder på konto). För att kunna analysera stora mängder textuell data måste man därför först försöka strukturera den på olika sätt för att sedan kunna utvinna information ifrån den. Språkteknologi eller datorlingvistik (eng. *language technology*) är det forskningsområde inom vilket man utvecklar tekniker för automatisk analys av text. Vi har i detta arbete begränsat oss till texter på engelska, inhämtade från Twitter. De metoder som vi använder är relativt generella men måste vanligtvis anpassas beroende på vilken källa och vilket språk som är av intresse. En viktig sak att notera är att språkbruket på sociala medier såsom Twitter skiljer sig mycket från traditionell nyhetstext, vilket är vad de flesta metoder inom språkteknologi tidigare utvecklats för.

Det finns ett flertal tillgängliga verktyg, program och programmoduler för att göra textanalys. I SIA-projektet har vi konstruerat en så kallad NLP[4]-komponent, det vill säga en komponent som utför olika typer av textanalys. Denna NLP-komponent är i sin tur uppbyggd av olika tillgängliga java-bibliotek för textanalys. I arbetet med Twitter-data har vi utgått från java-biblioteket GATE[5], och mer specifikt delkomponenten TwitIE [10], vilken är speciellt utvecklad för att göra textanalys på tweets. Exempel på grundläggande textanalys som utförs av TwitIE är att dela upp texten i meningar och tokens (bokstavssekvenser som ofta motsvarar ord), för vilka ordklass och grundform tas fram.

Genom att analysera texten som finns i tweets kan vi exempelvis identifiera och markera delar av texten som är speciellt intressanta. Vi kallar generellt dessa delar för extraktioner. GATE och TwitIE, liksom många andra tredjepartsbibliotek, erbjuder en specifik typ av extraktioner som innebär att man kan identifiera entiteter, det vill säga namngivna objekt, så som personer (Barack Obama), platser (Stockholm, Spanien) och organisationer (FOI, Apple Inc.).

[4] NLP – natural language processing
[5] https://gate.ac.uk/

För att möjliggöra utvinning av annan information i texten har vi implementerat ett regelspråk som kallas för ”object patterns”. Idén liknar den som beskrivs i en artikel av Hearst [33] och kan liknas vid att använda reguljära uttryck på ordnivå istället för bokstavsnivå (vilket traditionella reguljära uttryck annars appliceras på). I varje regel går det också att använda sig av entiteter och extraktioner från andra regler, vilket gör språket uttrycksfullt och utbyggbart.

Vid användning av dessa object patterns skriver en användare mer eller mindre enkla regler som fångar speciellt intressanta uttryck och fraser. Det är en regelbaserad metod, och har därmed både för- och nackdelar: det är å ena sidan ibland enklare för en människa att förstå resultatet från en regelbaserad metod än vid användning av maskininlärning, men det krävs manuellt arbete från användaren för att skriva reglerna. Vidare krävs också ofta manuellt arbete för att skapa de träningsdata som annars behövs vid maskininlärning.

Tanken med regelspråket är att det ska gå relativt snabbt att börja studera något fenomen som är av intresse. Vi har testat regelspråket och tagit fram regler för ett par olika fall som bland annat relaterar till vilseledning. I nästa avsnitt beskrivs ett specifikt exempel. Utöver det har vi också tagit fram några enkla regler för att fånga en del mer generella fraser. Tanken med dessa regler är att ge en överblick av innehållet i texterna som analyseras.

Som ett första användningsområde för de utvecklade textanalysteknikerna har vi börjat med att försöka fånga explicita textuella uttryck för att någon försöker vilseleda, ljuga, eller sprida propaganda. På detta stadium handlar det inte om att automatiskt försöka identifiera lögner eller vilseledningar i text, utan om att automatiskt identifiera tweets där någon uttryckligen anklagar någon för att fara med osanning eller på annat sätt föra någon bakom ljuset. Tanken är att om någon skriver att något är lögn alternativt osanning så är det antingen det, eller så far vederbörande med osanning (medvetet eller omedvetet). Båda dessa fall är av intresse.

För att fånga explicita osanningar listade vi först engelska ord som vi ansåg hörde hit, som t.ex: chicanery, cunning, deceit, deceitfulness, deceive, deception, defamation, hoax, hypocrisy, hypocrite, lie, lier, misdirect, misdirection, misinformation, propaganda, untrue, untruth, untruthful, vilify och vilification. Vi skrev sedan regler som använder sig av dessa begrepp. Tanken här är inte att försöka fånga allt, utan att snarare försöka visa på att detta kan ge resultat med relativt liten insats. I Figur 1 syns en del av resultatet då vi tillämpade dessa regler på en liten mängd tweets nedladdade med sökord rörande Ukraina-krisen.

Resultaten från textanalysen kan dels nyttjas i den visuella datainteraktionen som beskrivs vidare i Kapitel 4.4, och dels i det fristående prototypverktyget PhraseBrowser som inte beskrivs närmare i denna rapport.

ukraine propaganda: 55	putin propaganda: 3
us propaganda: 23	obama's lie: 3
obama lie: 19	russia's subterfuge: 3
obama deception: 11	ukraine lie: 2
russian deception: 9	euromaidan propaganda: 2
lie mr. president: 8	slander russia: 2
obama lies: 5	propaganda re #ukraine: 2
moscow propaganda: 4	obama easter lie: 2
kremlin propaganda: 4	cold war propaganda: 2
hypocrisy russia: 3	russia lies: 2
russia propaganda: 3	

Aggressive Russian Nationalism Unleashed: In **Moscow Propaganda** War Even Opposition Is Singing Kremlin Tune: http://t.co/Q69xvjlpJS

RT @Roger_Moorhouse: Similar methods followed in Latvia and Estonia. **Russian deception** operations - #maskirovka - have a long, inglorious …

RT @PZilgalvis: Anne Applebaum - A fearful new world, imperiled by **Russia's subterfuge**: a new kind of war, a new kind of invasion. http://…

Lavrov and **Putin propaganda** machine are on overdrive today- preparing for Russian troops to move into East #Ukraine?

