

Examensarbete

15 högskolepoäng, Grundnivå

Att hitta en nål i en höstack: Metoder och tekniker för att sälla och gradera stora mängder ostrukturerad textdata

Finding a Needle in a Haystack: Methods and techniques for
screening and grading large amounts of unstructured textual
data

Albin Carlson

Emeli Pettersson

Sammanfattning

Big Data är i dagsläget ett populärt ämne som kan användas för en mängd olika syften. Bland annat kan det användas för att analysera data på webben i hopp om att identifiera brott mot mänskliga rättigheter. Genom att tillämpa tekniker inom områden som *Artificiell Intelligens (AI)*, *Information Retrieval (IR)* samt *data-visualisering*, hoppas företaget Globalworks AB kunna identifiera röster vilka uttrycker sig om förtryck och kränkningar i social media. Artificiell intelligens och informationshämtning är dock breda områden och forskning som behandlar dem kan finnas långt tillbaka i tiden. Vi har därför valt att utföra en systematisk litteraturstudie i syfte att kartlägga existerande forskning inom dessa områden. Med en litterär sammanställning bistår vi med en ontologisk överblick i hur ett system som använder dessa tekniker är strukturerat, med vilka metoder och teknologier ett sådant system kan utvecklas, samt hur dessa kan kombineras.

Abstract

Big Data is a popular topic these days which can be utilized for numerous purposes. It can, for instance, be used in order to analyse data made available online in hopes of identifying violations against human rights. By applying techniques within such areas as *Artificial Intelligence (AI)*, *Information Retrieval (IR)*, and *Visual Analytics*, the company Globalworks Ltd. aims to identify single voices in social media expressing grievances concerning such violations. Artificial Intelligence and Information Retrieval are broad topics however, and have been an active area of research for quite some time. We have therefore chosen to conduct a systematic literature review in hopes of mapping together existing research covering these areas. By presenting a literary compilation, we provide an ontological view of how an information system utilizing techniques within these areas could be structured, in addition to how such a system could deploy said techniques.

Tillkännagivande

Vi vill rikta ett stort tack till vår handledare Gion Koch Svedberg. Tack för allt stöd du gett oss under arbetets gång. Tack för att du hjälpt oss att konkretisera våra tankar när vi svävat ut, och ett sista TACK för att du, alltid i positiv och glad anda, pushat oss framåt vid stunder där det gått trögt.

Ordlista

NLP - Diverse tekniker vilka bearbetar naturligt språk i form av text och tal, främst för att få datorer att tolka det mänskliga språket. Har många område för användning, som exempelvis språköversättning, tal till text, och vise versa.

Web crawler/spider - Ett script eller program vilket används i syftet att spara ner data från diverse webbportaler eller via API:er på ett automatiserat eller manuellt vis.

Human-in-the-Loop - Konceptet att en människa tar del i akten av att exempelvis träna en maskininlärningsalgoritm. En människa i inlärningsprocessen.

Noise - Redundant data, eller "brus", vilken inte kan användas i syftet att extrahera information. Förekommer i en mängd olika former såsom *HTML*-taggar och reklam. Kan även te sig i form av stoppord som *the*, och *a* i engelskspråkig text, vilka utgör föga nytta för algoritmer som behandlar ostrukturerat språk. Det är av stor vikt att försöka reducera detta brus i högsta möjliga mån då resultatet av algoritmer som bearbetar datan påverkas.

Semantic analysis - Refererar inom lingvistik till processen att härleda mening utifrån syntaktiska strukturer i skriven text och tal. Datorer förstår inte skriftspråk på samma vis som en människa gör. De är inkapabla att resonera kring innebörden av ord och meningar utifrån erfarenheter och kontext. Därför krävs det att semantiken i datan analyseras och regelverk konstrueras.

Sentiment analysis/Opinion mining - En process vilken försöker analysera uttryckt sentiment från människor givet ett specifikt ämne. Kan bland annat användas som ett verktyg för att analysera positiv eller negativ respons utifrån kundrecensioner för en given produkt.

Innehållsförteckning

1	Introduktion	1
1.2	Tidigare forskning	2
1.2.1	Skrapning	2
1.2.2	Behandling	2
1.2.3	Lagring	3
1.2.4	Visualisering för insiktsgenerering	3
2	Problemformulering	4
2.1	Syfte och mål	5
2.2	Avgränsningar	5
3	Metod och metoddiskussion	6
3.1	Metodbeskrivning	6
3.2	Datainsamling	7
3.2.1	Databaser	7
3.2.2	Sökning och sällning av litteratur	8
3.3	Metoddiskussion	9
3.3.1	Experiment	9
3.3.2	Fallstudie	9
3.3.3	Design and Creation	10
4	Resultat	11
4.1	Litteraturstudie	11
4.1.1	Litteratursökning	11
4.2	Sammanfattning av fynd i litteraturen	14
4.2.1	Skrapning av data	15
4.2.2	Behandling av data	16
4.2.2.1	Maskininlärning	23
4.2.3	Lagring av data	27
4.2.4	Visualisering för insiktsgenerering	27
5	Analys	30
5.1	Visualisering för insiktsgenerering	32
5.2	Behandling av data	32
5.2.1	Maskininlärning	32
5.3	Lagring av data	33
5.4	Skrapning av data	33
6	Diskussion	34

6.1 Etiska aspekter	34
7 Slutsatser och vidare forskning	36
8 Referenser	37

1 Introduktion

I det moderna samhället genereras dagligen enorma mängder digital data som en följd av människors konstanta uppkoppling till internet via portabla enheter. År 2017 uppmättes antalet unika mobilanvändare i hela världen uppnå 5 miljarder och siffrorna beräknas fortsätta stiga till hela 5.9 miljarder fram till år 2025 [1]. Big Data är en term som anses synonym med denna explosionsartade datagenerering, och kan definieras som data vilken vuxit sig så pass stor att den blivit problematisk att hantera med traditionella medel [2]. Företag kan potentiellt använda sig av Big Data för att driva deras verksamhet framåt [3], då den personliga information som genereras från de miljardtals internetanvändarna kan användas för att bland annat identifiera trender inom en uppsjö av områden [4].

Globalworks AB är ett icke-statligt företag vilket ämnar att skrapa och bearbeta textuell data via sociala medier, bloggar, forum och nyhetsartiklar i syftet att försöka finna indikationer på brott mot mänskliga rättigheter på arbetsplatser i högriskländer som exempelvis Kina, Thailand och Vietnam. I rapporten “*Wasting time, wasting youth*” [5] presenterar bolaget sin proposition med vilken de skall försöka uppnå sin vision. De har utvecklat verktyget *social@risk*TM, vilket kan komplettera fysisk och socialt orienterad företagsbesiktning (eng. *auditing*), och detta utan de medföljande konsekvenserna en sådan kan medföra. Repressalier kan förekomma för intervjuade arbetare vid traditionella, fysiska besiktningar. Genom att skrapa och analysera data från sociala medier efter klagomål från anställda, kan detta åstadkommas anonymt utan att arbetares identitet avslöjas i processen. Utifrån rapporten framgår det att en rad olika bekymmer existerar för arbetare i Kinesiska fabriker, och detta trots tidigare försök att eliminera dem. Elva typfall presenteras, vilka har funnits med Globalworks informationssystem:

“ [...] 1. Klagomål relaterade till rekrytering, 2. Diskriminering, 3. Förtryck från ledning, 4. Klagomål relaterade till löner, 5. Överdriven övertid, 6. Ineffektiva medel för rapportering av klagomål, 7. Tvångsarbete, 8. Studentarbetare/interner, 9. Klagomål relaterade till inkvartering, 10. Psykologisk stress, 11. Yrkesrelaterad ohälsa och brist av säkerhet.” [5].

I dagsläget står Globalworks inför problemet att försöka fullfölja sin vision om att tillgodose möjligheten att proaktivt identifiera brott mot mänskliga rättigheter i globala försörjningskedjor. För att kunna fullfölja visionen erfordras en undersökning av dagens “*state-of-the-art*” för forskning inom områden vilka relaterar till förmågan att skrapa data, samt filtrera och analysera denna. Målet är att, i en så hög grad som möjligt, försöka automatisera extraktionen av relevant information utifrån stora mängder data. Den extraherade informationen skall agera underlag för en manuell analys utförd av en expertpanel i hopp om att dra konkreta slutsatser rörande brister och kränkningar av mänskliga rättigheter på arbetsplatser.

Följande delkapitel ämnar måla upp bakomliggande information för ämnet vilket arbetet skall beröra. Syftet är orientera läsaren i problemområdet på en övergripande nivå för att denne på ett begripligt sätt skall kunna bli införstådd med de mer djupgående beskrivningarna som återfinns i efterföljande kapitel. Inledningen syftar även att måla upp det huvudsakliga syftet med arbetet, samt dess målsättning. Vidare kommer även de aktuella forskningsfrågorna beröras, och slutligen de avgränsningar som gjorts.

Resterande text i arbetet är strukturerad enligt följande: Metodavsnittet ämnar att lyfta det metodval som gjorts med tillhörande metoddiskussion. Resultatavsnittet visar först en sammanställning av de kvantitativa resultat som framkommit vid sökning av litteratur. Sedan presenteras en sammanställning av funna artiklar tillsammans med dess respektive huvudområde. Eftersom resultaten är omfattande presenteras endast en sammanställning av dessa i resultatdelen. Den detaljerade dokumentationen finns att tillgå i bilagor. Resultaten av litteraturstudien skrivs sedan ut i textform i syfte att beskriva de fynd som gjorts. Därefter följer en analys av resultaten samt en diskussion. Arbetet avslutas med att presentera slutsatser och vidare forskning inom ämnet.

1.2 Tidigare forskning

Vi har, givet ett antal analyserade artiklar [6-11], identifierat fyra huvudområden vilka inkapslar dataflödet i ett informationssystem som i någon mån extraherar, bearbetar, och visualiserar data. Dessa fyra områden ligger som grund till strukturen av denna uppsats och kommer att återkopplas till framöver i mer ingående detalj. Följande underrubriker kommer att diskutera dem övergripande.

1.2.1 Skrapning

För att kunna behandla stora mängder data vilka gjorts tillgängliga online, är det först nödvändigt att dessa sparas ner, eller "skrapas" från webben. Detta åtagande kan utföras med på en mängd olika sätt, till exempel med hjälp av en så kallad *web crawler*, eller *spider* [12].

API:er (eng. *Application Programming Interface*) [13] är en annan metod vilken kan användas i samma syfte. API:er är onlinetjänster som aktörer kan tillhandahålla allmänheten med syftet att bland annat dela med sig av sin offentliga data. Ett exempel på en sådan aktör är microbloggportalen *Twitter* [14], vars API har visat sig vara ett värdefullt verktyg för att extrahera information utifrån naturlig språktext [15]. Den offentliga datan från *Twitter* har visat sig användbar för forskare som exempelvis undersöker ämnet *sentimentanalys* [16].

1.2.2 Behandling

Huvudområdet *behandling* innefattar tekniker vilka bearbetar data på något vis. Denna bearbetning kan ske under samtliga delmoment i ett systems informationsflöde. Exempelvis kan behandlingsalgoritmer användas i kombination med skrapning för att filtrera oanvändbar data med hjälp av tekniker som *web page segmentation* [17], [18] och *noise removal* [19], [20].

Efter det att data skrapats kan den användas i syftet att utvinna information, vilket övergripande refereras till som *informationsinhämtning* (eng. *Information Retrieval*) [20]. För att extrahera information ur rådata måste datan bearbetas på olika sätt. För bearbetning av ostrukturerad text används ofta tekniker inom området *NLP* (eng. *Natural Language Processing*) [21]. Ett exempel på en teknik inom NLP är *POS*-taggning (eng. *Part-of-Speech tagging*), vilken används för att identifiera semantiska element som verb och adjektiv i text [8].

En annan metod vilken drar nytta av NLP, och som med tiden blivit mycket populär, är *sentimentanalys*. Med sentimentanalys kan respons för en given produkt och/eller ämne graderas i syfte att exempelvis ge ett företag insikt om hur de skall kunna förbättra produkter, få nöjdare kunder, förutspå marknaden, et cetera [7].

1.2.3 Lagring

För att kunna bearbeta stora, ostrukturerade dataset på ett effektivt sätt är det först lämpligt att lagra dessa i en databas. Beroende på en rad olika faktorer är valet av databas dock inte alltid givet för ett system, ett fenomen känt som *polyglot persistence* [22, kap.13]. Om mängden data är tillräckligt stor kan den behöva lagras i ett kluster av datorer [22, kap.13], något som traditionella relationsdatabaser inte utvecklats för. Traditionella relationsdatabaser är nödvändigtvis inte heller det bästa alternativet för data som skrapats från sociala medier, då sådan data är ostrukturerad i sin natur [4]. Informationssystemets behov bör analyseras och den databas som bäst tillmötesgår dessa behov bör väljas.

För stora datamängder används i dagsläget ofta databasmodeller vilka faller under den myntade termen *NoSQL*. Dessa varianter innefattar i huvudsak *graph*-, *document*-, *column-family*- och *key-value-databaser* [22], [23].

1.2.4 Visualisering för insiktsgenerering

Det ultimata målet av bearbetningsprocessen är att generera ny kunskap och nya insikter utifrån den information som har utvunnits ur den skrapade datan. Till exempel kan data analyseras på ett explorativt vis i hopp om att finna nya korrelationer eller mönster [24, pp.5]. För att uppnå detta krävs det att information presenteras på ett korrekt vis i rätt kontext, något som kan åstadkommas genom att visualisera datan [4] (eng. *visual analytics*). Med hjälp av visuell analys går det bland annat att se förändringar över tid och hitta relationer mellan delar i informationen. Datavisualisering kan lämpligen utföras med hjälp av verktyg, något som på senare tid blivit allmänt genomförbart tack vare moderna, kosteffektiva datorer [24, pp.7]. Presentationen av information utförs oftast i hopp om att kunna generera nya insikter utifrån den. Individer som undersöker informationen måste förstå kontexten och de eventuella samband som förekommer [11].

2 Problemformulering

Tidigare forskning inom ämnen såsom sentimentanalys (SA) och tekniker för förbehandling existerar redan. Dock finns det få tydliga litterära sammanställningar som beskriver de metoder som kan användas för att skapa ett informationssystem som inte helt och hållet kretsar kring sentimentanalys. Vid de efterforskningar vi gjort kring ämnet kunde det inte heller hittas någon litteratur som tydligt beskriver hur processen att gå från rådata till att generera nya insikter går till. Det kan bero på kontexten i den problemställning som ska lösas, då sentimentanalys bygger på att utröna om information i exempelvis en produktrecension är positiv eller negativ. Därför räcker det att analysera användares bedömning av produkten för sig.

Globalworks siktar på att utröna djupare information än enbart sentiment utifrån den data de samlar in från sociala medier. Företaget har målet att skapa en djupare förståelse och kunskap för vad som debatteras kring på sociala medier gällande förändringar som sker på olika arbetsplatser. Förändringarna kan innebära exempelvis ett ökat antal repressalier eller en försämring av arbetsvillkor hos ett företag. För att skapa förståelse krävs det att innehållet i det som skrivs analyseras på ett djupare plan i syftet att sedan använda den informationen för att dra slutsatser om förändringarna som sker. Företagets nuvarande system är inte färdigutvecklat och saknar automatiserade steg som sträcker sig från data till *relevant* information. Dessa steg sker i dagsläget genom manuell hantering av data. Den mängd information som systemet hittar skapar problem för de experter som ska läsa och sälla ut relevant data. Detta eftersom rådatan inte bara är stor till mängden, utan även innehåller en del brus i form av reklam och annonser. Därför behöver datan skrapas, behandlas och kategoriseras på ett annorlunda sätt, jämfört med hur detta görs i dagsläget, i syfte att reducera mängden irrelevant data. Globalworks informationssystem innehåller i dagsläget funktioner som behöver förbättras och byggas ut för att på ett effektivt sätt analysera de trender och mönster som fångas upp i den skrapade datan, för att sedan korrelera resultaten. Baserat på problemformuleringen ämnar detta arbete att besvara följande forskningsfrågor:

På vilka sätt är det möjligt att behandla och gradera text efter innehåll baserat på de senaste teorierna inom "Finding a Needle in a Haystack"-problemet, i syfte att främja Globalworks arbete att lokalisera brott mot mänskliga rättigheter på arbetsplatser?

Givet huvudfrågans bredd har tre delfrågor formulerats för att denna lättare skall kunna besvaras:

- a. Vilka metoder och tekniker används enligt litteraturen idag för att behandla ostrukturerad textuell data?
- b. På vilka sätt kan man kombinera de metoder som finns i litteraturen gällande insamling och bearbetning av stora mängder data i syfte att underlätta för experter att sälla ut den mest relevanta informationen inom deras område?
- c. Vilka delar av processen kan tänkas automatiseras och vilka kräver experthantering?

2.1 Syfte och mål

Baserat på tidigare forskning och ovanstående problem, konstateras det att "Finding a Needle in a Haystack"-problemet är spretigt eftersom problemet går att angripa med hjälp av flertalet olika metoder och tekniker. Därför syftar detta arbete att, genom en systematisk litteraturstudie, samla in och bygga kunskap om de metoder och tekniker som används i informationssystem i dagsläget. Globalworks har en vision om att försöka automatisera delar av systemet till högsta möjliga grad. Detta med hjälp av automatisk klassificering av innehåll, automatisk identifiering av diskussionsämnen samt automatisk analys av förändringar över tid.

Med en litterär sammanställning kommer vi att bistå Globalworks med en ontologisk överblick i hur ett system för Information Retrieval (IR) är strukturerat, med vilka metoder och teknologier ett sådant system kan utvecklas, samt hur dessa kan kombineras. Målet med arbetet är en kartläggning av relevanta metoder och teknologier som Globalworks kan använda sig av för framtida utveckling av deras system. Önskemålet från företaget är att studien ska vara förutsättningslös och oberoende av hur företaget arbetar i dagsläget, vilket även kan gynna andra forskare i framtiden då studien kartlägger befintliga metoder och tekniker inom fältet.

