Kieliteknologia analytiikan tukena sotilas- ja viranomaistyössä

Viljami Venekoski ja Jouko Vankka

Abstract

In the digital age, many government agencies need to gather, analyze and make informed decisions based on constant streams of new digital data. To this aid, developments in the field of language technology have presented organizations with tools to efficiently analyze very large textual data sets. The contemporary methods allow organization to dynamically retrieve relevant information from sources such as social media or professional communication channels, enchancing overall situation awareness and decision making procedures. These tools enable a more informed approach to analytics and intelligence, providing possibilities to develop new procedures in information-reliant organizations such as the military and government infrastructuce management. This article serves as an introduction to some of the most common methods used in contemporary language technologies and particularly their application in relation to governmental and military contexts. Furthermore, potential applications of language technologies in intelligence and analytics domains are discussed.

Johdanto

Valtaosa nykypäivän kommunikaatiosta organisaatioiden sisällä sekä toimijoiden välillä tapahtuu erilaisia viestintäteknologioita hyödyntäen. Vaikka esimerkiksi sotilasympäristössä voidaan käyttää määrämuotoista viestiprotokollaa, organisaatioiden tavallisimpia viestiväyliä ovat verrattain informaalit pikaviestimet ja sähköposti. Lisäksi sosiaalisen median kommunikaatioväylistä on tullut yhä kiinteämpi osa paitsi organisaatioiden julkisuusviestintää niin myös yksi tavallisimmista kanssakäymismuodoista yksityishenkilöiden välillä.

Elektronisissa viestiväylissä käytävän kommunikaation primaarisena tarkoituksena on välittää informaatiota toimijalta toiselle. Viestinnän digitaalisuus ja

tästä johtuva automaattinen tallentuminen tietojärjestelmiin tarjoaa kuitenkin mahdollisuuden analysoida organisaatioiden toimintaa laajemmin. Automatisoidun sisältöanalyysin avulla voidaan esimerkiksi huomata, että viranomaisten keskustelussa on edellisen tunnin aikana keskusteltu poikkeuksellisen paljon tietystä turvallisuusuhasta, mikä voidaan välittää tietona johtamisjärjestelmälle tai korkeamman tason päättäjille (ks. Puuska ym. 2016). Vastaavasti tekstipohjaisen sotilasviestinnän automaattisesta analysoimisesta saatavan informaation on tulkittu tukevan johtamisjärjestelmiä (Medina 2008). Toisaalta kartoittamalla julkista viestintää mm. sosiaalisessa mediassa voidaan saada selville, mitkä aiheet yksilöitä puhuttavat ja esimerkiksi millaisia tunteita näihin liitetään tai millainen on niin sanottu kansalaisten yleinen mielipide (Liu 2012; Lazer ym. 2009). Vastaavasti tiedustelu nojaa pitkälti ulkoisten toimijoiden viestinnän analyysiin, mikä voi tapahtua myös julkisten viestiväylien välityksellä (Omand ym. 2012). Avoimen datan tiedustelu (*open source intelligence, OSINT*) ja tiedustelu sosiaalisessa mediassa (*social media intelligence, SOCMINT*) ovatkin tärkeä osa modernia tiedustelutoimintaa; arviolta 80–95% kaikesta tiedustelutiedosta saadaan tai olisi saatavissa julkisilta kanavilta (Pallaris 2008). Some-palveluiden käyttäjät raportoivat matalan tason havaintoja ympäristöstään sosiaaliseen mediaan, eli toisin sanoen käyttäjät ovat joukkoistaneet (engl. *crowdsourced*) osan tiedustelutoiminnasta itselleen lataamalla salienttia informaatiota toimintaympäristöstään julkisille viestiväylille (Stottlemyre 2015). Täten sosiaalista mediaa analysoimalla myös toimijat, jotka suorittavat operatiivista tiedustelua, voivat saada laajemman kirjon matalan tason havaintoja käyttöönsä hyödyntämällä automaattista julkisten viestiväylien analyysiä.

Laajamittainen tekstidatan analytiikka on tullut mahdolliseksi modernien kieliteknologioiden myötä. Ne mahdollistavat viestiliikenteen ja muun kielellisen informaation automaattisen analysoimisen riippumatta siitä, missä muodossa viestintä tapahtuu (Hurwitz ym. 2015). Toisin sanoen samoihin periaatteisiin pohjautuvia menetelmiä voidaan käyttää niin formaalin viranomaisviestinnän valvonnassa kuin sosiaalisen median kartoittamiseen. Viime vuosina on tapahtunut erityistä kehitystä kieliteknologian menetelmissä, jotka pystyvät mallintamaan viestinnän merkityssisältöjä eli *semantiikkaa*. Kun viestit pystytään esimerkiksi luokittelemaan sisällöltään irrelevanteiksi tai potentiaalisesti kiinnostaviksi, manuaalisen tiedustelutyön määrää voidaan vähentää merkittävästi. Merkitysten mallintaminen mahdollistaa myös nk. tuntemattomien tuntemattomien löytämisen – kun tiedetään, että tekijä nimeltään *A* on kiinnostava, voidaan tietokoneavusteisesti selvittää, mitkä viestinnässä ilmenevät toistaiseksi tuntemattomat tekijät ovat merkitykseltään samankaltaisimpia *A*:n kanssa. Erityisesti tiedustelukontekstissa samankaltaisuuksien estimointi suhteessa tunnettuihin uhkiin tai muihin käsitteisiin voi tukea informaation

sisällyttämistä osaksi laajempaa toiminta-avaruutta (Biermann ym. 2004). Toisaalta, automaattiset menetelmät voivat aina erehtyä, joten tuleekin harkita onko virhemarginaali hyväksyttävä ja automatisaatio luotettavampi kuin rajallisen kognitiivisen kapasiteetin omaava ihminen (ks. esim. Vachon ym. 2011).

Tässä artikkelissa esitellään ensiksi joitakin kieliteknologian menetelmiä lyhyesti, mitä seuraa lyhyt esittely menetelmien mahdollisista sovellutusalueista erityisesti suhteessa viranomaistoimintaan ja sotilaalliseen kontekstiin. Artikkelin tarkoituksena on tutustuttaa lukija kieliteknologian käsitteistöön ja sovellusmahdollisuuksiin sekä tarjota aiheesta kiinnostuneelle lukijalle lähtökohta alan kirjallisuuteen käytettyjen lähteiden kautta.

Kieliteknologian menetelmiä

Tässä luvussa esitellään lyhyesti modernin kieliteknologian sovellusten kannalta olennaisimpia menetelmiä sekä käsitteistöä. Esille nostetaan erityisesti analytiikalle ja tiedustelutoiminnalle olennaiset tekstien numeeriset esitykset eli vektorirepresentaatiot. Lisäksi esitetään suomen kielen ja pikaviestimien erityishaasteita automaattiselle kielianalyysille.

Tekstin vektorirepresentaatiot

Luonnollisissa tilanteissa tuotettua tekstiä voidaan käyttää raakadatana kieliteknologisissa analyyseissa. On olemassa laaja kirjo erilaisia menetelmiä ja analyyseja, jotka soveltuvat tekstidatalle. Kenties tavallisimmin menetelmät perustuvat samankaltaisuuksien etsimiseen joko sanojen tai erityisesti *dokumenttien* välillä. Dokumentilla tarkoitetaan jotakin rajattua tekstikokonaisuutta, kuten kirjaa tai yhtä chat-viestiä. Jotta koneelliset analyysit ovat mahdollisia, tekstidata on yleensä muutettava merkkijonoista numeeriseen muotoon. Tällöin dokumentista muodostetaan numeerinen *representaatio*, joka puolestaan on tavallisimmin *vektori* eli *vektorirepresentaatio*. Tekstin esittäminen vektorimuodossa mahdollistaa erilaisten laskennallisten menetelmien soveltamisen tekstiaineistoille, mutta toisaalta ei ole itsestäänselvää, mikä on paras tapa vektorisoida kielellistä dataa. Tämän vuoksi huomattava osa edellisvuosien kieliteknologian tutkimuksesta on koskenut eri vektroisointimenetelmiä sekä tapoja arvioida kuinka hyvin vektorirepresentaatiot vangitsevat luonnollisen kielen ominaisuuksia kuten merkitystä (ks. esim. Turney & Pantel 2010; Venekoski ym. 2016; Baroni ym. 2014; Hill ym., 2015).

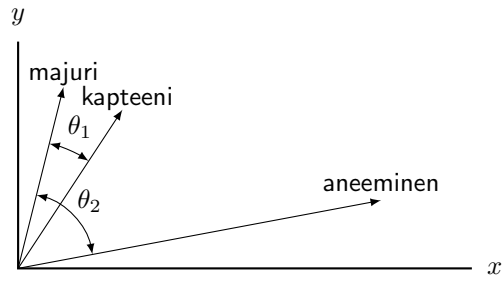
Tyypillisesti tekstien dokumenttivektorit koostuvat elementeistä, jotka kuvaavat sanojen frekvenssiä kyseisessä dokumentissa tai vaihtoehtoisesti

abstraktimpia erottelevia piirteitä. Lähes kaikki tekstien vektoriesitykset pohjautuvat siihen oletukseen, että eri asioita tarkoittavat sanat esiintyvät erityyppisissä konteksteissa, eli niiden tilastollinen jakauma on keskenään erilainen (Turney & Pantel 2010). Vastaavasti merkitykseltään samankaltaisten sanojen ympäristön oletetaan olevan enemmän samankaltainen. Edellinen perustuu ajatukseen, että yksi sana saa merkityksensä kaikista niistä konteksteista, joissa se esiintyy (nk. merkityksen jakaumahypoteesi, *distributional hypothesis of semantics*; Harris 1954). Tähän oletukseen nojaten on kehitetty valtaosa nykyisin käytössäolevista menetelmistä sanojen ja tekstien numeeristen representaatioiden muodostamiseksi.

