Tuomas Nivala

TWITTERIN TUOTTAMAN DATAN TEKSTIANALYTIIKKA JA SOVELLETTAVUUS JULKISHALLINNOSSA

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2013

TIIVISTELMÄ

Nivala, Tuomas
Twitterin tuottaman datan tekstianalytiikka ja sovellettavuus julkishallinnossa
Jyväskylä: Jyväskylän yliopisto, 2013, 96 s.
Tietojärjestelmätiede, pro gradu-tutkielma
Ohjaaja: Eetu Luoma

Tässä pro gradu-tutkielmassa tarkastellaan sosiaalisen median mikroblogipalvelu Twitterin tuottaman datan tekstianalytiikkaa sekä arvioidaan tämän sovellettavuutta julkishallinnon palvelukseen. Tutkimuksen tarkoituksena on selvittää, mitä Twitter-datan tekstianalytiikalla voidaan tutkia, millaisia menetelmiä näissä tutkimuksissa on käytetty ja millaisia tuloksia on saatu. Julkishallinnon osalta mielenkiinnon kohteena on se, kuinka näitä menetelmiä voidaan käyttää julkishallinnon organisaatioiden tekstianalytiikassa.

Twitter-datan tekstianalytiikan menetelmien osalta tutkielmassa on tehty kirjallisuuskatsaus olemassa olevaan tutkimukseen. Empiirisessä osuudessa on suoritettu puolistrukturoidut teemahaastattelut aihepiiristä julkishallinnon kohdeorganisaatioiden edustajien kanssa. Nämä organisaatiot olivat Kansaneläkelaitos (KELA) ja Terveyden ja hyvinnoinnin laitos (THL).

Tutkielman tuloksina havaitaan Twitter-datan tekstianalytiikkaa voitavan käyttää hyvin laaja-alaisesti erilaisissa tutkimuksissa. Tekstianalytiikan menetelmien todettiin soveltuvan erittäin hyvin Twitterin tekstidatan hyödyntämiseen lukuun ottamatta poliittista tutkimusta. Julkishallinnon todetaan hyötyvän potentiaalisesti lukuisin tavoin sosiaalisen median luoman datan seurannasta tekstianalytiikan keinoin. Sen sijaan Twitter-datan tekstianalytiikan menetelmien soveltuvuutta julkishallinnon oman tekstimuotoisen datan käsittelyyn ei voida tämän tutkielman perusteella arvioida.

Avainsanat: Twitter, tekstianalytiikka, julkishallinto, KELA, THL

ABSTRACT

Nivala, Tuomas
Text analytics of Twitter-generated data and applicability for public governance
Jyväskylä: University of Jyväskylä, 2013, 96 p.
Information Systems Science, Master's Thesis
Ohjaaja: Eetu Luoma

In this Master's Thesis, examination has been done on the use of text analytics on Twitter-generated data and on the applicability of these methods for public governance. The purpose of the study is to define what types of research can be done based on Twitter data text analytics, what are the methods that has been used and what kind of results have been achieved. Further regarding public governance, interest is focused on how these methods could be applied.

Regarding Twitter data text analytics methods, literature research was done on existing research literature. In the empirical part of the study semi-structured theme interviews were done with the representatives from two different organizations of the public governance. These organizations were the Social Insurance Institution of Finland (KELA) and National Institute for Health and Welfare (THL).

As the result of the study, wide applicability of the text analytics methods on Twitter data was discovered. Twitter data text analytics methods can be efficiently used for variety of research topics although political research remains a challenging topic. Results indicate also that public governance can potentially benefit in various ways from the surveillance of social media data. However, the applicability of Twitter data-based text analytics methods for public governance's own textual data cannot be evaluated on the basis of this study.

Keywords: Twitter, text analytics, public governance, Social Insurance Institution of Finland (KELA), National Institute for Health and Welfare (THL)

SISÄLLYS

TIIVISTELMÄ 2

ABSTRACT 3

SISÄLLYS 4

1 JOHDANTO 5
1.1 Johdatus tutkielmaan 5
1.2 Tutkimuksen käsitteistöä 7

2 TUTKIMUKSEN TEKEMINEN 9
2.1 Tutkimusongelma ja tutkimuskysymykset 9
2.2 Tutkimusmenetelmät ja tutkimuksen kulku 10

3 TWITTER, TEKSTIANALYTIIKKA JA TWITTER-DATAN ANALYSOITAVUUS 16
3.1 Twitter 16
3.2 Twitterin käytön tyypillisiä piirteitä 17
3.3 Tekstianalytiikka ja sosiaalinen media 21
3.4 Twitter-datan analysoitavuus 24

4 TWITTER-DATAN TEKSTIANALYTIIKKA 30
4.1 Trendien ja uutisaiheiden havaitseminen 30
4.2 Epidemioiden seuranta 34
4.3 Sävyanalyysiin perustuvia tutkimuksia 39
4.4 Poliittinen tutkimus 44
4.5 Joukkoistettu aistinta ja seuranta 47
4.6 Kirjallisuuskatsauksen tulokset ja yhteenveto 53

5 EMPIIRINEN TUTKIMUS 60
5.1 Empiirisen tutkimuksen tulokset 60
5.2 Empiirisen tutkimuksen johtopäätökset 73

6 TUTKIELMAN TULOKSET JA JOHTOPÄÄTÖKSET 81

7 YHTEENVETO 86

LÄHTEET 88

LIITE 1 HAASTATTELUKYSYMYKSET 95

1 JOHDANTO

Tässä luvussa on esitetty ensimmäisen alaluvun osalta lyhyt johdatus tutkielman aihepiiriin. Toisessa alaluvussa on esitelty tutkielman kannalta keskeinen käsitteistö.

1.1 Johdatus tutkielmaan

Viime vuosien aikana eräs merkittävimmistä suuntauksista paitsi informaatioteknologian alalla myös monilla muilla aloilla on ollut jatkuva tietomäärien kasvu. Erilaisten datavarantojen määrä, koko, datan kertymisnopeus ja monimuotoisuus kasvavat edelleen ripeästi. Yleisesti ottaen globaalin tiedon valtava määrä ja kertymisnopeus mahdollistavat myös ennaltanäkemättömiä mahdollisuuksia tiedon analysoinnin kannalta. Sosiaalisen median kasvu on keskeinen osa tätä ilmiötä ja on erityisesti johtanut tekstimuotoisen datan määrän merkittävään kasvuun. Tämän datan hyödyntämiseksi tekstianalytiikan menetelmät ovat keskeisessä asemassa.

Twitter on tällä hetkellä toiseksi suosituin sosiaalisen median palvelu reilulla 500 miljoonalla käyttäjällään. Mikä tekee Twitteristä poikkeuksellisen kiinnostavan palvelun on sen käytäntö tiedon julkisuuden kanssa. Käyttäjien lähettämät viestit, twiitit (engl. tweet), ovat julkista tietoa ja tätä dataa voidaan kerätä Twitterin tarjoamien ohjelmointirajapintojen (engl. Twitter API, Application Programming Interface) avulla. Twiitit varustetaan myös viestin lähettämisajankohdan tiedolla sekä geolokaatiotiedolla. Koska aika- ja paikkatiedot sekä viestien tekstisisällöt ovat saatavilla julkisesti, voidaan tekstianalytiikan menetelmiä käyttää twiiteistä koottuihin otoksiin. Tällä on huomattavaa tutkimuksellista sovellettavuutta ja merkitystä.

Tämän tutkielman ensimmäinen tutkimusongelma oli perehtyä näihin menetelmiin, niiden käyttöön eri tutkimuskohteissa sekä saatuihin tuloksiin. Toinen tutkimusongelma koski näiden menetelmien sovellettavuutta julkishallinnon palvelukseen. Tutkimusongelmat ovat tärkeitä ja ajankohtaisia,

sillä johtuen sosiaalisen median ripeästä kehityksestä on todennäköistä, ettei julkishallinnon taholta tämän avaamiin mahdollisuuksiin ole tartuttu optimaalisesti. Lisäksi erilaiset tekstianalytiikan menetelmät ovat edelleen kehitysvaiheessa ja parhaiden toteutustapojen löytämiseksi vertailevaa tutkimusta oli perusteltua tehdä. Tutkimusongelmaan perehtyminen tuottaa myös paljon muuta informaatiota, kuten kartoitusta julkishallinnon organisaatioiden tekstianalytiikan käytöstä ja mahdollisista tulevaisuuden käyttöskenaarioista.

Täsmällisesti muotoiltuna tutkielman tutkimuskysymykset olivat:

Tutkimuskysymys 1. Miten Twitteristä saatavaa tekstimuotoista dataa voidaan käyttää tekstianalytiikan avulla?

Tutkimuskysymys 2. Ovatko Twitter-datan tekstianalytiikan menetelmät, työkalut ja mallit sovellettavissa julkishallinnon tarpeisiin?

Ensimmäisen tutkimuskysymyksen osalta on tarkasteltu kysymystä kolmesta eri näkökulmasta. Näitä olivat se, mihin käyttökohteisiin tekstianalytiikkaa on tutkimuksissa käytetty. Toinen näkökulma tarkasteli käytettyjä menetelmiä ja tekniikoita. Kolmas näkökulma otti huomioon saavutetut tulokset. Näin on menetelty, jotta voitaisiin mahdollisimman hyvin arvioida käyttökohteita, menetelmiä ja tuloksia tarkasteltaessa sovellettavuutta julkishallinnon kontekstiin.

Toisessa tutkimuskysymyksessä on pyritty tarkastelemaan sekä Twitter-dataan perustuvien tekstianalytiikan menetelmien soveltuvuutta julkishallinnon omaan tekstimuotoiseen dataan että sosiaalisen median tuottamaan dataan.

Tutkimusmenetelminä tutkielmaa tehtäessä olivat teoreettisen osuuden osalta kirjallisuuskatsaus ja empiirisen osuuden osalta puolistrukturoitu teemahaastattelu kahden eri julkishallinnon organisaation asiantuntijatahojen kanssa. Kirjallisuuskatsauksessa tutustuttiin hyvin laaja-alaisesti Twitter-dataa hyödyntäviin tekstianalyyttisiin tutkimuksiin. Luvut 3 ja 4 tässä tutkielmassa käsittelevät kirjallisuuskatsauksen satoa. Empiirisen osuuden osalta suoritettuja teemahaastatteluita edelsivät esittelytilaisuudet, joissa tutkielman tekijä kertoi kirjallisuuskatsauksen aikana esiintulleiden menetelmien käytöstä, tekniikoista ja tuloksista. Tämän jälkeen suoritettiin haastattelut, joissa haastateltat reflektoivat näkemyksiään tekstianalytiikan soveltamisesta kohdeorganisaatioissa.

Soveltuvuutta julkishallinnon tarpeisiin on arvioitu sekä tulkitsevasti haastatteluiden perusteella että formaalimmin asettamalla kolme arviointiperustetta. Näitä olivat observatiivinen, analyyttinen ja deskriptiivinen arviointiperuste (Hevner ym., 2004, s.86). Arviointiperusteita on käytetty paitsi haastatteluiden osalta myös kirjallisuuskatsauksen aikana arvioitaessa erilaisten menetelmien soveltuvuutta julkishallinnon palvelukseen. Lisäksi kirjallisuuskatsauksen jälkeen on luotu hypoteesit, joiden oikeellisuutta on

arvioitu empiirisen vaiheen jälkeen. Luvussa 5 on esitelty ensin haastattelututkimuksen tuloksia ja tämän jälkeen johtopäätöksiä.

Tutkimuksen tavoitteena oli kuvata laaja-alaisesti tekstianalytiikan menetelmien käyttöä Twitter-pohjaisen datan hyödyntämisessä sekä löytää potentiaalisia sovellutuskohteita näille julkishallinnon kontekstissa. Näissä tavoitteissa on onnistuttu. Laaja selvitys eri käyttökohteista, menetelmistä ja näillä saaduista tuloksista on esitelty kirjallisuuskatsauksen yhteydessä. Soveltamiskohteita koskien julkishallinnon omaa tekstimuotoista dataa ei ole kyetty esittämään, koska kohdeorganisaatioilla ei ole tässä määrin tekstimuotoista dataa. Sen sijaan sosiaalisen median tuottaman datan osalta julkishallinnolla vaikuttaisi olevan potentiaalisesti huomattavia määriä sovellutuskohteita aina tutkimuksellisista käyttökohteista erilaisiin seurantajärjestelmiin. Näiden osalta tarkempi analyysi on suoritettu luvussa 6.