RT @OSCE_RFoM: The Kidnapping Of Journalists Is The Latest Step In **Ukraine Propaganda** Wars http://t.co/GL4BlSIfw4 via @maxseddon @buzzfeed

@StateOfUkraine schools/kindergartens were even closed today! a new low in **euromaidan propaganda**. a disgraceful lie.

Figur 1: En del av resultatet från en körning av regler för detektion av explicita osanningar och propaganda på en liten mängd tweets nedladdade med sökord rörande Ukraina-krisen. Överst några fraser och antalet gånger de förekom, nederst några tweets som dessa

4.2 Detektion av automatiserat beteende

Ett vanligt förekommande fenomen vid påverkansoperationer i sociala medier är användningen av så kallade botar (eng. *bots*). Enkelt uttryckt kan detta sägas vara automatiserade konton som exempelvis kan användas för att:

- automatiskt reagera på vissa nyckelord genom retweets eller att svara med fördefinierade budskap,
- sprida virus eller annan skadlig kod, eller
- spamma Twitter-hashtags med tweets i syfte att få dessa att bli s.k. ”trending topics” eller att dränka relevanta diskussioner såsom hashtags om politiska protester med helt irrelevanta tweets (det senare har exempelvis förekommit för att tysta mexikanska [22], syriska [62] och ryska [30] aktivister).

Inom befintlig datavetenskaplig forskning finns det ett relativt stort antal forskningsartiklar som fokuserar på problemet med att upptäcka automatiserat beteende. Majoriteten av dessa artiklar är dock inriktade på det relativt avgränsade delproblemet med att upptäcka spam eller spammare i sociala medier. Algoritmer avsedda att detektera spam bygger ofta på olika typer av *kontoegenskaper* [14] (t.ex. antal följare), *innehållsbaserade attribut* [4] (t.ex. likhet mellan meddelanden skickade från kontot eller antal spamrelaterade ord i meddelandena) och/eller *nätverksbaserade egenskaper* [16] [11] [23] (t.ex. hur stor spridning ett meddelande får eller till vilken grad någon svarar på de meddelanden som ett konto skickar).

Att sprida spam kan i vissa informationsoperationssammanhang vara ett effektivt sätt att vilseleda genom att dränka annan information med orelaterat innehåll. På det stora hela finns det dock också flera andra typer av informations- och påverkansoperationer som använder sig av automatiserat beteende utan att för den skull sprida spam. Detta gör att det generella problemet med att upptäcka automatiserat beteende i sociala medier i allmänhet, och vid användning för påverkansoperationer i synnerhet, blir extra intressant.

För detektion av automatiserade Twitter-konton från ett tekniskt perspektiv rekommenderas den intresserade läsaren att läsa artikeln ”Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?” [14]. Som titeln antyder skiljer dessa författare inte bara mellan människor och botar, utan också den tredje kategorin ”cyborgs”, vilken definieras som ett mellanting där en bot och människa hjälps åt. Ett typexempel på detta är en människa som registrerar ett konto och som interagerar med sina vänner, men däremellan använder automatiserad programvara för att med regelbundna mellanrum skicka ut automatiska tweets (t.ex. nyhetsartiklar med ett visst innehåll). Baserat på deras analys av en stor mängd Twitter-data från människor, botar och cyborgs tar författarna fram ett automatiserat klassificeringssystem som försöker avgöra om ett konto är mänskligt

eller automatiserat baserat på en rad olika egenskaper. Utdata från detta system indikerar att ungefär 50 % av alla Twitter-användare i deras datamängd är mänskliga, 40 % är cyborgs och 10 % är botar. Om detta stämmer innebär det att endast hälften av alla Twitter-konton utgörs av användare som inte använder automatiserad programvara, vilket tydligt visar på vikten av att kunna skilja mänskliga från icke-mänskliga användare.

Så vitt vi vet finns det tidigare inget vetenskapligt arbete som tar upp problemet med att detektera (storskalig) användning av automatiserade konton för politiska och militära informationsoperationer. I vårt arbete [52] föreslår vi därför den metodik som översiktligt sammanfattas nedan:

Första steget i metoden är att samla in data som är relevant för den aktuella kontexten, i syfte att skapa en så kallad träningsmängd. I vårt fall har vi än så länge bara applicerat metoden på Twitter-data som relaterar till den pågående konflikten mellan Ryssland och Ukraina, men syftet är att metoden ska vara generisk och lätt att applicera på andra datamängder (eller rent utav andra typer av sociala medier). Givet antagandet att en relativt stor del av alla meddelande och konton som används för att skicka dessa härstammar från automatiserad aktivitet så är nästa steg i metoden att för ett stort antal relevanta attribut markera de mest extrema som härstammande från potentiella bot-konton. För att exemplifiera detta kan vi ta attributet *antal-skickade-meddelanden-per-timme*. Undersöks detta attribut för alla konton i den valda datamängden så kommer vi se en fördelning där väldigt många ligger nära värdet 0, men det även finns många konton som har betydligt högre värde än så. För de mest extrema värdena där konton under en längre tidsperiod har ett genomsnitt på hundratals eller rent av tusentals meddelanden så går det med ganska hög säkerhet säga att detta är ett resultat av automatiserat beteende. Poängen här är att inte göra detta för endast ett attribut utan många, där vi bland annat inkluderar många olika attribut som har att göra med likhet mellan skickade meddelanden, de ämnen det skrivs om och de mottagare man adresserar, men också den regelbundenhet med vilka meddelande skickas. Efter att på detta sätt ha märkt upp konton i träningsmängden som potentiella bot-konton utifrån ett antal olika attribut görs sedan en samlad bedömning av huruvida ett konto ska taggas som ett bot-konto eller ej. I vår första implementation av metoden har alla konton som för minst ett attribut identifierats som avvikande taggats som botar, medan övriga konton i träningsmängden taggas som manuella konton. Mer avancerade och alternativa angreppssätt är dock tänkbara framöver.