2.2 Avgränsningar

Globalworks informationssystem innefattar insamling, behandling, lagring och visualisering av stora textuella datamängder. I detta arbete kommer endast de olika områdena diskuteras övergripande vad gäller datainsamling, där det nämns i korthet hur datainsamling går till. De metoder och tekniker som analyseras i arbetet kommer att vara avgränsade till att endast behandla ostrukturerad textdata då det är den typen av data som finns tillgänglig på sociala medier och chattforum [6]. Detta görs i syfte att arbeta fram en översikt för hur de metoder och teknologier som finns kan implementeras i framtiden för att effektivisera Globalworks process i att sälla och gradera textuell data baserat på textinnehåll. Med detta görs också en avgränsning i att istället för att ta fram en fullständig prototyp, kommer det här arbetet endast att bistå Globalworks med en litteratursammanfattning och ontologisk överblick inom området för Information Retrieval/ Information Extraction. Tillämpning och därmed kontexten av detta arbete kommer därför att avgränsas till Globalworks verksamhet inom riskbedömning av arbetsvillkor i högriskländer. I sammanfattningen av fynd i litteraturen görs en avgränsning vid Machine Learning (ML) då ML är ett stort område i sig. Därför kommer vi inte djupdyka inom ämnet utan endast beskriva de funktioner som är nödvändiga för arbetet och hur dessa kan nyttjas av Globalworks.

3 Metod och metoddiskussion

I följande kapitel beskrivs den metod som använts vid insamling av den litteratur arbetet innefattar. Kapitlet behandlar även den datainsamling- och sällningsprocess som tillämpas, samt en diskussion av alternativa metoder vilka hade passat arbetet men valts bort.

3.1 Metodbeskrivning

Forskningsfrågorna adresseras genom en systematisk litteraturstudie där vetenskapliga artiklar är den enda källan till data. Syftet med en systematisk litteraturstudie är att på ett väldokumenterat och strukturerat sätt identifiera, utvärdera och tolka den litteratur som är relevant för de forskningsfrågor som ska besvaras. Vetenskapliga artiklarna granskas i syftet att kartlägga den forskning som finns kring nuvarande metoder och tillvägagångssätt. Flertalet artiklar analyserades och delades in i deras tillhörande områden. Detta resulterade i formationen av den struktur som används i litteraturstudien för att förklara hur ett informationssystem är uppbyggt. Eftersom forskningsområdet för arbetet varken är nytt eller snävt är det av största vikt att även granska det snarlika området för sentimentanalys för att avgöra om den forskning som framkommit där även kan användas i Globalworks informationssystem.

Den systematiska litteraturstudie som utförts i denna uppsats är strukturerad efter det förslag som Kitchenham presenterar i "Procedures for Performing Systematic Reviews" [25]. Processen hos en systematisk litteraturstudie består av tre huvudsteg där det första utgår på att planera studien. Målet med planeringen är att identifiera behovet av litteraturstudien samt att planera hur sökningen skall utföras för att skapa ett protokoll för dokumentering. Steg två innefattar att välja det område arbetet ska behandla och hitta relevanta nyckelord att söka på, söka efter artiklar, kontrollera artiklarnas kvalitet och välja ut de artiklar som är mest relevanta för forskningsfrågorna för att sedan dokumenteras i enlighet med steg tre. För att välja ut de mest relevanta artiklarna för forskningsfrågorna sållas artiklarna i enlighet med [26], där vi även lagt till två extra steg för att säkerställa artiklarnas relevans för ämnet.

Resultatet av den systematiska litteraturstudien framställs på så sätt att den kvantitativa data, vilken är de resultat sökningarna genererar, analyseras på ett kvalitativt sätt. Detta resulterar i en kartläggning av de områden artiklarna behandlar och ger litteraturstudien den strukturen som är nödvändig för att beskriva ett informationssystem.

De artiklar litteraturstudien består av beskriver ingående de olika algoritmer och tekniker som arbetet behandlar, medan litteraturen i inledningen ligger på en mer övergripande nivå. Genom att använda ett systematiskt tillvägagångssätt kommer störst fokus att läggas på litteraturen som ingår i litteraturstudien då resultatet baseras på den.

Det finns både för och nackdelar med att göra en systematisk litteraturstudie för att framställa den här typen av arbete. Att utföra en systematisk litteraturstudie är omfattande och tidskrävande, vilket leder till att majoriteten av tiden för arbetet går åt till att läsa och analysera artiklar. Därmed minimeras tiden

för skrivandeprocessen. Dock levererar den sållningsprocess som används en säkerhet i att de dokument som används i litteraturstudien är relevanta för forskningsfrågorna. Därför är den här metoden den mest relevanta i relation till utformningen av forskningsfrågorna och levererar med säkerhet trovärdig och relevant litteratur.

Forskningsfrågans delfrågor *a*, *b* och *c* ses som delar i processen att besvara huvudfrågan. Fråga *a* kommer att besvaras med hjälp av litteraturstudien. Fråga *b* kommer att besvaras delvis av litteraturstudien och delvis av den kartläggning av metoder som kommer att göras baserat på litteraturstudien. När kartläggningen är gjord kommer en summering visa var i systemet dessa metoder kan implementeras och hur de kan kombineras med varandra. Resultatet av fråga *b* kommer sedan att ge en tydlig bild gällande om hypotesen i fråga *c* kan bekräftas eller förkastas.

3.2 Datainsamling

3.2.1 Databaser

Då forskningsfrågorna riktar sig till att granska metoder och tekniker som finns i litteraturen, och på vilka sätt dessa kan kombineras, har vi valt att uteslutande använda oss av akademisk litteratur. Litteraturen består främst av journal- och konferensartiklar publicerade i vetenskapliga tidsskrifter. Journal- och konferensartiklar har hög trovärdighet och studier som riktar sig till informationssystem publiceras ofta i formen av journal- och konferensartiklar.

Vi har i enlighet med kriterierna för arbetet använt oss primärt av ACM Digital Library (Association for Computing Machinery), IEEE Transactions och Google Scholar, där ACM Digital Library har valts som primär databas. Detta eftersom den är lätt att söka i och innehåller journal- och konferensartiklar som ligger på den detaljnivå som behövs för att utföra litteraturstudien. Det finns dock nackdelar med att använda ACM Digital Library då sökalgoritmen ibland ändrar årtal vid sökning. Genom att läsa på och förstå hur sökningen fungerar kunde problemet lösas då det framkom att databasen visar resultat från den tidigast publicerade artikeln. Problemet åtgärdades genom att vid varje sökning säkerställa att det skett en sökning från det år som ingår i de uppställda inkluderings- och exkluderingskriterierna.

Även om ACM Digital Library är en stor databas var det svårt att hitta artiklar som behandlar ämnen om hur ett system bör applicera visualiseringstekniker för att visa datan systemet hittat. Därför gjordes valet att gå vidare till IEEE Transactions som även den är en lätt databas att använda och innehåller artiklar inom datavetenskapliga ämnen. IEEE Transactions användes som komplettering till det material som hittats i ACM Digital Library och de inkluderings- och exkluderingskriterier som användes där applicerades även för sökningar i IEEE Transactions. Sökningarna genererade många dubletter men även nya artiklar hittades gällande att användare lättare ska kunna förstå informationen som genereras i ett IR-system. Andra kompletterande artiklar har även hittats via Google Scholar, där sökningen bestod av söksträngar innehållande

den information som önskas. Artiklar från Google Scholar valdes ut baserat i huvudsak på rubrikens relevans men även på hur många gånger artikeln citerats.

3.2.2 Sökning och sällning av litteratur

Artificiell Intelligens (AI) och IR har varit ett populärt ämne i över 30 år och det finns mycket litteratur som sträcker sig långt bak i tiden. Dock gjordes valet att endast använda artiklar som är publicerade från och med år 2008 för att säkerställa att den information som samlas in är state-of-the-art. Genom att sälla bland information från de senaste 10 åren upplevdes detta kunna uppnås.

Som första steg i processen för datainsamling identifierades en rad nyckelord med hjälp av [3] vilka ansågs vara de mest relevanta för sökningen av användbara artiklar. Detta eftersom författarnas informationssystem innehåller delar liknande de som ingår i Globalworks informationssystem. Efter identifieringen av nyckelord sammanställdes även synonymer till dessa.

Till en början gjordes sökningarna endast på ett sökord, eller en synonym till detta. Dock resulterade detta i ett för stort antal träffar. Därför erfordrades ytterligare filtrering genom att kombinera söktermer med varandra, vilket resulterade i att sökresultaten blev färre. Sällningen fortsatte sedan genom att leta relevanta rubriker för ämnet. Sökorden kombinerades med hjälp av de booleska uttrycken AND och OR, där AND visade sig vara mest effektivt. OR användes endast vid de tillfällen där AND gav för få antal träffar inom ämnet.

Genom att följa "The Three-Pass Approach" [26] tillsammans med en uppsättning regler sällades artiklarna till en början ut och de mest användbara artiklarna sparades ner. Reglerna innefattar att titta på relevanta rubriker, relevanta abstrakt, relevanta fulltexter samt att eliminera dubletter. Ytterligare två steg för sällning applicerades baserat på 1). Innehåll av tekniker som ingår i de system artiklarna beskriver och 2). Relevans för Globalworks. Anledningen till att avvika från metoden [26] och läsa artiklarna på nytt, ur olika perspektiv, var för att säkerställa att de innehåller information relevant till både litteraturstudien och Globalworks.

Sökningarna gjordes med hjälp av följande sökord och kombinationer av dessa: *unstructured text data* eller *pre-processing + metadata analysis*, *text analysis*, *text extraction*, *segmentation*, *morphology analysis*, *semantic analysis*, *similarity analysis*, och *sentiment analysis*. Sökresultat, regler och artiklar dokumenterades enligt den matris som skapades inför sökningarna (se tabell 1 under 3.1 Litteraturstudie). Vid kompletterande sökningar tillkom sökord om *insights*, *outliers*, *knowledge extraction*, *natural language processing* och *natural language understanding*.

Vid sökningar där inga relevanta artiklar hittades användes den mest relevanta artikeln till att söka mer litteratur inom området baserat på artikelns referenser. Tekniken kallas "Backwards Snowballing" [27] och med hjälp av denna hittades ytterligare 3 relevanta artiklar att använda i arbetet.

3.3 Metoddiskussion

Diskussionen ämnar att ta upp andra metoder som kan vara användbara för att besvara forskningsfrågorna men som valts bort av olika anledningar.

3.3.1 Experiment

Ett experiment går ut på att undersöka relationen mellan orsak och verkan, där målet är att bevisa eller motbevisa den kausala länken mellan beroende och oberoende variabler [28]. Att utföra experiment för att bekräfta eller förkasta den hypotes som ställs i delfråga c är ett alternativt tillvägagångssätt som kan ge ett ingående svar på om det går att automatisera delar i Globalworks informationssystem eller inte.

Genom att göra antagandet att det går att automatisera vissa delar av Globalworks informationssystem går det att framställa en hypotes som exempelvis säger att *“Genom att byta ut inputen från expert X mot regelverket Y kommer resultatet förbli detsamma”*. Inputen från expert X i detta fall kan vara de nyckelord som matas in manuellt i systemet innan sökning. Regelverket Y i detta fall är det regelverk som programmerats till ett neuralt nätverk, där regelverket söker i en text efter de inprogrammerade nyckelorden. Enligt hypotesen innebär detta att experten X och inputen Y är de oberoende variablerna. Resultaten är den beroende variabeln eftersom målet då är att se om resultaten förändras, hur de förändras och varför [28]. Resultaten framställs sedan genom observationer och mätningar som visar eventuella skillnader mellan tester på originalsystemet och tester på den modifierade delen av systemet.

Fördelarna med att använda experiment som metod i detta fall är att metoden är accepterad och väletablerad, samt att experiment anses vara den mest vetenskapliga metoden för den här typen av arbete. För det här arbetets del innebär det att det skulle behöva utföras experiment på varenda metod eller teknik som föreslås i resultaten i syfte att komma fram till den som fungerar bäst för att kunna ersätta en expert i ett tidigt skede. Därför kommer det vara svårt att hålla alla beroende och oberoende variabler under konstant kontroll. Experimenten kan heller inte utföras utan att göra noggranna efterforskningar som beskriver de metoder och tekniker som finns tillgängliga att modifiera. Därför är experiment som metod istället att föredra vid framtida arbete, snarare än en del av det här arbetet. Andra nackdelar med att utföra experiment är att det urval av testfall som kan testas är begränsat och eftersom resultaten endast kommer att vara giltiga för dessa fall är slutresultatet inte generaliserbart. Det är inte heller lätt att hitta rätt urval av testfall.

3.3.2 Fallstudie

Att utföra en fallstudie är ett alternativt tillvägagångssätt som hade passat arbetet då en fallstudie utgår ifrån ett specifikt fall och studerar detta på djupet. Målet med en fallstudie är att få en detaljerad insikt i hur det specifika fallet fungerar i dess naturliga miljö för att förstå komplexa relationer och processer [28]. Detta kan uppnås genom att kombinera en fallstudie med andra tillvägagångssätt för att generera data, såsom intervjuer, observationer eller undersökningar. Att utföra en

fallstudie med dokument och intervjuer som underlag hade kunna hjälpa till att sälla ut de specifika metoder och tekniker som bör granskas i studien. Detta för att på så sätt inkludera de experter som använder Globalworks system i syfte att utröna vad experterna anser fungerar bra, samt mindre bra gällande de metoder och tekniker som används i systemet.

Forskningsfrågorna skulle sedan besvaras genom att göra en jämförande studie baserat på metoder som finns i andra informationssystem hämtade ur litteraturen. Dock skulle en jämförande studie behöva utföras för varje enskilt fall för att sedan göra kopplingar mellan resultaten och metoderna. Detta skulle vara tidskrävande att utföra och resultaten skulle bli svåra att generalisera [28]. Sett till utformningen av forskningsfrågans delfrågor, är inte en fallstudie ett optimalt tillvägagångssätt för att utreda vilka metoder som används idag. Globalworks experter har dessutom endast kunskap om ett fåtal av de metoder som beskrivs i litteraturen. Därför är det viktigt att göra en kartläggning av metoder och tekniker som finns innan en jämförande studie kan utföras.

3.3.3 Design and Creation

Design and Creation-processen består av 5 steg: Awareness, Suggestion, Development, Evaluation och Conclusion [28] och kan implementeras på olika sätt beroende på vilket resultat som önskas. Vid användning av Design and Creation hamnar litteraturstudien under steget *Awareness*, eftersom steget innebär att forskaren skapar sig en förståelse för problemet, vilket är det vi vill uppnå med detta arbete. Summeringen hamnar under steget *Suggestion* eftersom det lämnas ett förslag på metoder och lämpliga områden att implementera dessa i, baserat på den litteratur som finns. Trots att arbetet endast uppfyller två av stegen i Design and Creation-processen platsar arbetet in i Globalworks utvecklingskedja, då deras mål är att utföra de tre sista stegen i processen; *Development*, *Evaluation* och *Conclusion*. Det här arbetet skulle därför kunna ses som en del i den Design and Creation-process Globalworks kommer att använda sig av för att implementera de förändringar som föreslås i detta arbete. Dock anser vi att den här metoden är överflödigt för ändamålet att besvara forskningsfrågorna och har därför valt att inte använda oss av Design and Creation som huvudsaklig metod.

4 Resultat

I detta kapitel presenteras det slutgiltiga resultatet av den litteraturstudie vilken arbetet berör. Initialt presenteras de kvantitativa resultaten av sökprocessen med relaterade tabeller, samt förklaringar för deras struktur. Slutligen framförs en summering av resultatet och inblickar i de av litteraturstudien funna artiklarna.

4.1 Litteraturstudie

Litteraturstudien ämnar att ge en överblick över forskningen inom ämnen som kan vara relevanta för Globalworks, och visa vilka metoder och tekniker som används idag för att behandla ostrukturerad textuell data, samt hur dessa kombineras med varandra.

Litteratursökningen visar sammanställd information utifrån sökningar via både ACM Digital Library och IEEE Transactions, se tabell 1,2 och 3. För att ge en inblick i hur de slutgiltiga artiklarna sållats ut, presenteras de regler som använts i sållningsprocessen (se avsnitt 2.2.2) samt mängden kvarstående artiklar efter de olika reglerna tillämpats. Sökningen från Google Scholar ingår inte i den totala sammanställningen då databasen endast användes för kompletterande sökningar, och inte dokumenterades på samma sätt som sökningarna i ACM Digital Library och IEEE Transactions. En tabell över de artiklar som hittades via Google Scholar redovisas i bilaga D.

För en mer detaljerad insyn i hur artiklarna har dokumenterats finns sammanställningarna bifogade i bilagorna A-D. Att presentera de detaljerade resultaten i bilagor fyller syftet att den som är extra intresserad skall kunna använda bilagorna vid sidan av läsningen. Detta för att läsaren själv ska kunna dra paralleller mellan den detaljerade dokumentationen och informationen som finns att tillgå i litteraturstudien.

4.1.1 Litteratursökning

Tabell 1: Total sammanställning av artikelsökning till systematisk litteraturstudie

Inkluderings/ Exkluderingskriterier	Kvarvarande artiklar
Sökträffar i ACM och IEEE baserat på publiceringsform och publiceringsår	3766
Lästa rubriker	1088
Relevanta rubriker	151
Innehåll abstrakt	96
Innehåll fulltexter	47
Eliminering av dubletter	43

Tabell 2: Visar kombination av sökord samt sökresultat från ACM.

Sökord för ACM	Databas	Sökt i	Resultat	Valda artiklar
"pre-processing" OR "metadata analysis"	ACM	Fulltext	579	2
"pre-processing" AND "text analysis"	ACM	Fulltext	3	0
"text extraction"	ACM	Fulltext	74	5
"pre-processing" AND "segmentation"	ACM	Fulltext	57	2
"morphology analysis"	ACM	Fulltext	1	0
"pre-processing" OR "semantic analysis"	ACM	Fulltext	470	5
"similarity analysis"	ACM	Fulltext	69	0
"pre-processing" AND "sentiment analysis"	ACM	Fulltext	8	1
"unstructured text data" OR "metadata analysis"	ACM	Fulltext	105	3
unstructured text data AND text analysis	ACM	Fulltext	6	0
unstructured text data AND segmentation	ACM	Fulltext	45	1
unstructured text data AND semantic analysis	ACM	Fulltext	491	5
"sentiment analysis"	ACM	Fulltext	1174	5
"text" AND "outliers"	ACM	Fulltext	40	1
TOTALT			3122	30

Tabell 3: Visar kombination av sökord samt sökresultat från IEEE i detalj

Sökord för IEEE	Databas	Sökt i	Resultat	Valda artiklar
((pre-processing) AND text analysis)	IEEE	Fulltext	278	8
((knowledge extraction) AND information retrieval) AND smart systems)	IEEE	Metadata	38	1
(insight-based evaluation)	IEEE	Metadata	16	1
((knowledge indexing) AND text based knowledge)	IEEE	Metadata	258	2
((Recurrent network) AND temporal knowledge)	IEEE	Metadata	54	1
TOTALT			644	13

4.2 Sammanfattning av fynd i litteraturen

Kapitlet ämnar sammanställa de fynd som hittats i litteraturen och att ge en djupare inblick i hur olika metoder och tekniker används inom de identifierade områdena. I tabell 4 visas de huvudområden som identifierats tack vare litteraturen, vilka delområden dessa kan delas in i samt vilka artiklar som tillhör respektive område. Artiklarna i tabell 4 refereras till med hjälp av det ID-nummer som tilldelats respektive artikel vid mappning (se bilaga A-D). De artiklar som har lyfts fram i litteraturstudien visas även med tillhörande referensnummer enligt strukturen **artikelID:[referensnr]**. Detta för att läsaren skall kunna dra paralleller mellan resultattabellerna, de detaljerade beskrivningarna i litteraturstudien och referenslistan.