Tavat muodostaa vektorirepresentaatiota tekstistä voidaan karkeasti jakaa kahteen tyyppiin: sanojen esiintymisjakaumaa (1) laskeviin ja (2) ennustaviin menetelmiin (Baroni ym. 2014). Laskevat menetelmät ovat perinteisempiä ja toimiviksi todettuja menetelmiä, mutta niiden soveltamista rajoittaa se, etteivät ne pysty mallintamaan sanojen merkitystä eli *semantiikkaa* kovin tarkasti. Eräs tällainen menetelmä, tf–idf statistiikka (*term frequency–inverse document frequency*) laskee kuinka monta kertaa kukin aineiston eli *korpuksen* sana esiintyy kussakin dokumentissa, ja sanojen frekvenssit kerrotaan painokertoimella sen mukaan kuinka monessa dokumentissa kukin sana esiintyy (Spärck Jones 1972). Tällaisen esityksen johdosta eri sanoja eri eli oletettavasti erilaisia merkityssisältöjä käsittelevistä dokumenteista muodostetaan erilaisia vektoriesityksiä, jolloin eri sisällöt pystytyään erottelemaan matemaattisesti toisistaan. Sanojen käänteinen painotus puolestaan vähentää hyvin yleisten sanojen painoarvoa erottelevana tekijänä – jos jokin sana esiintyy kaikissa dokumenteissa yhtä monta kertaa, se ei toimi erottelevana piirteenä. Menetelmä toimii hyvin, kun tehtävänä on luokitella dokumentteja eri kategorioihin käsiteltävien aihealueiden mukaan (Venekoski ym. 2016), mutta sen sovellettavuus rajoittuu pitkälti luokittelutehtäviin.

Neuroverkkomallit

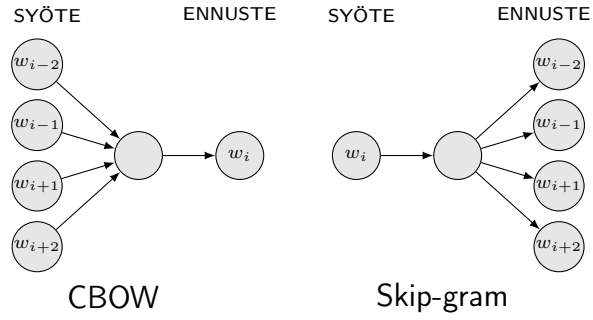
Ennustavat menetelmät, erityisesti nk. neuraaliset kielimallit, ovat saaneet laajaa huomiota viime vuosina. Näillä tarkoitetaan koneoppimiseen pohjautuvia laskennallisia malleja, joissa datasta rakennetaan informaatiota neuroverkon avulla. Ihmisaivojen tavoin neuroverkkomalleissa matalan tason data syötetään eteenpäin kerrosten läpi ylemmille verkon tasoille, jolloin sen sisältämä informaatio muotoutuu jokaisen kerroksen jälkeen tapahtuvassa painotuksessa kompleksisempaan muotoon. Esimerkiksi ihmisen visuaaliset aistimukset lähtevät silmän aistisolujen aktivaatioinformaatiosta, joka neuraalisissa rakenteissa muotoutuu muoto- ja väriaistimuksiksi ja lopulta havainnoksi



Kuva 1. Samankaltaisuudet vektoriavaruudessa. Koska $\theta_1 < \theta_2$, sanavektori $\mathbf{w}_{\text{majuri}}$ on samankaltaisempi vektorin $\mathbf{w}_{\text{kapteeni}}$ kanssa kuin vektorin $\mathbf{w}_{\text{aneeminen}}$.

näkökentässä olevasta objektista (ks. esim. Kalat 2009). Samalla tavoin neuraalisissa kielimalleissa syötetty teksti muodostuu neuroverkon kerroksien läpikäynnin myötä kompleksisemmaksi vektorirepresentaatioksi, josta pystytään päättelemään myös tekstin merkitystä koskevaa informaatiota. Neuraalimallit luovat vektoriesityksen jokaiselle aineistossa esiintyvälle sanalle (ja osa myös dokumenteille), jotka yhdessä muodostavat *vektoriavaruusmallin* aineistosta. Vektorimallilla voidaan esimerkiksi arvioida sanojen samankaltaisuutta sen perusteella, kuinka lähellä toisiaan sanoja vastaavat vektorit sijoittuvat avaruudessa. Kuva 1 havainnollistaa, kuinka sanavektorien välisiä samankaltaisuuksia voidaan laskea ottamalla kosini vektorien välisestä kulman avulla, ts. kosinisamankaltaisuudella (engl. *cosine similarity*).

Vektoriavaruusmalleja voidaan luoda useilla erilaisilla menetelmillä, mutta neuraalisista kielimalleista kenties eniten puhutuimpia ovat nk. *word embedding* -mallit, erityisesti word2vec (Mikolov ym. 2013a, Mikolov ym. 2013b). Näiden mallien tavoitteena on oppia ennustamaan todennäköisin sana, kun mallille annetaan kohdesanaa ympäröivät sanat. Esimerkiksi lauseessa ”*kissa söi hiiren*” mallin tulisi osata ennustaa sanojen ”*kissa*” ja ”*hiiren*” pohjalta sana ”*söi*”. Malli käy läpi sille annetun aineiston tyypillisesti useita kertoja ottaen tarkasteluun määrätyn *ikkunan* tekstiä (esim. 5 peräkkäistä sanaa) kerrallaan. Parhaiten mallin ennustamistavoitteeseen sopivat vektorirepresentaatiot muodostetaan tämän oppimisprosessin kuluessa. Lopputuloksena saadaan vektoriavaruusmalli, joka koostuu vektoreista kaikille aineiston sanoille. Menetelmällä luotujen sanavektorien sijainti avaruudessa simuloi tosimaailman merkityssuhteita; samankaltaiset sanat ovat lähellä toisiaan, ja erityisesti *samansuhteisilla* sanapareilla on samansuuntainen relaatio vektoriavaruudessa. Ilmiötä on havainnollistettu Kuvassa 2. Vastaavia word2vec-mallin kaltaisia neuraalisia kielimalleja



Kuva 2. Havainnollistus word2vec -mallien toimintaperiaatteesta. CBOW -malli ennustaa sanaa w_i tämän kontekstin perusteella, kun taas Skip-gram -malli ennustaa ympäröivää kontekstia sanan w_i perusteella. Ennuste muotoutuu, kun syötteen vektoriesitys painotetaan uudelleen neuroverkon piilokerroksessa (keskimmäinen solu). Muokattu lähteestä Mikolov ym. (2013c).

ovat esimerkiksi GloVe (Pennington ym. 2014) sekä fastText (Bojanowski ym. 2016).

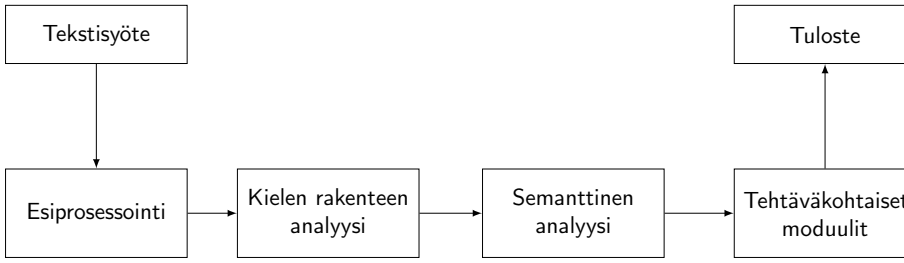
Edellä kuvattu word2vec –malli muodostaa numeerisen esityksen sanoista, mutta ei tekstidokumenteista. Laajemmille teksteille onkin kehitetty Paragraph Vector –menetelmä (Le & Mikolov, 2014), joka toimii samalla periaatteella kuin word2vec, mutta jossa yhdeksi konteksti- tai kohde-elementiksi asetetaan dokumentin indeksi, josta tarkastelun kohteena oleva teksti on peräisin. Dokumenttien vektoriesitys voidaan muodostaa myös laskemalla keskiarvovektori kaikkien dokumentissa esiintyvien sanojen word2vec-vektorien pohjalta. Vaikka jälkimmäinen menetelmä on yksinkertainen, sen on todettu toimivan hyvin tavallisimmissa kieliteknologian sovelluksissa. Tämä on havaittu esimerkiksi tutkimuksessa, jossa Venekoski ym. (2016) arvioivat eri dokumenttien vektorirepresentaatioita luokittelutehtävässä, kun pohja-aineistona käytettiin internetkeskustelujen Suomi24-korpusta (Aller Media Oy, 2014). Myös lukuisia muita menetelmiä dokumenttiesitysten rakentamiseksi on kehitetty (ks. esim. De Boom ym. 2015; Kiros ym. 2015; Lebret & Collobert 2014). On kuitenkin huomattava, että useista sanoista koostuvien tekstien mallintaminen on huomattavasti vaikeampaa kuin yksittäisten sanojen, sillä pidempiin teksteihin sisältyy enemmän informaatiota, joka on myös käsitteellisesti kompleksisempaa kuin yksittäisten sanojen merkityssisällöt.

Suomen kieli

Suomi on paljon taivuttava eli *agglutinoiva* ja *morfologisesti kompleksinen* kieli, mikä tuottaa erityisiä haasteita kieliteknologian menetelmille. Valtaosa menetelmistä on kehitytetty englannin kielelle, eivätkä ne tällöin toimi suoraan tai yhtä hyvin suomenkielisellä datalla. Mikäli menetelmät hyödyntävät informaatiota kielen rakenteesta, kuten sanajärjestystä tai taivutusta, onkin sovellettava kielispesifejä työkaluja. Suomessa on huomattava kielitieteen ja -teknologian tutkimuksen perinne, joten kielianalyysiin sopivia työkaluja on tyypillisesti saatavilla (ks. esim. Pirinen 2015; Haverinen ym. 2013).