Saaduilla tuloksilla voi myös olla huomattavaa merkitystä otettaessa seuraavia askelia kohti tehokkaampaa julkishallinnon suorittamaa sosiaalisen median datan hyödyntämistä. Mahdollisia jatkotutkimuksen aiheita voisi olla muiden julkishallinnon organisaatioiden tekstianalytiikan tarpeiden kartoittaminen tai toimivien järjestelmien kehittäminen eri julkishallinnon instansseille. Näistä konkreettisimmat voisivat olla terveyteen ja sosiaalisiin muutoksiin liittyvät tutkimukset THL:n osalta sekä yleisemmällä tasolla jonkinlaisen sosiaalisen median seurantajärjestelmän luominen julkishallinnon taholta.

1.2 Tutkimuksen käsitteistöä

Tässä alaluvussa on kuvattu tutkielman keskeinen käsitteistö.

Sosiaalinen media on elektronisten työvälineiden ja internetin sovelluksien käyttöä entistä tehokkaampaan informaation ja kokemusten jakamiseen ja niistä keskustelemiseen toisten ihmisten kanssa (Moturu, 2009).

Twitter on sosiaalisen median mikrobloggaus-palvelu, jonka tärkeimpiin käyttömahdollisuuksiin kuuluu lyhyiden 140-merkkisten viestien, ”twiittien”, lähettäminen ja vastaanottaminen. Nämä twiitit ovat pääosin tekstimuotoista julkista tietoa.

Twitter-datasta tai *Twitterin luomasta datasta* puhuttaessa tutkimuksessa tarkoitetaan sitä kaikkea tietomäärää, joka on saatavissa Twitterin ohjelmointirajapintojen (Twitter API, engl. Application Programming Interface) kautta.

Tekstianalytiikalla (engl. *text analytics, text mining*) tarkoitetaan tietämyksen löytämistä ja hankkimista tekstimuotoisista datavarannoista (Stavrianou, Andritsos & Nicoloyannis, 2007). Tekstianalytiikan tehtävä on löytää syvällisempää tietoa tekstidatan joukosta erilaisin menetelmin. Tällaisia menetelmiä ovat esimerkiksi erilaiset koneoppimisen muodot.

Sävyanalyysillä ja rinnasteisesti *mielipidelouhinnalla* (engl. sentiment analysis) tarkoitetaan erilaisten menetelmien soveltamista analysoitavana olevan tekstin tunteellisen kontentin selvittämiseksi. Sävyanalyysi on positiivisten ja negatiivisten tunteiden, mielipiteiden ja arvioiden tunnistamista (Wilson, Wiebe & Hoffman, 2005).

Kansaneläkelaitos (KELA) on itsenäinen julkisoikeudellinen laitos, jonka hallintoa ja toimintaa valvovat eduskunnan valitsemat valtuutetut. Kansaneläkelaitoksen tehtäviä ovat sitä koskevien säädösten mukaiset sosiaaliturvaa koskevat tehtävät. Tämän lisäksi KELA:n tehtävänä on tiedottaa etuuksista ja palvelutoiminnasta, harjoittaa etuusjärjestelmien ja oman toimintansa kehittämistä palvelevaa tutkimusta, laatia tilastoja, arvioita ja ennusteita sekä tehdä ehdotuksia toimialaansa koskevan lainsäädännön kehittämisestä (”Laki Kansaneläkelaitoksesta 17.8.2001/731”. Kelan lakipalvelu www-sivusto.
<http://www.edilex.fi/kela/fi/lainsaadanto/20010731.> 20.10.2013).

Terveyden ja hyvinvoinnin laitos (THL) on sosiaali- ja terveysministeriön hallinnonalalla toimiva tutkimus- ja kehittämislaitos. THL palvelee valtion ja kuntien päättäjiä, järjestöjä, tutkimusmaailmaa ja kansalaisia. Terveyden ja hyvinvoinnin laitoksen (THL) tarkoitus on vaikuttaa edistävästi suomalaisten terveyteen ja hyvinvointiin. Lisäksi tehtävinä ovat sairauksien ja sosiaalisten ongelmien ehkäisy sekä sosiaali- ja terveyspalveluiden kehittäminen. THL toteuttaa tehtäväänsä tutkimuksen, seurannan ja arvioinnin, kehittämistyön, asiantuntijavaikuttamisen ja viranomaistehtävien sekä kansainvälisen yhteistyön avulla (”Terveyden ja hyvinvoinnin laitos – Organisaatio”. Thl.fin www-sivusto.
<http://www.thl.fi/fi_FI/web/fi/organisaatio> 20.10.2013).

2 TUTKIMUKSEN TEKEMINEN

Tässä luvussa käsitellään tutkimuksen tekemiseen liittyviä aiheita. Näitä ovat tutkimusongelman tunnistaminen, tutkimuskysymyksien määrittely, tutkimusaiheen rajaus, tutkimuksen tekotapa, käytetyt menetelmät ja aineiston hankinta sekä tulosten analysointitapa. Ensimmäisessä alaluvussa on kuvattu tutkimusongelmaa ja tutkimuskysymyksiä. Toisessa alaluvussa on kuvattu tutkimusmenetelmien käyttö, tutkimusaineiston hankinta ja tutkimuksen kulku.

2.1 Tutkimusongelma ja tutkimuskysymykset

Tutkimuskysymysten muotoilussa nousi esiin kaksi aihealuetta. Ensimmäisenä mielenkiinnon kohteena oli olemassaoleva tutkimus Twitteristä saatavilla olevan tekstimuotoisen datan tekstianalytiikasta. Twitterin suhteellisesta nuoruudesta huolimatta sen tuottamaan tekstimuotoiseen dataan kohdistuvaa tekstianalytiikkaa on jo käytetty lukuisissa tutkimuksissa. Tutkimusongelmaksi ja mielenkiinnon kohteiksi muodostui se mitä Twitter-datan tekstianalytiikalla voidaan tehdä, millaisilla menetelmillä ja millaisia tuloksia tällä on saavutettu.

Kirjallisuuskatsauksen osalta tämä jakoi tarkastelunäkökulmat kolmeen osaan. Ensinnäkin haluttiin selvittää minkälaista tutkimusta Twitter-datan tekstianalytiikalla oli tehty sekä minkälaista tutkimusta potentiaalisesti vastaavin metodein kyettäisiin tekemään. Toiseksi haluttiin perehtyä menetelmiin, joilla tutkittavaa ilmiötä oltiin lähestytty. Kolmanneksi haluttiin selvittää kuinka hyvin eri asiayhteyksissä menetelmien oltiin todettu toimivan. Näiden kolmen eri näkökulman kautta voitiin luoda kokonaisvaltainen näkemys, jonka perusteella oli mahdollista tehdä olettamuksia ja hypoteeseja menetelmien mahdollisesta siirtämisestä toiseen kontekstiin – tässä tapauksessa julkishallinnon organisaatioiden palvelukseen.

Toinen tutkimuskysymyksien muotoiluun vaikuttava tekijä oli näiden työkalujen, menetelmien, mallien ja lähestymistapojen sovellettavuus julkishallinnon oman tekstimuotoisen datan analytiikkaan sekä mahdollisesti

sosiaalisen median tuottaman tekstidatan analytiikkaan julkishallinnon taholta. Toisena tutkimusongelmana tässä tutkimuksessa on, voidaanko Twitter-datan tekstianalytiikasta tehdyistä tutkimuksista siirtää toimintamalleja tai valmiita työkaluja julkishallinnon tekstianalytiikan palvelukseen. Tämän osalta teeman mukaisia kysymyksiä ovat esimerkiksi se, voidaanko jotain tiettyä mallia käyttää sellaisenaan tai modifioituna kohdeorganisaatioiden tekstianalytiikassa. Tässä tapauksessa siirrettävä asia – Hevner ym. (2004) termein IT-artifakti - voi olla algoritmi, metodi, menetelmä, työkalu tai mikä tahansa käytetty mallintamis- tai lähestymistapa. Toisin sanoen, lähes mitä vain mitä voidaan oppia esimerkkitapauksista kirjallisuuskatsauksen pohjalta ja siirtää julkishallinnon kontekstiin. Mikäli jotain menetelmiä voidaan siirtää julkishallinnon organisaatioiden toimintaan, niin millaisia nämä käyttöskenaariot ovat ja kuinka hyvin omaksutut mallit toimivat? Tutkimuksen tältä osin asiaa tarkastellaan sovelletusta design science-näkökulmasta (Hevner, March, Park & Ram, 2004 ja Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007) sekä haastattelutuloksia tulkiten.

Näiden tutkimusongelmien pohjalta voidaan määritellä tutkimuksen tutkimuskysymykset.

Tutkimuskysymys 1. Miten Twitteristä saatavaa tekstimuotoista dataa voidaan käyttää tekstianalytiikan avulla?

- Millaisia käyttökontekstit ovat
- Millaisilla työkaluilla, menetelmillä ja lähestymistavoilla
- Millaisia tuloksia tekstianalytiikkaan pohjautuvissa tutkimuksissa on saatu ja onko tämä ollut hyödyllistä

Tutkimuskysymys 2. Ovatko Twitter-datan tekstianalytiikan menetelmät, työkalut ja mallit sovellettavissa julkishallinnon tarpeisiin?

- Tekstianalytiikan soveltaminen organisaation omaan tekstimuotoiseen dataan
- Tekstianalytiikan soveltaminen sosiaalisen median tekstimuotoiseen dataan

2.2 Tutkimusmenetelmät ja tutkimuksen kulku

Tutkimuksen teoreettinen vaihe toteutettiin käyttämällä tutkimusmenetelmänä kirjallisuuskatsausta. Kirjallisuuskatsaus toteutettiin pääosin Websterin ja Watsonin (2002) ohjeistuksien mukaisesti. Toisaalta lähteinä kirjallisuuskatsauksen suorittamisen osalta olivat myös Hirsjärven, Remeksen ja Sajavaaran (2010) sekä Chris Hartin (1998 & 2001) ohjeistukset. Aihepiiri jaettiin suunniteltujen tutkimuskysymysten perusteella teema-alueiksi. Nämä teema-alueet käsittelivät pääosin Twitteristä saatavilla olevan tekstimuotoisen datan

tekstianalytiikkaa. Myös muita teemoja käytettiin lähteiden etsinnässä ja esimerkiksi julkishallinnon IT-tarpeita ajatellen pyrittiin löytämään teoreettisia lähteitä. Näiden merkitys kuitenkin jäi olemattomaksi kirjallisuuskatsauksen sisällön osalta. Toisaalta julkishallinnon kannalta arvioita saatiin tutkimuksen toisessa empiirisessä vaiheessa kohdeorganisaatioiden edustajilta. Tutkimuskysymysten perusteella luotujen teema-alueiden pohjalta suoritettiin tutkittavan kirjallisuuden etsintä.

Websterin ja Watsonin (2002) mukaan kirjallisuuskatsauksen tulisi luoda vankka pohja tutkittavan ilmiön tietämyksen lisäämiseksi. Tämän vuoksi kirjallisuuskatsauksessa on käsitelty laaja-alaisesti Twitteriä ja tekstianalytiikan käyttökohteita sekä analysoitu Twitter-datan ominaisuuksia tekstianalytiikan hyödyntämisen näkökulmasta. Kirjallisuuskatsaus rajattiin kuitenkin siten, että pääasiassa oltiin kiinnostuneita tieteellisessä tutkimuksessa tehdyistä yrityksistä hyötyä Twitter-datasta. Konteksti oli tekstianalytiikan käytettävyys koskien lähinnä yhteiskunnallista tai kansanterveydellistä hyötyä tuottavia tutkimuksia. Toisin sanoen, kirjallisuuskatsauksessa on keskitytty enemmänkin näitä käsitteleviin tai sivuaviin tutkimuksiin eikä niinkään esimerkiksi liiketaloudellisia analyysitarpeita koskeviin tutkimuksiin. Näin esimerkiksi tekstianalytiikan käyttö markkinoinnin kohdentamiseen on pitkälle sivuutettu.