Tabell 4: Sammanställning av vilka artiklar i tabellerna 5,6, 7 och 8 som tillhör vilket område.

Huvudområde	Tekniker	Artiklar artikelID:[referensnr]
Skrapning	Fokuserad crawling, gömd crawling, inkrementell crawling, distribuerad crawling	2:[17], 12:[18], 13:[31], 22, 26:[69], 29:[64], 30:[51], 32:[4], 39:[57], 42, 44:[40], 45, 48:[12], 51:[46], 60:[30]
Behandling	Parsing, tokenization, segmentering, lemmatisering, stemming, POS-tagging, named entities, feature selection, TF-IDF, maskininlärning	1, 2:[17], 3, 4, 5:[49], 7:[55], 8, 9:[54], 10:[8], 11:[52], 12:[18], 13:[31], 14:[19], 15, 16, 17:[59], 18:[60], 19, 20:[50], 21, 22, 23, 24:[47], 25:[6], 26:[69], 27:[62], 28:[66], 29:[64], 30:[51], 31, 32:[4], 33, 34:[65], 35:[45], 36:[56], 37, 38:[43], 39:[57], 40, 41, 42, 44:[40], 45, 47:[2], 49:[58], 50:[10], 51:[46], 52:[41], 53:[61], 54:[11], 55:[7], 57:[53], 58:[63], 60:[30], 61:[48]
Lagring	Relationsdatabas, icke-relationell databas	6:[9], 32:[4], 35:[45], 47:[2], 49:[58], 54:[11]
Visualisering & Insiktsgenerering	Trender, mönster, korrelationer, system	6:[9], 16, 18:[60], 20:[50], 22, 26:[69], 29:[64], 32:[4], 43:[68], 45, 46:[67], 47:[2], 49:[58], 51:[46], 56, 59

4.2.1 Skrapning av data

För att data ska kunna sparas ned från webben krävs det i första hand att identifiera den data som är relevant för ändamålet. Detta kan göras med hjälp av att skrapa/crawla/indexera hemsidor. Att tillämpa dessa tekniker innebär att på ett systematiskt sätt arbeta sig igenom innehållet i en webbsida för att utröna vilken typ av information sidan innehåller. Detta för att sedan ladda ner innehållet som är relevant sett till ändamålet. Att crawla en webbsida tillhör den automatiserade delen av IR och det finns olika strategier för att uppnå målet: Fokuserad crawling, gömd crawling, inkrementell crawling och distribuerad crawling [12].

Saini och Arora nämner i en kartläggning av webcrawling att *Fokuserad crawling* innebär att fokusera sin crawler genom att ge den specifika riktlinjer att följa. När dessa riktlinjer är uppfyllda så skrapas data ner. Detta tillvägagångssättet går att kombinera tillsammans med andra tillvägagångssätt som är baserade på nyckelord, exempeldokument, ontologibaserat eller data mining-baserat. Dessa sätten är beroende av inputen av den information som önskas för att kunna skrapas ner [12].

Gömd crawling (eng. *Hidden crawling*) är ett sätt att söka igenom den dolda webben (eng. *deep web eller hidden web*) med hjälp av ett sökformulär istället för att använda hyperlänkar. Även detta tillvägagångssättet kan användas som grund till andra grenar inom gömd crawling såsom trädbaserade tillvägagångssätt, domänspecifika tillvägagångssätt och säkerhetsbaserade tillvägagångssätt [12].

Inkrementell crawling utförs genom att arbeta sig igenom URLer (eng. *Uniform Resource Locators*) inkrementellt i syfte att återbesöka sidor och prioritera URLer utefter det. Det finns grenar inom inkrementell crawling som kan användas, som exempelvis tillvägagångssätt för data mining och ett sätt som innebär att återkomma till en sida för att uppdatera den inom specifika tider för att se till att den data som hämtas alltid är uppdaterad och färsk [12]. Att använda inkrementell crawling kan dock vara problematiskt vid crawling av exempelvis forum då det finns censureringar som inte accepterar vissa inlägg, samt administratörer som tar bort inlägg som inte är godkända enligt forumets regler.

Distribuerad crawling innebär att en samling datorer söker igenom hyperlänkar med hjälp av sökmotorer i syfte att indexera data. Distribuerad crawling kan användas som bas till andra grenar inom kategorin för distribuerade tillvägagångssätt, som exempelvis map reduced-baserade, model-baserade, data mining-baserade samt peer-to-peer (P2P) tillsammans med hashtabeller [12].

Eftersom sociala medier blir mer komplexa med tiden på så sätt att användare ger konstant input i form av exempelvis tweets, inlägg på Facebook, stora blogginlägg, med mera, behöver crawling av dessa ständigt optimeras. Med det ökade antalet användare och företag som vill marknadsföra sig, ökar även riktad reklam och annonser. Reklam och annan irrelevant information i rådatan ses som brus (eng. *noise*) och rådatan behöver därför helst rensas innan den laddas ner. Detta kan göras med hjälp av att dela upp innehållet på en webbsida i olika segment för att sedan ladda ned det segment av text, eller bild som önskas. En webbsidas segment presenteras exempelvis i ett DOM-träd (W3C-skapad

datastruktur utifrån HTML eller XML [29]), där HTML-taggar och XML-taggar delas in i dess respektive område [30].

Med tiden har det utvecklats många algoritmer för att segmentera hemsidor. Pappas *m. fl.* [17] presenterar en algoritm, kallad *SD-algorithm (Style-Density Tree algorithm)*, vilken analyserar både visuella och icke-visuella delar med hjälp av ett DOM-träd. Algoritmen använder sig av korpus, vilka kan byggas både på manuellt och automatiskt vis. Pasternack och Roth [18] presenterar ett kombinerat tillvägagångssätt, *Maximum Subsequence Segmentation*, vilket försöker lösa problem med existerande segmenteringstekniker. Många av dessa tekniker kräver expertkunskap, arbetar på en specifik typ av layout template, och/eller kräver mycket processorkapacitet. Den presenterade algoritmen är semi-supervised och är således inte helt automatiserad. Joshi och Liu [31] presenterar ett automatiserat tillvägagångssätt som grundar sig runt DOM analys och NLP tekniker, med resonemanget att när segmentering utförs bör algoritmen kunna resonera kring HTML-innehållet likt hur en människa skulle resonera. Algoritmen kan segmentera både text- och bilddata.

API:er är en annan princip vilken kan användas i syfte att skrapa data. Ifall en portal vilken används för skrapning tillhandahåller ett sådant kan data enkelt skrapas. Det är dessutom möjligt att göra detta med automatiska scripts. Genom att kommunicera med API:er kan data presenteras via responsmeddelanden i ett strukturerat format. Det kan dock finnas restriktioner för användningen av API:er. För att API-ägarens servrarna skall hålla sig stabila brukar det finnas begränsningar för hur många anrop som får göras per sekund [14]. Detta kan visa sig vara problematiskt i situationer där data ämnas att skrapas så snabbt som möjligt på ett automatiskt vis.

Twitter's API för crawling har visat sig vara populärt inom forskningsområden som använder NLP-tekniker då datan som finns i dessa sammanhang är ostrukturerad och förekommer i stora kvantiteter [15], [16], [32-40].

4.2.2 Behandling av data

Att behandla data i syfte att utröna nödvändig information är ett måste om det ska gå att hitta en nål i en höstack. Globalworks fokuserar på ostrukturerad text i dokument. Ostrukturerad text är svår att extrahera och kategorisera, därför behöver datan behandlas med hjälp av flertalet NLP-tekniker [7], [41], [42]. Studier visar att förbehandling av data är nödvändigt för att utvinna den information som önskas utifrån ostrukturerad textdata [41], [43], [44]. Mhatre *m. fl.* påvisar i sin studie att en kombination av hantering av slang, lemmatisering och eliminering av stopword gav bäst resultat [44]. Dock finns även studier som visar på att det vid behandling av ostrukturerad textdata, i syfte att utföra *sentimentanalys*, inte kräver behandlingstekniker i lika stor omfattning och att det därför går att bortse från vissa steg i behandlingsprocessen [45]. Experiment utförda av Jiangqiang and Xiaolin visar även att förbehandling som innefattar att ta bort stopword, siffror och URL:er inte ger någon förbättrad prestanda av klassificeringen, men bidrar genom att minska brus i datan väsentligt [39].

De finns flertalet verktyg tillgängliga för att behandla data. Exempel på dessa är: NLTK, OpenNLP, CoreNLP, FudanNLP, Gensim, LTP och NiuParser [7]. De två sistnämnda verktygen behandlar det kinesiska språket.

Med hjälp av de tekniker som beskrivs i det här avsnittet, behandlas data i flera olika steg i syfte att reducera mängden brus för att sedan utvinna den mest relevanta informationen.

Tokenization och segmentering

Tokenization är en del i processen att rensa textdata och innebär att bryta ner stora texter till paragrafer, fraser, ord eller tecken i ord. Avgränsning för var meningar ska delas upp kan variera men ofta avgränsas meningar vid exempelvis blanksteg eller indrag i texten. Andra avgränsande markörer är punkter och kommatecken men även vissa stoppord kan anses vara markörer för avgränsning. När texter delas in i paragrafer eller meningar kallas det ofta för talindelning eller segmentering [41], [48]. Tokenization kan lätt förväxlas med segmentering. Att segmentera text innebär även det att dela upp textstycken i mindre delar. Skillnaden mellan dessa är att tokenization är fokuserat på själva uppdelningsprocessen som bryter ner en mening till ord, eller ord till tecken med hjälp av blanksteg i texten, medan segmentering fokuserar på paragrafer och meningar. Tokenization kan dock inte användas för att behandla meningar skrivna på kinesiska. Det kinesiska språket skrivs med symboler och det finns inget mellanrum mellan ord i en mening. Därför behöver texten istället segmenteras parallellt med användning av ett lexikon för att förstå vad symbolerna betyder. För att underlätta detta finns, som redan nämnt, verktyg att ta hjälp av [7].

Samonte *m. fl.* [49] delar in sin textdata i tokens i syfte att extrahera känslor utifrån datan (eng. *opinion mining*). Genom att först rensa datan och dela in den i tokens blir det lättare att kunna hitta de nyckelord som symboliserar positiva eller negativa känslor i texten. Dock har de inte tagit bort avgränsande markörer såsom utropstecken eller frågetecken då dessa tecken kan vara till hjälp att analysera den känsla användaren vill förmedla i texten [49]. Vid hämtning av data går det även att använda tokenization för att bryta ner de HTML-taggar webbsidan innehåller till listor av taggar och ord. Att bryta ner text i tokens tjänar sitt syfte i att endast skicka den mest nödvändiga inputen till de efterföljande stegen i behandlingsprocessen. [18], [50] utförde tokenization i detta syfte för att sedan transformera datan med hjälp av stemming. Dock nämner [50] att det var svårt att utföra tokenization på grund av den ostrukturerade datan som finns i bloggar och artiklar. Att utföra tokenization på stora mängder ostrukturerad text kan enligt [50] ge en fingervisning på vilka ord som är mest förekommande i ett textstycke. De tokens som återstår efter tokenization och segmentering fungerar som input till algoritmer som eliminerar stoppord och utför stemming och lemmatisering.

Eliminering av stoppord

Stoppord är en form av brus och utgörs av ord som inte ger någon lexikal vikt i en mening, exempelvis prepositioner och pronomen. Dessa ord anses inte vara nyckelord och tas därför bort från den skrapade datan, i syfte att rensa och minska mängden data som skrapats, och därmed göra datan mer lätthanterlig [10], [43]. Lätthanterlig innebär att det blir lättare för algoritmer att processa datan. Om det

finns experter som analyserar data i systemet kan det vara fördelaktigt att spara en kopia av dokumenten för framtida bruk. Detta då den processade datan kan vara svårläslig för de experter som gå tillbaka och titta i texten.

Genom att eliminera stoppord ökar även precisionen för de algoritmer som används vid exempelvis klassificering av data och träning av ML-algoritmer [45], [47].

Det traditionella tillvägagångssättet för att eliminera stoppord görs med hjälp av en ordlista, även kallat korpus, där de ord som ska tas bort är fördefinierade. Texten som skrapats jämförs med ordlistan och därefter tas alla irrelevanta ord bort. Det finns fördefinierade korpusar att använda som innehåller de vanligaste stopporden. Dock kan det, beroende på hur känslig domänen är, även behövas en domänspecifik korpus som komplement för att säkerställa att rätt stoppord tas bort [51]. Vijayarani och Ilamathi [10] nämner andra metoder för att eliminera stoppord. Bland annat en metod vid namn "Zipf's law", där ord kan tas bort med hjälp av Term Frequency (TF). Denna metod mäter förekomsten av ord där de ord som har (baserat på satta tröskelvärden) hög TF respektive låg TF plockas bort.

Det finns verktyg att implementera och använda sig av i detta syfte. Artikel [43] och [44] använder verktyget NLTK (eng. *Natural Language Tool Kit*), som är en samling av open source- verktyg, där det bland annat finns fördefinierade korpusar på flertalet olika språk att använda sig av vid eliminering av stoppord.

Lemmatisering och stemming

Lemmatisering och stemming är den del i behandlingsprocessen som innefattar att rätta till ordböjningar till dess originalform. Stemning och lemmatisering är relativt lika i sitt sätt att behandla ord. Dock ligger skillnaden i att lemmatisering lägger mer fokus på den morfologi som ligger bakom ordet och att ersätta ordet med dess rätta synonym [10], [44]. Ordet "behandling" byts då ut till dess originalform, det vill säga att "behandla" är ordets lemma. Stemning fokuserar på att hitta ordets rotform i syfte att reducera dimensionen på data. Roten av exempelvis "skriva", "skrev" eller "skrivning" blir då ordet "skriv".

Majoriteten av de funna artiklarna som innefattar behandling av data använder stemming för att lokalisera ordens rotform. Vijayarani och Ilamathi nämner i sin undersökning att det finns olika syften med att använda stemming; Trunkering, statistisk stemming och mixade tillvägagångssätt. För trunkering används algoritmerna Porter's stemmer, Lovin's stemmer, Paice/Husk's stemmer och Dawson's stemmer, där Porter's stemmer är den mest populära att använda [10], [44]. Nackdelen med Porter's stemmer är att den är kontextberoende, vilket kan leda till felstemning [41]. Vid statistisk stemming används algoritmerna *N-gram*, *HMM* och *YASS* och dessa är språkoberoende. De mixade stemmers som behandlas i [10] använder sig både av modalitet och morfologisk indelning av ord och kräver att det finns en stor korpus att tillhandahålla för att dessa typerna ska fungera.

[6], [47] använde stemming och lemmatisering för att sedan med hjälp av en egen algoritm expandera roten av ordet och hitta ordens synonym. På så sätt kunde till exempel [6] och [47] identifiera positiva och negativa sentiment, tillhörande ordets rotform.

Part-of-Speech tagging (POS-tagging)

POS-tagging innebär att markera ut de ord i en mening som motsvarar ett verb, adjektiv, substantiv och så vidare, i syfte att förstå vilken roll ordet har i sin kontext. Tag exemplet: “*The sailor dogs the hatch*”. Ordet “dogs” kan ha olika innebörd och i det här fallet betyder “dogs” inte “hundarna” och är således inte ett substantiv, utan ett verb. Genom POS-tagging kan man berätta i vilken situation ordet används, och på så vis förstå vilket ord det är. Genom att förstå ordets position och samtidigt förstå ordets kontext i en mening går det att effektivisera algoritmerna för klassificering eftersom det i POS-taggen går att utmärka de nyckelord som ligger i fokus för texten [47].

Dey och Haque [47] utförde en studie i syfte att påvisa hur olika NLP-tekniker fungerar på ostrukturerad textdata vid behandling av blandade språk (eng. *cross lingual*), såsom Engelska-Hindi. För att utföra experimentet i studien används Stanford Parser ¹, där POS-tagging ingår som ett steg i att dela orden i meningar. Resultaten för POS-tagging visade att identifikation av engelska ord gav bäst precision [47 pp. 109]. I studien framgick det även att Stanford Parser kan identifiera engelska verb och adjektiv trots att orden är felstavade. Identifikationen av substantiv, i meningar med blandade språk, visade sig dock vara opålitlig då majoriteten av de ord som inte var på engelska identifierades som substantiv.

POS-tagging används ofta vid sentimentanalys för att analysera känslor i meningar skrivna av internetanvändare. Enligt [7] och [8] är verb, adverb, adjektiv och substantiv de ord som uttrycker känslor bäst. Därför används POS-tagging i syfte att hitta dessa ord, för att sedan kunna avgöra polariteten i meningen [8].

Andra fall där POS-tagging används är i system som tar emot frågor på ett naturligt språk. Systemet tar då emot frågan och bryter ner den för att identifiera ordens olika POS-positioner. Dessa taggar används sedan av en så kallad “query generator” för att skapa de frågor som skickas vidare till systemets sökmotor [11].