I nästa steg används den tidigare (semi-automatiskt) märkta träningsmängden för att träna upp en så kallad maskininlärningsbaserad klassificerare som lär sig skilja på botar och mänskliga konton. Syftet med detta steg är att datorn själv lär sig att hitta mer komplexa mönster som kan användas för att skilja botar från människor. Resultatet från denna process är en klassificerare som sedan kan användas för att appliceras på nya konton som hämtas in och därmed direkt kan kategoriseras som antingen botar eller mänskliga konton. En sådan klassificerare har implementerats

i våra prototypverktyg, vilket gör att alla inhämtade Twitter-konton som har en tillräckligt stor mängd associerade tweets kan förses med en tagg som anger huruvida detta konto tros vara automatiserat eller inte. Dessa bedömningar är inte perfekta utan behäftade med en viss osäkerhet. Preliminära tester visar att bedömningarna har god *precision* men sämre *recall*, det vill säga att en stor del av de konton som pekas ut som automatiserade är det, men att en relativt hög andel av faktiska automatiserade konton inte identifieras.

4.3 Detektion av kapade konton

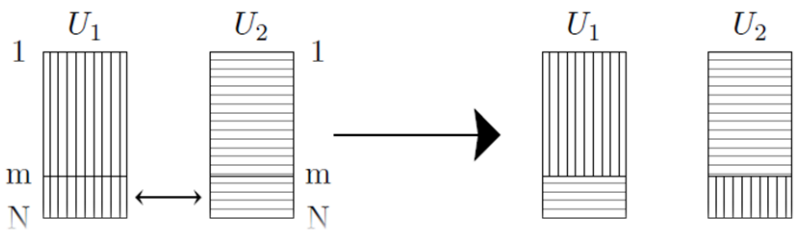
Det har gjorts relativt omfattande forskning avseende att utveckla algoritmer och system som kan lära sig att känna igen spam och andra typer av automatiserat beteende. Detta gör att Twitter och andra sociala medietjänster ofta är relativt snabba med att stänga ned konton som används för att sprida uppenbart spam. Mindre nogräknade aktörer som vill sprida sitt budskap, desinformation eller skadlig kod till en större mängd mottagare använder därför ofta kapade konton som en ersättning för, eller ett komplement till, användandet av automatiserade konton skapade för detta ändamål. Genom att kapa existerande konton ges en möjlighet till att utnyttja de relationer och den trovärdighet eller auktoritet som de rättmätiga ägarna till dessa konton har. Ett välkänt exempel på en sådan kontokapning är det tillfälliga övertagande av ett Twitter-konto tillhörande Associated Press som gjordes 2013. Kapningen uppmärksammades relativt snabbt, men inte snabbare än att meddelandet: ”Breaking: Two Explosions in the White House and Barack Obama is injured” hann skickas ut till nästan två miljoner följare och spridas vidare som en löpeld med tusentals retweets och stora temporära konsekvenser för den amerikanska börsen. Hackare tillhörande *Syrian Electronic Army* som stödjer Syriens president Bashar al-Assad har också tagit på sig flera kapningar av andra Twitter-konton, däribland CBS News och människorättsorganisationen Human Rights Watch.

Relaterat till problematiken med kapade konton är också fenomenet med att människor mer eller mindre informerat övertygas eller luras till att ”låna ut” sina konton. Ett exempel på detta är terrororganisationen IS som utvecklat appen ”The dawn of glad tidings”. Vid nedladdningen av denna ger användaren tillstånd till appen att autonomt publicera tweets från användarens Twitter-konto, vilket möjliggör storskaliga och kostnadseffektiva påverkanskampanjer som når en stor målgrupp.

För att möjliggöra detektion av kapade konton på Twitter har FOI tagit fram algoritmer [53] som i mångt och mycket bygger på metoden COMPA [18]. Något förenklat kan denna metod sägas bestå av en träningsfas där första steget är att etablera en normalmodell för hur ett visst användarkonto vanligtvis beter sig och sedan flagga upp avvikelser från denna normalmodell som potentiella kapningar. Normalmodellen etableras genom att beräkna hur ofta olika attribut förekommer i

den observerade träningsfasen, till exempel hur ofta det aktuella användarkontot skickat tweets en viss timme på dagen, hur ofta ett visst språk har använts i användarkontots skickade tweets, eller hur frekvent användaren använt en viss hashtag. Genom att beräkna motsvarande värden för nya tweets från användarkontot och att jämföra dessa mot den existerande normalmodellen erhålls ”anomalipoäng” för varje attribut. Dessa aggregeras sedan ihop till ett viktat medelvärde. Om denna aggregerade anomalipoäng visar sig vara större än ett fördefinierat tröskelvärde flaggas kontot som potentiellt kapat, medan det för anomalipoäng som är lägre än tröskelvärdet förutsätts vara normalt.

Ett uppenbart problem med den föreslagna metoden är att sätta ett ”korrekt” tröskelvärde. Om det sätts alltför högt så finns det risk att faktiska kapningar inte kommer att upptäckas, medan alltför låga tröskelvärden riskerar att leda till en ansenlig mängd falsklarm. För att på ett bättre sätt kunna studera effekterna av olika tröskelvärden och att jämföra olika varianter av algoritmer mot varandra har vi tagit fram en utvärderingsmetodik i vilken vi skapar ”artificiella” kapningar genom att slumpmässigt ”växla” eller ”korsa” tweets skrivna av en användare med tweets från en annan användare (Figur 2).



Figur 2: Artificiell "kapning" av två konton, där kapningen sker vid tidpunkt *m*.

Om den utvärderade algoritmen ger utslag före det att någon växling av tweets har skett så är detta ett tecken på att algoritmen eller systemet är för känsligt (genererar falsklarm) medan det är för okänsligt om det går lång tid från det att två konton har korsats till det att algoritmen eller systemet reagerar. På detta sätt kan vi utan tillgång till stora mängder faktiska kapningar få en ganska bra bild av hur väl olika algoritmer fungerar. Genom att jämföra olika algoritmer och val av tröskelvärden för när ett konto ska flaggas som en potentiell kapning har vi följaktligen kunnat identifiera lämpliga tröskelvärden som ger en god balans mellan upptäcktssannolikheten och antalet falsklarm.

Exempel på tillämpningar för den utvecklade förmågan att identifiera kapade konton är att använda den för att identifiera potentiella kapningar av Twitter-konton som tillhör politiker, nyhetsmedia, etc. Kapningar av dessa konton kan få

stora effekter på samhället om de används för att sprida desinformation och felaktiga uppgifter som inte upptäcks i tid.