Pikaviestimissä ja sosiaalisessa mediassa tekstien kieliasu on merkittävästi asiatekstejä vaihtelevampaa. Englanninkielisen aineistoja tutkittaessa on havaittu, että kieliteknologian sovellusten tarkkuus putoaa keskimäärin noin 10 %-yksikköä, kun pohja-aineistona käytetään sosiaalisen median tekstejä uutisteksteihin verrattuna (Foster ym. 2011). Suomen kielessä ongelma on tätäkin suurempi, sillä eri puhekieliset muodot voidaan kirjoittaa niitä vastaavalla sanan ”kirjakielisestä” muodosta eroavalla kirjoitusasulla (Koskenniemi ym. 2012), kun taas esimerkiksi englannissa eri tavoin lausuttu sana kuitenkin kirjoitetaan tyypillisesti samalla tavalla. Tämän sekä mittavan taivutuksen vuoksi suomenkielisissä aineistoissa on huomattavasti enemmän uniikkeja sanoja tai *merkkijonoja*, minkä vuoksi tietyn merkkijonon esiintymisjakaumaan perustuvat menetelmät voivat tuottaa heikompilaatuisia tuloksia. Toisaalta tekstissä voi esiintyä homonyymejä eli identtisiä merkkijonoja, jotka viittaavat eri käsitteisiin (esim. ”*kuusi*”, joka voi viitata puuhun tai numeroon). Suomenkielinen tekstiaineisto voidaan kuitenkin *lemmatisoida* eli muuttaa jokainen sana tämän merkitystä vastaavaan perusmuotoon, jolloin kieliteknologian menetelmiä voidaan soveltaa paremmin (Venekoski ym. 2016). Lemmatisointi voi kuitenkin olla haasteellista erityisesti some- ja pikaviestinteksteille, joissa kielen kieliopillinen rakenne on osittain muttei täysin puhekielistä.

Vaikka harvinaiset tai muodoltaan kompleksisemmat kielet kuten suomi voivat tuottaa haasteita tavallisimmille kieliteknologian menetelmille, tekstidatan kieli ei estä analytiikkaa. Valtaosa nykypäivän menetelmistä on kieliriippumattomia tai kasvavissa määrin yhä robustimpia rakenteen vaihtelulle kielien välillä (ks. esim Bojanowski ym. 2016). Voimme siis soveltaa hyvin tuloksin englannin kielisellä aineistolla tutkittua menetelmää esimerkiksi suomen, ruotsin tai venäjän aineistoille – jopa ilman, että ymmärrämme kohdeaineiston lingvistisestä rakenteesta mitään. Analyytiikka, tiedustelu ja muu tietokoneavusteinen päätöksenteko luonnollisesti vaatii, että informaatio on päätöksentekijän ymmärrettävissä, mutta tällöinkään kohdeaineiston kieli ei välttämättä ole este, sillä aineisto tai tulokset voidaan kääntää automaattisesti



Kuva 3. Tyypillinen tekstianalyysin arkkitehtuuri. Muokattu lähteestä Koskenniemi ym. (2012).

päätöksentekijän ymmärtämälle kielelle. Automaattinen kääntäminen vapauttaa ihmiskääntämiseen kohdistettuja resursseja ja voi tehdä johtopäätöksistä luotettavampia varsinkin, jos kyseessä on harvinainen lähdekieli, jonka osaaminen on ammattilaispopulaatiossa rajoittunutta. Muun muassa edellisten syiden vuoksi esimerkiksi DARPA on tehnyt merkittävän panostuksen LORELEI (Low Resource Languages for Emergent Incidents) –projektiin, joka tähtää harvinaisten kielten automaattiseen kääntämiseen ja kieliriippumattoman analytiikkaprosessien kehittämiseen (Onyshkevych 2014).

Sovellukset

Kieliteknologian menetelmiä on olemassa laaja kirjo, ja useita eri menetelmiä käytetäänkin yhdessä käytännön sovelluksissa. Esimerkiksi kielellisen raakadatan esiprosessointi ja kielen rakenteen analyysit (katso esim. Pirinen & Cliath 2015; Haverinen ym. 2013) ovat esiaskeleita ennen muita tekstianalyysin vaiheita. Kuvassa 3 esitetään tyypillinen kieliteknologiasovelluksen ja siihen liittyvän analyysiprosessin rakenne. Kaikki vaiheet eivät ole jokaiselle sovellukselle välttämättömiä; esimerkiksi tekstin esiprosessointi ja kielen rakenteen eli *morfosyntaksin* analysointi tyypillisesti parantavat myöhempien analysointivaiheiden tuloksia, mutta eivät ole välttämättömiä askeleita sovellusarkkitehtuurissa. Kuvassa mainittuna semanttisena analyysinä voidaan pitää esimerkiksi tekstin tai sanojen vektorisoimista tai eksplisiittisempää merkityssisällön tulkintaa. Tehtäväkohtaisia moduulit puolestaan vaihtelevat merkittävästi riippuen halutusta sovellusalueesta. Tyypillisimpinä kielianalyysin tehtävinä voidaan pitää tekstien luokittelua (Özgür ym. 2005; Sun ym. 2003; Venekoski ym. 2016), tunneanalyysiä (engl. *sentiment analysis*, esim. Balahur

ym. 2014; Hirschberg & Manning 2015; Gamon 2004; Liu 2012), nimettyjen entiteettien tunnistamista (*named entity recognition*, esim. Sun ym. 2003) ja informaation ekstraktointia (Bradford 2006; Duma & Menzel 2016), konekäänTämistä (Mikolov ym. 2013; Hill ym. 2014) sekä tekstin tuottamista ja tiivistämistä (Khan ym. 2015).

Osa edellisistä tehtävistä liittyy toisiinsa ja monet voivatkin toimia toisilleen komplementaarisina. Esimerkiksi tekstin kääntämiinen tietokoneavusteisesti antaa mahdollisuuden hyödyntää muussa analytiikassa erikielisiä tietolähteitä, vaikka analyytikko ei osaisi datan alkuperäisiä kieliä. Konekääntämiseen onkin tehty mittavia panostuksia niin yksityisellä puolella (esim. Google Translate) sekä sotilaskontekstissa (ks. DARPA LORELEI, Rolston & Kirchhoff 2016; Onyshkevych 2014). Puolestaan tekstejä voidaan ensin luokitella eri aiheisiin, jonka jälkeen luokille voidaan suorittaa tunneanalyysiä, jolloin voidaan saada tietoa, millä tavoin kyseisten luokkien aiheisiin suhtaudutaan. Edellisen kaltaiset järjestelmät ovat tyypillisiä kaupallisen puolen analytiikkaratkaisuja, joiden avulla voidaan päätellä, miten mediassa suhtaudutaan tuotettuihin palveluihin. Vastaavasti tällaiset menetelmät voivat antaa tutkimus- ja tiedustelulaitoksille mahdollisuuden arvioida ”yleistä mielipidettä” (Liu, 2012; Hirschberg and Manning, 2015). Informaation ekstraktointi puolestaan voi tehostaa esimerkiksi massiivisten tesktitietokantojen kuten tietovuotojen analyysiä erityisesti silloin, kun ei tarkalleen tiedetä, mitä tietokannasta etsitään; kun mielenkiintona ovat nk. *tuntemattomat tuntemattomat*. Vaikka organisaatiot itse eivät hyödyntäisi kieliteknologian menetelmiä, toimijoiden tulisi tiedostaa, että vastapuolen kilpailevat toimijat voivat hyödyntää samoja tekniikoita, sillä edellämainittujen menetelmien implementaatiot ovat verrattain yksinkertaisia ja niihin on olemassa tyypillisesti julkisesti saatavilla olevia ohjelmistoja.

Kieliteknologiat liittyvät olennaisesti kasvavaan kognitiiviseen tietojenkäsittelyn (engl. *cognitive computing*) ja big data –analytiikan alueisiin (Hurwitz ym. 2015). Arviolta 80% organisaatioiden keräämästä datasta on järjestämätöntä (engl. *unstructured*), eikä suurta osaa tästä informaatiosta ole aiemmin voitu hyödyntää. Kieliteknologian ja ”älykkään” tai ”kognitiivisen” analytiikan ratkaisut kuitenkin saattavat datassa piilevän tiedon päätöksentekijöiden saataville. Nykyisten menetelmien myötä tiedon järjestämättömyys ei ole rajoite, eikä olemassaolevaa kielidataa ole syytä jättää muun analytiikan ulkopuolelle. Kognitiivisen järjestelmien etuna on se, että ne tyypillisesti kehittyvät automaattisesti uutta tietoa kohdatessaan, eikä kehittyneitä järjestelmiä tarvitse ohjelmoida uudelleen vastaamaan tehtävästä riippuen. Analytiikkamenetelmien kehitys ja automatisoituminen tuovat datasta saatavan informaation eri alojen asiantuntijoiden käyttöön, mikä tukee suoriutumista varsinaisessa asiantuntijatyössä. Jotkin uusimmista analytiikkajärjestelmistä toimivatkin luonnollisen kielen

käyttöliittymällä, ja onkin esitetty, että seuraava askel tekoälyn kehityksessä on ohjelmistojen sekä ohjelmointikielien täydentäminen luonnollisella kielen käyttöliittymillä (Mikolov ym. 2015).