Named Entities

Named Entity Recognition (NER) är även känt som *entity identification*, *entity chunking* och *entity extraction*. Detta är ett delmoment vid behandling av information, där NER innebär att lokalisera samt klassificera namngivna entiteter i ostrukturerad textdata till fördefinierade kategorier. Dessa kategorier kan te sig i form av exempelvis personnamn, organisatoriska namn, platser, kvantiteter, monetära värden och procentenheter. Meningen “Berit köpte mjukvaran *social@risk*TM från Globalworks AB i december 2018.” skulle då efter att NER applicerats se ut så här:

[Berit]_{Person} beställde mjukvaran *social@risk*TM från [Globalworks AB]_{Organisation} i [december]_{Tid} [2018]_{Tid} .

NER används i många syften, exempelvis vid sentimentanalys [7], besvara frågor om tider, platser, personer och företag [11], segmentering av text [52], samt vid summering av text [53]. Vid sentimentanalys kan NER användas i syfte att hitta

¹ <https://nlp.stanford.edu/software/lex-parser.html>

potentiella källor till åsikter. Analysen sker på meningsnivå, vilket innebär att en algoritm arbetar sig igenom en mening i ett dokument för att utföra diverse operationer på denna, inklusive NER [7]. NER kan även utföras på en mer detaljerad nivå i dokumentet där entiteterna representerar produktanvändares åsikter vid exempelvis produktrecensioner. Dessa entiteter är explicita och flertalet korresponderande åsikter kan knytas an till entiteten [7].

Dey och Haque förklarar dock i sin artikel att det är svårt att göra NER inom sentimentanalys av ostrukturerad text, då texten ofta är felstavad och ibland inte använder versaler vid namn. Även om versaler används kan det vara så att det inte är ett namn på en person eller organisation, utan en namngiven funktion såsom "Bluetooth", vilket definieras som namn på person när NER genomförs [47]. I det nämnda exemplet går det att se att *social@risk*TM inte är en identifierad entitet. Det beror i detta fallet på att namnet på mjukvaran inte börjar med en versal. Problemet går att kringgå genom att programmera regler eller skriva en egen korpus för lokalisering av entiteter [47].

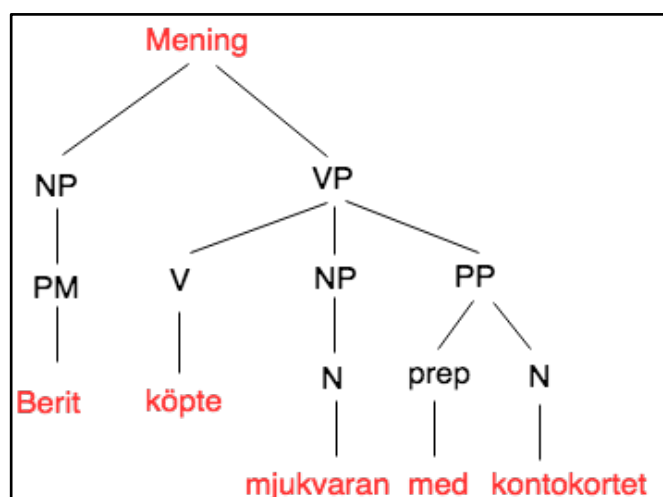
Israel *m. fl.* [54] använder *General Architecture for Text Engineering (GATE)* tillsammans med *Multi-language Predicate Arguments Extractor (MultiPAX)* för att utföra NER i syfte att utföra semantisk analys på multi-fokuserade dokument, genom att värdera meningar utefter hur många entiteter de innehåller och tilldela meningen det värdet. De utför sedan en normalisering av värdet genom att dela antal entiteter med totalt identifierade entiteter [54].

Vid summering av text används NER i syfte att filtrera ut de ord som summeringen ska bestå av. Meenaa och Gopalanib [53] har publicerat ett ramverk för ett verktyg som utför automatisk summering av texter där *named entities* utgör steg nummer två i att avgöra om ordet ska tillhöra summeringen eller inte. Ramverket är oberoende av domäntyp men använder kunskap som finns tillgänglig hos olika korpus.

Det finns verktyg att använda för att utföra NER, [7] nämner bland annat *OpenNLP* och *CoreNLP*. Som tidigare nämnt finns också verktyget *GATE* att tillhandahålla i detta syfte.

Parsing - Syntaxanalys

Eftersom de dokument som skrapas ner kan innehålla flera typer av språk så behöver texten parsas. Att parse ett dokument innebär att bryta ner dess struktur till individuella komponenter [20]. Parsing och POS-tagging kan verka lika på så sätt att båda tillvägagångssätten bidrar med detaljerad information om ord i en mening. Skillnaden är att POS-tagging bidrar med lexikal information om ord, medan parsing erhåller syntaktisk information. Det finns två olika tekniker för parsing: *dependency parsing* och *constituency parsing*. Dependency parsing innebär att sammankoppla individuella ord med deras relationer och constituency parsing innebär att iterera igenom en text för att bryta ned texten i delar [46]. Vid parsing produceras ett träd som representerar den grammatiska strukturen, tillsammans med motsvarande förhållanden och beståndsdelar i en given mening [7] (se fig.1).



Figur 1: Exempel av ett constituency-baserat parserträd med dess indelningar och beroenden.

Parsning bidrar alltså med en struktur som är mer rik på information jämfört med POS-tagging. Syntaktisk parsning är enligt Sun *m. fl.* en central del då parsning hjälper till att medla mellan lingvistiska uttryck och innebörden i en mening [7]. POS-tagging och parsing används ofta i kombination med varandra i syfte att bryta ned och analysera ord samt dess innebörd som utgör en text [7], [46], [47].

Att välja features

Att välja features (eng. *Feature Selection*) är det steg i behandlingsprocessen som går ut på att sälla bort redundanta och irrelevanta features som påverkar datan negativt, utan att förstöra den semantiska innebörden i en mening. Genom att utföra feature selection minskar dimensionerna på datan som ska användas. Detta är inte bara viktigt för att säkerställa att de features som ska användas är signifikanta, utan bidrar även till att tiden för att analysera datan minskar [56]. Att välja ut de features som ska användas sker oftast innan datan matas in till en algoritm för inlärning. Det finns flertalet existerande metoder för att utföra feature selection: *Information Gain*, *Mutual Information* eller *Document Frequency*, där feature väljs ut baserat på deras uträknade feature-värde. De features med högst värde prioriteras för användning [56].

Ribeiro *m. fl.* [55] har skapat algoritmen *Omega*, vilken simultant utför övervakad (eng. *supervised*) diskretisering av data och feature selection. Algoritmen väljer features genom att sälla bort de features som endast genererar ett intervall och behåller resten då dessa features oftast är oberoende av klasser. Vid utvecklandet av *Omega* utfördes tester i syfte att jämföra algoritmens funktion av diskretisering och feature selection med redan existerande algoritmen *Chi2* och *Relief*. Resultaten visar att *Omega* producerade ett träd med minst antal noder och lägst felfrekvens.

Vid träning av klassificerare väljs features ut i syfte att minska dimensionerna på feature-vektorn [43], [56]. Gayathri och Marimuthu [56] klassificerar text baserat på feature selection med syftet att träna klassificerarna *k-nearest neighbor (KNN)* och *Support Vector Machine (SVM)*. Detta görs med hjälp av TF-IDF, vilket i detta syfte innebär att en faktor tilldelas termer i en mening för att symbolisera hur viktig termen är. Resultatet visar att ju mer data som kommer in

till klassificerarna desto mer opålitliga blir resultaten. Därför är det viktigt att minska antalet features till de mest relevanta för att klassificerarna ska kunna bibehålla sin optimala prestation. [43] utförde experiment där målet var att förbättra klassificering med hjälp av olika metoder för förbehandling. Metoderna som används vid experimenten är: eliminering av stoppord och stemming tillsammans med feature selection. Resultaten visar att feature selection har en positiv inverkan på klassificerarens prestation tillsammans med eliminering av stoppord och stemming.

Det finns andra tillvägagångssätt för användning av feature selection. GeethaRamani *m. fl.* [57] gör försök att identifiera känslor i nyhetsartiklar och använder sig av *Correlation Based Feature Selection*. Vilket innebär att de features som väljs ut ska ha hög korrelation till dess tillhörande klass, men inte ha någon korrelation mellan varandra. Experiment utfördes där författarna skapade en feature-vektor med ord vilka representeras genom en binär sträng. Orden hämtades från *WordNet-Affect* vilket är en samling ord tillhörande kategorier som exempelvis känslor och beteenden. Om det upptäckts ord i ett textstycke som inte finns i samlingen letas ordets rotform upp och läggs till, vilket resulterar i att dimensionerna på vektorn blir stora. Genom att applicera feature selection minskade dimensionerna på vektorn från 6542st till 68st.

Term Frequency - Inverse Document Frequency

TF-IDF är ett statistiskt tillvägagångssätt att vikta termer i ett dokument som återfinns i en samling av dokument i syfte att se hur ofta ord förekommer. De ord som är oftast förekommande får således den största vikten [42], [58]. TF-IDF består av två olika delar där Term Frequency (TF) är det första steget som görs för att i sin tur applicera Inverse Document Frequency (IDF). TF är den algoritm som räknar ut hur frekvent en term förekommer i *ett dokument*, medan IDF är den algoritm som normaliserar vikten av termerna genom att räkna hur frekvent ord förekommer i en hel *samling* av dokument. TF-IDF värdet är specifikt för det dokument som beräknas [42]. Uträkningen för TF-IDF går till enligt följande [42], [58]:

$$Tfidf(t, f, d) = tf(t, d) * idf(t, d) \quad \text{alternativt} \quad Tfidf(t, d) = tf(t, d) \times idf(t)$$

TF-IDF kan användas i flera olika steg i behandlingsprocessen, exempelvis vid filtrering av stoppord inom fält som klassificering och summering av text [10], [54]. Djellali [59] använder TF-IDF i syfte att indexera termer och sedan träna en modell för klustring av data. Dock nämner författaren att det semantiska förhållandet mellan ord i en mening försvinner vid användandet av TF-IDF då uträkningen endast värdesätter termer. Författaren presenterar även en lösning för problemet. TF-IDF kan även användas i syfte att identifiera och extrahera nyckelord [11], [60]. Genom att vikta de mest värdefulla orden och ge dessa som input till metoden för dimensionsreduktion kunde [60] extrahera de 10.000 nyckelord som finns i applikationens kluster.

4.2.2.1 Maskininlärning

Inom kategorin databehandling existerar det en specifik familj av tekniker vilka har visat sig prestera utmärkt för en mängd olika problem inom områden såsom sjukvård, datavetenskap och finans. Maskininlärning (ML) är en underkategori inom Artificiell Intelligens (AI) vilken fokuserar på möjlighörandet för datasystem att lära sig hur en specifik uppgift bör lösas på automatiskt vis utifrån given data [4]. För ML existerar det en del olika klasser som exempelvis *supervised learning*, *unsupervised learning*, och *deep learning* [4], [7].

Supervised Learning

Supervised learning är en klass inom ML vilken drar generaliseringar utifrån data som fördefinierats av människor [4]. Genom att träna algoritmer utifrån exempel kan således algoritmer inom den här klassen ta lärdom av dessa för att sedan förutspå framtida, okända dataset och situationer. Där ML används i syftet att förutspå en kontinuerlig serie av värden kallas problemet för *regression*, medans det rör sig om *klassificering* när diskreta värden bearbetas.

För supervised learning finns det algoritmer som visat sig populära. *Naive Bayes* klassificerare är en teknik vilken ofta används i samband med klassificering av internettrafik [4]. Santos och Ladeira [45] applicerar Naive Bayes på ett dataset som behandlats med diverse förbehandlingstekniker i syftet att utföra sentimentanalys. Författarna beskriver att den initialt gav ett bra resultat med det träningsdataset som användes. Dock gav den ett sämre resultat i senare skeden som följd av att antalet positiva revyer blev fler i antalet än andra. När detta händer tenderar algoritmen att klassificera fler instanser som positiva. Med hjälp av så kallad "oversampling" och "undersampling" kunde man överkomma denna problematik.

Även Nhlabano och Lutu [43] beskriver hur Naive Bayes algoritmen kan användas som ett klassificeringbaserat tillvägagångssätt för sentimentanalys. Författarna argumenterar för sitt användande av Naive Bayes som en följd av dess förmåga att vara lättlärd. En rad olika förbehandlingstekniker användes även i denna studie, där resultatet av Naive Bayes-algoritmen undersöktes närmare för att se hur behandlingsteknikerna påverkar algoritmen. Författarna konstaterade att det mest exakta resultatet kunde uppnås då antalet dimensioner reducerats med hjälp av feature selection, i samband med stop word removal och stemming [43]. Naive Bayes är populär som följd av dess enkelhet att implementera och använda som motsats till andra klassificeringstekniker [61]. Algoritmen har dock kritiserats för att vara överlag dålig, och beskrivs göra grova antaganden rörande den bearbetade datan. Genom att föreslå diverse heuristiska lösningar för de problem som algoritmen står inför påvisar författarna i [61] att Naive Bayes kan tävla med state-of-the-art klassificerare såsom SVMs.

Jianqiang och Xiaolin analyserar även effekterna av olika behandlingstekniker för sentimentanalys. Detta gjordes för två olika klassificeringsproblem och för fyra olika klassificerare, däribland Naive Bayes [39]. Den bearbetade datan hämtades från Twitter. I sammanfattningen konstaterade författarna att filtreringen av URLs, stopword och nummer hade en minimal

påverkan på prestandan hos algoritmen, men att det trots detta fortfarande är lämpligt att utföra filtrering för att reducera brus och öka precision.

En annan teknik inom supervised learning är *Support Vector Machines* (SVMs), vilka har visat sig användbara i syftet att utföra sentimentanalys och klassificeringar. Sharma och Dey föreslår en hybridimplementation för sentimentanalys baserad på en SVM, där de förbättrar algoritmen genom att träna den med hjälp av diverse "bagging"- och "boosting"-tekniker [62]. Detta tillvägagångssätt visade sig utklassa det resultat som producerats av ensamstående SVMs vid klassificering.

Santos och Ladeira [45] applicerar i sin artikel bland annat en SVM för sentimentanalys på ett dataset av recensioner, vilka förbehandlats i syfte att minska brus bestående av internetslang, förkortningar och stavningsfel. SVM visade sig prestera bäst av de olika teknikerna, vilket kungjordes med en *5-fold cross validation*, en statistisk teknik vilken kan användas för att gradera resultatet av maskininlärningstekniker som har ett begränsat dataset.

Gayathri och Marimuthu [56] har i sin studie undersökt både *K-Nearest Neighbour* (KNN) och SVMs för textklassificering. Författarna diskuterar bland annat de största nackdelarna med SVMs, där de menar på att användandet av SVMs är krävande rent beräkningsmässigt. Något som blir extra tydligt när dataset växer i storlek. Dock påstår författarna att SVMs vanligen sägs vara den mest exakta algoritmen för textklassificering. SVMs är även mindre känsliga för brus i jämförelse med andra klassificeringsalgoritmer såsom Naive Bayes eller Random Forest [39].

Utöver Naive Bayes och SVMs kan även tekniken KNN användas för supervised learning. KNN beskrivs som ett tillvägagångssätt som är lätt att implementera och kommer med en hög grad av effektivitet för många klassificeringsproblem. Den blir dock beräkningsmässigt tung i kombination med att den påverkas negativt precisions-mässigt när mängden data och dimensioner växer [56].

Unsupervised Learning

Grundkonceptet bakom unsupervised learning är *klustering*, vilket innebär att data kategoriseras utefter upplevd liknelse av attribut i relation till annan data [4]. Klustering är en teknik vilken grupperar liknande dokument, men skiljer sig från kategorisering genom att dokument klustras i farten istället för genom fördefinierade ämnen [11]. En enkel klustringsalgoritm skapar en vektor av "topics" för varje analyserat dokument och mäter vikter för hur väl de olika dokumenten passar ihop med varandra. Klustering kan användas för tusentals olika dokument.

Det första steget för klustering är att transformera de dokument som skall behandlas, vilka typiskt förekommer i textformat, till en lämplig representation för klustringens ändamål [11]. Därefter måste diverse förbehandlingssteg tillämpas, såsom eliminering av stopppord, stemming och filtrering av domänord. Detta är viktigt för att reducera antalet dimensioner, ett centralt problem för textbearbetningsalgoritmer, eftersom beräkningstiden för nämnda algoritmer skalar dåligt i relation med att antalet dimensioner växer. Genom att länka olika dokument efter liknande attribut kan vi effektivt finna relationer i data vilken vi möjligtvis inte funnit med traditionella sökmedel.

Klustring kräver ingen träning av fördefinierad data, därav namnet “unsupervised” och är en kraftfull teknik för att extrahera förutsägbar information som gömmer sig i ett dataset. Den hjälper till att fokusera på relevant information i en korpus [59]. Metoden kan hjälpa till att identifiera homogena grupper av objekt baserat på värden hos attributen (dimensionerna) objekten besitter. Klustring har på senare tid blivit alltmer kritiserad som dataanalytisk metod [59]. De flesta klustringsmetoderna är känsliga för outliers, brus, presentationsordning, och växande dimensioner. Djellali presenterar en lösning i [59] för att tillmötesgå dimensionsproblemen som förekommer vid stora mängder textuell data. Författaren tillämpar en “wrapper model” som filtrerar ut relevanta och lämpliga variabler som kan reducera dimensionerna på datan tillsammans med algoritmer för brusreduktion.

Traditionella textmining- tekniker såsom klustring och topic modelling kan inte användas trivialt vid processande av livestreamad data, utan används därför vanligtvis för att ge meningsfulla insikter i syfte att förbättra exempelvis sökfrågor eller att analysera ett statistiskt textkorpus [51].

Schubert *m. fl.* [51] föreslår en approach för att identifiera trender i storskalig data, där de tillämpar en trestegsapproach för att identifiera uppdykande ämnen vilka de senare bearbetar med klustring. Huang *m. fl.* [65] presenterar en topic detection- metod baserad på textklustring och topic model- analys. Traditionella topic detection-metoder lämpar sig dock inte för gles microbloggtext, såsom twitterdata, något som den presenterade metoden ämnar ändra.

Deep Learning / Neural Networks

Under många år har maskinlärningstekniker såsom *SVM* och *logistic regression* tillämpats som förstahandsval för många problem inom NLP [46]. Detta hoppas kunna förändras med ett nytt upptåg av tekniker för *deep learning*, vilka har givit imponerande resultat inom områden såsom exempelvis mönsterigenkänning och *speech recognition* [7], [46]. *Neurala nätverk* (NN) som är baserade på vektorrepresentationer har visat sig producera bra resultat för diverse NLP-problem, såsom NER och POS-tagging, mycket tack vare uppfinnandet av *word embeddings* och tillämpandet av *multi-level automatic feature representation learning* [7], [46].