Kirjallisuuskatsausta tehtäessä keskityttiin erittäin vahvasti nimenomaan Twitteristä saatavan tekstidatan tekstianalytiikkaan. Näin ollen materiaali, joka käsittelee yleisemmällä tasolla sosiaalista mediaa on jäänyt vähemmälle huomiolle. Päähuomio on hyvin painokkaasti asetettu koskemaan Twitteriä, Twitteristä saatavaa tekstimuotoista dataa ja lähestymistavan osalta nimenomaan tekstianalytiikkaa analyysimenetelmänä. Tämä osaltaan sulkee pois muun Twitteriä koskevan tutkimuksen (kuten käyttäjien muodostamien verkostojen tutkimisen).

Huolimatta Twitteriin liittyvän tekstianalytiikan tutkimusalan verrattain nuoresta iästä lähteitä löytyi runsaasti. Tutkija on lähteiden osalta pyrkinyt paitsi saamaan aikaiseksi kokonaisvaltaisen poikkileikkauksen erilaisista sovellutuskohteista myös rajaamaan kirjallisuuskatsauksen antia siten, että esimerkiksi hyvin samankaltaisia tutkimuksia ei ole liikaa korostettu. Tärkeäksi on koettu luoda kattava esitys erilaisista sovellutusalueista sekä lähestymistavoista Twitter-datan tekstianalytiikkaan.

Kirjallisuuskatsaus suoritettiin siten, että lähteiden alustavan etsintävaiheen jälkeen kunkin lähteen käsittelyssä nousi tärkeiksi seikoiksi kolme asiaa. Näitä olivat tutkimuskohteen ja soveltamisalan huomioiminen, käytetyt menetelmät ja lopulta se minkälaisia tuloksia tutkimuksissa oltiin saatu. Näiden kolmen näkökulman selvittäminen kustakin esimerkkitutkimuksesta oli perusteltua, jotta käytettyjä malleja voidaan menestyksekkäästi arvioida annetuin design science-arvosteluperustein (Hevner ym., 2004) sekä pohtia näiden sovellettavuutta julkishallinnon kontekstiin.

Kirjallisuuskatsauksen lähdeaineiston kerääminen tapahtui periaatteessa kahdessa osassa. Ensimmäisessä osassa suoritettiin alustava lähteiden etsintä. Lähdeaineistoa läpikäydessä tapahtui automaattisesti

lähdemateriaalin etsinnän toinen vaihe, jolloin mielenkiintoisiksi todettujen tutkimusten perusteella löydettiin uusia lähteitä. Teema, jonka pohjalta kirjallisuuskatsauksen alustava materiaalihaku suoritettiin, olivat tutkimukset joita oli suoritettu Twitterin tuottaman tekstidatan analyysistä tekstianalytiikan keinoin.

Materiaalia kirjallisuuskatsaukseen etsittiin hyödyntäen pääasiallisesti Jyväskylän yliopiston kirjaston JYKDOK-hakukantaa ja Google Scholar-palvelua. Etenkin Twitter-datan käsittelyn osalta Google Scholar antoi parhaat tulokset. Lisäksi hakuja suoritettiin Nelli-portaalin kautta tekemällä hakuja laaja-alaisesti pikahaku-toiminnon avulla. Tämän avulla oli mahdollista suorittaa hyvin laajamittaisia hakuja ripeästi. Aihepiirin taustojen osalta ei siis pyritty keskittymään aihealue-kohtaisiin hakuihin vaan hakuja suoritettiin laaja-alaisesti kaikilta aihepiirien alueilta. Tämä johti siihen, että materiaaliin valikoitui runsaasti myös materiaalia, jota ei voinut tämän tutkimuksen osalta käyttää. Tällaista olivat esimerkiksi Twitterin tuottamaa dataa toisenlaisesta analyysinäkökulmasta (esim. graphit) lähestyvät tutkimukset. Jossain määrin hakuja suoritettiin myös Tampereen yliopiston julkaisuarkiston TamPubin kautta. Lopulliseen tutkimukseen materiaalia kertyi kuitenkin TamPubista olemattomasti.

Käytettyjä hakusanoja oli huomattava määrä, sillä kirjallisuuskatsauksessa pyrittiin etsimään lähteitä monista eri näkökulmista ja monilta eri sovellutusalueilta. Täten Twitter-datan ja tekstianalytiikan teemaa lähestyttiin lukuisilla eri hakusanoilla. Tyypillisiä hakuja olivat esimerkiksi Twitter, Twitter analytics, Twitter semantics, text analytics, data mining, twitter big data, twitter analytics public governance ja sosiaalinen media. Hakutuloksia tarkasteltaessa löytyi myös etsittyjen aiheiden kannalta mielenkiintoisia lähteitä, joiden perusteella kirjallisuuskatsausta voitiin viedä eteenpäin myös perehtymällä kyseisten lähteiden lähdeluetteloon. Etsimällä lähteitä myös merkittävimpien aihepiiriä koskevien artikkeleiden lähdeluetteloiden perusteella saavutettiin myös alustava näkemys aihepiirin keskeisistä ja eniten käytetyistä artikkeleista.

Tutkimuksen empiirisessä vaiheessa toteutettiin puolistrukturoidut (Järvinen & Järvinen, 2011) teemahaastattelut (Hirsjärvi & Hurme, 2000) KELA:n ja THL:n asiantuntijoiden kanssa. Teemahaastatteluiden osalta seurattiin pitkälle Hirsjärven ja Hurmeen (2000) ohjeistuksia. Puolistrukturoitu haastattelu valittiin haastatteluiden menetelmäksi, jotta riittävä joustavuus aihepiirien ja mahdollisen uuden, yllättävän tiedon saamisen suhteen voitiin säilyttää.

Tutkimuksen empiisen vaiheen alku toteutettiin valmistamalla kohdeorganisaatioiden henkilöitä esitelmällä ja koulutustilaisuudella kirjallisuuskatsauksen tuloksista. Esitelmät ja koulutustilaisuudet toteutettiin etukäteen valmistellun esityksen avulla. Tilaisuudessa tutkimuksen tekijä esitteli tutkimuskysymykset ja lähestymistavan sekä alusti aihepiiriin. Tämän jälkeen käsiteltävänä oli Twitterin tekstimuotoisen datan analysoitavuus sekä sen hyvät ja huonot puolet. Tutkija esitti myös kirjallisuuskatsauksen aikana

esiintulleita tuloksia, esitti esimerkkejä tekstianalytiikan käytöstä selvittäen toimintaperiaatteita sekä antoi alustavia arvioita tekijöistä, jotka saattaisivat olla merkityksellisessä asemassa pohdittaessa menetelmien mahdollista hyödyntämistä julkishallinnon kontekstissa. Esitellyt tulokset koskivat Twitter-datan analysoitavuutta, hyödyntämistä erityyppisten tutkimusalojen saroilla sekä konkreettisia tuloksia. Tutkija kuvasi myös seikkaperäisesti erilaisten IT-artifaktien toimintaa sekä selitti käytettyjen mallien ja menetelmien toimintaperiaatteet. Lisäksi tutkija esitti omia arvioitaan Twitter-datan tyypillisistä hyvistä piirteistä pyrkien herättämään keskustelua sekä ajatuksia Twitterin tuottaman tekstidatan ja kohdeorganisaation tekstidatan sekä tekstianalytiikan yhtäläisyyksistä ja eroista. Kirjallisuuskatsauksessa olleista tutkimuksista esiteltiin taudinseurantamenetelmien osalta Lampos ja Cristianin (2010) flunssapisteisiin perustuva menetelmä ja Paul ja Dredzen (2011a ja 2011b) ATAM- ja ATAM+-mallit. Uutisaiheiden ja trendien etsinnän osalta käsiteltiin Mathioudakisin ja Koudasin (2010) TwitterMonitor-malli. Sävyanalyysin osalta Bollen ym. (2011) mielialatutkimus esiteltiin kuten myös Mitchell ym. (2013) onnellisuutta koskeva tutkimus. Lisäksi esimerkkinä sivuttiin sävyanalyysiin perustuen tehtyjä tutkimuksia Twitter-datan korreloinnista osakekurssien kanssa. Tämän lisäksi esiteltiin New Yorkin alueella suoritettu sävyanalyysitutkimus (Bertrand, Bialik, Virpee, Gros & Bar-Yam, 2013). Lopulta käsiteltiin EMOTIVE-projektin tutkimusta (EMOTIVE-projekti, 2013a) esimerkkinä Twitter-dataan perustuvan valvontajärjestelmän kehittämisestä.

Koulutustilaisuuden jälkeen suoritettiin itse haastattelut, joissa haastateltavat pääsivät muun muassa arvioimaan Twitter-datan tekstianalytiikassa käytettyjen IT-artifaktien sovellettavuutta oman organisaationsa tekstidataan. KELA:n tapauksessa esitelmä pidettiin Jyväskylässä KELA:n tiloissa 13.9.2013 ja haastattelut aihepiiriin liittyen suoritettiin kaksi viikkoa myöhemmin 27.9.2013. THL:n osalta esitelmä ja haastattelu suoritettiin samassa yhteydessä THL:n tiloissa Helsingissä 2.10.2013.

Haastattelutilaisuudessa tutkija esitti vastaukset tallenteelle nauhoittaen 12 kysymystä. Osa kysymyksistä oli rakennettu hyödyntäen Hevner ym. (2004) esittämiä design science-arviointiperusteita ja osa koski kohdeorganisaatioiden tekstidatan ja tekstianalytiikan nykytilaa ja tulevaisuuden visioita. Haastattelutilanteissa esitetyt kysymykset on esitetty tämän tutkimuksen yhteydessä liitteessä 1. Kysymykset 2 - 11 on rakennettu siten, että on mahdollista saada relevanttia tietoa design science-perusteiselle arviointianalyysille (Hevner ym., 2004). Toisin sanoen haastateltavia on pyydetty arvioimaan esiteltyjen menetelmien toimintaa esitellyissä yhteyksissä (observatiivinen arviointiperuste) sekä arvioimaan menetelmien ominaisuuksia (analyyttinen arviointiperuste). Lisäksi kysymyksissä on pyritty saamaan esiin informaatiota koskien potentiaalisia käyttöskenaarioita, joissa haastateltavien mielestä menetelmiä voitaisiin käyttää (deskriptiivinen arviointiperuste). Haastattelutilanteen tulokset ja haastateltavien vastauksien sisällön perusteella tehty analyysi löytyvät empiirisen osion tuloksia käsittelevästä luvusta 5.

Haastattelumateriaalin keräämisen jälkeen tulokset litteroitiin eli vastaukset kirjoitettiin puhtaaksi. Viimeisessä vaiheessa suoritettiin haastatteluiden analyysi käymällä vastaukset läpi. Näiltä osin analyysiote pohjasi sovellettuna Hevner ym. (2004, s.86) esittämiin design science-arviointiperusteisiin sekä tulkitsevaan analyysiin vastauksista. Toisaalta analyysissä tuli ottaa huomioon kirjallisuuskatsauksen pohjalta tehdyt hypoteesit, joten haastatteluissa saatua informaatiota verrattiin myös näihin.

Sovelletulla design science-näkökulmalla tarkoitetaan tässä yhteydessä sitä, että koska tutkielman yhteydessä varsinaisesti uutta IT-artifaktia ei luoda tai kokeilla uudessa kontekstissa, ei myöskään kaikkia Hevnerin ym. (2004) arviointiperusteita voida käyttää. Lisäksi arvioinnin kohteena ei tämän tutkielman osalta ole yksi tietty IT-artifakti vaan kokonainen tutkimuksen ala, jonka menetelmien soveltuvuutta toiseen kontekstiin pyritään hahmottamaan. Koska kyse on alustavasta kartoituksesta kohdeorganisaatioiden kohdalla on arviointiperusteita käytetty siinä määrin kuin se on mahdollista. Tutkittavaa kohdetta voidaan Hevner ym. (2004, s.86) mukaan arvioida observatiivisella, analyyttisellä, kokeilevalla, testaavalla ja kuvailevalla lähestymistavalla. Näistä kokeileva ja testaava arviointiperuste edellyttää, että jokin konkreettinen malli, menetelmä tai muu IT-artifakti toteutetaan, jotta sitä voitaisiin testata.