4.4 Visuell datainteraktion

Föregående delkapitel (kapitel 4.1, 4.2 och 4.3) beskriver olika sätt att automatiskt analysera och klassificera Twitter-meddelanden i jakten på användare och meddelanden som skapats för att användas i olika typer av påverkans- och informationsoperationer. Dessa automatiserade processer kan snabbt appliceras på stora mängder data på ett sätt en människa aldrig skulle klara av. Men processernas underliggande regler måste tas fram utifrån någon slags förståelse för vad det är för fenomen som eftersöks och därmed begränsas de till att fånga just det.

Ett bredare och mer förutsättningslöst tillvägagångssätt är att förse analytiker med ett verktyg för att visuellt interagera med ett stort underlag av data från sociala medier. Många miljoner tweets och Twitter-användare kan åskådliggöras visuellt från en mängd olika perspektiv. Det finns flera syften med ett sådant verktyg:

- *Interagera med data*: Genom att först ge analytikern en visuell översikt av en större mängd data kan bredare mönster och potentiellt intressanta tidsperioder hittas. Med hjälp av att zooma in och filtrera direkt i de olika diagrammen och visuella presentationerna, kan dessa potentiellt intressanta områden undersökas mer noggrant från flera olika perspektiv. Steg för steg kan analytikern borra sig ner i allt mer detaljrika data. Målet är att till slut hitta konkreta fall av vilseledning.
- *Identifiera nya hypoteser*: När nya mönster, avvikelser eller beteenden har identifierats kan dessa omsättas till hypoteser som kan läggas till det batteri av regler för automatisk analys av vilseledning som sedan appliceras på större datamängder och på så sätt berikar underlaget med metadata.
- *Manuell kategorisering*: När data görs visuellt greppbara uppkommer ett behov att markera utsnitt och direkt ge dem etiketter. Det kan röra sig om att tagga användare som analytikern misstänker är automatiserade, eller tweets som anses ta ställning i en viss fråga. Det är relaterat till att identifiera nya hypoteser, men handlar vanligtvis om ett försteg i den processen.
- *Sätta automatiskt genererad metadata i en kontext*: Det är inte bara rådata från Twitter som kan åskådliggöras och interageras med – den automatiska analysen kan införlivas i de visuella representationerna. Målet med att integrera metadata från den automatiska analysen är att stödja analytikerna att hitta i och förstå underlaget. Därtill ger kombinationen data och metadata möjlighet att värdera kvaliteten på den automatiska analysen.

Det finns ett antal forskningsområden som mer eller mindre inkluderar hur visuella representationer av dataunderlag bör utformas och analyseras, som exempelvis informationsvisualisering, visual analytics och business intelligence. Vi väljer att kalla vår ansats *visuell datainteraktion* för att tydliggöra att det både handlar om att ta fram lösningar för att visualisera och interagera med data. Att presentera samma underliggande data på olika sätt kan göra ett till synes omöjligt problem mer hanterbart. Att dessutom snabbt kunna interagera med dataunderlaget gör att en analytiker kan växla mellan flera perspektiv för att systematiskt sätta sig in i materialet.

Det blir allt billigare att tekniskt samla in och lagra stora mängder data. När en människa ska ta till sig information från de växande dataunderlagen måste det till effektiva sätt att presentera och interagera med dem eftersom människans kognitiva kapacitet inte utvecklas på samma sätt som tekniken. Viktiga data måste framhävas och den mänskliga användaren (i vårt fall analytikern) måste stödjas i att gå från översikt till detaljer och navigera bland data med snabba svarstider. Några exempel på kända problem inom visualisering av stora mängder data är:

- Dataöverbelastning (eng. *data overload*): användarens förmåga att ta till sig information är lägre än vad mängden data kräver [59].
- Förändringsblindhet (eng. *change blindness*): användaren går miste om förändringar i systemet [43].
- Förlorat visuellt moment (eng. *lost visual momentum*): användaren tappar tempo vid övergångar mellan olika delar av systemet [61] [57].
- Tillkomst av sekundära uppgifter: användaren tappar fokus på huvuduppgiften på grund av sekundära uppgifter [59].
- Automatiseringsöverraskningar (eng. *automation surprises*): när användaren lämnats utanför av systemets automatik och får svårare att hantera uppkomna situationer utanför systemets förmåga [36][60].

Syftet med visuell datainteraktion är att försöka tackla dessa problem genom ett systematiskt designarbete. En längre beskrivning av det systematiska designarbetet är utanför ramen för denna rapport men några av de ingående principerna inkluderar följande:

- Att först ge först en *översikt* av data, därefter effektiva möjligheter till att *zooma och filtrera* för att sen ge *detaljer vid behov* [47].
- Hitta olika *referensramar* som är relevanta för underliggande data, sätt data i en *kontext* och *framhäv* förändring och relevans [59].
- Minimera grafik som inte representerar verkliga data och var konsekvent i designen [54].

- *Dynamiska frågor*: Låt användaren interagera med data direkt i diagrammen och ge omedelbar återkoppling [46].

Det finns ett flertal välkända datorverktyg[6] för visuell datainteraktion som används inom vitt skilda områden där stora mängder data genereras. Eftersom datakällan sociala medier snabbt blivit populär, har ofta de mer kända verktygen specifika moduler som hanterar denna typ av data. När problemet snävas in på visuell datainteraktion av stora mängder data från sociala medier för analys av påverkansoperationer blir forskningsunderlaget genast tunnare. Den ansats vi använt oss av är därför en kombination av existerande produkter och egen utveckling och anpassning. I vår nuvarande implementation och det exempel som beskrivs härnäst har vi använt oss av analysramverket TIBCO Spotfire[7].