Sosiaalisen median analytiikka

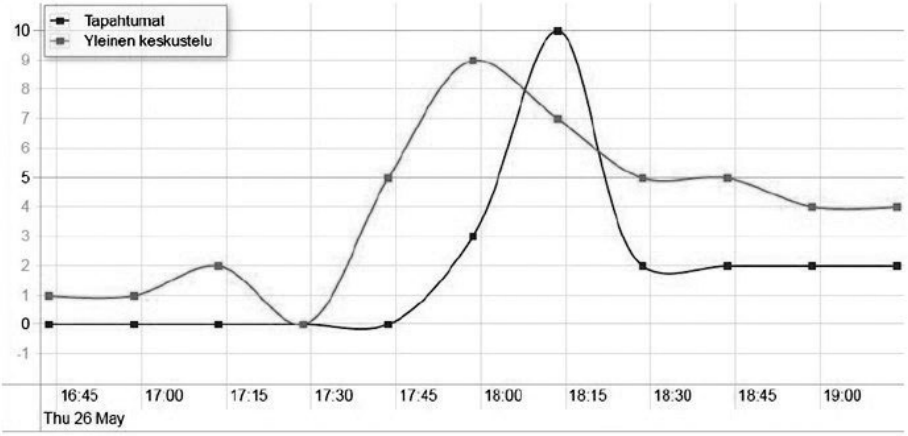
Sosiaalisesta mediasta on muotoutunut merkittävä osa yksilöiden ja myös organisaatioiden vuorovaikuttamista. Siinä missä yksittäiset toimijat voivat jakaa ja ottaa kantaa heitä koskeviin ajankohtaisiin ilmiöihin, samanmieliset toimijat voivat myös organisoitua ja vaihtaa informaatiota keskenään. Informaation tuottaminen ja ympäristön havainnointi on joukkoistunut sosiaalisen median käyttäjille, ja palvelujen sisältöjä analysoimalla yksittäisten käyttäjien havainnot voidaan välittää tiedustelujärjestelmille. Tiedustelutoiminnan tarkoituksellinen joukkoistaminen tulee kuitenkin erottaa julkista dataa hyödyntävästä some-analytiikasta; ensimmäisessä organisaatio pyytää ulkopuolisia tahoja kuten yksityishenkilöitä välittämään tietoa organisaatiolle (esim. vihjepuhelimet), kun taas jälkimmäisessä dataa voidaan kerätä julkisista lähteistä ilman sisällöntuottajien eksplisiittistä lupaa (Stottlemyre 2015). Määritelmällisesti avoimeen dataan pohjautuva tiedustelu eli OSINT kohdistuu aineistoihin, joiden alkuperäinen tarkoitus ei ollut tukea tiedustelutoimintaa tai kansallista turvallisuutta (Stottlemyre 2015). Sosiaalisen median sisältöjen kerääminen tiedustelutarkoituksessa voidaan kuitenkin nähdä julkisessa keskustelussa yksityisyydensuojan rikkomisena tai moraalisesti harmaana toimintana, vaikka juridisia esteitä sille ei olisi – johtuen pitkälti Edward Snowdenin paljastuksista koskien Yhdysvaltain kansallisen turvallisuusviraston NSA:n toimintaa (Lahneman 2016). Toisaalta suuri osa sosiaalisen median analytiikasta ei kohdistu tiettyihin yksilöihin, vaan populaatioiden mielipiteisiin ja käyttäytymiseen sekä näitä koskeviin ilmiöihin. Some-analytiikan primaarisena tarkoituksena voidaan pitää tarkkailtavien ilmiöiden ymmärtämistä; tieto sosiaalisista verkostoista, vuorovaikutuksesta ryhmien sisällä ja välillä, sekä ilmiöiden tai liikkeiden käsitteellis-rakenteellisen dynamiikan ymmärtäminen voi osoittautua arvokkaaksi ilmiön ymmärtämisen kannalta. Viranomaistoiminnalle tällainen analytiikka on relevanttia erityisesti suhteessa turvallisuusuhkien kuten onnetomuuksien ja ekstremismin torjumisessa (Omand ym. 2012).

Julkiset sosiaalinen median palvelut ovat osa kansalaisten arkea sekä kasvavasti viranomaistoimintaa. Organisaatiot ylläpitävät julkisuuskuvaansa tyypillisesti sosiaalisen median välityksellä sekä seuraavat heitä koskevaa mielipiteenvaihtoa (Zeng ym. 2010). Siinä missä kaupalliset toimijat voivat pyrkiä vaikuttamaan kuluttajien mielipiteisiin perinteisellä mainonnalla ja lisääntyvästi tuottamalla some-sisältöä, joka esittää omat tuotteet positiivisessa valossa,

poliittisten toimijoiden voidaan olettaa haluavan vaikuttaa yksilöiden mielipiteisiin vastaavilla tavoilla (Liu 2012; Zeng ym. 2010). Tällaisille toimijoille on relevanttia saada tietoa toimiensa vaikuttavuudesta, jotta tulevia toimia voidaan kohdistaa tarkemmin. Some-analytiikka onkin olennainen osa useiden erilaisten organisaatioiden vaikuttavuusarviointiprosessia. Vastaavasti osa toimijoista pyrkii muovaamaan mielipiteitä mm. trollaamalla ja levittämällä keinotekoisia mielipiteitä sekä disinformaatiota. Erityisesti edellisten kaltaisten spesifien aktiviteettien havaitsemiseksi onkin kehitetty erilaisia kieliteknologiaan pohjautuvia ratkaisuja (ks. Liu 2012), jotka ovat kiinnostavia myös viranomais- ja sotilastoiminnalle.

Monet aiemmista tekstiaineistojen analytiikkaa käsittelevistä tutkimuksista ovat nojanneet määrämuotoiseen aineiston analyysiin ja automaattiset ratkaisut on voitu tehdä havaitun rakenteen pohjalta (esim. Sun ym., 2003). Automaattinen analytiikka on tyypillisesti luotettavaa, kun tiedetään, että tekstit noudattavat tiettyä rakennetta ja analyysimenetelmät voidaan räätälöidä huomioimaan tämän rakenteen erityispiirteet. Erityisesti sotilaallisessa kontekstissa myös viestien kielellinen (tai *syntaktinen*) rakenne voi olla osittain formaali ja noudattaa asetettua protokollaa (Medina, 2008), samaa oletusta ei voida yhtälailla tehdä esimerkiksi sosiaalisen median tekstidatasta. Osa somepalveluista perustuu kuvien jakamiseen ja keskustelupalstoillakin on tyypillistä liittää viestiin kuvia, mikä tarkoittaa että tällaiseen ympäristöön kohdistuvan analytiikan tulisi voida käsitellä myös ei-tekstuaalista sisältöä. Useissa mikrotekstipalveluissa, kuten Twitterissä, erikoismerkkejä voidaan käyttää osoittamaan viesti tietylle käyttäjälle (@M.Virtanen) tai asettamaan viestille aihetunniste (#maanpuolustus), mutta tekstin kieliopillinen rakenne voi olla hyvinkin vaihteleva. Vaikka sosiaalisen median teksteille onkin kehitetty erityisiä menetelmiä, jotka normalisoivat tekstin muodon tunnistaen linkkejä, hashtageja tai muita erityispiirteitä (Foster ym. 2011), vastaava rakenteellinen analyysi ei suoraan ekstraktoi informaatiota tekstien merkityssisällöstä. Tämän vuoksi useat modernit kieliteknologiat ovat keskittyneet yleispäteviin ratkaisuihin, jotka pystyvät estimoimaan tekstien sisältöä riippumatta niiden kontekstisidonnaisesta rakenteesta (Mikolov ym. 2013a; Bojanowski ym. 2016). Toisaalta, viestikanavan tyypillisiä piirteitä voidaan hyödyntää analytiikassa; hashtagejä voidaan hyödyntää viestien luokittelussa ja puheenaiheiden tutkimuksessa siinä missä viestien kohdistuksia voidaan hyödyntää sosiaalisten verkostojen analysoimisessa.

Kuten mainittua, kieliteknologian sovellukset mahdollistavat julkisen mielipiteen analysoimisen dynaamisesti. Esimerkiksi Brigadir ym. (2015) demonstroivat, kuinka muutokset Skotlannin itsenäisyyteen sekä Yhdysvaltain esivaaleihin liittyvien Twitter-keskustelujen dynamiikassa näkyivät



Kuva 4. Esimerkki kieliteknologian ratkaisusta havainnollistaa viestiväylällä käytävää keskustelua automaattisesti luokitellun aiheen pohjalta.

muutoksena kielimallin sanavektoreissa eli mallinnetuissa käsitteissä. Tutkimuksissa käytettyjä menetelmiä hyödyntämällä voidaan esimerkiksi havaita ilmiöihin liittyvää sosiaalis-käsitteellistä koheesiota tai ”kuplautumista” ryhmien välillä, mikä puolestaan mahdollistaa automaattisen tavan havaita merkitsevän poikkeavasti viestiviä ryhmittymiä suuresta viestijäjoukosta. On kuitenkin huomioitava, että tälläisessä kielimallin tulkinnassa tehdään huomattava oletus simuloidun mallin ja reaalimaailman vastaavuudesta, mikä ei välttämättä ole oikea. Vaikka laskennallisessa mallissa tai muussa autonomisessa järjestelmässä havaittaisiin muutos, malli ei ole ekvivalentti reaalimaailman kanssa. Erityisesti merkitystä mallintavien teknologioiden validiteettia voidaan toistaiseksi pitää avoimena kysymyksenä. Puolestaan kielellisen datan tunnevalenssin analysoimista (*sentiment analysis*) voidaan pitää vakiintuneena, validina menetelmänä, joka voi tuoda merkittävää lisäarvoa päätöksentekoprosesseihin, joissa informaatiota ihmisten subjektiivisesta kokemuksesta on arvokasta (Zeng ym. 2010; Liu, 2012).