Täten arviointiperusteina on kyetty käyttämään Hevner ym. (2004, s.86) esittämien lähestymistapojen mukaisesti observatiivista, analyyttista ja deskriptiivistä näkökulmaa. Sivutuotteena haastattelututkimuksessa saadaan myös tietoa julkishallinnon kohdeorganisaatioiden tämän hetkisestä tekstianalytiikan käytöstä ja tekstimuotoisten datavarantojen määrästä ja –laadusta.

Arviointiperusteista observatiivisessa arvioinnissa (Hevner ym., 2004, s.86) tarkastellaan IT-artifaktin käyttöä erilaisissa käyttötilanteissa. Tässä tutkimuksessa arvointiperustetta on lähestytty siten, että esittelytilaisuudessa käytettyjä malleja on esitelty eri käyttötilanteissa. Tämän jälkeen haastattelutilaisuudessa on haastateltavilta saatu näkökulma mallin toimivuuteen.

Analyyttisen arvioinnin (Hevner ym., 2004, s.86) osalta arviointia suoritettiin staattisesta ja arkkitehtuurisesta näkökulmasta. Toisin sanoen arviot liittyvät käytetyn IT-artifaktin ominaisuuksiin (kuten kompleksisuuteen) ja IT-artifaktin soveltuvuuteen osaksi kohdeorganisaation IT-arkkitehtuurisia ratkaisuja.

Deskriptiivinen arviointiperuste (Hevner ym., 2004, s.86) oli toteutetun empiirisen tutkimuksen osalta hedelmällisin. Haastateltavien tehtävänä oli kuvailla erilaisia käyttöskenaarioita, joissa IT-artifaktit olisivat käyttökelpoisia. Tätä kautta voitiin saada informaatiota erilaisista skenaarioista, joissa tekstianalytiikan menetelmiä voitaisiin haastateltavien mielestä hyödyntää.

Empiirisessä vaiheessa suoritetun teemahaastattelun kysymyksistä osa oli myös rakennettu siten, että haastateltavilta saadaan vastauksia

arviointiperusteita vastaaviin teema-alueisiin. Näin teoriassa voidaan saada uutta, asiantuntija-arvioihin perustuvaa tietoa menetelmien soveltuvuudesta. Valitettavasti empiirisessä vaiheessa kuitenkin hyvin nopeasti selvisi vähäisen tekstimuotoisen datan ja tekstianalytiikan olemassaolo kohdeorganisaatioissa. Tämä vaikeutti Hevnerin ym. (2004) esittämien arviointiperusteiden suoraviivaista käyttöä. Tältä osin haastattelutilanne myös muutti tutkimuksen painotusta hieman. Koska oli selvää, ettei Twitter-datan tekstianalytiikan menetelmiä voitu soveltaa kohdeorganisaatioiden olemattomaan tekstimuotoiseen dataan menetti tämä puoli merkitystään. Näin ollen potentiaalisten käyttökohdeskenaarioiden kartoitus tulevaisuudessa korostui haastattelussa. Koska organisaatioiden omaa tekstimuotoista dataa oli vähän painottui haastattelu myös suunniteltua enemmän siihen, miten kohdeorganisaatiot – ja laajemmin julkishallinto – voivat käyttää Twitterin ja sosiaalisen median tuottamaa tekstimuotoista dataa tekstianalytiikan avulla.

Lopulta kirjallisuuskatsauksen ja koulutustilaisuuksien sekä teemahaastatteluiden ja kohdeorganisaatioiden edustajien kanssa suoritettujen keskusteluiden perusteella voitiin luoda kokonaisnäkemys, joka vastaa toiseen tutkimuskysymykseen. Toisin sanoen voitiin esittää arvioita ja näkemyksiä siitä, missä määrin Twitter-datan tekstianalytiikassa käytettyjä menetelmiä (IT-artifakteja) voidaan siirtää tai modifioida kohdeorganisaation tekstianalytiikkaan.

3 TWITTER, TEKSTIANALYTIIKKA JA TWITTER-DATAN ANALYSOITAVUUS

Tässä luvussa esitellään kirjallisuuskatsauksen osalta Twitteriä palveluna, tekstianalytiikkaa ja sen soveltamista Twitter-kontekstissa sekä käsitellään Twitterin käyttöä erilaisissa yhteyksissä. Tekstianalytiikan osalta luvussa tarkastellaan yleisemmin sosiaalisen median ja tekstianalytiikan soveltamisen aihepiiriä sekä etenkin Twitter-datan analysoitavuuteen liittyviä piirteitä.

3.1 Twitter

Twitter on sosiaalisen median mikroblogipalvelu, joka on perustettu vuonna 2006. Käyttäjiä Twitterillä tätä kirjoitettaessa on globaalisti yli 500 miljoonaa tehden siitä tällä hetkellä toiseksi suurimman sosiaalisen median palvelun Facebookin jälkeen. Joillakin maantieteellisillä alueilla ja demografisesti edustettuna Twitter on jo suurin sosiaalisen median palveluntarjoaja. Palvelu perustuu lyhyiden, maksimissaan 140-merkkisten viestien eli twiittien (engl. tweet) käyttöön. Näitä käyttäjät voivat lähettää ja edelleen jakaa sähköpostilla, sms-viesteillä tai suoraan älypuhelimista erilaisten sovellusten avulla.

Mikä tekee Twitteristä poikkeuksellisen kiinnostavan palvelun on kuitenkin sen käytäntö tiedon julkisuuden kanssa. Twiitit ovat periaatteessa julkista tietoa ja näistä voidaan muodostaa laajoja tekstimuotoisen datan varantoja tutkimusta varten. Tiedon julkisuus on poikkeava käytäntö moniin muihin sosiaalisen median palveluihin verrattuna ja niinpä tätä voidaan hyödyntää monella tavalla.

Twitterissä olevaa tekstimuotoista dataa voidaan kerätä Twitterin ohjelmointirajapintojen avulla (engl. Application Programming Interface) avulla. Näitä ohjelmointirajapintoja ovat esimerkiksi Twitter Rest API, Twitter Search API ja Twitter Streaming API. Quincey ja Kostkova (2010) kuvaavat

tyypillisen tavan päästä käsiksi Twitterin dataan. Twitterin API:en (Application Programming Interface) kautta on mahdollista hakea vapaasti tietoa twiiteistä. Niin kutsutulla Rest API:lla päästään käsiksi keskeiseen Twitter-dataan, kuten käyttäjäprofiilien informaatioon. Search API:lla voidaan sen sijaan tehdä reaaliaikaisia etsintöjä esimerkiksi hakusanoihin perustuen. Myös muita parametrejä haulle voidaan antaa, kuten esimerkiksi tuloksena palautettujen twiittien määrä. Hauissa löydetyt twiitit yhdistettynä käyttäjätietoihin (kuten paikkatiedot) ja aikamerkintään (milloin twiitti julkaistiin) palautetaan haussa atom- tai json-muodossa (xml- tai JavaScript Object Notation-muodossa). Tästä palautetusta datasta voidaan edelleen erotella mielenkiinnon kohteina olevat keskeiset asiat eri ohjelmointikielien avulla.

Datan tekstimuotoisuudesta johtuen on tekstianalytiikan soveltamisella keskeinen osa. Tämä mahdollistaa hyvin monenlaista tutkimusta esimerkiksi mielipiteiden kartoituksessa ja erilaisessa tulevien asioiden ennakoinnissa. Tutkimusta on tehty esimerkiksi politiikan (Tumasjan, ym., 2010) ja osakemarkkinoiden (Bollen, Mao & Zeng, 2011) saralla. Samoin Twitteriä voitaisiin periaatteessa käyttää erilaisten kriisitilanteiden ennakoivaan toteamiseen. Sävyanalyysia (engl. sentiment analysis) soveltamalla on ollut esimerkiksi mahdollista tehdä kollektiivista mielialaa koskevia tutkimuksia (Bollen, Pepe & Mao, 2011).

Twitter-dataan perustuva tutkimus ja Twitterin tekstimuotoiseen dataan sovelletut tekstianalytiikan menetelmät hakevat vielä jossain määrin muotojaan. Ongelmina on esimerkiksi tekstidatan määrä ja menetelmien skaalautuvuus (Aggarwal & Zhai, 2012). Mikäli kuitenkin oikeita indikaattoreita etsitään tekstianalytiikan avulla tulisi esimerkiksi tulevaisuuden ennakoinnin tietyillä toimialoilla olla mahdollista Twitter-datan pohjalta. Esimerkiksi Asur ja Huberman (2010) ovat tutkineet kuinka Twitter-viestintää analysoimalla voidaan menestyksellisesti ennakoida elokuvien lippumyyntimenestystä. Tulevaisuuden ennakoiminen hyvin monilla erilaisilla aloilla on hyvinkin mahdollista, mikäli relevantteja tekijöitä osataan etsiä tekstimuotoisen datan joukosta.

Kaiken kaikkiaan voidaan todeta, että Twitter on paitsi sosiaalisen median palveluna viestinnän väylä on se myös tutkijoiden kannalta potentiaalisesti hyvin rikkaan ja mielenkiintoisen informaation sijainti- ja levittämisväylä.

3.2 Twitterin käytön tyypillisiä piirteitä

Jotta voidaan arvioida, miten tekstianalytiikkaa voidaan hyödyntää Twitter-datan analytiikassa on perusteltua tarkastella miten ja millä tavalla Twitteriä ylipäänsä käytetään. Tässä alaluvussa käsitellään Twitterin käyttöä, twiittien tyypillistä luonnetta ja käyttäjien demografisia piirteitä.

Javan, Songin, Fininin ja Tsengin (2007) tutkimuksen mukaan ihmiset käyttävät Twitteriä pääosin puhuakseen päivittäisistä aktiviteeteistaan

ja etsiäkseen sekä jakaakseen informaatiota. Suurin osa twiiteistä käsittelee päivittäisiä rutiineja tai mitä ihmiset ovat juuri parhaillaan tekemässä (Java ym., 2007). Tyypillisiä twiiteissä esiintyviä aktiviteettejä ovat keskustelu, linkkien ja informaation jakaminen sekä uutisaiheista tiedottaminen ja niiden kommentointi. Javan, Songin, Fininin ja Tsengin (2007) mukaan Twitterin käyttäjät voivat toimia eri rooleissa paitsi informaation lähteinä myös sen etsijöinä sekä levittäjinä. Täten Twitter on paitsi eräs sosiaalisen median palvelu myös sangen optimaalinen informaation lähde. Evan Williams, eräs Twitterin perustajista, onkin todennut:

> "What we have to do is to deliver to people the best and freshest most relevant information possible. We think of Twitter as it's not a social network, but it's an information network. It tells people what they care about as it is happening in the world."
> —Evan Williams

Twitter-käyttäjät toimivat Java ym. (2007) kuvaamissa rooleissa kommentoidessaan tapahtumia ja uutisia reaaliaikaisesti. Tästä seurauksena merkittävistä tapahtumista syntyvä twiittien virta voi toimia merkittävänä uutislähteenä (Castillo, Mendoza & Poblete, 2011). Tämä on usein nopein tapa saada informaatiota kehittymässä olevasta uutisesta, tapahtumasta tai tilanteesta (Mills, Chen, Lee & Rao, 2009). Perinteisistä informaation lähteistä Twitterin erottaa myös käyttäjien usein twiiteissään esiin tuoma henkilökohtainen mielipide tai tunnetila twiitin kontenttia koskien (Phuvipadawat ja Murata, 2010). Vertailtaessa Twitteriä ja perinteistä mediaa havaittiin, että Twitter voi olla erityisen hyvä media aiheille, joilla on vähäinen huomio perinteisessä uutismediassa (Zhao, Jiang, Weng, He & Lim, 2011). Yleisesti ottaen Twitterin käyttäjät ovat verrattain vähän kiinnostuneita maailmanlaajuisista uutisista mutta auttavat kuitenkin aktiivisesti levittämään uutisdataa tärkeistä tapahtumista twiittaamalla (Zhao ym., 2011). Nämä tosin saattavat poiketa perinteisen uutisoinnin valtavirrasta sen mukaan mitä käyttäjät kokevat tärkeäksi jakaa. Kenties kuvaavaa onkin, että 160 miljoonan twiitin koeaineistolla suoritetun tapahtumia ja uutisaiheita koskevan tutkimuksen mukaan julkisuuden henkilöiden kuolemat ovat nopeimmin leviäviä uutisia Twitterissä (Petrović, Osborne & Lavrenko, 2010).