Word embeddings, eller *distributional representations*, bygger på idén att ord med liknande innebörd tenderar att återfinnas i liknande kontext [46]. En embedding förekommer i formen av en vektor, och används för att analysera egenskaperna hos grannarna till ett ord. Detta kan vara till nytta när man vill fånga upp information om hur meningar är strukturerade, eller att finna likheten mellan ord. Word embeddings används oftast som det första steget i en deep learning modell.

Den andra prominenta faktorn vilken gör neurala nätverk fördelaktiga är *multi-level automatic feature representation learning*, vilket innebär ett oberoende av någon form av feature engineering [7]. Detta leder till att kravet för människor att besitta kunskap inom den domän man arbetar inom inte blir lika högt [66]. Traditionella maskininlärningsalgoritmer för NLP-problem baserar sig ofta mycket på handtillverkade features, vilka är tidskrävande att producera och är dessutom ofta inkompleta [7]. Vidare argumentation för användandet av NN är att de även

effektivt skalar uppåt ur ett prestandaperspektiv när mängden data ökar [66]. Bland de mest populära deep learning-teknikerna i dagsläget finner vi *Convolutional neural networks* (CNNs), *recurrent neural networks* (RNNs) och *recursive neural networks*.

Som en följd av populariseringen av word embeddings och dess förmåga att representera ord i ett distribuerat utrymme, uppstod ett behov av en funktion vilken extraherar features av högre nivåer [46]. CNNs visade sig vara effektiva för att lösa detta problem. Övergripande fungerar ett CNN som så att data processas genom diverse lager av bearbetningskärnor, eller *convolutional filters*, vilka utför specifika beräkningar på den inmatade datan [46]. Efter bearbetning skickas den bearbetade datan vidare till nästa lager vilket i sin tur utför bearbetning. Oftast används hundratals olika bearbetningskärnor i ett nätverk.

Çano och Morisio [66] nämner att CNNs presterat extremt väl för bildanalys, och presenterar i deras artikel ett CNN för sentimentanalys. Detta CNN tränas med hjälp av annoterade word embeddings, och uppnår en precision på hela 91.2%, en förbättring om man jämför med exempelvis SVMs vilka tidigare nämns uppnå 75-83% (90% i en boostad SVM presenterad av Haddi *m. fl.* [63]) .

CNNs är effektiva för att finna semantik inom en begränsad omgivning. De är dock tunga modeller som kräver ett stort träningsset [51]. Problem uppstår därmed när det inte finns en stor mängd data att träna på. Ett annat problem är att de är inkapabla att bevara en sekventiell ordning i deras representationer, något som andra modeller som *Recurrent Neural Networks* (RNNs) hanterar bättre.

RNNs är en typ av neuralt nätverk vilket bygger på idén att processa sekventiell information, vilket i kontexten innebär att olika iterationer av bearbetningsprocesser bygger på de från föregående iterationer [46]. På så vis kan man konstatera att RNNs har effektivt minne från tidigare bearbetningar. RNNs används effektivt för diverse NLP-koncept som *language modeling*, *machine translation*, och *speech recognition*, och är väl lämpat för detta ändamål då innebörden av ord och meningar i naturligt språk ofta bygger på semantisk mening och struktur från tidigare ord och meningar [7], [46]. Utöver att vara strukturellt kompatibla med naturligt språk är även RNNs användbara för att bearbeta data av olika storlek. De kan på ett dynamiskt sätt bearbeta alltifrån väldigt långa meningar, till paragrafer och kompletta dokument [7], [46]. Utöver förmågan att vara användbara för översättning och taligenkänning passar sig RNNs även väl för uppgifter som sentimentklassificering, POS-tagging och subjektivitetsdetektering [7], [46].

Wang *m. fl.* presenterar i deras artikel ett adaptivt RNN vilket bygger på "kapslar" som representerar varje sentimentkategori givet ett analysproblem [64]. Det föreslagna nätverket ger en state-of-the-art prestation för sentimentanalys, och kräver ingen form av lingvistisk kunskap.

Borges *m. fl.* beskriver att trots att ickeinjära metoder såsom NN och SVMs ofta ger de mest precisa resultaten för de flesta förutsägelser gällande okänd data, är det inte realistiskt att använda dem i situationer där validering erfordras som en följd av att de är black-box tekniker [67]. Som en följd av detta har neurala nätverk varit begränsade för användning inom exempelvis äldre system. För att lösa detta problem försöker man använda sig av white-box modeller.

Recursive Neural Networks är en annan modell vilken passar för att bearbeta naturlig språkdata. Den används vanligen som en parser, och är väl anpassad för att traversera diverse hierarkiska strukturer, såsom trädstrukturer vilka bland annat kan representera hierarkin för grammatik i naturlig språktext [46].

4.2.3 Lagring av data

För att lagra stora mängder ostrukturerad data krävs en lämplig databas, vilket nämnts tidigare. I dagsläget är det vanligt att använda databaser vilka faller under termen NoSQL. Detta har blivit populärt som följd av dessa databasers förmåga att effektivt skalas upp i kombination med att tillhandahålla flexibilitet och snabbhet [4]. En nackdel är dock att dessa databaser inte direkt uppfyller egenskaperna *ACID* (eng. *atomicity*, *consistency*, *integrity*, och *durability*), utan dessa måste programmeras in manuellt.

Ett annat behov som databaser behöver tillmötesgå är egenskaperna *MAD* (eng. *magnetic*, *agile*, och *deep*). Med detta vill det sägas att databasen bör vara kapabel till att använda data från en mängd olika källor samt att på en strukturell och fysisk nivå kunna synkronisera med denna data [2].

Det är fortfarande möjligt att använda relationella databaser i syftet att lagra stora mängder data. I sådana situationer är det dock lämpligt att strukturera den lagrade datan på ett plan lämpligt för den relationella databasen. I [9] struktureras 50 miljoner olika instanser av användarbeskrivningar av textuell natur i en relationell databas. Detta uppnåddes genom att strukturera upp den skrapade datan efter en mängd olika index.

Det existerar ramverk som kan underlätta uppsättningen av ett databassystem. Ett exempel på ett sådant är Hadoop, vilket kan användas för att starta upp ett ekosystem av kluster [45].

Valet av databas kan vara användbart för hur data ämnas visualiseras i senare skeden. Författarna av [58] använder i sin artikel NoSQL-databasen *Neo4J* i syftet att lagra ostrukturerad textdata från bland annat Wikipedia för att sedan visualisera denna i hopp om att finna nya associationer i datan.

4.2.4 Visualisering för insiktsgenerering

Det slutgiltiga målet med behandlingsprocessen av data vilken skall analyseras är ofta att försöka dra nya insikter utifrån den. För att analytiker skall kunna resonera kring denna data måste den därmed visualiseras i någon mån. Visualiseringar bygger på features och indexering av nyckelord för att bygga grafiska representationer av dokument [11]. Valet av visualiseringstyp är av stor vikt. Det är möjligt att presentera data i råformat, till exempel som ett JSON-objekt eller som en vanlig tabell. Dessa visualiseringar blir dock orimliga att använda för tolkning i samband med att mängden data växer. *Visual analytics* är ett koncept vilket innefattar vetenskapen om att på ett analytiskt vis resonera kring interaktiva visuella gränssnitt [4]. Genom att visualisera data i samband med att tillhandahålla diverse interaktionsmöjligheter, kan stora mängder data effektivt reduceras till mer lättförstådda beståndsdelar [11]. Det finns en del olika grundtekniker för att visualisera data i olika dimensioner, som Arjun *m. fl.* diskuterar [4]:

1. *Data maps* - En form av visualisering som vanligen struktureras som en hybrid av kartografi och statistik. Kan exempelvis användas för att dokumentera regionala kvantitativa värden som medelålder per region.
2. *Time series* - Teknik vilken målar upp skillnader i mätvärden över ett tidsspänn, kan bland annat användas för att visa kurser för aktier.
3. *Space-time narrative* - En multivariat (beroende av ett flertal variabler) representation vilket introducerar *rymd* som en ytterligare dimension för en *time series*.
4. *Relational graphics* - Visuell representation som inte nödvändigtvis är bunden till någon specifik typ av variabel. Här undersöks relationen mellan två eller fler kvantiteter. Ett exempel på en instans där en sådan relation representeras kan vara förhållandet mellan rökning och dödsfall över ett tidsspänn och för en viss region. Med en sådan representation kan till exempel effekten av en nyinförd reglering av rökning utvärderas [4].

När en visuell presentation tillämpats är även valet av analytisk metod viktig för hur beslutstagare bör resonera kring denna. För visualisering av stora mängder data har koncepten *ADV* (eng. *Advanced Data Visualization*) och *visual discovery* visat sig vara de koncept med den största potentiella tillväxten som förvalda metoder för dataanalys [2]. *ADV* är en datadriven, explorativ approach som applicerar metoder för dataanalys i kombination med interaktiv visualisering. Den lämpar sig i situationer då analytikern inte har mycket kunskap om den analyserade datan.

För att bättre kunna visualisera data är det användbart att förstå vilka aspekter av visualiseringen som bäst bidrar till insiktsgenerering. I en studie undersöker Guo *m. fl.* hur användare kommer fram till nya insikter utifrån interaktioner med visuella användargränssnitt [68]. De undersöker även vilka designfaktorer som potentiellt har en negativ påverkan på denna process. Genom att bland annat analysera loggar vilka registrerat interaktionerna mellan användare och gränssnitt hoppades författarna svara på frågan. Mätningar av denna kvantitativa data pekade på att specifika delar i interaktionen stod för de nya insikterna, något som senare även kunde bekräftas med kvalitativ analys och videoövervakning [68].

För att visualisera data är det oftast enklast att använda ett redan utvecklat verktyg. Databasen *Neo4J* tillhandahåller exempelvis ett sådant verktyg [58]. Verktöget använder ett så kallat *query language* vid namnet *Cypher*, vilket kan liknas med det som används med SQL. *Cypher* kan med hjälp av olika nyckelord och regler användas för att utföra sporadiska förfrågningar till databasen för att experimentera fram olika representationer och perspektiv. I [58] används ett exempel på en situation där *Neo4J* används för visualisering, där symptom och medicin, vanligt förekommande mellan två sjukdomar, visualiseras och resoneras kring.

Utöver *Neo4J* existerar fler verktyg för ändamålet att visualisera data. I [60] presenteras ett ytterligare verktyg, *ViTA-SSD*, vilket arbetar på semi-strukturerad data i syftet att utföra explorativ visuell analys på relationell data. Genom att presentera både information rörande metadata för de analyserade dokumenten i

kombination med diverse visualiseringar kan nya mönster lätt identifieras. Systemet utför dimensionsreduktion på de dokument vilka den analyserar. Därefter används en metod för att snabbt klustra datan. Användaren för systemet kan här påverka klustringen med olika inställningar för avstånd et cetera. I artikeln studerades även den upplevda användarvänligheten för verktyget, vilken visade positiv respons från testarna.

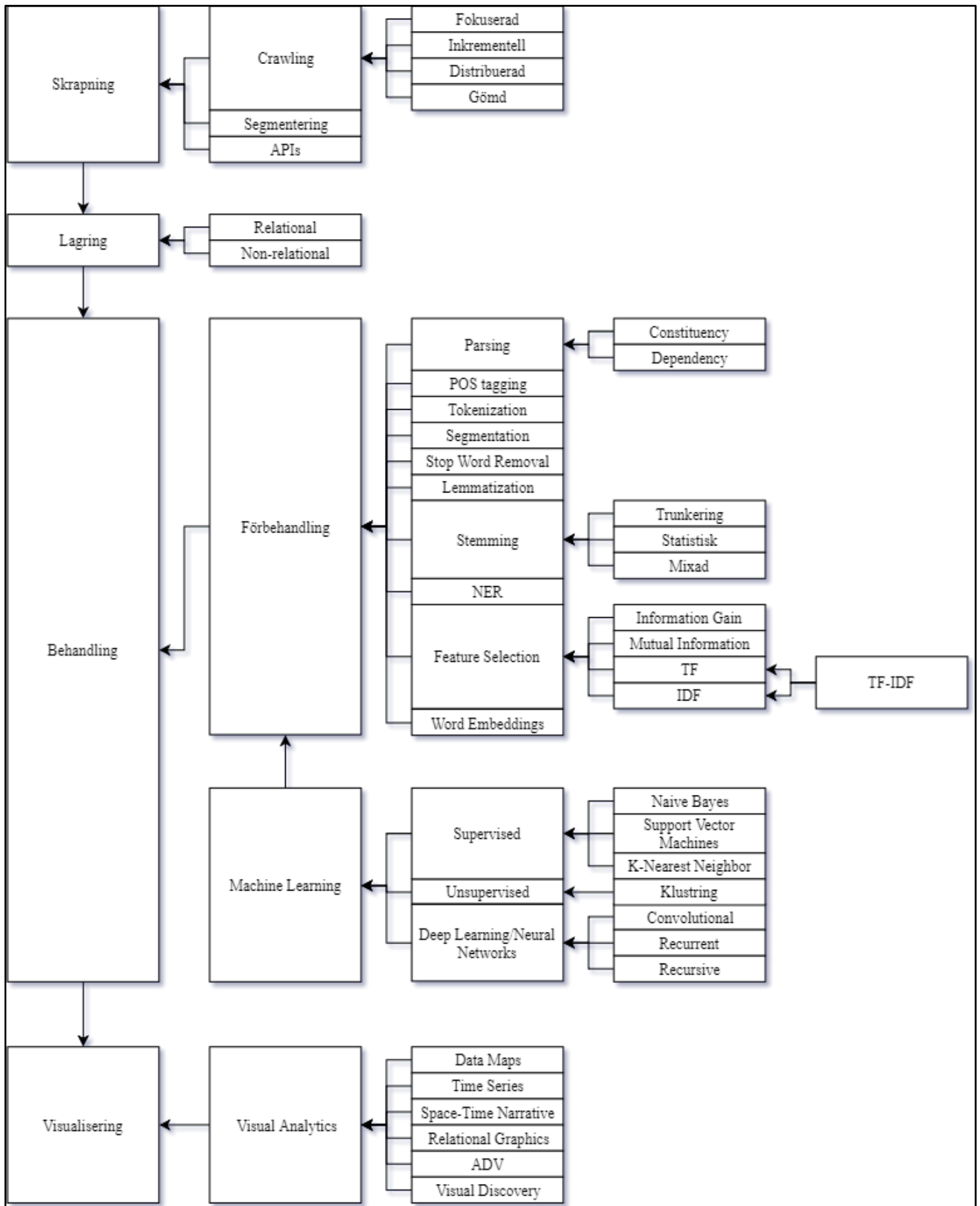
En i dagsläget populär metod för bearbetning av data för insiktsgenerering är sentimentanalys, eller *opinion mining* som det även kallas. Det finns en del olika verktyg för detta ändamål. I [69] presenteras *iFeel*, ett verktyg på webben vilket används för att utföra sentimentanalys på olika sätt. I verktyget tillämpas 8 olika tekniker för sentimentanalys på samma gång, bland dessa återfinns *PANAS-t*, *Emoticons*, *SentiWordNet*, *Happiness Index*, *SASA* och *SenticNet*. Meningar vilka skall klassificeras matas in i ett användarvänligt typsnitt och resultatet presenteras på ett intuitivt vis. Med *iFeel* kan användare skapa en övergripande bild av olika tekniker och jämföra dessa för en mer omfattande bild av sentiment rörande given input.

5 Analys

Många av de artiklar som hittades berör ämnet sentimentanalys då det är ett populärt ämne som ligger rätt i tiden. Globalworks har dock föga intresse att utröna positiva, negativa och neutrala känslor i en text då detta endast är en yttlig beskrivning av känslor. Företagets intresse ligger djupare på så sätt att de behöver kunna *förstå* kontexten i en situation som beskrivs av användarna i publika forum i syfte att djupare kunna förstå den situation de befinner sig i. I många fall är det inte möjligt att identifiera uttryckta kränkningar med enbart sentimentanalys, speciellt i situationer där uttryckt missnöje inte kan ges explicit. Exempelvis kan anställda i Kina inte uttrycka direkt missnöje utan att "censurera" vad de tycker genom att använda andra termer, eller kodord, som inte reflekterar den faktiska innebörden av de ord som står i en mening [5]. Några av de grundläggande tekniker som används vid sentimentanalys skulle dock kunna användas i Globalworks system, eftersom det är tekniker som ingår inom området för NLP generellt. Därför fattade vi beslutet att behålla dessa artiklar.

Vid analys av [6-11] som nämns i avsnitt för tidigare forskning, kunde vi se ett mönster i att det generellt existerar fyra huvudområden som summerar dataflödet genom ett informationssystem: Skrapning av data, behandling av data, lagring av data samt visualisering av data i syfte att generera nya insikter. Som tidigare nämnt ligger dessa fyra områden till grund för strukturen av detta arbete. För att få en överblick av de tekniker som tagits upp inom respektive huvudområde i litteraturstudien visas ett diagram (se fig. 2). I vissa artiklar nämns andra verktyg och tekniker som inte visas i diagrammet då informationen om dessa är knapphändig, exempelvis gällande semi-supervised learning.

Resterande delar i kapitlet ämnar att analysera fynden inom respektive huvudområde. Vi har i detta kapitel valt att i omvänd ordning resonera kring de fyra huvudområdena eftersom vi anser att det viktigaste delmomentet i hela processen är den slutgiltiga presentationen av datan, då det är denna som hjälper Globalworks experter att komma till nya insikter. Vi har utifrån litteraturstudien kommit fram till slutsatsen att diverse bearbetningsalgoritmer är känsliga för olika varianter av brus och struktur på data, vilket i sin tur kan påverka den data som visualiseras. Därmed börjar vi med att analysera ämnet för visualisering i syfte att utefter den analysera vilka tekniker som bör användas vid förbehandling, bearbetning och lagring.



Figur 2: Sammanställning av de tekniker som presenteras i resultaten, med tillhörande huvudområde.