Keskustelukanavilla ilmenevän merkityssisällön suhteellisen vaihtelun vertailu dynaamisessa ikkunassa tarjoaa mahdollisuuden lähes reaaliaikaiseen tilannekuvaan tarkkailtavasta ilmiöstä (Omand ym. 2012). Puuska ym. (2016, ks. alla) kehittivät tilannekuvakonseptin kyberturvallisuusaiheisen viranomaisviestinnän kontekstiin, mutta samaa periaatetta voidaan soveltaa laajemmin muille viestintäkanaville. Kuvaasa 4 on esitetty hypoteettinen tilannekuva, joka havainnollistaa viestiväylällä käytävää keskustelua. Tilannekuva on muodostettu siten, että kanavan viestit on automaattisella järjestelmällä luokiteltu joko

”Tapahtumiksi” tai ”Yleiseksi keskusteluksi” ja näiden kahden keskusteluluokan frekvenssit esitetty aikajanalla. Tällaisessa tilannekuvassa muutokset kahden luokan suhteellisesta jakaumassa kertovat siitä, että viestiväylällä keskustellaan poikkeuksellisen paljon (tai vähän) tapahtumista, eli jokin tapahtuma on tuolla ajanhetkellä ollut keskustelijoille olennainen. Vastaavia järjestelmiä on kehitetty erityisesti pelastustoimen tueksi, sillä hätätilat kuten suuronnettomuudet tai luonnonkatastrofit näkyvät kyseisiä aihealueita koskevana kasvaneena keskusteluaktiviteettina sosiaalisessa mediassa, minkä lisäksi tapahtumapaikoilla läsnäolevat ihmiset tyypillisesti raportoivat havaintojaan mikroblogeihin tahi muihin some-palveluihin, toimien ensikäden informaationlähteinä tapahtuman laadusta sekä välittäen esimerkiksi geospatiaalista informaatiota tapahtuma-alueesta (Yin ym. 2012). Tällöin hätätilaan liittyvän julkisen keskustelun analyysi tukee pelastustoimen ja muiden viranomaistahojen tilannetietoisuutta, antaen mahdollisuuden informoidumpaan päätöksentekoon (Yin ym. 2012). Vastaavasti muissa konteksteissa tiedustelutoimintaa suorittava toimija voi ennalta määrittää tälle relevanttia merkityssisältöjä ja tarkkailla niiden suhteellista frekvenssiä viestiväylillä; poikkeukselliset tapahtumat ilmenevät tavallisesti käsitteellisenä piikkinä sosiaalisessa mediassa ennen kuin tieto niistä leviää journalistisessa mediassa (Omand ym. 2012).

Some-analytiikka sotilasympäristössä

Sotilaat sekä monet muut viranomaiset ovat työnkuvansa puolesta tekemisissä arkaluontoisen informaation kanssa. Tietoturvallisuuden koulutuksesta ja ohjeistuksesta huolimatta yksittäiset henkilöt saattavat ladata arkaluontoista tai muutoin salassapidettävää informaatiota julkisille viestikanaville. Eräs tällainen tapaus nousi julkisuuteen, kun venäläinen sotilas Sanya Stokin oli ladannut kuvapalvelu Instagramiin kuvia itsestään sekä sotilasvarustuksesta, mukaanlukien BUK-ilmatorjuntajärjestelmästä (Szoldra, 2014). Kuviin oli liitetty ns. geotageja eli geospatiaalista metadataa, joka paljasti Stokinin ottaneen kuvia Ukrainan alueella. Vaikka Venäjän asevoimat ja osa mediasta kiistänyt kuvien aitoudein, tapaus toimii esimerkkinä tilanteesta, jossa julkisesti saatavilla olevasta datasta on voinut saada taktisesti revelevanttia informaatiota.

Kielellisten aineistojen automaattinen tunneanalyysi on osoittanut, että relevantin informaation ei tarvitse olla eksplisiittisesti mainittuna datassa, jotta informaatio voitaisiin saada selville analytiikalla. Osa sosiaalisen median analytiikasta kuten mielipiteiden louhinta perustuu oletukseen, että informaatiota yksilöistä voidaan päätellä tämän käyttäytymisestä some-palveluissa. Vastaavasti on kehitetty sovelluksia, joiden avulla pystytään ennustamaan

tarkkailtavien yksilöiden psyykkistä oireilua kuten masentuneisuutta tai itsetuhoisia ajatuksia näiden kirjoittamien tekstien ja muun verkkoaktiviteetin pohjalta (Rosa ym. 2016; Wang ym. 2013). Menetelmät perustuvat automatisoituun analysiin, joka tyypillisesti huomioi myös epäsuoria emotionaalisesta valenssista kertovia piirteitä. Esimerkiksi masentuneet henkilöt käyttävät enemmän yksikön ensimmäistä persoonataivutusta ja vähemmän monikon persoonaa suhteessa kontrollipopulaatioon, ja toisaalta heidän viestintä tapahtuu iltaisin myöhempään kellonaikaan kuin ei-masentuneiden (Wang ym. 2013). Psyykkisen tilan tarkkailusta voi olla merkittävää hyötyä myös sotilaskontekstissa, sillä useat sotilaiden työtehtävät ovat psyykkisesti kuormittavia, mutta vaativat stressitekijöiden sietokykyä eli resilienssiä. On mahdollista ennaltaehkäistä psyykkistä oireilua ja tukea yksikön suorituskykyä, mikäli pystytään ajoissa havaitsemaan riskiryhmiin kuuluvia yksilöitä tai muutoksia psyykkisestä hyvinvoinnista kertovassa käytöksessä. Vastaavastaa analysoimalla sotilaiden kirjoittamia palveluksenaikaisia tekstejä voitaisiin mahdollisesti arvioida heidän taistelutahtoaan tai muuta emotionaalista reaktiivisuutta, mikä puolestaan edesauttaa suorituskyvyn ylläpitoa.

Sosiaalisesta mediasta on muodostunut tärkeä työkalu nykypäivän ekstremistisille liikkeille. Verkon julkiset ja usein vähemmän valvotut alustat helpottavat ekstremististen liikkeiden toimintaa, mukaan lukien uusien jäsenten rekrytointia, kouluttamista, ideologian levittämistä sekä materiaalisten resurssien hankintaa (Hale 2012). Verkkosivustojen automatisoitu louhinta ja sisältöanalyysi mahdollistaa ääriliikkeiden tai näiden toimintaan viittaavan aktiviteetin tunnistamisen, jolloin asiantuntija voidaan ohjata tarkastelemaan ekstremismiin viittaavia sisältöjä. Tällaisessa sovelluksessa kyseessä on usein luokittelutehtävä, jossa tekstidokumentin sisältö arvioidaan koneellisesti joko ekstremismiin viittaavaksi tai irrelevantiksi tekstiksi. Esimerkiksi Yang ym. (2011) kehittivät kielen vektorimallintamiseen ja koneoppimiseen pohjautuvan menetelmän, jolla he pystyivät tunnistamaan vihapuhetta sisältäviä viestejä äärioikeistolaisilta sivuilta parhaimmillaan yli 90% tarkkuudella. Toisessa tutkimuksessa Sun ym. (2003) kehittivät järjestelmän, joka kykeni tunnistamaan yksittäisten terroritekojen tekijän, kohteen ja uhrin kohtalaisella tarkkuudella terroritekoja käsittelevästä aineistosta. Vastaavia menetelmiä on kehitetty useita ja voidaan todeta, että ekstremismin automaattinen tunnistaminen koneellisesti on kiistämättä mahdollista – kysymykseksi muodostuu tunnistuksen tarkkuus ja erityisesti sovellusten kehittäminen sopiviksi kansallisesta turvallisuudesta vastaavien viranomaisten työympäristöihin.

Informaation louhinta samankaltaisuuksilla

Sotilastiedustelussa on tavanomaista pyrkiä mallintamaan tarkkailtavien toimijoiden käyttäytymistä vertaamalla havaittua toimintaa kyseisen tai vastaavan populaation tyypilliseen aktiviteettiin (Biermann ym. 2004). Kuitenkaan uusista kohteista ei välttämättä ole valmista mallia tai tietokantaa, jota vasten uutta informaatiota voitaisiin verrata, joskin arvioita samankaltaisuuksista toisten kohteiden toimintamalleihin voidaan toki tehdä. Uuden tilanteen ensimmäisenä haasteena voidaan pitää alustavan tiedustelutiedon keräämistä ja tilannekuvan muodostamista tämän pohjalta. Alustava tilannekuva voidaan saavuttaa analysoimalla kohteen populaatiota koskevaa tietoa olemassaolevista tietokannoista, mutta nykypäivänä myös kohdepopulaation käyttämästä tai tätä koskevasta sosiaalisesta mediasta (Onyshkevych 2014). Esimerkiksi aihemallinnusmenetelmät (engl. *topic modelling*) voivat jakaa tällaisen tekstidatan temaattisesti samankaltaisiin klustereihin, jotka muodostavat yhdessä toimijoita koskevan käsitteellinen kentän, mikä puolestaan voi tuottaa informaatiota erityisesti tarkkailtavien toimijoiden kulttuurista ja motivaatiosta. Tällaista laskennallista mallia voidaan verrata muihin, toisen kontekstin datasta tuotettuihin malleihin.

Joissakin tapauksissa olemassaolevia malleja ja tietokantoja kohteiden toiminnasta on, mutta ne päivittyvät tiheästi tai huomattavalla määrällä uutta informaatiota. Esimerkiksi on mahdollista, että julkisuuteen vuotaa viranomaisiakin kiinnostava massiivinen tietokanta (kuten Wikileaksin vuodot). Erittäin suurien tietokantojen manuaalinen analysoiminen veisi suuria määriä henkilötyövoimaa, joten automaattiset analysointiratkaisut ovat usein tarpeellisia. Kieliteknologia tarjoaa esimerkiksi mahdollisuuden tunnistaa automaattisesti toimijoita tekstitietokannasta (*named entity recognition*, NER) ja toisaalta mallintaa samankaltaisuuksia havaittujen toimijoiden välillä. Tällöin olemassaolevan tiedon perusteella kyetään löytämään tunnettujen kohteiden kanssa samankaltaisia eli tiedustelijalle relevantteja kohteita sekä näitä koskevia, inhimillistä tarkastelua vaativia dokumentteja.