Phuvipadawatin ja Muratan (2010) tutkimuksessa huomioidaan twiiteissä olevan kaksi tärkeää sisällöllistä aspektia. Näitä ovat ensiksikin tunteellisesti värittynyt sisältö ja toiseksi faktatiedot. Toisin sanoen ihmiset esittävät mielipiteitään sekä tunteitaan twiittien asiayhteyksien lomassa. Tätä voidaan hyödyntää sävyanalyysilla (engl. sentiment analysis). Tunnekontentti korostuu twiiteissä esimerkiksi eri symbolien (hymiöt, huutomerkit) avulla sekä käyttämällä vahvoja tunnelatauksia omaavia ilmauksia ja sanoja (Phuvipadawat ja Murata, 2010). Faktatietokontentti taasen välittyy tekstidatan, linkkien sekä tiedon lähteeseen viittaavien elementtien kautta. Phuvipadawat ja Murata (2010) toteavat tekstimuotoisen datan olevan hyvin tärkeää uutisten löytämisen kannalta. Tekstiosuudesta ovat analysoitavissa uutisen keskeiset

tiedot, kuten mitä tapahtuu, missä ja milloin. Käyttäjät myös usein helpottavat tekstiosuuden analysointityötä varustamalla twiittejään hashtag-symbolilla (#) ja keskeisellä avainsana-kuvauksella twiitin sisällöstä (Hu & Liu, 2012). Tämä osaltaan tekee tekstidatasta tietyllä tavalla rakenteellisempaa ja helpottaa analysointia.

Twitterin käyttöä voidaan myös tarkastella kuluttajien välisen tiedonsiirron välineenä. Jansen, Zhang, Sobel ja Chowdury (2009) ovat tutkimuksessaan tarkastelleet Twitteriä elektronisena kuluttajilta kuluttajille tapahtuvan markkinoinnin mediana - vapaasti suomenkieliseksi käännettynä siis puskaradiona. Tämä word of mouth-tyyppinen tieto voi olla tärkeää esimerkiksi yrityksille, jotka markkinoivat tuotteita Twitterin välityksellä. Voidaan puhua elektronisesta word of mouth-markkinoinnista, jonka päätehtävinä on tarkkailla kuluttajien kulutustottumuksia kuluttajilta kuluttajille tapahtuvan viestinnän avulla ja kyetä vaikuttamaan näihin markkinoinnin keinoin (Jansen ym., 2009). Tässä yhteydessä on mahdollista käyttää tekstianalytiikkaa esimerkiksi twiittien tuotekohtaisten tietojen sävyn ja positiivis-negatiivisen kontentin löytämiseen. Jansenin, Zhangin, Sobelin ja Chowduryn (2009) tutkimuksen otoksessa 19% twiiteistä pitää sisällään jonkin brandin mainninnan. Näistä brandia koskevista twiiteistä lähes 20% kuvasi jotain tuntemuksia. Näistä tuntemuksista 50% oli positiivisia ja 33% negatiivisia. Toisin sanoen sävyanalyysillä voitiin saada tuloksia tuotemerkkejä koskien. Koska Jansen ym. (2009) havaitsivat Twitterissä käytetyn kielen olevan hyvin samantapaista kuin arkikielenkäytön tulivat he johtopäätökseen siitä, että nykyisin Twitter on eräänlainen elektorinen vastine kuluttajien puskaradiolle (Jansen ym., 2009). Mikäli twiittien mielipiteitä kuvastava kontentti on lähimaillakaan Jansenin ym. (2009) otoksen määriä tulisi olla hyvin mahdollista soveltaa tätä elektonista puskaradiota myös julkishallinnon osalta. Tämän olettamuksen perusteella Twitterin tuottama tekstidata voisi parhaimmillaan tarjota reaaliaikaisen palautekanavan esimerkiksi hallinnollisia uudistuksia koskien.

Suomen osalta aktiivisten Twitter-käyttäjien määrä tällä hetkellä on sangen riittämätön tilastollisessa mielessä kuvaamaan koko kansakuntaa populaationa. Twitter ei itse julkaise aktiivisesti suomalaisia käyttäjiä koskevia tilastoja. Suomenkielisten twiittien julkaisemista seuraamalla on voitu yksityisten tahojen tutkimuksissa kuitenkin päätyä arvioon käyttäjämääristä. Esimerkiksi helmikuussa 2013 arvioitiin aktiivisia suomalaisia Twitter-käyttäjiä olevan noin 26000 (Twittercensus, 2013). Vaikka arvioon tuleekin suhtautua varauksella lienee se oikeassa suuruusluokassa.

Mahdollisen tulevan ja laajamittaisemman Twitter-dataan pohjautuvan tutkimustarpeen osalta on tärkeää hahmottaa joitain Twitterin käyttäjiin liittyviä demografisia taustatekijöitä. Koska Twitter-data tuskin koskaan on tilastotieteelliseltä kannalta täydellinen läpileikkaus populaatiosta on tärkeää tietää, missä potentiaaliset eroavaisuudet ilmenevät. Twitter-käyttäjien demografisista ominaisuuksista ei tutkimusta löytynyt Suomen osalta, joten kirjallisuuskatsauksessa on näiltä osin huomioitu Yhdysvaltoja

koskeva tutkimus. Tätä tutkimusta silmällä pitäen voitaneen tehdä joitain alustavia arvioita myös suomalaisten Twitterin käyttäjien demografisista ominaisuuksista tulevaisuudessa Twitterin käytön yleistyttyä.

Mislove, Lehmann, Ahn, Onnela ja Rosenquist (2011) käsittelevät tutkimuksessaan yhdysvaltalaisten Twitterkäyttäjien demografisia ominaisuuksia. Mielenkiinnon kohteena oli, ovatko Twitterkäyttäjät otoksena kelvollinen ja kuvaava otos yhteiskunnasta ja mikäli ei, niin mitkä demografiset tekijät ovat yli- ja aliedustettuina Twitterkäyttäjien joukossa. Tutkimuksen tarkoituksena oli näitä säännönmukaisuuksia löytämällä helpottaa tulevaisuudessa Twitter-dataan perustuvan analytiikan käyttöä siten, että nämä demografiset vääristymät voidaan huomioida.

Tutkimusaineistona käytettiin 1,755 miljardia twiittiä 55 miljoonalta käyttäjältä vuosilta 2006-2009. Karsittuna esimerkiksi paikkatietojen saatavuuden avulla materiaali kutistui siten, että edustettuina oli noin 3 miljoonaa käyttäjää, toisin sanoen reilu 1% Yhdysvaltojen väestöstä. Tutkimuksessa keskityttiin tarkastelemaan sitä, missä Twitterin käyttäjät maantieteellisesti sijaitsevat, kuinka paljon naisia ja miehiä käyttäjien joukossa on ja millainen jakauma käyttäjien joukossa etnisesti on. Mislove ym. (2011) mainitsevat myös, että ihannetilanteessa muitakin ominaisuuksia tulisi huomioida. Tällaisia olisivat esimerkiksi sosio-ekonominen status, käyttäjien koulutustaso sekä työpaikkatietoihin liittyvä data. Twitterkäyttäjien maantieteellinen sijoittuminen saatiin selvitettyä sangen suoraviivaisesti. Sukupuolen osalta tutkijat selvittivät 5836 etunimen kirjaston avulla käyttäjätietojen pohjalta käyttäjien sukupuolta. Vastaavasti etnisyyttä pyrittiin arvioimaan käyttäjien antamien sukunimi-tietojen pohjalta. Luonnollisesti näitä lähestymistapoja voidaan kritisoida mahdollisesta alttiudesta virheille, mutta otoksen huomioon ottaen tulosten täytynee olla vähintäänkin hyvin suuntaa-antavia.

Johtopäätöksenä tutkimuksesta Mislove ym. (2011) toteavat, että asutuimpien alueiden käyttäjät ovat yliedustettuina verrattuna kansakuntaan populaationa. Tämä ero oli vielä oletettuakin suurempi ja osoittaa, että potentiaalisissa Twitter-datan analytiikkaan perustuvissa tuloksissa voi ilmetä vääristymää koskien hyvin asuttuja ja vähän asuttuja seutuja. Lisäksi havaittiin miesten olevan keskimäärin aikaisempia Twitterin käytön omaksujia. Toisin sanoen miesten julkaisemat twiitit olivat keskimäärin yliedustettuina, joskin ajan myötä tämä ero pienenee myös naisten omaksuessa Twitterin välineenä. Etnisten tietojen pohjalta ei voitu tehdä suoria johtopäätöksiä mutta suuntaviivaa antavana havaittiin esimerkiksi kaukasialaisten henkilöiden yliedustus suurissa kaupungeissa. Tutkimuksesta ei käy ilmi käyttäjien ikään perustuvaa ryhmittymistä. Kuitenkin myös riittävän kuvaavan otoksen saaminen kaikista ikäryhmistä on koettu joissakin tutkimuksissa ongelmalliseksi (Paul & Dredze, 2011b).

Suomeen ja suomalaisten tulevaisuuden Twitter-käyttäjien kohdalle sovellettuna tutkimuksen tulokset etnisyyden osalta eivät liene kovinkaan mielenkiintoisia – yhtymäkohtia Yhdysvaltojen tilanteeseen ei

juurikaan ole. Sen sijaan voitaneen varovaisesti olettaa, että samankaltaiset trendit asutuskeskusten ja sukupuolijakauman kohdalla toteutuvat. Toisin sanoen Suomea koskevassa Twitter-datan analytiikassa tulisi mahdollisesti huomioida alusta pitäen nämä yli- ja aliedustukset populaatiossa, jotta tilastollisesti päteviä arvioita koko populaatiosta voitaisiin tehdä.

Organisaatioiden käytössä Twitter on usein jätetty yksisuuntaisen viestinnän työvälineen asemaan. Lovejoy, Waters ja Saxtonin (2011) mukaan voittoon pyrkimättömät organisaatiot eivät hyödynnä Twitteriä lähellekään optimaalisella tavalla sidosryhmien osalta. Lovejoy ym. (2011) mukaan sosiaalista mediaa käytetään edelleen pääosin yhdensuuntaisena viestintäväylänä. Hyödyntämällä Twitteriä paitsi viestinnässä myös tekstianalytiikan osalta vuorovaikutteisemmin olisi mahdollista rakentaa kokonaan uusia julkishallinnon alaisia palvelukonsepteja. Twitteriä kyettäisiin käyttämään esimerkiksi tulevaisuudessa monisuuntaisessa viestinnässä terveydenhuollon parissa potilaiden ja terveydenhuollon instanssien välillä. Alustavia kokeiluja tällä saralla on jo tehty (Hawn, 2009). Viestinnän suhteen vastavuoroisuuden ja vuorovaikutuksen määrä voisi siis olla merkittävämpi sen sijaan, että pitäydytään perinteisessä yhdensuuntaisessa viestinnässä. Tässä voidaan ajatella myös Twitter-datan tekstianalytiikalla olevan sijansa työkaluna, joka voi tuottaa merkityksellistä informaatiota takaisin päin paitsi sidosryhmiltä myös muilta käyttäjiltä. Konkreettisena esimerkkinä julkishallinnon organisaatioiden kannalta voisi olla palautteen saaminen epäsuorasti tekstianalytiikkaan pohjautuvan palautejärjestelmän kautta. Tällaiselle voisi olla käyttöä esimerkiksi julkishallinnon laajan mittakaavan muutoshankkeissa.