5.1 Visualisering för insiktsgenerering

Området visualisering för insiktsgenerering sträcker sig långt bortom vad vi initialt trodde det skulle göra. Det berör inte enbart de visualiseringar som presenteras av den bearbetade datan, utan även om förståelsen av den. Hjärnan och människan måste kunna tolka den information som visas till den grad att genereringen av nya insikter och idéer möjliggörs. Baserat på information från tidigare forskning samt vår litteraturstudie, kan vi därmed konstatera att den slutgiltiga visualiseringen av datan bör vara interaktiv för optimal användarvänlighet. I det bästa av fall bör det även vara lämpligt att ha möjligheten att experimentera med olika vyer och perspektiv, då det är det utforskande momentet som tillgängliggör ökade chanser för insiktsgenerering. I nuläget ser det dock inte ut som att det existerar fullt utvecklade tillvägagångssätt för att tillgängliggöra möjligheten för analytiker att utföra visuell utforskning av datan i hoppen om att komma fram till meningsfulla och relevanta resultat [60]. Därför bör olika verktyg och ramverk kombineras för att försöka uppnå önskad effekt och möjlighet för interaktion.

5.2 Behandling av data

Förbehandlings tekniker används för att behandla den ursprungliga, ostrukturerade textdatan i syftet att eliminera brus, extrahera viktiga features, finna semantiska likheter och syften, med mera. Det är viktigt att använda korrekta förbehandlings tekniker för senare bearbetningsalgoritmer då dessa oftast är känsliga för brus. Vilka av de tekniker som listas ovan som appliceras i ett IR-system är individuellt baserat på uppgiften systemet ska lösa. För ostrukturerad textdata rekommenderas det i flertalet artiklar att majoriteten av de tekniker som presenteras i resultaten appliceras i syfte att skicka ren och rätt data till algoritmerna som sköter klassificering och inlärning.

I de situationer där ett fast beslut tagits rörande de slutgiltiga behandlings- och visualiseringsteknikerna för datan, kan det vara lämpligt att bearbeta den skrapade datan med diverse förbehandlingsalgoritmer *innan* den lagras in i databasen. Syftet med detta är för att eventuellt spara utrymme i längden, och framförallt beräkningstid för framtida queries och bearbetningar. Ett sådant beslut bör noggrant övervägas då urvalet av användbara tekniker för senare bearbetning påverkas av den typen av förbehandling som utförs.

5.2.1 Maskininlärning

Tekniker inom området maskininlärning har visat sig prestera fenomenalt för de flesta problem rörande bland annat klassificering och analys, där tekniker som SVMs har regerat som mästare. Användandet av maskininlärningsalgoritmer är i stor grad baserad på förekomsten av annoterad data för träning. I de flesta situationer saknas det sådan data vilket leder till att tekniker inom klassen *unsupervised learning* och *semi-supervised learning* är de mest troliga kandidaterna för användning där stora mängder ostrukturerad data skrapas.

I resultatdelen har vi undersökt ett fåtal klasser inom området maskininlärning såsom supervised learning, unsupervised learning och deep

learning. Vi har dock inte presenterat en komplett bild över dessa. Som en följd av bristfällig information i de av litteraturstudien funna artiklarna har vi valt att inte inkludera någon information om dessa. För den intresserade rekommenderar vi artiklarna skrivna av Ali *m. fl.* [4], Sun *m. fl.* [7] samt Young *m. fl.* [46] bland andra.

5.3 Lagring av data

Valet av databas är generellt beroende på systemkrav och attribut hos den data som skall skrapas. Det är inte omöjligt att lagra ostrukturerade dataset i relationella databaser, det beror på hur man resonerar kring datan man lagrar och hur man strukturellt väljer att lagra den. Man kan fortfarande indexera användares posts från exempelvis twitter med hjälp av ID, datum och innehåll i form av namngivna entiteter, för att på så vis få det att fungera i en relationell databas. Här framstår dock ett problem om hur väl en relationell databas kan användas för att tillmötesgå diverse komplexa queries. Beroende på restriktioner i den relationella modellen kan även flexibiliteten i de queries man kan använda begränsas. Detta blir mer problematiskt vid beaktandet av att system oftast har krav på attribut som snabbhet. En icke-relationell databas kan tillmötesgå komplexa queries bättre. I dokumentdatabaser kan exempelvis data lagras i komplexa nästlade strukturer kallade *aggregat* [22]. Alltså är det förslagsvis bättre att använda icke-relationella databaser för stora mängder ostrukturerad data.

5.4 Skrapning av data

Att "hitta en nål i en höstack" symboliserar i Globalworks fall att försöka hitta de enskilda röster som uttrycker åsikter eller klagomål, via sociala medier, kring arbetsplatser i högriskländer. För att göra detta möjligt är det nödvändigt att skrapa den data som ska analyseras. Crawlingen i sig är en automatiserad del inom IR och det finns flertalet tillvägagångssätt att skrapa data beroende på vilket mål som ska uppnås. Det är därför viktigt att analysera syftet gällande skrapningen innan valet av tillvägagångssätt görs. I resultatkapitlet för skrapning lyfts endast huvudområdena kring ämnet. Saini och Arora [12] nämner mer ingående information om vilka tillvägagångssätt som kan tillämpas och artikeln kan ge Globalworks en bra insyn i hur de kan resonera kring problemet.

6 Diskussion

De tillämpningar som används vid IR är utspridda och majoriteten av de artiklar som hittades behandlar området för sentimentanalys. Det går att se i dokumentationen av artiklarna (se bilaga A-D) där det finns en tydlig skillnad mellan antalet artiklar funna. Antalet behandlingstekniker är fler än de artiklar som är funna gällande skrapning, lagring och visualisering. Majoriteten av dessa artiklar fokuserar nämligen på dessa tekniker i syftet att utföra sentimentanalys.

Många av de artiklar som behandlar sentimentanalys handlar om hur det i ostrukturerad textdata går att hitta positiva, negativa eller neutrala sentiment för att sedan kategorisera in sentimenten. Som tidigare nämnt har tyvärr detta påverkat våra resultat negativt då Globalworks behöver hjälp med hur de ska förstå djupet i texten snarare än positivitet och negativitet. Dock kan det finnas viss nytta för Globalworks att använda sig av sentimentanalys i ett tidigt skede av kategorisering. Sentimentanalys kan genomföras med hjälp av en "lexicon based approach" som betygsätter vissa ord som negativa eller positiva, för att sedan jämföra det slutgiltiga sentimenten utav alla orden givet en eller flera posts. Genom att skapa ett eget lexikon innehållande de ord som ska sökas och betygsättas skulle Globalworks kunna utföra en tidig analys av sentiment.

Eftersom olika discipliner korsar varandra var det svårt att utföra litteraturstudien då avgränsningarna för var områden börjar och slutar är otydlig. Olika områden använder sig av olika vetenskapstraditioner, metoder och teknologier. Detta i sin tur resulterar i att olika begrepp används för mer eller mindre samma saker, vilket var svårt att veta då vi saknade den bakomliggande kunskap som krävdes vid sökning av litteratur. Därför borde vi, innan arbetet påbörjades, införskaffat oss mer djupgående information inom respektive område för att lättare kunna utröna vilka artiklar som innehåller information som är relevant samt att på ett djupare plan kunna analysera den information artiklarna innehåller.

Att utföra en systematisk litteraturstudie anses dock fortfarande vara det bästa tillvägagångssättet för att införskaffa den kunskap som krävs och samtidigt få en överblick i hur området för IR ser ut för att sedan kunna ge Globalworks en ontologisk överblick av de metoder och tekniker som används enligt litteraturen idag.

Rörande reflektioner för eventuell subjektivitet har vi inte kunnat frånga att vara subjektiva, då vi sållade ut våra relevanta artiklar i litteraturstudien baserat på den subjektiva bedömningen av vad som är relevant för Globalworks eller ej. Överlag presenteras den relevanta litteraturen dock på ett objektiva sätt och den generalisering av metoder och tekniker som visas i fig. 2 är en sammanställning av det som reflekteras i litteraturen.

6.1 Etiska aspekter

Globalworks samlar in data som är tillgänglig publikt, vilket innebär att alla med tillgång till internet kan läsa informationen. Att samla in den här typen av data kan väcka frågor kring den etiska aspekten gällande lagring av insamlad data och anknytning till personer [4], [42]. Att samla in data som är publik är inget ovanligt

och sker ofta i syfte att skräddarsy reklamannonser som riktar sig till individer. Den data Globalworks samlar in används endast till att lokalisera åsikter och arbetsrelaterade problem inom mänskliga rättigheter, där den insamlade datan inte knyts an till specifika personer. För att försäkra sig om användarnas integritet gör Globalworks inte heller några försök i att matcha insamlad data med andra källor. Den enda individuella informationen som lagras av en användare är dess kön. Personens kön lagras i syfte att kunna dra slutsatser gällande kränkande behandling av könsrelaterad karaktär.

Trots att Globalworks endast lagrar användarens kön kan detta vara ett problem enligt den GDPR-lag (*General Data Protection Regulation*) som trädde i kraft den 25:e maj 2018. Enligt GDPR måste användare av sociala medier godkänna att den data som publiceras får skrapas. Detta är i Globalworks fall en stor nackdel då det skulle utsätta användare som postar på internet i direkt fara. Det finns dock punkter i GDPR-lagen som kan ses som undantag gällande denna regel; Artikel 6(1) - "Laglig behandling av personuppgifter" listar sex punkter, varav minst en av dessa måste vara uppfyllda för att få lov att bearbeta data [70 pp.36]. Globalworks uppfyller i dagsläget minst två av dessa punkter. Det är dock av största vikt för företaget att noggrant se över GDPR-lagen för framtida skrapning av data.

7 Slutsatser och vidare forskning

Vi har utfört en systematisk litteraturstudie i syfte att ta reda på vilka metoder och tekniker som finns i dagsläget för att komma närmre en lösning till det "hitta en nål i en höstack"-problem Globalworks står inför. Därför utformades forskningsfrågan till att besvara vilka metoder och tekniker som finns i litteraturen i dagsläget, hur dessa kan kombineras för att underlätta för experter att sälla ut den mest relevanta datan, och till vilken grad det går att automatisera delar i Globalworks process.

Resultaten visar att det finns fyra huvudområden i ett IR-system: Skrapning av data, behandling av data, lagring av data och visualisering för insiktsgenerering. Inom dessa huvudområden presenteras den funna litteraturen i en sammanställning (se tabell 4) för att sedan presenteras i djupare detalj i en litteraturstudie. De metoder och tekniker funna i litteraturen presenteras med hjälp av ett diagram (se fig. 2).

Teknikerna som tas upp i litteraturstudien kan kombineras på olika sätt beroende på den uppgift som ska lösas. I Globalworks fall är det lämpligen värt att lägga större vikt på tekniker för klustring och visualisering för att kunna se den data som samlats in från olika perspektiv.

Gällande automatisering av delar i Globalworks system framkommer det att delen för crawling för tillfället är olämplig att automatisera då det är experter som innehar domänkunskap i ämnet. Efter crawlingen är slutförd går det att automatisera delar som exempelvis förbehandling och brusreducering samt bearbetning och klustring. Att automatisera generering av insikter är inget som rekommenderas då det i Globalworks fall krävs en djupare förståelse för människor, språk och kontext som inte en dator kan bidra med för tillfället.

Vidare arbete för Globalworks kan vara att implementera ett urval av de metoder och tekniker som presenteras i litteraturstudien i syfte att experimentera fram de metoder som ger bäst precision när det gäller att hitta en nål i en höstack.

8 Referenser

- [1] GSM Association. (2018) *The Mobile Economy 2018*. [Online]. Available: <https://www.gsma.com/mobileeconomy/wp-content/uploads/2018/02/The-Mobile-Economy-Global-2018.pdf>
- [2] N. Elgendy and A. Elragal, "Big Data Analytics: A Literature Review Paper," in *Advances in Data Mining. Applications and Theoretical Aspects*, vol. 8557, P. Perner, Ed. Cham: Springer International Publishing, 2014, pp. 214–227.
- [3] A. Bartusiak and J. Lässig, "Semantic Processing for the Conversion of Unstructured Documents into Structured Information in the Enterprise Context," in *Proceedings of the 12th International Conference on Semantic Systems - SEMANTiCS 2016*, Leipzig, Germany, 2016, pp. 125–128.
- [4] A. Ali, J. Qadir, R. ur Rasool, A. Sathiaselan, A. Zwitter, and J. Crowcroft, "Big data for development: applications and techniques," *Big Data Analytics*, vol. 1, no. 1, Dec. 2016.
- [5] S. Brehm and H. Magnusson. (2017, Aug.). *Wasting time, wasting youth*. Globalworks. Lund, Sweden. [Online]. Available: <http://globalworks.se/wp-content/uploads/2018/01/Dell-Report-Wasting-time-wasting-youth.pdf>
- [6] L. Dey and S. M. Haque, "Opinion Mining From Noisy Text Data," in *Proceedings of the second workshop on Analytics for noisy unstructured text data - AND '08*, Singapore, 2008, pp. 83-90.
- [7] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, Jul. 2017.
- [8] X. Dai, I. Spasic, and F. Andres, "A Framework for Automated Rating of Online Reviews Against the Underlying Topics," in *Proceedings of the SouthEast Conference on - ACM SE '17*, Kennesaw, GA, USA, 2017, pp. 164–167.
- [9] K. Inui *m. fl.*, "Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents," in 2008 *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Sydney, Australia, 2008, pp. 314–321.
- [10] D. S. Vijayarani and J. Ilamathi, "Preprocessing Techniques for Text Mining - An Overview," *International Journal of Computer Science & Communication Networks*, vol.5, no.1, pp.10 (no. 7-16), Mar. 2015.

- [11] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, Aug. 2009.
- [12] C. Saini and V. Arora, "Information retrieval in web crawling: A survey," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2016, pp. 2635–2643.
- [13] T. D. Nguyen, A. T. Nguyen, H. D. Phan, and T. N. Nguyen, "Exploring API Embedding for API Usages and Applications," in *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, Buenos Aires, 2017, pp. 438–449.
- [14] Twitter, Inc., *Docs - Twitter Developers*. Internet: <https://developer.twitter.com/en/docs.html>, 2018 [Dec. 05, 2018].
- [15] M. Youness, E. Mohammed, and B. Jamaa, "Twitter Data Classification Using Big Data Technologies," in *Proceedings of the 2018 International Conference on Internet and e-Business - ICIEB '18*, Singapore, Singapore, 2018, pp. 124–129.
- [16] L. Branz and P. Brockmann, "Sentiment Analysis of Twitter Data: Towards Filtering, Analyzing and Interpreting Social Network Data," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems - DEBS '18*, Hamilton, New Zealand, 2018, pp. 238–241.
- [17] N. Pappas, G. Katsimpras, and E. Stamatatos, "Extracting informative textual parts from web pages containing user-generated content," in *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies - i-KNOW '12*, Graz, Austria, 2012, p. 1.
- [18] J. Pasternack and D. Roth, "Extracting article text from the web with maximum subsequence segmentation," in *Proceedings of the 18th international conference on World wide web - WWW '09*, Madrid, Spain, 2009, p. 971.
- [19] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, New York, New York, USA, 2010, p. 441.
- [20] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, "The Information Retrieval Process," in *Web Information Retrieval*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–26.

- [21] M. Lease, “Natural language processing for information retrieval: the time is ripe (again),” in *Proceedings of the ACM first Ph.D. workshop in CIKM on - PIKM '07*, Lisbon, Portugal, 2007, p. 1.
- [22] P. J. Sadalage and M. Fowler, *NoSQL distilled: a brief guide to the emerging world of polyglot persistence*, Upper Saddle River, NJ: Addison-Wesley, 2013.
- [23] G. Harrison, *Next Generation Databases*, Berkeley, CA: Apress, 2015.
- [24] R. Mazza, *Introduction to information visualization*. London: Springer, 2009.
- [25] B. Kitchenham, “Procedures for performing systematic reviews,” Keele, UK, Kelle University, Tech. Report. TR/SE-0401, ISSN:1353-7776, July 2004.
- [26] S. Keshav, “How To Read a Paper”, *ACM SIGCOMM Computer Communication Review*, vol. 37, no 3, July 2007.
- [27] C. Wohlin, “Guidelines for snowballing in systematic literature studies and a replication in software engineering,” in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14*, London, England, United Kingdom, 2014, pp. 1–10.
- [28] B. J. Oates, *Researching information systems and computing*. Los Angeles, CA, USA: Sage, 2006.
- [29] P. Le Hégarret, R. Whitmer and L. Wood, *Document Object Model (DOM)*, (2009) [Online]. Available: <https://www.w3.org/DOM/> . Accessed on: Dec. 9, 2018.
- [30] A. S. Vargas. *Web page segmentation, evaluation and applications*. Web. Université Pierre et Marie Curie - Paris VI, 2015.
- [31] P. M. Joshi and S. Liu, “Web document text and images extraction using DOM analysis and natural language processing,” in *Proceedings of the 9th ACM symposium on Document engineering - DocEng '09*, Munich, Germany, 2009, p. 218.
- [32] G. Laboreiro, L. Sarmiento, J. Teixeira, and E. Oliveira, “Tokenizing micro-blogging messages using a text classification approach,” in *Proceedings of the fourth workshop on Analytics for noisy unstructured text data - AND '10*, Toronto, ON, Canada, 2010, p. 81.

- [33] Q. You, “Sentiment and Emotion Analysis for Social Multimedia: Methodologies and Applications,” in *Proceedings of the 2016 ACM on Multimedia Conference - MM '16*, Amsterdam, The Netherlands, 2016, pp. 1445–1449.
- [34] S. Asur and B. A. Huberman, “Predicting the Future with Social Media,” in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, Washington, DC, USA, 2010, pp. 492–499.
- [35] V. Peña-Araya, M. Quezada, and B. Poblete, “Galean: Visualization of Geolocated News Events from Social Media,” in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '15*, Santiago, Chile, 2015, pp. 1041–1042.
- [36] F. S. Relucio and T. D. Palaoag, “Sentiment analysis on educational posts from social media,” in *Proceedings of the 9th International Conference on E-Education, E-Business, E-Management and E-Learning - IC4E '18*, San Diego, California, 2018, pp. 99–102.
- [37] S. Canuto, M. A. Gonçalves, and F. Benevenuto, “Exploiting New Sentiment-Based Meta-level Features for Effective Sentiment Analysis,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*, San Francisco, California, USA, 2016, pp. 53–62.
- [38] W. Wolny, “Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms,” *25th International Conference on Information Systems Development (ISD2016 POLAND)*, Katowice, Poland, 2016 pp.476-483.
- [39] Z. Jianqiang and G. Xiaolin, “Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” *IEEE Access*, vol. 5, pp. 2870–2879, 2017.
- [40] F. Liu, F. Weng, and X. Jiang, “A Broad-Coverage Normalization System for Social Media Language,” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 1035–1044, Jeju, Republic of Korea, 8-14 July 2012.
- [41] A. S. Nayak, A. P. Kanive, N. Chandavekar and R. Balasubramani , “Survey on Pre-Processing Techniques for Text Mining,” *International Journal Of Engineering And Computer Science*, Jun. 2016.
- [42] F. Provost and T. Fawcett, *Data Science for Business*. CA, USA: O'Reilly Media Inc, 2013.