Kieliteknologian menetelmiä käsittelevässä luvussa kuvattiin sanojen merkitystä estimoivan word2vec-kielimallin toimintaperiaate. Taulukoissa 1 ja 2 on listattu poimintoja hakusanojen stuxnet ja hamas kanssa samankaltaisimmista sanoista kolmessa eri word2vec-mallissa. Mallit on rakennettu erikielisten ja laadullisesti erilaisten aineistojen pohjalta: ensimmäisen ainestoina on Suomi24-tekstikorpus (Aller Media Oy, 2014), toisessa venäjänkielinen Wikipedian sekä viimeisessä GoogleNews-uutisportaalin tekstejä. Taulukoista voidaan intuitiivisesti havaita, että word2vec-menetelmä kykenee estimoimaan sanojen merkityksen samankaltaisuutta riippumatta siitä, minkä kielistä aineistoa menetelmälle syötetään.

Taulukko 1. Hakusanan "*stuxnet*" kanssa samankaltaisimmat sanat eri kielimalleissa.

#	stuxnet		
	Suomi24	ru.wikipedia	GoogleNews
1	ydinlaitos	duqu	Stuxnet_malware
2	ydinterrorismi	вирус (*virus*)	cyber_superweapon
3	iaea	антивирусных (*antivirus*)	worm_propagation
4	ydinvoimala	кибератаки (*verkkohyökkäyksen*)	Conficker.c
5	ydinenergia	вредоносного (*haitallinen*)	Stuxnet_worm
10	ydinreaktori	malware	Downadup_worm
15	reaktori	вирусом (*virus*)	Koobface_variant
25	voimalaitos	антивирусный (*viruslääke*)	Intrusion_prevention
50	tvon	хакеров (*hakkereita*)	AutoRun
70	iea	руткит (*rootkit*)	Alureon
90	säteilyvuoto	вредоносной (*haitallinen*)	Intrusion_detection
100	megawatti	conficker	ActiveX_vulnerability

Taulukko 2. Hakusanan "*hamas*" kanssa samankaltaisimmat sanat eri kielimalleissa.

#	Hamas		
	Suomi24	ru.wikipedia	GoogleNews
1	gaza	хамаса (*Hamas*)	Fatah
2	palestiinalainen	фатх (*Fatah*)	Palestinian_Authority
3	hizbollah	палестинцев (*palestiinalaiset*)	Palestinian
4	hizbollahin	хезболла (*Hezbollah*)	Gaza
5	fatah	хезболлы (*Hezbollah*)	Hezbollah
10	terroristijärjestö	оон (*PLO*)	Abbas
15	terroristi	арафата (*Arafat*)	Israel
25	israel	баргути (*Barghouti*)	Netanyahu
50	al-aksan	нецарим (*Netzarim*)	Aksa_Martyrs_Brigades
70	intifada	дженин (*Jenin*)	Hezbullah
90	palestinalaisten	шхем (*Sikem*)	Corporal_Shalit
100	tulitus	боевикам (*militantteja*)	Mofaz

Kolmea mallia vertailemalla voidaan havaita, että saman hakusanan kanssa samankaltaisimmat sanat ovat temaattisesti erilaisia eri malleissa: stuxnet–sana tuottaa Suomi24-mallissa erityisesti ydinvoimaan liittyvää käsitteistöä, kun taas GoogleNews-mallissa samankaltaisimmat sanat liittyvät enemmän kyberrikollisuuteen. Malleista GoogleNewsin aineisto on käsitelty NER-työkalulla, joten moniosaiset nimet tai käsitteet ovat tallentuneet alaviivoin eroteltuina merkkijonoina. Tämän vuoksi esimerkiksi hamas–hakusanalla 100 samankaltaisimman sanan joukossa on Hamasiin jollakin tavalla assosioitavia henkilöitä kuten Isrelin pääministeri *Netanyahu* tai Hamasin vangitsema israelilainen sotilas

Taulukko 3. Analogioita Suomi24-kielimallin pohjalta. Taulukossa on annettu 3 samankaltaisinta sanaa D kullekin analogialla. cos(θ) on sanavektorin $\mathbf{w}_D$ ja yhtälöstä $\mathbf{w}_B$ - $\mathbf{w}_A$ + $\mathbf{w}_C$ saatavan vektorin välisen kulman kosiniarvo.

A	B	C	D	$\cos(\theta)$
usa	cia	venäjä	kgb	.568
			kreml	.543
			fsb:n	.532
espanja	franco	italia	benito	.355
			mussolini	.328
			mussolin	.326
leopard	panssarivaunu	hornet	hävittäjä	.624
			ilmavoimat	.617
			pommikone	.563

Corporal_Shalit. Samankaltaisuushakujen perusteella voidaan myös tehdä se havainto, että kaikista samankaltaisimmat sanat ovat enemmän synonyymisiä hakusanan kanssa, mikä ei informaation louhimisen kannalta ole erityisen mielenkiintoinen havainto. Kun puolestaan haetaan kaukaisempia sanoja kuten 100., 300. tai 500. samankaltaisin sana, sanat ovat edelleen assosioituneet hakusanaan, mutta kuvaavat vaikeammin määriteltävissä olevia semanttisia suhteita. Juuri tällaiset kauemmat assosiaatiot voivat paljastaa uutta informaatiota tuntemattomista kohteista, jotka ovat assosioituneet tunnettuun kohteeseen *hakusana*. Vaikka merkityksen mallinnus näyttää uskottavalta kaikissa kolmessa mallissa, on huomattava, että mallinnuksen laatua tulisi arvioida joko semantiikan arviointiin kehitetyillä mittareilla tai lopullisen sovelluksen praktisella arvolla.

Tekstin merkityssisältöjä mallintavien menetelmien pohjalta voidaan arvioida myös kompleksisempia kohteiden välisiä suhteita kuin ainoastaan samankaltaisuutta. Niinkutsutussa analogiatehtävässä ”*A ja B ovat kuten C ja D*” selvitetään, millä kielimallin sanalla *D* on samankaltaisin vektorirepresentaatio siihen vektoriin, joka saadaan lineaarisella algebralla $\mathbf{w}_B - \mathbf{w}_A + \mathbf{w}_C \approx \mathbf{w}_D$ (ks. Mikolov ym. 2013c). Tässä tehtävässä voidaan tarkastella sellaisia tuntemattomia kohteita D, jotka liittyvät kohteeseen C samalla tavalla kuin A liittyy B:hen. Analogiatehtävää on havainnollistettu Taulukossa 3. Analogiat ovat siis eräs yksinkertainen tapa hyödyntää mallinnettua merkityssisältöä, joka mahdollistaa tuntemattoman informaation löytämisen. Toisaalta nykyisellään analogioita voidaan pitää huomattavasti samankaltaisuuksien tutkimista heikompana menetelmänä, sillä käytännössä kielimallien tuottamat analogiat eivät vastaa ihmiskäyttäjien intuitiivista oletusta analogioista. Tätä voidaan pitää

käsitteiden merkityssisältöjen mallinnuksen epätarkkuutena, tai toisaalta sinä, että tutkitut menetelmät mallintavat kieltä laadullisesti eri tavalla kuin ihmiset.

Vaikka samankaltaisuuksia estimoimalla voitaisiin saada selville aiemmin tuntematottomia toimijoita, on huomioitava, että automaattiset informaationlouhimismenetelmät eivät välttämättä paljasta spesifiä, taktista informaatiota toimijoista (Pallaris 2008). Yleisesti ottaen voidaan todeta, että kielipohjaiset analyysit ovat tarkoituksenmukaisia, kun pyrkimyksenä on arvioida kohteiden olemassaoloa, suhteita tai yleistä motivaatiota – datan korrelaatioista tai informaation löytymisestä ei voi päätellä kausaatiota. Organisaatioita koskevaa taktista informaatiota ei välttämättä ole saatavilla julkisilla kanavilla, mutta vähemmän ammattimaiset tahot tai yksilöt saattavat herkemmin jakaa myös taktisesti relevanttia informaatiota julkisilla viestintäväylillä. Täten yksilöiden toiminnasta kiinnostuneet tiedustelua harjoittavat tahot (kuten yritykset tai viranomaiset kuten poliisi) voivat hyötyä avoimen datan kieliteknologiapohjaisesta tiedustelusta enemmän verrattuna sotilasorganisaatioon.

Tutkimus: Viranomaisviestinnän kielianalyysi

Informaatio- ja kommunikaatioteknologia on kasvavissa määrin integroitunut osaksi myös sotilas- sekä erityisesti viranomaistoiminnan toimintaympäristöä. ICT-työkaluja käytetään pääasiassa informaation välitykseen, mutta kieliteknologia mahdollistaa niiden käytön myös informaation lähteenä. Erityisesti sotilasympäristössä viestintä yksilöiden välillä voi nojata pikaviestimien tai muiden viestiväylien kautta tapahtumiin lyhyisiin tilannepäivityksiin eli nk. mikroteksteihin (Rosa & Ellen 2009). Toisaalta julkisten viranomaisorganisaatioiden välinen koordinointi voi olla haastavaa, sillä päätöksenteko saattaa nojata puutteelliseen informaatioon ja useisiin subjektiivisiin tulkintoihin tilanteesta. Tämän vuoksi viranomaisviestintää ja muuta matalamman tason toimintaa dynaamisesti tarkkailevat kognitiiviset järjestelmät voisivat tuoda merkittävän objektiivisen ja koordinoivan lisän päätöksentekijöiden tueksi (ks. Hurwitz ym. 2015).