Tässä alaluvussa käsiteltiin joitakin Twitterin käytön tyypillisiä piirteitä. Ihmisten todettiin käyttävän Twitteriä keskusteluun ja linkkien, informaation ja uutisten jakamiseen (Java, Song, Finin & Tseng, 2007). Käyttäjien jakamat twiitit ovat luonteeltaan informatiivisia sekä todellisen tiedon suhteen että käyttäjien mielipiteitä ja tunteita heijastelevana (Phuvipadawat ja Murata, 2010). Johtuen näistä piirteistä voidaan Twitter nähdä myös väylänä, jossa käyttäjät kertovat mielipiteitään tuotteista ja palveluista (Jansen ym., 2009). Demografisten tekijöiden kannalta merkille pantavaa on lähinnä hyvin asuttujen alueiden yliedustus Twitter-datassa ja kenties miesten aiempi Twitterin omaksuminen (Mislove ym., 2011).

3.3 Tekstianalytiikka ja sosiaalinen media

Tässä tutkielmassa keskitytään Twitterin tuottaman tekstimuotoisen datan tekstianalytiikan menetelmien tarkasteluun. On kuitenkin perusteltua tarkastella tekstianalytiikan soveltamista laajemminkin sosiaalisen median näkökulmasta. Tässä alaluvussa on tarkasteltu tekstianalytiikan soveltamista yleisesti sosiaalisen median kontekstissa.

Tekstianalytiikalla (engl. text analytics, text mining) tarkoitetaan tietämyksen löytämistä ja hankkimista tekstimuotoisista datavarannoista (Stavrianou, Andritsos & Nicoloyannis, 2007). Tekstianalytiikan tehtävä on löytää kaavamaisuuksia datasta eli siis tuottaa syvällisempää tietoa tekstidatan merkityksestä (Aggarwal & Zhai, 2012). Tyypillisiä tekstianalytiikan työvälineitä ovat esimerkiksi ohjattu ja ohjaamaton koneoppiminen, erilaiset tilastolliset ja todennäköisyyksiin perustuvat lähestymistavat ja erilaiset luokittelu- sekä indeksointimenetelmät.

Tekstianalytiikka on saanut paljon huomiota viime aikoina, koska sosiaalisen median tuottama strukturoimaton tekstimuotoinen data on määrällisesti kasvanut valtavasti (Aggarwal & Zhai, 2012). Ripeästi kehittyneet IT-teknologiat esimerkiksi tiedon siirtämisen, käsittelyn ja tallentamisen osalta ovat vaikuttaneet tähän ratkaisevasti. Sosiaalinen media avaa mahdollisuuden tutkia tekstimuotoista dataa täysin uudessa mittaluokassa. Tästä johtuen datan määrä on haaste tekstianalytiikan soveltamiselle (Barbier & Liu, 2011). Sosiaaliselle medialle tyypillinen tekstin muoto asettaa myös omia vaateitaan tekstianalytiikalle. Esimerkiksi tekstiosuuksien lyhyys asettaa haasteita tekstianalytiikan soveltamiselle (Hu & Liu, 2012). Toisaalta sosiaalisen median palveluille tyypillinen vapautunut kielen käyttäminen, epätyypillisten ilmauksien käyttö ja huumori aiheuttavat tulkinnallisia ongelmia automaattisen analytiikan kannalta (Kaufman & Kalita, 2010). Lisäksi datassa sisältää runsaasti välimerkkejä sekä semanttisen tulkinnan kannalta merkityksettömiä ja usein toistuvia stop word-sanoja (Barbier & Liu, 2011). Tämä luo osaltaan mahdollisuuksia hyvin monien eri alojen tutkimukselle mutta myös tarpeen kehittää edelleen tekstianalytiikan tekniikoita (Aggarwal & Zhai, 2012).

Tekstimuotoisen datan runsas määrä ja saatavuus johtavat siihen, että tekstianalytiikkaa tulee kyetä tekemään myös entistä dynaamisemmalla (Aggarwal & Zhai, 2012) ja skaalautuvammalla tavalla (Hu & Liu, 2012). Lisäksi sosiaalisen median osalta Aggarwalin ja Zhain (2012) mukaan tarvitaan enenevissä määrin yhdistelmää tekstianalytiikasta ja multimediasisältöjä ja konteksteja ymmärtävästä analytiikasta. Myös Hu ja Liu (2012) huomioivat, että tulevaisuudessa on tarpeellista löytää tehokkaampia keinoja tekstianalytiikan ja muiden mediamuotojen analytiikan yhdistämiselle. Tällä hetkellä esimerkiksi Twitter-datan osalta tämä voidaan tehdä lähinnä tunnistamalla tekstianalytiikan avulla mahdolliset ulkoisiin lähteisiin ja kuviin viittaavat linkit.

Aggarwalin ja Zhain (2012) mukaan keskeisiä avainongelma-alueita ja toisaalta soveltamistapoja tekstianalytiikalle sosiaalisen median kontekstissa ovat muun muassa seuraavat.

- Syvällisemmän informaation luominen tekstidatan entiteettejä ja niiden välisiä suhteita kuvaamalla
    - Tarvitaan uusia mallinnustekniikoita
- Tekstidatan yhteenvetotekniikoiden kehittäminen

- Ohjaamattoman oppimisen tekniikoiden, ryvästämisen (engl. clustering) ja aihepiirianalyysin (engl. topic modeling) hyödyntäminen ja kehittäminen.
    - Laajasti sovellettavissa
    - Ei tarvita harjoitusaineistoa kuten ohjatussa oppimisessa
- Tekstidatan indeksointimenetelmien kehittäminen
- Ohjatun oppimisen tekniikoiden kehittäminen
- Kielten välisen tekstianalytiikan kehittäminen
- Käytetyn kielen kielellisten ominaisuuksien tuottamien ongelmien huomiointi
- Todennäköisyyksiin perustuvien analytiikan menetelmien kehittäminen

Aggarwalin ja Zhain (2012) mukaan haaste on myös parantaa analysoitavan tekstin ymmärrettävyyttä. Analytiikkaa ajatellen tämä tarkoittaa siirtymistä enemmän tekstin semantiikkaa ja sisältöä arvioivaan lähestymistapaan. Tämä taasen vaatii uusia tapoja mallintaa oleellista kontenttia hyvin lyhyistä ja arkikielellä kirjoitetuista tekstipätkistä. Aggarwal ja Zhai (2012) korostavat myös kontekstin parempaa huomioimista. Koska tekstidata on yleensä jollain tavalla liitoksissa erilaisiin konteksteihin tulee nämä saada tavalla tai toisella paremmin huomioitua analytiikassa. Tekstianalytiikan rinnakkaisajamisen merkitys kasvanee Aggarwalin ja Zhain (2012) arvion mukaan tulevaisuudessa. Datan suuresta määrästä johtuen voi olla tarpeen kehittää menetelmiä, joilla tekstianalytiikkaa voidaan ajaa rinnakkain. Tällaista voisi olla esimerkiksi ohjattujen ja ohjaamattomien oppimisen algoritmien rinnakkaisajo.

Hu ja Liu (2012) huomioivat viestien lyhyyden aktivoivan ihmisiä osallistumaan ja ottamaan kantaa. Tämän vuoksi lyhyet viestit ovat tehokkaita sosiaalisessa mediassa. Tämä asettaa haasteita tekstianalytiikan osalta esimerkiksi tekstien luokittelulle ja sävyanalyysille. Koska pidemmässä tekstissä on paljon enemmän sanoja, samat menetelmät suoraan siirrettynä eivät välttämättä päde. Toisena keskeisenä erona Hu ja Liu (2012) mainitsevat sosiaalisen median tekstidatan luonteen. Käytetty kieli vaihtelee siten, että tekstin taso on hyvin vaihtelevaa. Koska käyttäjiä on hyvin erilaisia, vaihtelee käytetty kieli ja ilmaisutapa sekä sisältö paljon voimakkaammin kuin normaalissa tekstidatasssa. Tämä vaikeuttaa esimerkiksi tekstin semanttista arvottamista**.**

Usein käyttäjät voivat myös leikitellä sanoilla ja ilmauksilla tai jopa mahdollisesti keksiä kokonaan uusia ilmauksia (Hu & Liu, 2012). Tällaisia voisivat olla esimerkiksi ”How r u? tai ”I h8 dis phone”. Ihmiset kykenevät intuitiivisella tasolla tulkitsemaan nämä, mutta tekstianalytiikan kannalta tämä on hyvin haastavaa. Toisaalta sosiaalisen median tekstidatassa voi olla tekstianalytiikkaakin helpottavia ylimääräisiä tekijöitä (Hu ja Liu, 2012). Esimerkiksi Twitterin tapauksessa tämmöinen on hashtagien käyttö, joka tietyissä tilanteissa helpottaa tekstianalytiikkaa.

Yhteenvetona sosiaalisen median dataan tekstianalytiikkaa sovellettaessa huomioitavia tekijöitä ovat siis yksittäisten tekstiosuuksien lyhyys ja arkityyppinen kielellinen ilmaisu (Hu & Liu, 2012) sekä runsas stop word:ien määrä datassa (Barbier & Liu, 2011). Koska sosiaalinen media kokonaisuudessaan on merkittävä ja kasvava strukturoimattoman tekstimuotoisen datan lähde on syntynyt vaatimukset siitä, että tekstianalytiikkaa tulisi kyetä soveltamaan dynaamisemmin, skaalautuvammin ja reaaliaikaisemmin. Entiteettien tunnistaminen kohdetekstistä ja niiden välisten merkityssuhteiden kuvaaminen korostuvat kuten myös erilaiset inhimillistä päätöksentekoa helpottavat yhteenvetotekniikat (Aggarwal & Zhai, 2012). Lisäksi tekstianalytiikan suhteen tekstiä koskevan kontekstin huomioiminen korostuu. Yleisesti ottaen tekstianalytiikan menetelmien tulisi siis olla mahdollisimman yksinkertaisia ja skaalautuvia. Mikäli Twitterin tapauksessa seurataan reaaliaikaisesti twiittivirtaa korostuu sen merkitys, että dataa haetaan vain kerran ja että käytetty malli on rakenteeltaan yksinkertainen.

3.4 Twitter-datan analysoitavuus

Tässä alaluvussa käsitellään Twitter-datan analysoitavuuteen liittyviä tekijöitä. Pyrkimyksenä on kirjallisuuskatsauksen lähteiden avulla arvioida, millaista Twitterin tuottama tekstimuotoinen data analysoitavuuden kannalta on.

Twitteristä saatavassa tekstidatassa on sekä hyviä että huonoja puolia analysoitavuuden kannalta. Huonoja asioita ovat twiittien verrattain pieni koko (140 merkkiä) ja monimerkityksiset viestit, kuten ihmisten käyttämä ironia ja huumori. Tämä omalta osaltaan usein pakottaa käyttämään jonkin asteista käsiteanalyysia tai mielipidelouhintaa (engl. sentiment analysis) osana analytiikkaa. Ehdottomia hyviä puolia verrattuna esimerkiksi hakukonepalveluiden tai muiden sosiaalisen median palveluiden tuottamaan dataan on datan julkisuus ja se, että käyttäjien ilmoittamat maantieteelliset sijainnit eli geolokaatiotiedot ovat saatavilla (Ji, Chun & Geller, 2012). Huono puoli geolokaatiotietojen osalta on, että näiden tietojen paikkansapitävyys ei usein ole kovinkaan varmaa ja kuvitteellisia, ei-todellisia paikkatietoja käytetään yleisesti. Mobiilikäyttäjiltä geolokaatiotiedot kuitenkin saadaan täsmällisesti. Vaikka oikea paikkatieto olisikin annettu, vaikuttaa siltä, että käyttäjillä on taipumus usein ilmoittaa paikkatietonsa sangen karkealla tasolla, esimerkiksi harvoin tarkemmin kuin yksittäisen kaupungin tarkkuudella (Hecht, Hong, Suh & Chi, 2011). Toisin sanoen paikkatietojen granulariteetti jää usein karkeaksi, mikä on havaittu ongelmalliseksi. Tämä vaikeuttaa käytännössä esimerkiksi tautien edistymisen visualisointeja karttapohjalla. Lisäksi Twitter-dataa analysoitaessa on otettava huomioon, että Twitterin käyttäjät muodostavat suurelta osin homogeenisen ryhmittymän, joka ei ole välttämättä tiukan tilastotieteellisesti kuvaava otos esimerkiksi jonkin tietyn alueen koko väestöstä. Tähän vaikuttaa luonnollisesti Twitterin käyttösuosio alueellisesti, mutta myös tyypillisesti käyttäjät ovat ryhmittyneet siten, että

edustavaa tietoa ei saada esimerkiksi kaikista ikäryhmistä (Paul & Dredze, 2011b).