- [43] V. V. Nhlabano and P. E. N. Lutu, *Impact of Text Pre-Processing on the Performance of Sentiment Analysis Models for Social Media Data*, Department of Computer Science University of Pretoria, Pretoria, South Africa, 2018.
- [44] M. Mhatre, D. Phondekar, P. Kadam, A. Chawathe, and K. Ghag, "Dimensionality reduction for sentiment analysis using pre-processing techniques," in *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, 2017, pp. 16–21.
- [45] F. L. dos Santos and M. Ladeira, "The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset," in *2014 Brazilian Conference on Intelligent Systems*, Sao Paulo, Brazil, 2014, pp. 50–54.
- [46] T. Young, D. Hazarika, S. Poria, and E. Cambria, *Recent Trends in Deep Learning Based Natural Language Processing*, arXiv:1708.02709 [cs], Aug. 2017.
- [47] L. Dey and S. K. M. Haque, "Studying the effects of noisy text on text mining applications," in *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data - AND '09*, Barcelona, Spain, 2009, p. 107.
- [48] A. Stavrianou, P. Andritsos, and N. Nicoloyannis, "Overview and semantic issues of text mining," *ACM SIGMOD Record*, vol. 36, no. 3, p. 23, Sep. 2007.
- [49] M. J. C. Samonte, H. I. B. Punzalan, R. J. P. G. Santiago, and P. J. L. Linchangco, "Emotion detection in blog posts using keyword spotting and semantic analysis," in *Proceedings of the 3rd International Conference on Communication and Information Processing - ICCIP '17*, Tokyo, Japan, 2017, pp. 6–13.
- [50] R. McArthur, "Uncovering deep user context from blogs," in *Proceedings of the second workshop on Analytics for noisy unstructured text data - AND '08*, Singapore, 2008, pp. 47–54.
- [51] E. Schubert, M. Weiler, and H.-P. Kriegel, "SigniTrend: scalable detection of emerging topics in textual streams by hashed significance thresholds," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, New York, New York, USA, 2014, pp. 871–880.
- [52] R. R. Pillai and S. M. Idicula, "Linear text segmentation using classification techniques," in *Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in India - A2CWic '10*, Coimbatore, India, 2010, pp. 1–4.
- [53] Y. K. Meena and D. Gopalani, "Domain Independent Framework for Automatic Text Summarization," *Procedia Computer Science*, vol. 48, pp. 722–727, 2015.

- [54] Q. Israel, H. Han, and I.-Y. Song, "Semantic analysis for focused multi-document summarization (fMDS) of text," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing - SAC '15*, Salamanca, Spain, 2015, pp. 339–344.
- [55] M. X. Ribeiro, A. J. M. Traina, and C. Traina, "A new algorithm for data discretization and feature selection," in *Proceedings of the 2008 ACM symposium on Applied computing - SAC '08*, Fortaleza, Ceara, Brazil, 2008, p. 953.
- [56] K. Gayathri and A. Marimuthu, "Text document pre-processing with the KNN for classification using the SVM," in *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, Coimbatore, Tamil Nadu, India, 2013, pp. 453–457.
- [57] R. GeethaRamani, M. N. Kumar, and L. Balasubramanian, "Identification of emotions in text articles through data pre-processing and data mining techniques," in *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, Ramanathapuram, India, 2016, pp. 611–615.
- [58] R. Sadoddin and O. Driollet, "Mining and Visualizing Associations of Concepts on a Large-Scale Unstructured Data," in *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, Oxford, United Kingdom, 2016, pp. 216–224.
- [59] C. Djellali, "A new conceptual model for dynamic text clustering Using unstructured text as a case," in *Proceedings of the 2014 International C* Conference on Computer Science & Software Engineering - C3S2E '14*, Montreal, QC, Canada, 2008, pp. 1–7.
- [60] A. J. Soto, R. Kiros, V. Kešelj, and E. Milios, "Exploratory Visual Analysis and Interactive Pattern Extraction from Semi-Structured Data," *ACM Transactions on Interactive Intelligent Systems*, vol. 5, no. 3, pp. 1–36, Sep. 2015.
- [61] J. D. Rennie, L. Shih, J. Teevan, and D. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, 2003.

- [62] A. Sharma and S. Dey, "A boosted SVM based sentiment analysis approach for online opinionated text," in *Proceedings of the 2013 Research in Adaptive and Convergent Systems on - RACS '13*, Montreal, Quebec, Canada, 2013, pp. 28–34.
- [63] E. Haddi, X. Liu, and Y. Shi, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Computer Science*, vol. 17, pp. 26–32, 2013.
- [64] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment Analysis by Capsules," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, Lyon, France, 2018, pp. 1165–1174.
- [65] S. Huang, Y. Yang, H. Li, and G. Sun, "Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis," in *2014 Asia-Pacific Services Computing Conference*, Fuzhou, Fu Jian, China, 2014, pp. 88–92.
- [66] E. Çano and M. Morisio, "A deep learning architecture for sentiment analysis," in *Proceedings of the International Conference on Geoinformatics and Data Analysis - ICGDA '18*, Prague, Czech Republic, 2018, pp. 122–126.
- [67] R. V. Borges, A. d'Avila Garcez, and L. C. Lamb, "Learning and Representing Temporal Knowledge in Recurrent Networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2409–2421, Dec. 2011.
- [68] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw, "A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 51–60, Jan. 2016.
- [69] M. Araújo, P. Gonçalves, M. Cha, and F. Benevenuto, "iFeel: a system that compares and combines sentiment analysis methods," in *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, Seoul, Korea, 2014, pp. 75–78.
- [70] EUROPAPARLAMENTETS OCH RÅDETS FÖRORDNING (EU) "2016/679 om skydd för fysiska personer med avseende på behandling av personuppgifter och om det fria flödet av sådana uppgifter och om upphävande av direktiv 95/46/ EG (allmän dataskyddsförordning)" (2016) [Online]. Available: <https://publications.europa.eu>, Accessed on: Jan. 31, 2019.

Bilaga A

Detaljerade sökresultat från Acm Digital Library

Tabell 5: Visar en detaljerad sammanställning av artiklar och områden från artiklar som hittades i databasen ACM.

Artiklar från ACM						
ID	Typ /År	Författare	Titel	Tekniker	Anteckningar	Område
1	Konf./ 2017	A. Li and Y. Chen	Pre-processing Analysis for Chinese Text Sentiment Analysis	Vectorization: TFIDF and word2vec. Classification methods: Naive Bayes, CNN, LSTM, SVM,	Undersöker effekten av 8st pre-processing tekniker som behandlar korta kinesiska texter, i syfte att utföra semantisk analys på texter där segmentering är problematiskt, vilket det är i kinesiska texter.	Word level semantic analysis.
2	Konf./ 2012	N. Pappas <i>m. fl.</i>	Extracting informative textual parts from web pages containing user-generated content	SD Algorithm, DOM tree, SD-tree (HTML DOM tree),	Ett tillvägagångssätt för att upptäcka och extrahera informativa textdelar på webbsidor. Upptäcker även domäntyp. Både visuell och textbaserad information undersöks.	Scraping, pre-processing
3	Konf./ 2008	Y. Liu <i>m. fl.</i>	Real-time data pre-processing technique for efficient feature extraction in large scale datasets	Commentz-Walter (CW), Wu-Manber (WM) and BSS som bygger på Aho-Corasick (AC)	Ett förslag på en ny algoritm som parallellt matchar strängar på online och offline textdata. Resultaten är bättre än hos de state-of-the-art algoritmer som undersöks.	Word, sentence, doc - level
4	Konf./ 2011	T.Gottron <i>m. fl.</i>	Insights into explicit semantic analysis	Probabilistic Model of Term Weights on Explicit Semantic Analysis (ESA)	Presenterar grunden till ESA från ett teoretiskt perspektiv och visar en probabilistisk modell för vikten på olika termer, vilket avslöjar hur ESA faktiskt fungerar.	Semantic analysis
5	Konf./ 2017	M.J.C. Samonte <i>m. fl.</i>	Emotion detection in blog posts using keyword spotting and semantic analysis	Semantic Role Labeling (by the Cognitive Computation Group of the University of Illinois), RapidMiner, tokenization, Keyword spotting, Semantic Analysis	Genom att utföra semantisk analys på de 6 grundkänslorna som finns, syftar detta arbete att identifiera signifikanta mönster i de känslor som upptäcks. Olika algoritmer appliceras för att uppnå detta.	Pre-processing, Sentence level sentiment analysis

6	Konf./ 2008	K. Inui <i>m. fl.</i>	Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents	SVM-HMM algorithm to train an Event-time model, SVM-Multiclass package to train a Modality model, GRMM toolkit for conditional likelihood maximization includes Factorial CRFs (Conditional Random Fields), Tokenization, dependency parsing,	En ny applikation som ämnar att gräva fram erfarenheter utifrån användargenererat innehåll (User Generated Content på engelska) som innehåller information som; ämnen, erfarenheter, händelseuttryck, typ av händelse och pekar på källor. Skiljer sig mot sentimentanalys på så sätt att denna applikationen gräver djupare.	Some pre-processing, storage, Insights
7	Konf./ 2008	M.X. Ribeiro <i>m. fl.</i>	A new algorithm for data discretization and feature selection	C4.5 algorithm, data were discretized using three different methods: equal-sized (the continuous data was discretized in 10 intervals), Chi2 and Omega, Three different feature selection methods were applied to the datasets: Relief, Chi2 and Omega.	Ny algoritm (Omega) som används för att diskretisera data och välja ut features i syfte att förbehandla data som ska användas för att träna upp inlärningsalgoritmer.	Pre-processing
8	Konf./ 2009	H. Moeinzadeh <i>m. fl.</i>	Evolutionary-class independent LDA as a pre-process for improving classification	Class-independent LDA with Genetic Algorithm and Particle Swarm Optimization	Analyserar effekten av LDA och "genetic algorithm+particle swarm optimization i syfte att förbehandla datan innan den klassificeras.	Pre-processing
9	Konf./ 2015	Q. Israel <i>m. fl.</i>	Semantic analysis for focused multi-document summarization (fMDS) of text	General Architecture for Text Engineering (GATE) -> Tokenization, POS-tagging, Stanford Parser, NER. Multi-language Predicate Arguments Extractor (MultiPAX)	Ett försök att utmana state-of-the-art inom fokuserad summering av multitext-dokument (i.e. topic, query, question, category).	Pre-processing
10	Konf./ 2017	X. Dai <i>m. fl.</i>	A Framework for Automated Rating of Online Reviews Against the Underlying Topics	Tokenization, Stemming, Topic modelling: Latent Dirichlet Allocation (LDA), an unsupervised probabilistic method.Sentiment analysis: weighted word embeddings	Automatisk bedömning av en 5-stjärnig skala. Använder sig av förbehandlingstekniker, modellering av ämnen (LDA), klassificering av texter (platsidentifiering mm.) och sentimentanalys.	Pre-processing topic modelling, text classification, sentiment analysis (fine grained).

				method, Negation Handling, NLTK toolkit for POS-tagging, Google News Dataset (Word2vec Model), NLTK- toolkit and Gensim library.		
11	Konf./ 2010	R. R.Pillai and S. Mary Idicula	Linear text segmentation using classification techniques	Bayesian Network (Maximum A Posteriori (MAP)) and Decision trees, WEKA, C4.5 (J48) classifier, Topic Detection and Tracking (TDT) 2,	Domänoberoende metod för segmentering av ostrukturerad text.	Pre-processing
12	Konf./ 2009	J. Pasternack and D. Roth	Extracting article text from the web with maximum subsequence segmentation	Maximum subsequence segmentation; global optimization over token-level local classifiers. Tokenization. Porter Stemming, Naive Bayes, trigrams,	Maximum Subsequence Segmentation, en metod för att globalt optimera klassificeringsalgoritmer som ligger på lokal token-level i syfte att applicera den på nyhetssajter. Innehåller även en algoritm för att extrahera text från HTML-dokument.	Scraping/Text extraction from web pages, information retrieval. pre-processing
13	Konf./ 2009	P. M. Joshi and S. Liu	Web document text and images extraction using DOM analysis and natural language processing	Combination of HTML DOM analysis and Natural Language Processing (NLP does Named Entity Recognition (NER) with GATE).	Ett generiskt tillvägagångssätt som kombinerar HTML DOM-analys och NLP, i syfte att automatisera extraheringen av text och dess associerade bilder på olika webbsidor.	Scraping/pre-processing, Noise removal.
14	Konf./ 2010	C. Kohlschütter <i>m. fl.</i>	Boilerplate detection using shallow text features	1-fold cross validation, decision trees, linear support vector machines. Cross-Domain collection through CleanEval. Weka. measure classification accuracy by Precision, Recall, F1-Score, False Positive Rate and ROC Area under Curve (AuC). ZerO classifier.	Filtrering av Boilerplate (standardkod) från webbsidor med hjälp av en sk. billig metod som använder ytliga features. Likvärdig med state-of-the-art gällande reducerade "kostnader".	Information extraction. Noise removal.
15	Konf./ 2009	N. O'Hare <i>m. fl.</i>	Topic-dependent sentiment analysis of financial blogs	DiffPost algorithm to remove noise, N-word extraction, N-	Sentimentanalys av blogginslag rörande den finansiella sektorn. Riktat	Crawling, Sentiment analysis, topic

				sentence extraction, N-paragraph extraction. Machine learning techniques: multinomial naïve Bayes (MNB) classifier, Support vector machine (SVM) - Binary classification and 3-Point classification. WEKA was used in all experiments.	in sig på bloggare som bloggar mot företag och deras aktier. Tar även hänsyn till eventuella ämnesskiftningar.	shift handling.
16	Konf./ 2017	M. F. Hanafi <i>m. fl.</i>	SEER: Auto-Generating Information Extraction Rules from User-Specified Examples	Visual Annotation Query Language (VAQL),	En modell som fokuserar på att kombinera ML-tekniker och regelbaserade tillvägagångssätt i ett försök att underlätta den tidskrävande och komplicerade process som ingår i IR.	Pre-pocessing Information extraction. Insights
17	Konf./ 2008	C. Djellali	A new conceptual model for dynamic text clustering Using unstructured text as a case	Variable selection techniques and Fuzzy Adaptive Resonance Theory to increase productivity of knowledge extraction.	Ett schema för att förstärka den prediktiva naturen inom klustring.	Pre-processing, Data analysis. Clustering.
18	Konf./ 2015	A. J. Soto <i>m. fl.</i>	Exploratory Visual Analysis and Interactive Pattern Extraction from Semi-Structured Data	N/A	Ett textanalysverktyg för semi-strukturerade dokument. Siktar på att stötta användaren i dess väg att undersöka och hitta användbara mönster i semi-strukturerade dokument. På ett interaktivt sätt presenteras datan för användarna genom att visas i exempelvis kluster där användarna kan klicka sig till olika typer av information.	Feature learning and fast clustering. Visualization, Insights
19	Konf./ 2008	V. Jijkoun <i>m. fl.</i>	Named entity normalization in user generated content	NEN, NER, NE, trimming, joining and ngramming NEs, approximate name matching, identification of missing references and name disambiguation.	Föreslår 5 förbättringar av en Named Entity Normalization-algoritm i syfte att uppnå ett språkoberoende NEN-system som landar på 90% noggrannhet för det engelska språket.	Named entity recognition. Preprocessing.
20	Konf./	R. McArthur	Uncovering deep	Tokenization,	En algoritm som analyserar	Sense of self

	2008		user context from blogs	Hyperspace analogue to language (HAL), unstructured and noisy data (UAND) from TREC Bog track,	"sense-of-self" i inlägg på webben. Använder HAL.	analysis of user context, semantic analysis, Insights
21	Konf./ 2010	K. Murakami <i>m. fl.</i>	Statement map: reducing web information credibility noise through opinion classification	Validating credibility of statements. Biased noise detection from statements through viewpoint mining from the web. STATEMENT MAP, Information Retrieval (IR) and NLP. Semantic relations. Passage retrieval, linguistic analysis, structural alignment, semantic relation classification.	Målet är att identifiera trovärdigheten av källor på webben genom att kombinera IR och NLP-teknologier och fokusera på att organisera utlåtanden som hämtats från "Web by viewpoints". Identifierar semantisk klassificering innehållande 4 semantiska relationer i form av [Agreement], [Conflict], [Confinement] och [Evidence].	Credibility analysis. All
22	Konf./ 2015	S. Mishra <i>m. fl.</i>	Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization	Sentiment Analysis and Incremental Learning (SAIL), stochastic gradient decent (SGD),	Ett sentimentanalysverktyg som är GUI-baserat och innehåller förmågor som reducerar bl. a. kostnaden för algoritmen	All, Sentiment analysis tool.
23	Konf./ 2009	L. V. Subramaniam <i>m. fl.</i>	A survey of types of text noise and techniques to handle noisy text	SMS, Noise removal, Levenshtein distance, Word Error Rate (WER) and Sentence Error Rate (SER), Optical Character Recognition (OCR), Automatic Speech Recognition (ASR), Acoustic Measure, Language Model Measures, N-Best Measure, Linguistic Features, Statistical Machine Translation, BLEU Metric, NIST Metric, METEOR, Parallel Corpora, Information Retrieval (IR), Jaguar (search engine), Out of Vocabulary Words (E-Speak, new words, foreign words), Conditional Random Field (CRF), Query Suggestion, Language	En survey om noise i data och tekniker för att hantera detta.	Pre-processing