4.4.1 Exempel på identifiering av vilseledning och påverkansoperationer

Det första som görs i verktyget är att definiera vilken del av datamängden som ska hämtas från databasen. I Figur 3 visas resultatet av en sådan databashämtning med en tidsperiod på tre månader och med söktermer som försöker fånga Ukraina-krisen (överst till vänster). Dataunderlaget som hämtats är aggregerat på minutnivå, vilket innebär att alla värden relaterar till minut-utsnitt. Fyra tidslinjer ur olika perspektiv visar antalet tweets över tid i de valda tidsperioden. Höjden på staplarna visar antalet tweets. Det går tydligt att se att några luckor finns i vårt dataunderlag och hur en stor ”spik” av aktivitet startar under andra halvan av tidsperioden. Dygnsvariationen framgår också klart. Färgkodningen fångar olika aggregerade attribut för dessa tweets:

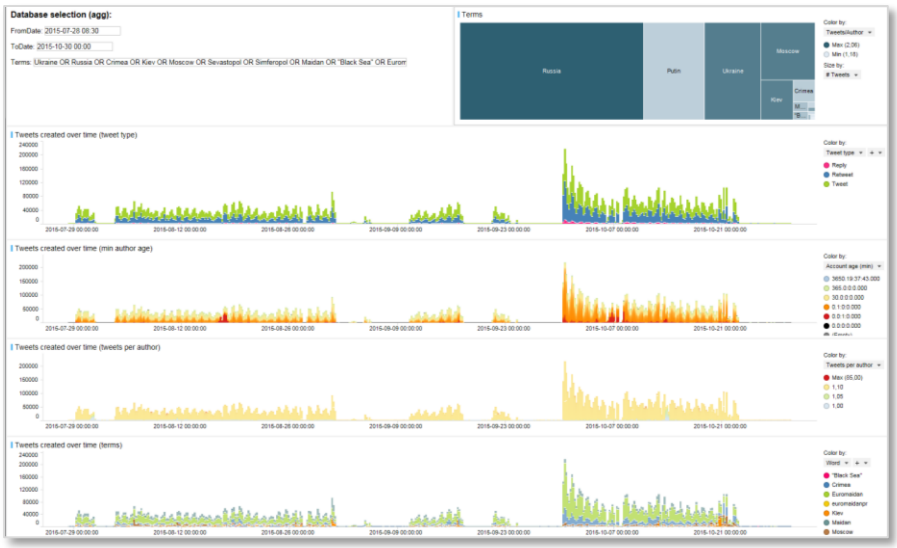
- Första tidslinjen visar typ av tweet. Grön står för vanlig ny tweet, blå för retweet och lila för svar på tweet. Vi ser hur svar är en mycken liten del av alla tweets i denna period.
- Andra tidslinjen visar åldern på konton. Mörkt rött står för minutgamla konton, och kallare färger allt äldre konton. Aggregeringen är, i det här fallet, kopplad till det *yngsta* kontot för varje minut men man kan enkelt ändra det till exempelvis medelvärde. Vi kan se framförallt två perioder på några dagar där mycket nya konton oftare har twittrat.
- Tredje tidslinjen visar genomsnittliga antalet tweets per konto. Kalla färger står för få tweets per användare och varmare färger för fler. Det är få mönster som kan urskiljas förutom en kort period närmare slutet på

[6] Några exempel på verktyg för visuell datainteraktion är: Spotfire, QlikView, Tableau, Kibana, Zoomdata, Scraawl, SAS Visual Analytics, Datawatch
[7] Mer information om TIBCO Spotfire finns här: http://spotfire.tibco.com

tidsperioden där plötsligt färre antal tweets skrivs i genomsnitt per användare.

- Fjärde tidslinjen visar fördelningen av tweets över olika söktermer. Varje färg representerar ett sökord. I vårt fall dominerar den ljusgröna färgen som står för ”Russia”.

Högst upp till höger i figuren visas ett så kallat treemap-diagram där varje fyrkant representerar en sökterm. Storleken står för antalet tweets som innehåller termen och färgkodningen står för hur många tweets per författare termen i genomsnitt har. Vi ser hur termerna ”Putin” och ”Ukraine” påträffas ungefär lika ofta i vårt underlag, men att ”Putin” återfinns i tweets där det är färre antal författare per tweet.



Figur 3: Tre månaders Twitter-data aggregerad på minutnivå.

I nästa steg ska vi titta närmare på en av perioderna då fler nya konton användes. Genom att markera den tidsperioden i den aggregerade tidslinjen hämtas underliggande rådata hem från databasen och presenteras i egna vyer. I Figur 4 visas urvalet från fem olika perspektiv:

1. Sökord
2. Typ av tweet
3. Tweets per användare
4. Genomsnittligt antal tweets per användare
5. Ålder på konto vid tweet-tidpunkten

Vi ville titta vidare på aktiviteten från de mycket nya kontona som syns som mörka element i botten på den nedersta tidslinjen. Genom att byta vy kan vi tydligare se dessa konton (Figur 5).

Figur 4: Drygt två dagars Twitter-data (med glapp där data fattas).

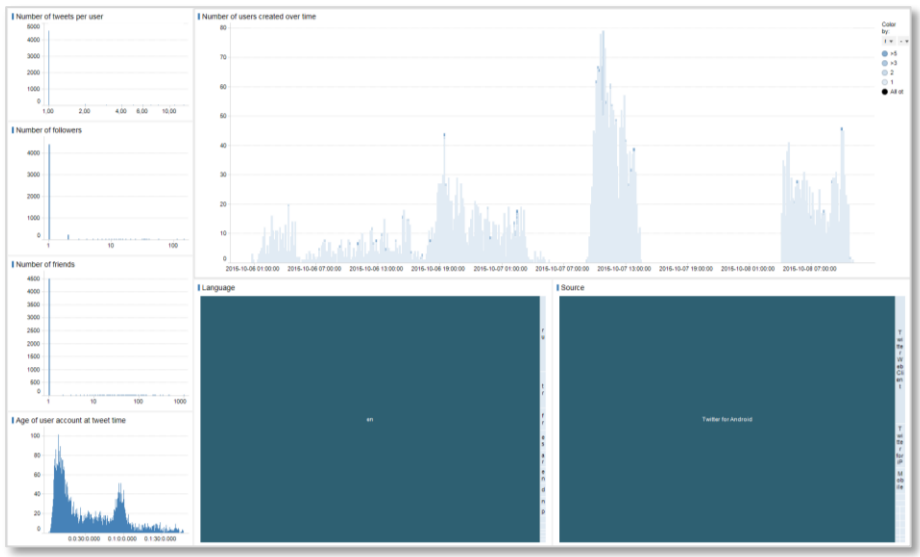
I Figur 5 visas histogram och treemaps för sju av tweetsens attribut (med början med de fyra mindre histogrammen):