Twiittien eräs ongelma analyysikohteena on se, että Twitter-datasta suuri osa on automaattisen luokittelun kannalta vaikeasti hahmotettavissa merkitystensä puitteissa. Ihmiset viljelevät usein monimerkityksisiä ilmauksia, huumoria ja ironiaa, joten tarvitaan menetelmiä päästä käsiksi twiittien todelliseen merkitykseen (González-Ibáñez, Muresan & Wacholder, 2011). Datan luotettavuus voidaan myös joissakin tapauksissa kokea ongelmalliseksi, kuten esimerkiksi käyttäjien ilmoittamien paikkatietojen osalta (Earle, Bowden & Guy, 2012).

Twiittien tulkittavuutta sävyjen kannalta voi lisätä twiittien sisältämät hymiöt. Muun muassa Go, Bhayni ja Huang (2009) esittelevät menetelmän, jossa hymiöitä käytetään koulutusdatana ohjatun oppimisen lähestymistavassa. Tutkimuksessa osoitetaan, että tällä tavalla ohjattu algoritmi saavutti korkean tarkkuuden arvioitaessa twiittien sentimenttiä. Barbosa ja Feng (2010) esittelevät tutkimuksessaan menetelmän, jolla twiittien syvempää merkitystä voidaan automaattisesti arvottaa. Aiemmin viestien sävyn analyysissa (engl. sentiment analysis) on paljon käytetty muun muassa n-grammeja merkityssuhteiden hahmottamiseen. Barbosa ja Feng kuitenkin huomioivat, että Twitter-viestien kohdalla twiittien lyhyt mitta, vain 140 merkkiä maksimissaan, asettaa rajoituksensa sisällön tulkintaan. Tutkimuksessa esitellään mahdollisena parempana tekniikkana uusi malli twiittien sävyn analyysiin. Mallissa hyödynnetään analyysitekniikoina algoritmin valvottua oppimista lähdeaineistosta ja luokittelumallia, jossa twiitit karkeasti arvioidaan subjektiivisiksi ja objektiivisiksi sekä positiivisiksi ja negatiivisiksi. Lopputuloksena Barbosa ja Feng (2010) esittävät menetelmän toimivan kohtuullisen hyvin ja olevan täten vaihtoehto perinteisille metodeille.

Toisaalta Zhang, Fuehres ja Gloor (2011) ovat tutkimuksessaan ottaneet kantaa siihen, että twiittien lyhyys omalta osaltaan auttaa kontentin tulkinnassa. Heidän mukaansa johtuen twiittien lyhyestä maksimimitasta (140 merkkiä) sekä käytetystä twiittien keskimitasta (11 sanaa O'Connor ym., 2010 mukaan) voidaan twiittien keskeinen sisältö sekä sentimentti löytää usein muutaman avainsanan avulla. Esimerkkinä tällaisista ilmauksista ja lauserakenteista voisi olla ilmaukset kuten ”I am feeling” ja ”I feel” yhdistettynä tiettyihin avainsanoihin ja –adjektiiveihin. Näiden perusteella voidaan hyvinkin lyhyistä twiiteistä hahmottaa keskeinen sisältö tehokkaasti ja tätä lähestymistapaa onkin joissain tutkimuksissa käytetty (esimerkiksi Bollen, Mao & Zeng, 2011. Toisin sanoen semanttisen analyysin kannalta twiittien lyhyys voi olla myös hyvä puoli.

Ei-täsmälliset twiittien geolokaatiotiedot ovat tutkimuksien kannalta ongelmallisia (Earle, Bowden & Guy, 2012). Hecht ym. (2011) on tutkinut perinteisten ainoastaan tietokonetta käyttävien Twitterin käyttäjien ilmoittamien paikkatietojen hyödyntämisen problematiikkaa. Perinteisesti on ajateltu paikkatietojen olevan suhteellisen puhdasta dataa ilman vääristeltyjä tietoja. Hecht ym. (2011) tutkimuksessa kuitenkin todetaan, että 34% käyttäjistä

ei antanut todellista paikkatietoa vaan tämän sijasta käytettiin kuvitteellisia paikan nimiä tai sangen yleisesti sarkastisia kommentteja. Tällaisia ovat esimerkiksi paikkatietoilmaukset kuten ”on the dark side of the moon” tai ”behind you”. Lisäksi suurin osa käyttäjistä ei kertonut paikkatietojaan sen tarkemmalla tasolla kuin missä kaupungissa asuvat. 18% käyttäjistä ei ilmoittanut paikkatietoja laisinkaan (Hecht ym., 2011). Vastaavasti Misloven ym. (2011) tutkimuksessa 75,3% julkisista käyttäjistä listasi vapaaehtoisesti paikkatietonsa. Tämänkin jälkeen jää esille kysymys käyttäjien ilmoittamien tietojen luotettavuudesta. Hecht ym. (2011) toteaakin perinteisten paikkatietojen analyysissä käytettyjen työkalujen olevan riittämättömästi varusteltuja tämän kaltaisen datan käsittelyyn. Tutkimuksessa testattiin koneoppimisen avulla voidaanko käyttäjien sijainti tällaisissa tapauksissa löytää pelkästään käyttäjän twiittien perusteella. Datan analyysitekniikoina käytettiin koneoppimista ja erilaisia luokittelutekniikoita esimerkiksi käyttäjien jakamiseen aktiivisiin ja passiivisiin. Tutkimuksen otoksena käytettiin 32 miljoona twiittiä noin 5 miljoonalta eri käyttäjältä. Näiden joukosta sattumanvaraistettiin 10 000 aktiivikäyttäjän otos, jota tarkasteltiin. Tutkimuksen tuloksena Hecht ym. (2011) toteaa, että aktiivisen käyttäjän maa ja osavaltio voidaan helposti ja verrattain tarkasti paikantaa pelkästään käyttäjän twiittien perusteella. Toisin sanoen periaatteessa käyttäjät paljastavat sijaintinsa twiiteissänsä joka tapauksessa. Tällä voi olla jatkossa käytännön sovellutuskohteita Twitter-datan analyysissä. Toisaalta Hecht ym. (2011) mainitsevat potentiaaliset yksityisyyteen liittyvät ongelmat menetelmän soveltamisessa. On kuitenkin otettava huomioon, että maa ja osavaltiokohtainen paikkatieto ei ole kovinkaan riittävä useiden sovellutuskohteiden kannalta.

Cheng, Caverlee ja Lee (2010) esittävät vastaavia tuloksia tutkimuksessaan, joka antaa tukea olettamukselle, että geolokaatiotietoja käyttämättömän aktiivisen Twitter-käyttäjän sijainti voidaan kohtuullisella tasolla ratkaista pelkästään käyttäjän twiittien perusteella. Tutkimuksessa kehitetään todennäköisyyksiin perustuva paikantamisjärjestelmä, jonka lähteenä ovat pelkästään käyttäjän twiitit. Menetelmä etsii vahvasti geolokaatioon sidottuja sanaja twiiteistä. Tämän pohjalta rakentuu todennäköisyyksiin perustuva esitys käyttäjän sijainnista. Tuloksena Cheng ym. (2010) esittävät menetelmän arvioivan sijaintitiedot sangen hyvin satojen twiittien perusteella. Noin 51% Twitter-käyttäjistä kyetään paikallistamaan 100 mailin tarkkuudella heidän todellisesta sijainnistaan.

Eräs vaihtoehto käyttäjien paikkatiedon ongelmaan on käyttäjän antaman paikkatiedon hyödyntäminen Google Maps-ohjelmointirajapinnan (Google Maps API) avulla koordinaattien saamiseksi. Näin on menetelty esimerkiksi Earlen ym. (2012) maanjäristys-kartoitustutkimuksessa. Vaihtoehtoisesti voidaan myös karsia epävarmat geolokaatiotiedot kokonaan pois tutkimusaineistosta. Mobiilikäytön ja täten täsmällisen geolokaatiodatan lisääntyessä tämä ongelma on kuitenkin vähenevissä määrin rajoittava. Tämä mahdollistaa kokonaan uusia sovellutusalueita, joissa geolokaatiotiedon täsmällisyys ja hieno granulariteetti on avainasemassa. Esimerkiksi Twitterin

käyttöä potentiaalisena infrastruktuurina älypuhelimien lähettämille sensorisille tiedoille on tutkittu (Demirbas, Bayir, Akcora, Yilmaz & Ferhatosmanoglu, (2010). Mallinnettaessa huomattavien sensorimäärien välittämää tietoa alueellisesti on erittäin tärkeää täsmällisten paikkatietojen saaminen.

Hashtagien (engl. hashtag, merkitään #-merkillä) avulla Twitterin tekstidatasta voidaan tehdä tekstianalytiikan tiedonhaun kannalta tietyllä tavalla rakenteellisempaa. Käyttäjät merkitsevät usein twiittejään hashtageilla, jotka sisältävät hyvin lyhyen kuvauksen aihepiiristä, jota twiitti koskee. Tämä mahdollistaa hakujen tekemisen hashtagien perusteella. Joissakin tutkimuksissa hashtagien etsintää hyödyntämällä on kyetty helpottamaan lähdeaineiston keräysvaihetta merkittävästi (Conover ym., 2011b). Hashtagien käyttö antaa jo itsessään karkean arvion twiitin sisällöstä vaikkei varsinaisesti paljastakaan paljoa twiitin sävystä. Lovejoy, Waters ja Saxton (2011) antavat esimerkiksi hakusanojen ”healthcare” ja ”#healthcare” käytön terveydenhuoltoa koskevan tutkimuksen kannalta. Ensimmäinen tuottaisi tuloksia, mutta nämä eivät välttämättä ole relevantteja terveydenhuollon kannalta. Ihmiset saattavat twiitata täysin toisesta aiheesta pelkästään mainiten terveydenhuollon. Sen sijaan toisen, hashtagillisen hakusanan käyttö tuottaisi relevantteja hakutuloksia, koska käyttäjät ovat twiitit näin merkinneet.

González-Ibáñez, Muresan ja Wacholder (2011) ovat tutkineet sarkasmin tunnistamisen ongelmallisuutta twiiteistä. Tutkimuksen päätarkoituksena oli hahmottaa sarkasmin erottamista aidoista positiivista tai negatiivistä sävyä sisältävistä twiiteistä. Tutkimuksessa hyödynnettiin lähdeaineistona kokoelmaa sarkastisista ilmauksista. Sarkastisten ilmauksien kirjasto kerättiin etsimällä twiittejä #sarcasm-tyyppisten hashtag-hakujen avulla. Näin saatiin koottua kokoelma sarkasmia sisältäviä twiittejä. Kokoelmaa käytettiin koulutusaineistona ja ohjattua koneoppimista hyödynnettiin tuloksien saamiseksi. Lopuksi tuloksia verrattiin ihmisten tekemiin tulkintoihin sarkasmista saman aineiston osalta. González-Ibáñez ym. (2011) mukaan koneoppimisen menetelmillä saadut tulokset eivät olleet hyviä. Toisaalta myöskin ihmiset kokivat sarkasmin arvioinnin hyvin ongelmalliseksi. Ihmisiltä saadun palautteen perusteella ongelmallisiksi koettiin se, että viestit eivät olleet yhteydessä mihinkään tiettyyn kontekstiin. Lisäksi ongelmalliseksi koettiin se, että usein ilmausten sarkasmin ymmärtämiseksi täytyi olla tietoinen jostain viime aikoina tapahtuneesta asiasta. Tämä ei ole kovinkaan rohkaisevaa automaattiseen pelkästään twiitteihin perustuvaan sarkasmin tunnistamiseen. Mikäli ihmisten oli vaikea tunnistaa tätä ilman konteksti- ja ajankohtaistietoja, on se automaattisesti koneoppimisella todella vaikeaa. Vaikuttaisi, että sarkasmin tunnistaminen vaatisi tekstitiedon lisäksi jotain ylimääräistä tietoa kontekstista, käyttäjästä ja aihepiirin viime aikaisista tapahtumista. Koska sarkasmin tunnistaminen ja täten käsittely tai seulominen on ongelmallista, vaikuttaa siltä, että tekstianalytiikan soveltamisessa suoraviivaisinta olisi pyrkiä välttämään sarkasmia. Tämä viittaisi siihen, että tekstianalytiikan

sovellettavuus Twitter-datan suhteen on ongelmattomampaa aineistossa, jossa sarkasmia käytetään vähemmän.