				Model (LM), Expectation Maximization, Text Classification, Summarization		
24	Konf./2009	L. Dey and S.K M. Haque	Studying the effects of noisy text on text mining applications	Tokenization, WordNet, POS-tagging, dependency analysis, Suggester (for spellchecking), keywords, NLP	En studie om effekterna som brusig textdata ger. Presenterar en analys om hur brus i form av inkorrekt engelska påverkar performance av NLP-verktyg och text mining-applikationer. Fokuserar på opinion mining.	pre-processing, Sentiment analysis.
25	Journ./2008	L. Dey and S.K M. Haque	Opinion mining from noisy text data	Tokenization, WordNet, POS-tagging, dependency analysis, Suggester (for spellchecking), keywords, NLP	Siktar på att extrahera åsikter hos kunder som bloggar och/eller ger feedback på en produkt. Fokuserar på flera nivåer gällande granularitet.	Alla
26	Konf./2014	M. Araújo m. fl.	iFeel: a system that compares and combines sentiment analysis methods	Detect sentiments in any form of text including unstructured social media data, SentiWordNet, Emoticons, PANAS-t, SASA, Happiness index, SenticNet,	Ett webbaserat verktyg för att mäta nivån av positiv och negativ påverkan, baserad på 8 olika tekniker inom sentimentanalys (PANAS-t, Emoticons, SentiWordNet, Happiness Index, SentiStrength, SASA, SenticNet) och jämför resultatet med deras kombinerade metod.	Alla, Sentiment analysis tool.
27	Konf./2013	A. Sharma and S. Dey	A boosted SVM based sentiment analysis approach for online opinionated text	Boosted SVM, vector space model (VSM), tokenization, stop word removal, stemming, AdaBoost,	Ett tillvägagångssätt för att klassificera sentiment med en boosted Support Vector Machine. Kombinerar integrerade samplingstekniker med en uppsjö av SVMs för att förbättra performance på att förutsäga sentiment. Resultaten gällande sentimentbaserad klassificering visar att flera SVMs tillsammans med bagging eller boosting ger bättre resultat än att endast använda en enstaka SVM.	Sentiment analysis.
28	Konf./2018	E. Çano and M. Morisio	A deep learning architecture for sentiment analysis	Sentiment analysis of long text documents, Bag-of-words, n-grams, convolutional	Presenterar ett tillvägagångssätt för att använda neurala nätverk vid sentimentanalys i stora	pre-processing, Sentiment analysis.

				neural network,	textdokument.	
29	Konf./ 2018	Y. Wang <i>m. fl.</i>	Sentiment Analysis by Capsules	Recurrent Neural Network (RNN), Glove, two-layer LSTM, two-layer GRU.	Presenterar ett RNN för sentimentanalys som uppnår state-of-the-art- performance.	Alla, Sentiment analysis.
30	Konf./ 2014	E. Schubert <i>m. fl.</i>	SigniTrend: scalable detection of emerging topics in textual streams by hashed significance thresholds	ELKI, hierarchical clustering with Ward linkage, language- specific stemming using the Xapian, stop word removing, Apache Lucene as backing index, Twitter's public streaming API at the "Spritzer", simple near-duplicate detector to remove obvious spam,	Trenddetektering för uppåtgående ämnen, inte bara för globalt uppåtgående ämnen.	Crawling, pre- processing, Trend detection.

Bilaga B

“Snowballade” referenser från ACM-artiklar

Tabell 6: Visar en detaljerad sammanställning av artiklar och områden från “snowballade” artiklar.

“Snowballade” referenser från ACM-artiklar						
ID	Typ /År	Författare	Titel	Tekniker	Anteckningar	Område
31	Konf./ 2005	Joel Martin <i>m. fl.</i>	Word alignment for languages with scarce resources	Sentence alignment, Stemming	Målet med ett sånt här system är att få koll på vilket ord (token) i en mening på ett språk motsvarar samma ord på ett annat språk. Fokuserar på språk som är knapphändiga, så som Engelska-Inuktitut, Engelsk-Hindi och Rumänsk-Engelska.	Cross-lingual, pre-processing
32	Journ./ 2016	Anwaar Ali <i>m. fl.</i>	Big data for development: applications and techniques	ML, DL, NoSQL databases for predictive analysis,	Teknologier och appliceringsområden för Big Data, i utvecklingssyfte. Tar upp saker som strukturerade, semistrukturerade och ostrukturerade inlärningstekniker, NLP, och mycket annat användbart.	All
33	Konf./ 2002	Mathias Creutz and Krista Lagus	Unsupervised Discovery of Morphemes	Minimum Description Length (MDL), Maximum Likelihood (ML),	Modellen är speciellt användbar för språk med rik morfologi, såsom finska. Två metoder applicerad där en är baserad på MDL och en är baserad på ML (Maximum Likelihood). Kvaliteten av segmenteringen mäts genom att jämföra med en existerande modell för analys. Experiment gjorda på Engelska och Finska korpus.	Pre-processing

Bilaga C

Detaljerade sökresultat från IEEE Transactions

Tabell 7: Visar en detaljerad sammanställning av artiklar och områden från artiklar som hittades i databasen IEEE

Artiklar från IEEE						
ID	Typ/ År	Författare	Titel	Tekniker	Anteckningar	Område
34	Konf./ 2014	Siqi Huang <i>m. fl.</i>	Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis	LDA, TF-IDF, Conduct clustering treatment, using K-means algorithm,	Upptäcker ämnen i microbloggar baserat på klustring- och topic model analysis med hjälp av förbehandling, LDA och topic modelling..	Content generation/pre-processing
35	Konf./ 2014	Fernando Leandro dos Santos and Marcelo Ladeira	The Role of Text Pre-processing in Opinion Mining on a Social Media Language Dataset	Hadoop, Mahout, terms standardization, Spell checking, Stemming with PTStemmer, Stop word removal, TF-IDF, Neural Network, Bayesian network, Naive Bayes and SVM Were tested, 5-fold cross validation,	Extraherar data från webben som består av reviews innehållande slangord, förkortningar och stavfel. Tester visar att förbehandling av data är av obetydande karaktär för den specifika uppgiften att recensera mobilapplikationer.	Pre-processing
36	Konf./ 2013	K. Gayathri and A. Marimuthu	Text document pre-processing with the KNN for classification using the SVM	Text Classification, Feature Selection, K-Nearest Neighbor, Support Vector Machine, IDF &TF	Studerar fördelar och nackdelar med KNN-klassificering och SVM-klassificering. Resultaten visar att KNN kan vara mindre noggranna än SVM.	Pre-processing
37	Konf./ 2017	Belainine Billal <i>m. fl.</i>	Semi-supervised learning and social media text analysis towards multi-labeling categorization	Comparison of Supervised methods : BR, CC, EnsembleML(CC), PCC(BCC), BaggingML(CC). Semi-supervised methods with Expectation-maximization algorithm (EM) and Classification maximization algorithm (CM) algorithms: BR, CC, EnsembleML(CC), PCC(BCC), DBPNN, DeepML, BaggingML (CC) NOTE: CC Stands for Classification Chain	En metod som kombinerar semi-övervakade metoder med graf-metoder för att extrahera ämnen på sociala nätverk med hjälp av en "multi-label classifier". Jämfört med baseline så ökade precisionen på klassificeringen.	Pre-processing
38	Journ./ 2018	V.V. Nhlabano and P.E.N. Lutu	Impact of Text Pre-Processing on the Performance of Sentiment Analysis	Stop word removal, stemming, Naive Bayes Classifier, Binary classification, Python	Presenterar resultaten av en studie som är utförd för att se vilken inverkan förbehandling av data (stemming, stop word	Pre-processing

			Models for Social Media Data	2.7, Natural Language ToolKit 3.2.5 (NTKL)	removal, feature selection) från sociala medier har. Resultaten visar att metoderna för att förbehandla data ökar precisionen hos metoder för sentimentanalys.	
39	Konf./ 2016	R. Geetha Ramani <i>m. fl.</i>	Identification of emotions in text articles through data pre-processing and data mining techniques	Pre processing phase: Conversion chinese-english, HTML tag removal, stop word removal, Stemming. Data Mining Phase: Feature vector formulation, Correlation based Feature selection, Ensemble Classification, Performance evaluation (10-fold cross validation). Prediction of sentiment by Functional Trees with Bagging.	En metodologi för förbehandling och mining av den behandlade datan. Diverse klassificeringsalgoritmer tillämpades för att klassificera positiva, negativa, mixade och neutrala sentiment. Resultaten visar att Funktionella Träd tillsammans med Bagging gav bästa möjliga resultat.	Scraping, pre-processing, content generation
40	Konf./ 2017	Mayuri Mhatre <i>m. fl.</i>	Dimensionality reduction for sentiment analysis using pre-processing techniques	Tokenization, Handling Expressive Lengthening, Emoticons Handling, HTML Tags Removal, Slangs Handling, Punctuations Handling, Stopwords Removal, Stemming and Lemmatization. Random Forest classifier for classification. Result was evaluated using 10 fold cross validation	Effekterna av förbehandling av data med metoderna nämnda i kolumnen för tekniker visar att data som är förbehandlad är lättare att undersöka än data som inte är förbehandlad.	Pre-processing
41	Journ. /2017	Zhao Jiangqiang and Gui Xiaolin	Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis	Replacing negative mentions, Removing URL links, Reverting words, Removing numbers, Removing stop words, Expanding acronyms , LL, NB, SVM, RF, N-grams,	Diskuterar effekterna av olika förbehandlingsmetoder för att klassificera sentiment. Testade 6 metoder för förbehandling på två typer av klassificeringsuppgifter och resultaten visade att Twitters klassificerare för sentiment förbättrades vid förbehandling av data men det skedde ingen förändring vid borttagning av URL-länkar, stoppord, mm. Naive Bayes och Random Forest classifiers verkar mer känsliga än Logistic regression och SVM när	Pre-processing

					förbehandlingsmetoder applicerades.	
42	Konf./ 2011	Chunye Wang <i>m. fl.</i>	Knowledge Extraction and Reuse within "Smart" Service Centers	N/A	En initial version av ett textanalyssystem som utvecklas och används av Cisco där målet är att optimera och effektivisera produktiviteten och effektiviteten hos servicecentret. Diskuterar behovet av att skapa fler "smarta" service centers och presenterar de forskningsgap som bekräftar detta.	Scraping, Pre-processing, Content generation, Insights
43	Journ. /2016	Hua Guo <i>m. fl.</i>	A Case Study Using Visualization Interaction Logs and Insight Metrics to Understand How Analysts Arrive at Insights	N/A	En utvärdering av ett system för visuell analys som är utvecklat för att analysera samlingar av dokument, strukturerade som spatiotemporala nätverk. För att skapa en förståelse för hur systemet används två olika tillvägagångssätt som innefattar analys av användning samt analys av interaktionsloggar. Målet är att påtala den utmaning som ligger bakom i att förstå hur design på en applikation påverkar generering av nya insikter.	Insights
44	Konf./ 2008	Ling Jiang <i>m. fl.</i>	Knowledge Indexing of Chinese Text Based Knowledge Element	Word segmentation, parsing, keyword extraction, guide information (KE name and KE information), initial knowledge elements, formalize knowledge elements	Research paper om Knowledge Indexing. Berättar om varför KI är bra och varför det borde appliceras vid textbaserad KE.	Scraping, Pre-processing, Content generation, Insights
45	Konf./ 2011	Vaishali Bhujade and N.J. Janwe	Knowledge Discovery in Text Mining Technique Using Association Rules Extraction	Transformation, filtration, stemming, stop word removal, feature selection and indexing by using TF-IDF,	Text mining-tekniker för att automatiskt extrahera associeringsregler från samlingar av textdokument. Tekniken kallas Extracting Association Rules from Text (EART) och fokuserar på orden och dess statistiska distribution i dokumentet. Består av 3 faser: förbehandling av text, association rule mining och visualisering.	Scraping, Pre-processing, Content generation, Insights
46	Journ. /2011	Rafael V. Borges <i>m. fl.</i>	Learning and Representing	RNN	En neural beräkningsmodell som genom ett neuralt	Insights

			Temporal Knowledge in Recurrent Networks		<p>nätverk kan lära sig och representera kunskap. Resultaten indikerar att verifiering och inläring kan integreras inom paradigmen för neurala beräkningar, vilket bidrar till utvecklingen av kunskapsbaserade system som kan ge prognoser samt tillhandahålla resultat som är tolkningsbara till den grad att de kan hjälpa forskare och ingenjörer att förbättra deras specifikationer.</p>	
--	--	--	--	--	--	--

Bilaga D

Detaljerade sökresultat från Google Scholar

Tabell 8: Visar en filtrerad sammanställning av artiklar och områden från artiklar som hittades i databasen Google Scholar.

Kompletterande artiklar från Google Scholar						
ID	Typ /År	Författare	Titel	Tekniker	Anteckningar	Område
47	Bok/ 2014	Nada Elgendy and Ahmed Elragal	Big Data Analytics: A Literature Review Paper	N/A	Analyserar olika verktyg och tekniker som kan appliceras på Big Data och visar vilka möjligheter som öppnas upp vid applicering av big data-analys inom diverse domäner.	Pre-processing, storage, analytics, Insights
48	Konf./ 2016	Chandni Saini and Vinay Arora	Information retrieval in web crawling: A survey	Automated, focused, distributed, incremental and hidden web crawlers	En review av strategier för att skrapa data. Tar upp olika sätt att skrapa på och hur dessa sätten kan appliceras baserat på problemställningen.	Crawling
49	Konf./ 2016	Reza Sadoddin and Osvaldo Driollet	Mining and Visualizing Associations of Concepts on a Large-Scale Unstructured Data	Point-wise Mutual Information (PMI), Z.score, TF-IDF, DisGeNET database, Neo4J	Ett ramverk för att beräkna och visualisera associationer mellan koncept genom att använda generiska mått för associeringar och den publika kunskapen som är tillgänglig via exempelvis Wikipedia.	Pre-processing, storage & visualization, Insights
50	Journ. /2015	Dr. S. Vijayarani and J. Ilamathi	Preprocessing Techniques for Text Mining - An Overview	Stemming, Stop words elimination, TF/IDF algorithms, Word Net, Word Disambiguation.	Tar upp olika förbehandlingstekniker som används vid text mining.	Pre-processing
51	Journ. /2017	Tom Young <i>m. fl.</i>	Recent Trends in Deep Learning Based Natural Language Processing	Natural Language Processing, Deep Learning, Word2Vec, Attention, Recurrent Neural Networks, Convolutional Neural Networks, LSTM, Sentiment Analysis, Question Answering, Dialogue Systems, Parsing, Named-Entity Recognition, POS Tagging, Semantic Role Labeling	Tar upp Deep Learning-modeller och metoder och ger en grundlig genomgång av dessa i ett försök att ge en förståelse över hur deep learning inom NLP sett ut förr, ser ut nu och kommer att se ut i framtiden.	All
52	Journ. /2016	Arjun Srinivas Nayak <i>m. fl.</i>	Survey on Pre-Processing Techniques for Text Mining	Tokenization, Stop-word Removal, Stemming, Porter, Krovetz.	Utvärderar algoritmerna Porter's och Kovertz i syfte att förbehandla data.	Pre-processing text mining

53	Journ. /2003	Jason D. Rennie <i>m. fl.</i>	Tackling the Poor Assumptions of Naive Bayes Text Classifiers	Weight-normalized Complement Naive Bayes (WCNB), Naive Bayes, SVM. TF-IDF	Föreslår lösningar för att tacka problem gällande Naive Bayes klassificerare. Tar upp både systemiska problem och problem som uppstår när texterna inte är rätt skrivna eller formaterade. resultaten visar att genom att implementera lösningarna så ligger Naive Bayes i klass med state-of-the-art klassificerare såsom SVM.	Pre-processing
54	Journ. /2009	Vishal Gupta and Gurpreet S. Lehal	A Survey of Text Mining Techniques and Applications	Transformation, filtration, stemming, stop word removal, feature selection and indexing by using TF-IDF,	En studie om text mining och dess tekniker.	All, including text mining
55	Journ. /2017	Shiliang Sun, Chen Luo and Junyu Chen	A review of natural language processing techniques for opinion mining systems	Opinion mining, Sentiment analysis, Natural language processing Deep learning, Machine learning	En review om NLP och dess tekniker som appliceras på opinion mining. Presenterar tekniker för förbehandling, olika nivåer inom detta och slutligen diskuteras utmaningar och problem inom området.	Pre-processing, Sentiment Analysis
56	Journ. /2012	Malcolm Clark <i>m. fl.</i>	Automatically structuring domain knowledge from text: An overview of current research	Supervised, unsupervised and weakly supervised learning methods	En översikt om automatiska metoder för att bygga strukturer för domänkunskap av textsamlingar. Ger en kort översikt gällande tekniker och tillvägagångssätt inom NLP, IR, och SW.	Övergripande, Insights
57	Journ. /2015	Yogesh Kumar Meenaa and Dinesh Gopalanib	Domain Independent Framework for Automatic Text Summarization	Segmentation, Stemming, Stop word removal, NER, Parsing, Content selection	Ett domänberoende ramverk för automatisk summering av text. Kategoriserar källtexten och sedan applicerar regler, metoder, vikter. Kan appliceras på både djupgående och övergripande summeringar.	Pre-processing
58	Journ. /2013	Emma Haddi <i>m. fl.</i>	The Role of Text Pre-processing in Sentiment Analysis	SVM, Unigrams, F-measure, FF, TD-IDF, FP	Utforskar rollen av förbehandling inom områden för sentimentanalys och rapporterar resultat från utförda experiment. Resultaten visar att med förbehandling med hjälp av SVM och rätt feature selection kan man öka resultatens noggrannhet..	Pre-processing, Sentiment Analysis
59	Journ. /1993	Mehdi Sagheb-Tehrani	The Technology of Expert Systems: Some Social Impacts	Decision making in ill structured specialized areas of narrow domain	Artikeln pekar ut att det finns ett behov av ett bredare perspektiv när man	All

					implementerar expertsystem gällande systemets påverkan på en organisation. Ger även en introduktion av teknologierna som finns i ett expertsystem och områden för vidare forskning.	
60	Journ./2015	Andrés Sanoja Vargas	Web page segmentation, evaluation and applications	Web page segmentation, Web applications, Evaluation, Web page analysis	Ger en djupare inblick i hur web page segmentation fungerar och hur det kan användas.	Crawling/scraping
61	Journ/2007	A. Stavrianou <i>m. fl.</i>	Overview and Semantic Issues of Text Mining	Stoplists, stemmer, tokenization, word sense disambiguation, noisy data, tagging, collocations, grammar/syntax, text representation, automated learning	Diskuterar olika problem inom text mining-området och berättar vilket tankesätt man bör ha vid implementering av dessa.	pre-processing