Maanpuolustuskorkeakoulun tutkimuksessa Puuska ym. (2016) tutkivat viranomaisten pikaviestiliikennettä kansallisessa kyberturvallisuusharjoituksessa. Harjoitukseen osallistuivat valtion kyberturvallisuudesta vastaavat viranomaisorganisaatiot, joiden tarkoituksena oli ratkaista simuloituja turvallisuuspoikkeamia normaalin toimintaprotokollansa mukaisesti. Erityisesti juuri kyberturvallisuuden kontekstissa uhkatilanteilta ja poikkeamilta suojautuminen on ensisijaisen tärkeää, sillä yhteiskunnan kriittinen infrastruktuuri on kasvavissa määrin riippuvainen sektoreita yhdistävien tietotekniikka- ja informaatiojärjestelmien toimivuudesta, ja yksittäisten kriittisten palveluiden

alasajo voi aiheuttaa merkittävää vahinkoa. Kyberturvallisuuteen liittyvä tiedustelutieto ja havainnot poikkeamista nousevat esiin kyberturvallisuuden parissa mutta usein eri organisaatioissa työskentelevien asiantuntijoiden kautta, mikä tekee havaintojen ja näitä koskevien toimenpiteiden koordinoinnista erityisen tärkeää (Jasper 2017).

Tutkimuksen kohteena olleeseen harjoitukseen osallistuneet viranomaiset koordinoivat toimintaansa harjoitukseen rakennetuilla oman organisaationsa sisäisillä sekä organisaatioiden välisillä viestintäkanavilla (Puuska ym. 2016). Harjoituksesta kerättiin yhteensä 3060 viestikanaville lähetettyä viestiä. Tutkijat luokittelivat nämä viestit manuaalisesti 6 eri kyberturvallisuusaiheiseen uhkaluokkaan (ks. MACCSA, 2013) sekä irrelevanteiksi. Tutkimuksen tavoitteena oli kehittää mahdollisimman tarkka ja reaaliajassa toimiva analysointityökalu, joka luokittelee automaattisesti jokaisen vastaavan pikaviestin yhteen kyberuhkaluokkaan. Tätä varten viestien sanat muutettiin perusmuotoisiksi hyödyntäen kielen rakenteen analysointityökalua, Finnish Dependency Parser -jäsennintä (Haverinen ym. 2013). Tätä laajennettiin sisältämään kyberterminologiaa kuten sanat '*spammaaja*' ja '*konffata*'. Perusmuotoisista viesteistä muodostettiin tf-idf vektorirepresentaatiot (ks. yllä), jotka edelleen syötettiin eri koneoppimisalgoritmeille. Algoritmit opetettiin neljällä viidesosalla viesteistä, jonka jälkeen algoritmit luokittelivat jäljelle jääneen viidesosan viesteistä uhkaluokkiin. Luokittelu toistettiin 5 kertaa siten, että jokainen viidesosa jätettiin pois opetusaineistoista kerran eli hyödynnettiin 5-kertaista *ristiinvalidointia* (engl. *cross-validation*), jotta tuloksista voitiin johtaa tilastollisesti luotettavampia päätelmiä. Paras algoritmi, usean luokan tukivektorikone (engl. *multiclass support vector machines, MSVM*), saavutti noin 75% luokittelutarkkuuden.

Aiemmissa tutkimuksissa on havaittu, että viestien luokittelu relevantteihin ja irrelevantteihin lisää viestiliikeennettä monitoroivien tahojen tilannetietoisuutta (Satterfield et al., 2011). Tässä mielessä kehitetyn luokittelujärjestelmän tarkkuutta voidaan pitää hyvänä, sillä aiemmassa tutkimuksessa (Catanzaro et al., 2006) havaittiin, että kun tarkkailijoille korostettiin visuaalisesti 75% kriittistä informaatiota sisältävistä viesteistä, tarkkailijoiden tilannetietoisuus parani ja he havaitsivat enemmän kriittistä informaatiota kuin silloin, kun kaikki tärkeät viestit korostettiin. Toisaalta tieto siitä, missä suhteessa ja kuinka paljon esimerkiksi viranomaiset keskustelevat tietystä "uhkakategoriasta" tai muusta toiminnanalaan kuuluvasta aiheesta, voidaan välittää osaksi laajempaa tilannekuva- tai johtamisjärjestelmää. Tällöin voidaan saada dynaamista tietoa siitä, mitkä aiheet koskettavat organisaation toiminnanalaa kullakin ajanhetkellä. Kappaleessa sosiaalisen median analytiikasta esitetty Kuva 4 demonstroi erästä mahdollista intuitiivista visualisointia, jolla kyseinen informaatio voidaan välittää osaksi laajempaa tilannekuvajärjestelmää.

Viranomaisasiantuntijat keskustelevat viestiväylillä tilanteista, jotka eivät rekisteröidy muihin numeerisiin valvonta- ja ylläpitojärjestelmiin, ja toisinaan tärkeä informaatio voi hukkua muiden viestien joukkoon. Viestiliikenteen ”hiljainen tieto” voidaan valjastaa hyötykäyttöön integroimalla kieliteknologian työkaluja kehitteillä oleviin viestiväyliin ja tilannekuvajärjestelmiin. Kyseisessä tutkimuksessa luotiin eräs yksi analytiikkakonsepti, mutta käytännön toteutuksessa vastaavaan ratkaisuun kuitenkin vaadittaisiin merkittävästi enemmän dataa kuin harjoituksessa saatavilla ollut 3060 viestiä, jotta menetelmää voidaan pitää luotettavana. Ongelman ratkaisuksi on esitetty yleispätevien kielimallien rakentamista ja niiden jalostamista erityisdomeenien kontekstiin (Duma & Menzel 2016), ja tämän vuoksi kyseistä Maanpuolustuskorkeakoulun tutkimusta jatketiin perustutkimukseksi koskien suomenkielisen sosiaalisen median kielidatan analysoimista (Venekoski ym. 2016).

Johtopäätökset

Tässä artikkelissa esiteltiin lyhyesti eräitä keskeisimpiä kieliteknologian menetelmiä ja niiden suhdetta sotilas- sekä viranomaistoiminnalle relevantteihin sovellusalueisiin. Viime vuosien tutkimus ja teknologioiden kehitys on mahdollistanut massiivistenkin tekstiaineistojen merkityssisältöjen analysoimisen tehokkaasti sekä riittävän luotettavasti. Kielimallien rakentaminen auttaa informaation organisoimista sekä erityisesti relevantin informaation löytämistä suuresta tekstimassasta. Toisaalta tekstiaineistoja tai tekstipohjaista viestintää voidaan automaattisesti kartoittaa luokittelemalla viestejä niiden merkityssisällön perusteella, mikä mahdollistaa automaattisten päätöksentekoa tukevien tilannekuvaratkaisujen implementoimisen esimerkiksi osaksi johtamisjärjestelmiä. Suuret määrät järjestämätöntä tekstidataa ovatkin mahdollisuus modernille tiedustelulle, jonka tulisi mahdollisuuksien mukaan hyödyntää julkisesti saatavilla olevia havaintoja ja muuta informaatiota sosiaalisesta mediasta.

Kieliteknologiat, koneoppiminen ja muu toisinaan ”tekoälyksi” kutsuttu data-analytiikka voidaan nähdä jatkumona digitalisaatiolle eli merkittävänä tietotekniikan ja tutkimuksen kehityssuuntana. Älykkäät teknologiat tulevat luultavimmin vaikuttamaan yhteiskunnan toimintaprosesseihin merkittävästi, joten erityisesti yhteiskunnan turvallisuudesta ja infrastruktuurista vastaavien viranomaistahojen tulisi olla edelläkävijöitä kyseisten teknologioiden saralla. Kirjoittajina toivomme, että tämä artikkeli on herättänyt lukijassa mielenkiinnon aihetta kohtaan, jolloin viiteluettelo tarjoaa monipuolisen katsauksen nykyaikaisia kieliteknologian sovelluksia koskevaan kirjallisuuteen.

Lähteet

Aller Media Oy (2014). The Suomi24 Corpus [tekstikorpus], versio 14/05/2015. Kielipankki. http://urn.fi/urn:nbn:fi:lb-201412171, (27.4.2015).

Balahur, Alexandra, Rada Mihalcea & Andres Montoyo (2014). Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications. *Computer Speech & Language*, 28(1), 1–6.

Baroni, Marco, Georgiana Dinu & German Kruszewski (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. Teoksessa *Proceedings of ACL 2014 (52nd Annual Meeting of the Association for Computational Linguistics)*. East Stroudsburg, PA, USA: Association for Computational Linguistics, 238–247.

Biermann, Joachim, Louis de Chantal, Reinert Korsnes, Jean Rohmer & Cagatay Ündeger (2004). From unstructured to structured information in military intelligence-some steps to improve information fusion. Technical report, DTIC Document.

Bojanowski, Piotr, Edouard Grave, Armand Joulin & Tomas Mikolov (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606* [ennakkojulkaisu]. https://arxiv.org/pdf/1607.04606v1.pdf, (23.11.2016).

Bradford, R. B. (2006). Relationship discovery in large text collections using latent semantic indexing. Teoksessa *Proceedings of the Fourth Workshop on Link Analysis, Counterterrorism, and Security*. Bethesda, Maryland.

Brigadir, Igor, Derek Greene & Pádraig Cunningham (2015). Analyzing discourse communities with distributional semantic models. Teoksessa *Proceedings of the ACM Web Science 2015 Conference*. New York, USA: ACM.

Catanzaro, Jean, Matthew Risser, John Gwynne & Daniel Manes (2006). Military situation awareness: Facilitating critical event detection in chat. Teoksessa *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(4), 560–564.