1. Antalet tweets per användare: En mycket tydlig fördelning där de allra flesta skickat ett enda tweet.
2. Antalet följare: Väldigt många konton utan någon följare.
3. Antalet vänner: Väldigt många konton utan någon vän.
4. Ålder på kontot vid tweet-tidpunkten: En ”spik” för mycket unga konton.
5. Den större tidslinjen visar när i tiden kontot skapades. En ”spik” mycket nyligen är markant.
6. Vår treemap till vänster visar fördelningen på språk för kontona. Flest är inställda på engelskt språk.
7. Vår treemap till höger visar fördelningen på Twitter-klient. Storlek står för antal användare och färg representerar antal tweets.

Den orimligt höga ”spiken” för mycket nya konton sticker ut tydligt. Genom att zooma in på denna topp får vi ett nytt utseende enligt Figur 6. Vi ser bland annat att så gott som samtliga 4 500 konton har samma språk och klient. Dessa konton

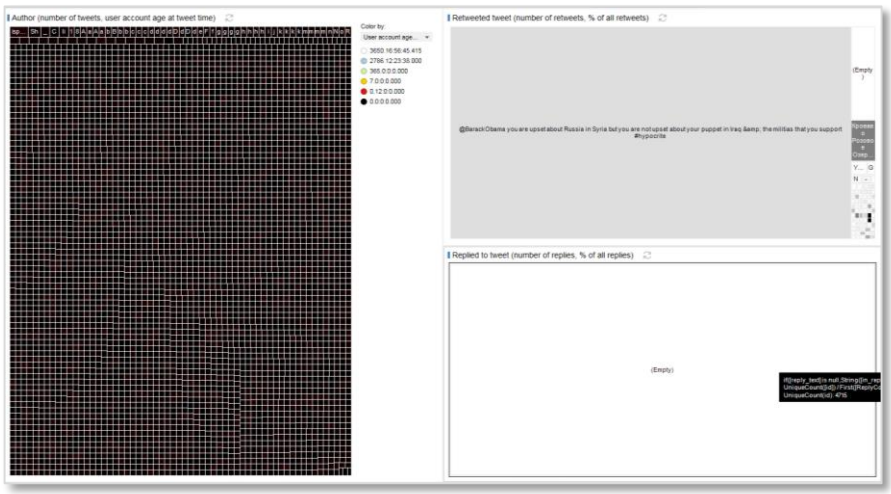
är uppenbarligen suspekta och vi väljer till sist att titta närmare på vad de har utfört under tidsperioden.

Figur 5: Användarnas statistiska fördelning för ett antal attribut.



Figur 6: De helt nyskapade kontonas fördelning på attribut.

I den sista figuren för vårt exempel (Figur 7) visas en treemap till vänster över samtliga 4 500 konton som vi filtrerade fram. Så gott som samtliga har samma storlek (de har skickat en tweet) och har samma färg (lika gamla). Ovan till höger finns ytterligare en treemap som visar vad (om något) som har retweetats av dessa användare. Det blir tydligt att nästan alla användare har retweetat en och samma Obama-kritiska tweet (stor fyrkant). Färgkodningen i detta fall står för hur stor del av tweetens alla retweets som skett inom vårt utvalda utsnitt: Vitt står för 0 % och svart står för 100 %.



Figur 7: De 4 500 lika gamla kontona som nästan samtliga reweetar en Obama-kritisk tweet.

Ovanstående är endast ett exempel på användning av verktyget för visuell datainteraktion, men förhoppningsvis ger det läsaren en idé om hur det snabbt går att analysera data på en grov aggregerad nivå, hitta olika former av avvikelser genom ren manuell inspektion, och sedan borra sig ned i underliggande data för att hitta olika typer av försök till vilseledning eller påverkansoperationer.

5 Slutsatser

I denna rapport har vi beskrivit hur vilseledning och påverkansoperationer i sociala medier allt viktigare medel i såväl politiska som militära konflikter. Statsaktörer såväl som terrororganisationer använder ofta Twitter och andra typer av sociala medier för att sprida propaganda, politiska budskap, och för att försöka påverka såväl medborgare, journalister och beslutsfattare. Utifrån perspektivet av denna utveckling blir förmågan att kunna detektera pågående vilseledningsförsök och påverkansoperationer allt viktigare.

I den forskning som bedrivits inom FoT-projektet Stödverktyg för Informationsfusion och Analys (SIA) har FOI tagit fram tekniker och verktyg för att identifiera indikatorer som kan tyda på någon form av vilseledning eller påverkansoperationer i sociala medier. Dessa innefattar bland annat olika former av automatisk textanalys, detektion av automatiserat beteende, och detektion av kapade konton. Därtill har olika typer av visuell datainteraktion tillämpats för att ge analytikern eller användaren en förbättrad förmåga att analysera möjliga vilseledningar. De tekniker och verktyg som presenterats i denna rapport ska ses som ett första steg mot att upprätta en förmåga där människa och maskin tillsammans kan identifiera påverkansoperationer i sociala medier. Det som presenterats i denna rapport är forskningsprototyper, inga skarpa system för operativ användning.

För de påverkansoperationer som bedrivs på sociala medier i dagsläget kan i många fall relativt enkla tekniker användas för att hitta tecken på dessa operationer: automatiserade konton skapas inte sällan i stora kvantiteter vid en och samma tidpunkt och twittrar inte sällan exakt samma budskap vid exakt samma tid på dygnet. I takt med att ”motmedel” i form av automatiserade och semi-automatiserade algoritmer och system för detektion av påverkansoperationer utvecklas kan vi dock förvänta oss att ”angriparen” blir alltmer sofistikerad i sina metoder. På detta sätt förutser författarna till denna rapport ett framtida ”arms race” där utvecklingen av metoder och tekniker för detektion av påverkansoperationer leder till mer sinnrika verktyg för en intelligent motståndares aktiva påverkansoperationer, som leder till förfinade metoder och tekniker för detektioner av påverkansoperationer, och så vidare. Med anledning av detta förefaller det rimligt att framtida framsteg och resultat inom forskning rörande detektion av påverkansoperationer inte kommer att offentliggöras i någon detaljerad grad. Detta är också anledningen till att de beskrivningar av de framtagna verktygen som görs i denna rapport är på en relativt övergripande nivå.

Att ta fram verktyg för att möjliggöra upptäckt och analys av potentiella vilseledningar och påverkansoperationer är ett viktigt steg på vägen, men för att ge någon reell nytta är det också viktigt att policy-beslut fattas om hur vi från svenskt håll vill eller ska bemöt påverkanskampanjer som tar plats i sociala medier.

Referenser

[1] Susan Abbasnejad och Ronak Moaf. Sociala medier i den svenska nyhetsrapporteringen om Iran. Examensarbete, Södertörns högskola, 2010.

[2] Spencer Ackerman. Jihadis' Next Online Buddy Could Be a Soldier. http://www.wired.com/dangerroom/2011/03/jihadis-next-online-buddy-could-be-a-soldier/, mars 2011. Wired, läst 18 januari 2013.

[3] Rebecca af Ugglas Åberg och Sofia Åkesdotter. Twitter som källa?: En totalundersökning av synliga källor i nyhetsartiklar under revolutionen i Egypten 2011. Examensarbete, Stockholms universitet, 2011.

[4] Amit A Amleshwaram, Narasimha Reddy, Sandeep Yadav, Guofei Gu och Chao Yang. Cats: Characterizing automation of twitter spammers. I: *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*, ss 1–10. IEEE, 2013.

[5] Jonas Andersson Schwarz, Johan Hammarlund, Magnus Kjellberg och Stefan di Grado. ”åsikter på sociala medier är inte den allmänna opinionen”. *Dagens Nyheter*, s 5 (Nyhetsdelen), 25 december 2014.

[6] Ilya Barabanov, Ivan Safronov och Elena Chernenko. Razvedka botom [Intelligence Using a Bot]. http://www.kommersant.ru/doc/2009256, augusti 2012. Kommersant, retrieved 7 November 2012.

[7] Vladimir Barash och John Kelly. Salience vs. commitment: Dynamics of political hashtags in Russian Twitter. Berkman Center Research Publication 2012-9, 2012.

[8] Arne Baudin, Anders Cedergren, Ove Pappila och Lars Ulfving. *Krigets traditioner*. Försvarshögskolan, Stockholm, 2011.

[9] J Bowyer Bell och Barton Whaley. *Cheating and deception*. Transaction Publishers New Brunswick, 1991.

[10] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard och Niraj Aswani. TwitIE: An open-source information extraction pipeline for microblog text. I: *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

[11] P.O. Boykin och V.P. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, April 2005.

[12] Marcel Broersma och Todd Graham. Social media as beat: tweets as a news source during the 2010 British and Dutch elections. *Journalism Practice*, 6(3):403–419, 2012.

[13] Adrian Chen. The agency. http://mobile.nytimes.com/2015/06/07/magazine/the-agency.html, juni 2015. The New York Times Magazine.

[14] Zi Chu, Steven Gianvecchio, Haining Wang och Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *Dependable and Secure Computing, IEEE Transactions on*, 9(6):811–824, 2012.

[15] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer och Alessandro Flammini. Political polarization on twitter. I: *International Conference on Weblogs and Social Media (ICWSM)*, 2011.

[16] Dave DeBarr och Harry Wechsler. Using social network analysis for spam detection. I: *Proceedings of the Third International Conference on Social Computing, Behavioral Modeling, and Prediction*, SBP'10, ss 62–69, Berlin, Heidelberg, 2010. Springer-Verlag.

[17] The propaganda war. *The Economist*, ss 28–29, 15 augusti 2015.

[18] Manuel Egele, Gianluca Stringhini, Christopher Kruegel och Giovanni Vigna. COMPA: Detecting compromised accounts on social networks. I: *Proceedings of the Network and Distributed System Security Symposium*, 2013.

[19] Martin J. Eppler och Ken W. Platts. Visual strategizing: The systematic use of visualization in the strategic-planning process. *Long Range Planning*, 42:42–74, 2009.

[20] Emma Fagerberg och Elsa Larsson. "Ölhallen"– en hjälp i nyhetsarbetet: En kvalitativ studie av några nyhetsjournalister och sociala medier. Examensarbete, Mittuniversitetet, 2010.

[21] Nick Fielding och Ian Cobain. Revealed: US spy operation that manipulates social media. http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks, mars 2011. The Guardian, läst 18 januari 2013.

[22] Klint Finley. Pro-government twitter bots try to hush mexican activists. http://www.wired.com/2015/08/pro-government-twitter-bots-try-hush-mexican-activists/, september 2015. Wired, läst 9 september 2015.

[23] M. Fire, G. Katz och Y. Elovici. Strangers intrusion detection – detecting spammers and fake profiles in social networks based on topology anomalies. *Human*, 2012.

[24] *Försvarspolitisk inriktning – Sveriges försvar 2016-2020*. Försvarsdepartementet, Regeringen, Prop. 2014/15:109, Stockholm, april 2015.

[25] *Handbok Informationsoperationer*. Försvarsmakten, Stockholm, 2008.

[26] *Försvarsmaktens Underrättelsereglemente (FM UndR)*. Försvarsmakten, Stockholm, 2010.

[27] *Militärstrategisk doktrin (MSD 12)*. Försvarsmakten, Stockholm, 2012.

[28] Perspektivstudien 2013. Försvarsmakten, beteckning FM2013-276:1, oktober 2013.

[29] *Operativ doktrin 2014 (OPD)*. Försvarsmakten, Stockholm, 2014.

[30] Ulrik Franke och Carolina Vendil Pallin. Russian politics and the internet in 2012. Teknisk rapport FOI-R–3590–SE, FOI, 2013.

[31] Ulrik Franke och Magnus Rosell. Prospects for Detecting Deception on Twitter. I: *2014 International Conference on Future Internet of Things and Cloud (FiCloud)*, ss 528–533. IEEE, 2014.