Onyshkevych, Byon (2014). Low Resource Languages for Emergent Incidents (LORELEI). http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents, (24.11.2016).

De Boom, Cedric, Steven Van Canneyt, Steven Bohez, Thomas Demeester & Bart Dhoedt (2015). Learning semantic similarity for very short texts. *arXiv preprint arXiv:1512.00765* [ennakkojulkaisu]. https://arxiv.org/pdf/1512.00765v1.pdf, (23.11.2016).

Duma, Mirela-Stefania & Wolfgang Menzel (2016). Data selection for it texts using paragraph vector. Teoksessa *Proceedings of the First Conference on Machine Translation*. Berlin, Germany: Association for Computational Linguistics, 428–434.

Foster, Jennifer, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Stephen Hogan, Joakim Nivre, Deirdre Hogan & Josef van Genabith (2011). #hardtoparse: POS tagging and parsing the twitterverse. Teoksessa *Proceedings of the 5th AAAI Conference on Analyzing Microtext*. Menlo Park, CA, USA: AAAI Press, 20–25.

Gamon, Michael. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Teoksessa *Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 841.

Hale, Chris (2012). Extremism on the World Wide Web: a research review. *Criminal Justice Studies*, 25(4), 343–356.

Harris, Zellig (1954). Distributional structure. *Word*, 10(2–3), 146–162.

Haverinen, Katri, Jenna Nyblom, Timo Viljanen, Veronika Laippala, Samuel Kohonen,

Anna Missilä, Stina Ojala, Tapio Salakoski & Filip Ginter (2013). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, 48(3), 493–531.

Hill, Felix, Kyunghyun Cho, Sébastien Jean, Coline Devin & Yoshua Bengio (2014). Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448* [ennakkojulkaisu]. https://arxiv.org/pdf/1412.6448v4.pdf, (24.11.2016).

Hill, Felix, Roi Reichart & Anna Korhonen (2015). SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4), 665–695.

Hirschberg, Julia & Christopher Manning (2015). Advances in natural language processing. *Science*, 349(6245), 261–266.

Hurwitz, Judith, Marcia Kaufman & Adrian Bowles (2015). *Cognitive Computing and Big Data Analytics*. Wiley.

Jasper, Scott (2017). US cyber threat intelligence sharing frameworks. *International Journal of Intelligence and CounterIntelligence*, 30(1), 53–65.

Kalat, James (2009). *Biological Psychology* (10th ed.). Wadsworth: Cengage Learning.

Khan, Atif, Naomie Salim & Yogan Jaya Kumar (2015). A framework for multi-document abstractive summarization based on semantic role labelling. *Applied Soft Computing*, 30, 737–747.

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba & Sanja Fidler (2015). Skip-thought vectors. Teoksessa C. Cortes, N. D. Lawrence, D. D. Lee, M. Suqiyama & R. Garnett (toim.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Neural Information Processing Systems Foundation (NIPS), Inc., 3294–3302.

Koskenniemi, Kimmo, Krister Lindén, Lauri Carlson, Martti Vainio, Antti Arppe, Mietta Lennes, Hanna Westerlund, Mirka Hyvärinen, Imre Bartis, Pirkko Nuolijärvi & Aino Piehl (2012). *Suomen kieli digitaalisella aikakaudella*. Berlin: Springer-Verlag.

Lahneman, William (2016). IC data mining in the post-Snowden era. *International Journal of Intelligence and Counter Intelligence*, 29(4), 700–723.

Lazer, David, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy & Marshall Van Alstyne. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915), 721–723.

Le, Quoc & Tomas Mikolov (2014). Distributed representations of sentences and documents. *arXiv preprint arXiv:1405:4053* [ennakkojulkaisu] https://arxiv.org/pdf/1405.4053v2.pdf, (24.11.2016).

Lebret, Rémi & Ronan Collobert (2014). N-gram-based low-dimensional representation for document classification. *arXiv preprint arXiv:1412.6277* [ennakkojulkaisu]. https://arxiv.org/pdf/1412.6277v2.pdf, (24.11.2016).

Liu, Bing (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1), 1–167.

Medina, Emily (2008). Military textual analysis and chat research. Teoksessa *2008 IEEE International Conference on Semantic Computing*. IEEE, 569–572.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* [ennakkojulkaisu]. https://arxiv.org/pdf/1301.3781v3.pdf, (24.11.2016).

Mikolov, Tomas, Armand Joulin & Marco Baroni (2015). A roadmap towards machine intelligence. *arXiv preprint arXiv:1511.08130* [ennakkojulkaisu]. https://arxiv.org/pdf/1511.08130v2.pdf, (24.11.2016).

Mikolov, Tomas, Quoc Le & Ilya Sutskever (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* [ennakkojulkaisu]. https://arxiv.org/pdf/1309.4168v1.pdf, (24.11.2016).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado & Jeffrey Dean (2013c). Distributed representations of words and phrases and their compositionality. Teoksessa *Advances in neural information processing systems 26 (NIPS 2013)*. Neural Information Processing Systems Foundation (NIPS), Inc., 3111–3119.

Multinational Alliance for Collaborative Cyber Situational Awareness (MACCSA) (2013). *Collaborative Cyber Security Situational Awareness (CCSA) Information Sharing Framework (ISF), Version 2.4.* https://www.terena.org/mail-archives/refeds/pdfjJz-1CRtYC4.pdf, (24.11.2016).

Omand, Sir David, Jamie Bartlett & Carl Miller (2012). Introducing social media intelligence (SOCMINT). *Intelligence and National Security*, 27(6), 801–823.

Özgür, Arzucan, Levent Özgür & Tunga Güngör (2005). Text categorization with class-based and corpus-based keyword selection. Teoksessa *Proceedings of the 20th International Conference on Computer and Information Sciences (ISCIS 2005)*. Berlin: Springer-Verlag, 606–615.

Pallaris, Chris. (2008). Open source intelligence: A strategic enabler of national security. *CSS Analyses in Security Policy*, 3(32), 1–3.

Pennington, Jeffrey, Richard Socher & Christopher Manning (2014). GloVe: Global vectors for word representation. Teoksessa *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*. Stroudsburg, PA, USA: Association for Computational Linguistics, 1532–1543.

Pirinen, Tommi (2015). Omorfi—free and open source morphological lexical database for finnish. Teoksessa B. Megyesi (toim.), *Proceedings of the Nordic Conference of Computational Linguistics, NODALIDA 2015*. Linköping, Sweden: Linköping University Electronic Press, 313–315.

Puuska, Samir, Matti Kortelainen, Viljami Venekoski & Jouko Vankka (2016). Instant message classification in Finnish cyber security themed free-form discussion. *International Journal On Cyber Situational Awareness*, 1(1), 1–4.

Rolston, Leanne & Katrin Kirchhoff (2016). *Collection of bilingual data for lexicon transfer learning*. UWEE Technical Report. Seattle, Washington, USA: University of Washington.

Rosa, Kevin Dela & Jeffrey Ellen (2009). Text classification methodologies applied to micro-text in military chat. Teoksessa *Proceedings of the 2009 International Conference on Machine Learning and Applications.* IEEE, 710–714.

Rosa, Renata, Demóstenes Rodríguez, Gisele Schwartz, Ivana de Campos Ribeiro & Graça Bressan. (2016). Monitoring system for potential users with depression using sentiment analysis. Teoksessa *Proceedings of the 2016 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 381–382.

Satterfield, Kelly, Victor Finomore, Courtney Castle & Joel Warm (2011). Evaluation tools to aid command and control operators in chat-based communication monitoring. Teoksessa *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Volume 55*. SAGE Publications, 480–484.

Spärck Jones, Karen. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1), 11–21.

Stottlemyre, Steven (2015). HUMINT, OSINT, or Something New? Defining Crowdsourced Intelligence. *International Journal of Intelligence and CounterIntelligence*, 28(3), 578–589.

Sun, Aixin, Myo-Myo Naing, Ee-Peng Lim & Wai Lam (2003). Using Support Vector Machines for Terrorism Information Extraction. Teoksessa *Proceeding of the First NSF/NIJ Symposium on Intelligence and Security Informatics 2003*. Berlin Heidelberg: Springer-Verlag, 1–12.

Szoldra, Paul (2014). A Russian Soldier's Instagram Posts May Be The Clearest Indication Of Moscow's Involvement In East Ukraine. *Business Insider*. www.businessinsider.my/russian-soldier-ukraine-2014-7/, (24.11.2016).

Turney, Peter & Patrick Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of artificial intelligence research*, 37(1), 141–188.

Vachon, François, Daniel Lafond, Benoît Vallières, Robert Rousseau & Sébastien Tremblay (2011). Supporting situation awareness: A tradeoff between benefits and overhead. Teoksessa *2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*. IEEE, 284–291.

Venekoski, Viljami, Samir Puuska & Jouko Vankka (2016). Vector space representations of documents in classifying finnish social media texts. Teoksessa G. Dregvaite & R. Damasevicius (toim.), *Proceedings of the 22nd International Conference on Information and Software Technologies, ICIST 2016*. Sveitsi: Springer International Publishing, 525–535.

Wang, Xinyu, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu & Zhana Bao (2013). A depression detection model based on sentiment analysis in micro-blog social network. Teoksessa J. Li ym. (toim.), *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining 2013*. Berlin: Springer-Verlag, 201–213.

Yang, Ming, Melody Kiang, Yungchang Ku, Chaochang Chiu & Yijun Li (2011). Social Media Analytics for Radical Opinion Mining in Hate Group Web Forums. *Journal of Homeland Security and Emergency Management*, 8(1), artikkeli 38.

Yin, Jie, Andrew Lampert, Mark Cameron, Bella Robinson & Robert Power (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, 27(6), 52–59.

Zeng, Daniel, Hsinchun Chen, Robert Lusch & Shu-Hsing Li (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, 25(6), 13–16.