[32] Lasse Granestrand. Resa i den twittrande klassens universum. *Dagens Nyheter*, s 16 (Söndagsdelen), 7 oktober 2012.

[33] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. I: *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, ss 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[34] Janerik Henriksson. Bloggare utreds för börsfiffel. TT, 6 augusti 2015.

[35] Alfred Hermida. Tweets and truth: Journalism as a discipline of collaborative verification. *Journalism Practice*, 6(5-6):659–668, 2012.

[36] Erik Hollnagel och Sidney Dekker. *Coping with computers in the cockpit*. Brookfield, VT : Ashgate, 1999. Includes index.

[37] Marie Kennedy. Gott om fejkade gilla på fejan. *Göteborgs-Posten*, ss 42–43, 5 oktober 2012.

[38] Emma Leyman och Ida Sönner. Från blogginlägg till nyhet: En kvantitativ studie om hur journalister använder sociala nätverk för att hitta nyheter. Examensarbete, Linnéuniversitetet, 2010.

[39] Alexander Linden och Jenny Sussin. How Crowdsourcing Can Reduce the Reliability of Social Media Analytics. Teknisk rapport, Gartner, Inc., februari 2014. G00260476.

[40] Ingmar Nevéus. De är soldater i kriget på nätet. *Dagens Nyheter*, ss 8–10 (Nyhetsdelen), 5 februari 2015.

[41] Myle Ott, Yejin Choi, Claire Cardie och Jeffrey T Hancock. Finding deceptive opinion spam by any stretch of the imagination. I: *Proceedings*

of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, ss 309–319. Association for Computational Linguistics, 2011.

[42] Ester Pollack. Sandy blåste liv i fejkade fotografier. *Svenska dagbladet*, s 12 (Kulturdelen), 8 november 2012.

[43] R. Rensink. Change detection. *Annual Review of Psychology*, 53:245–277, 2002.

[44] Daniel Sandström. Mediala snedsteg i sociala medier. *Sydsvenskan*, s A2, 2 september 2012.

[45] Alireza Shamsparto. Twitter och nyhetsrapportering: En studie om Twitters roll för journalism och kommunikation. Examensarbete, Södertörns högskola, 2011.

[46] Ben Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, november 1994.

[47] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. I: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, ss 336–, Washington, DC, USA, 1996. IEEE Computer Society.

[48] Ryan Singel. Juking Your Facebook 'Like' Stats Is as Easy as Sending a Message. http://www.wired.com/threatlevel/2012/10/facebook-likes-messages/, oktober 2012. Wired, läst 18 januari 2013.

[49] Maria Sköld. Medievanor ger ny politisk sprängkraft. *Göteborgs-Posten*, ss 40–41, 2 augusti 2014.

[50] Stefan Stieglitz och Linh Dang-Xuan. Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior. I: *System Science (HICSS), 2012 45th Hawaii International Conference on*, ss 3500–3509. IEEE, 2012.

[51] Mark Sweney. Facebook quarterly report reveals 83m profiles are fake. http://www.guardian.co.uk/technology/2012/aug/02/facebook-83m-profiles-bogus-fake, augusti 2012. The Guardian, läst 18 januari 2013.

[52] Christopher Teljstedt, Magnus Rosell och Fredrik Johansson. A Semi-Automatic Approach for Labeling Large Amounts of Automated and Non-Automated Social Media User Accounts. I: *Proc. The Second European Network Intelligence Conference, ENIC*. IEEE, 2015. In press.

[53] David Trång, Fredrik Johansson och Magnus Rosell. Evaluating Algorithms for Detection of Compromised Social Media User Accounts. I: *Proc. The Second European Network Intelligence Conference, ENIC*. IEEE, 2015. In press.

[54] Edward Tufte. *The visual display of quantitative information*. Graphics Press, 1983.

[55] Elin Turborn. Twitter har få användare – men stort genomslag. *Västerbottens-Kuriren*, s 10, 4 februari 2012.

[56] Falsk kommunikationsaktion mot Vattenfall. http://corporate.vattenfall.se/press-och-media/pressmeddelanden/2015/falsk-kommunikationsaktion-mot-vattenfall2/, 4 april 2015.

[57] D. Woods & J. Watts. *Handbook of Human-Computer Interaction*, kapitel How not to have to navigate through too many displays, ss 617–650. Elsevier, 1997.

[58] Gabriel Weimann. New terrorism and new media. Teknisk rapport, Wilson Center, 2014.

[59] D. Woods. Toward a theoretical base for representation design in the computer medium: Ecological perception and aiding human cognition. I: J. Flach, P. Hancock, J. Caird och K. Vicente, redaktörer, *Global perspectives on the ecology of human-machine systems*, band 1, ss 157–188. Hillsdale, Erlbaum, 1995.

[60] David Woods och Sidney Dekker. Anticipating the effects of technological change: A new era of dynamics for human factors. *Theoretical Issues in Ergonomics Science*, 1(3):272–282, 2000.

[61] David D. Woods. Visual momentum: A concept to improve the cognitive coupling of person and computer. *Int. J. Man-Mach. Stud.*, 21(3):229–244, september 1984.

[62] Jillian C. York. Syria's twitter spambots. http://www.theguardian.com/commentisfree/2011/apr/21/syria-twitter-spambots-pro-revolution, April 2011. The Guardian, läst 9 september 2015.

FOI är en huvudsakligen uppdragsfinansierad myndighet under Försvarsdepartementet. Kärnverksamheten är forskning, metod- och teknikutveckling till nytta för försvar och säkerhet. Organisationen har cirka 1000 anställda varav ungefär 800 är forskare. Detta gör organisationen till Sveriges största forskningsinstitut. FOI ger kunderna tillgång till ledande expertis inom ett stort antal tillämpningsområden såsom säkerhetspolitiska studier och analyser inom försvar och säkerhet, bedömning av olika typer av hot, system för ledning och hantering av kriser, skydd mot och hantering av farliga ämnen, IT-säkerhet och nya sensorers möjligheter.

FOI
Totalförsvarets forskningsinstitut
164 90 Stockholm

Tel: 08-55 50 30 00
Fax: 08-55 50 31 00

www.foi.se