Twiiteille tyypillinen tekstimuotoinen data on sangen kohinaista dataa (engl. noisy data). Tämä tarkoittaa, että tekstianalytiikan suoraviivaista hyödyntämistä vaikeuttavat tekstin sisältämät runsaat välimerkit ja epätavalliset ilmaisut. Tämän vuoksi Twitterin tuottamaa tekstidataa tulee usein siistiä ja esivalmistella ennen tekstianalytiikan soveltamista. Tyypillisesti tämä tarkoittaa välimerkkien ja erilaisten vähän kontenttia sisältävien ja usein toistuvien sanojen poistamista tekstistä tulkittavuuden parantamiseksi. Kehittyneempiäkin vaihtoehtoja twiittien esivalmistelluun on tutkittu.

Kaufman ja Kalita (2010) esittävät yksinkertaisen kaksiportaisen menetelmän twiittien tulkittavuuden parantamiseksi. Hän huomioi, että twiitit ovat luonteeltaan hyvin kohinaista dataa (engl. noisy data). Kieliopillisesti katsottuna twiiteissä käytetty kieli on myös hyvin vapaamuotoista, mikä hankaloittaa tulkintaa tekstianalytiikalla (Kaufman & Kalita, 2010). Koska twiittien tekstianalytiikassa on usein tarpeen käydä läpi huomattavia määriä twiittejä, on perusteltua esivalmistella näitä ja pyrkiä poistamaan kohinaa. Kaufman ja Kalita (2010) normalisoivat twiitit ensin poistamalla mahdollisimman paljon kohinaa tekstistä. Tämän jälkeen teksti käsitellään kielenkääntäjä-työkalulla, joka muuttaa viestit hyvin standardimuotoiseksi englanniksi parantaen täten twiittien käytettävyyttä tekstianalytiikan kohteena. Tuloksena on malli, jonka avulla twiitit voidaan muokata kieliopillisesti oikeaan ja tulkittavaan englanninkieliseen muotoon. Tämän kaltaisen kielen normalisointi-työkalun käyttö voi olla tietyin oletuksin hyvin tarkoituksen mukaista tekstianalytiikan kannalta, sillä twiittien sisältöä voidaan täten selkeyttää ja tekstianalytiikan osalta tulkittavuus paranee. On kuitenkin huomioitava, että tekstin normalisointi saattaa vaikuttaa myös twiittien sävyyn ja sisältöön. Täten riski vääristymiin tekstianalytiikan soveltamisessa on mahdollinen.

Kokonaisuutena Twitter-pohjaisen tekstimuotoisen datan hyviä puolia analysoitavuuden kannalta ovat datan julkisuus, aika- ja paikkamerkinnät sekä materiaalin suuri määrä. Täsmälliset paikkatiedot ovat mobiilikäyttäjien osalta erittäin hyvä puoli. Sen sijaan perinteisten tietokonekäyttäjien itsensä ilmoittamiin paikkatietoihin on suhtauduttava analytiikkaa suunniteltaessa varauksella. Vaihtoehdoiksi jää joko karsia nämä pois tai sitten vaihtoehtoisia menetelmiä paikantamiseen on käytettävä. On kuitenkin muistettava, että esimerkiksi käyttäjän ilmoittaman paikkatiedon avulla Google Maps-ohjelmointirajapinnasta saatavat koordinaatit eivät välttämättä ole lähelläkään käyttäjän todellista sijaintia viestin lähettämisen aikana.

Ongelmallisia asioita ja haasteita tekstianalytiikalle asettaa viestien lyhyys ja kohinainen luonne. Toisaalta viestien lyhyys voi olla myös hyvä puoli, sillä tämä saattaa parantaa viestien tulkittavuutta kontekstista riippuen. Lisäksi sarkasmin tunnistaminen automaattisesti twiiteistä vaikuttaa tällä hetkellä hyvin ongelmalliselta. Tekstidatan kohinan suhteen esivalmistelua ja siistimistä

voidaan käyttää twiittien ymmärrettävyyden parantamiseksi. Tämä kuitenkin pitää sisällään aina riskin alkuperäisen sävyn tai kontentin hukkaamisesta. Lopulta tilastotieteellisestä näkökulmasta katsottuna Twitter-data ei useinkaan tarjoa riittävän heterogeenistä ja demografisesti tasapainoista otosta populaatiosta.

4 Twitter-datan tekstianalytiikka

Tässä luvussa perehdytään Twitter-datan tekstianalytiikkaa koskeviin artikkeleihin. Artikkelit ovat pääosin peräisin vuosilta 2009 - 2013 ja hyödyntävät moninaisia tekniikoita datan analysoinnissa. Päähuomio on kiinnitetty kansanterveydellisestä tai yhteiskunnallisesta näkökulmasta katsottuna potentiaalisesti hyödyttäviin tutkimuskohteisiin. Monissa tutkimuksissa käytetään sangen saman tyyppisiä menetelmiä, joten suoraviivainen kategorioiminen käytetyiden tekniikoiden perusteella ei ole kovin järkevää. Täten artikkelit on tässä luvussa esitetty karkeasti jaoteltuna aihepiireittäin seuraavasti.

Ensimmäisessä alaluvussa käsitellään trendien ja uutisaiheiden havaitsemista Twitter-datasta. Toisessa alaluvussa käsitellään Twitter-datan analytiikkaa tautiepidemioiden ennakoinnissa ja seurannassa. Kolmannessa alaluvussa tutustutaan sävyanalyysiä soveltaviin tutkimuksiin. Neljäs alaluku käsittelee tekstianalytiikan soveltamista poliittisessa tutkimuksessa. Viides alaluku käsittelee erilaisiin seuranta-, kriisi- ja ennakoiviin järjestelmiin liittyvää tutkimusta.

4.1 Trendien ja uutisaiheiden havaitseminen

Alkavien trendien ja uutisaiheiden seuranta on hyvin tärkeä tutkimuksen kohde Twitter-datan tekstianalytiikassa (Mathioudakis ja Koudas, 2010). Tämä johtuu siitä, että sovellutusalueita on huomattavasti. Twitteristä saatavat ensimmäiset uutiset ovat usein nopein tapa saada informaatiota äkillisesti alkavasta tapahtumasta tai ilmiöstä. Lisäksi trendien ja tiettyjen twiiteissä esiintyvien puheenaiheiden löytäminen on hyvin hyödyllistä esimerkiksi markkinoinnin kannalta sekä etsittäessä tiettyjä indikaattoreita eri asioiden tulevaisuuden tilasta (Asur & Huberman, 2010). Twiitit heijastelevat sitä, mihin kuluttajat kiinnittävät huomionsa (Mathioudakis ja Koudas, 2010). Esimerkiksi

muodikkaiden trendien löytäminen antaa paremman mielikuvan mahdollisista kuluttajien valinnoista tulevaisuudessa. Toisaalta erilaisia alueellisiin trendeihin perustuvia suosittelujärjestelmiä voidaan kehittää perustuen ajalliseen, paikalliseen ja trendin teemaan pohjautuvaan analytiikkaan (Nagarajan, Gomadam, Sheth, Ranabahu, Mutharaju & Jadhav, 2009) ja trendi-ilmiöiden luokittelua helpottavaan taksonomiaan (Naaman, Becker & Gravano, 2011). Trendisanojen ripeällä löytämisellä on myös sovellutuksia esimerkiksi kriisien hallinnassa (Terpstra, de Vries, Stronkman & Paradies 2012). Usein ensimmäisenä paikalta, jossa jokin kriisi tapahtuu, saadaan tietoa juuri Twitterin välityksellä (Terpstra ym., 2012). Tämä tarkoittaa sitä, että trendisanojen ja ripeästi kehittyvien tapahtumien löytäminen on ensiarvoisen tärkeää esimerkiksi Twitter-pohjaisten varoitusjärjestelmien luomiseksi. Lisäksi Twitteristä on muotoutunut käyttökelpoinen työväline journalistiikan palvelukseen (Ahmad, 2010). Nousevien trendien ja uutisaiheiden seuranta ja havaitseminen muodostaa täten hyvin tärkeän tekstianalytiikan sovellutusalueen Twitter-dataa analysoitaessa.

Mathioudakis ja Koudas (2010) esittelevät artikkelissaan ”TwitterMonitor: Trend Detection over the Twitter Stream” trendien ja uutisaiheiden etsintään soveltuvan työkalun. Työkalu on TwitterMonitor, joka kykenee reaaliaikaiseen trendiseurantaan Twitterin twiittivirrasta sekä tarjoaa alkavan ilmiön parempaa ymmärtämistä tukevaa analytiikkaa. Käyttäjät voivat hyödyntää työkalua käyttöliittymän avulla etsimällä trendejä eri hakukriteerein sekä varustamalla potentiaalisia alkavia trendejä omilla kommenteillaan.

Mathioudakisin ja Koudasin (2010) toimintaperiaatteeltaan yksinkertainen TwitterMonitor havainnollistaa hyvin trendiseurannan järjestelmän toimintaa yksinkertaisimmillaan. Perusperiaatteeltaan sitä voinee pitää sangen ideaalisena trendiseurantaan. Suuria tekstimuotoisia datamääriä käsiteltäessä Mathioudakis ja Koudas (2010) toteavat järjestelmän skaalautuvuuden kannalta olevan tärkeää, että käytetty malli on mahdollisimman yksinkertainen. Toisin sanoen tämä tarkoittaa sitä, että tutkittavaa dataa ei jouduta jatkuvasti uudelleenprosessoimaan tai hakemaan toistuvasti Twitter API:sta. Esitetty malli näiltä osin vastaakin vaatimukseen yksinkertaisesta rakenteesta.

TwitterMonitor etsii twiittivirrasta sanaryppäitä – tiettyjen sanojen ilmentymiä, jotka ajallisesti ryöpsähtäen lisääntyvät yli normaalin esiintymiskeskiarvon. TwitterMonitorin Twitter API:a kuunteleva osa toimittaa twiittivirrasta dataa käsiteltäväksi. Äkillinen ja dramaattisesti lisääntynyt sanan esiintyminen twiiteissä indikoi usein sitä, että jotain on tapahtumassa. Toisin sanoen tämä usein tarkoittaa alkamassa olevaa trendiä tai uutisaihetta. Työkalu jakautuu yksinkertaisesti kolmeen funktionaaliseen osaan. TwitterMonitor suorittaa trendin tunnistuksen kahdessa osassa. Kolmannessa osassa tapahtuu edistyksellisempi trendiä koskeva analytiikka. Ensimmäinen osa, QueueBurst-algoritmi tunnistaa alkavia sanaryöpsähdyksiä. Mathioudakisin ja Koudasin (2010) mukaan QueueBurst käsittelee datan vain kerran tunnistamisen ollessa reaaliaikaista. Lisäksi tunnistamiseen on liitetty ominaisuuksia, jotka osaavat

tunnistaa sattumalta usein esiintyvät sanat ja poissulkea nämä. Algoritmi osaa myös välttää spam-twiittejä. Kun tunnistus on suoritettu eli on havaittu, että jokin tietty sana esiintyy poikkeuksellisen usein tietyssä ajassa (ja mahdollisesti maantieteellisesti rajattuna) aktivoituu työkalun toinen osa, GroupBurst.

GroupBurst:in tehtävänä on luokitella ilmiöön liittyviä sanoja. Tämä toteutuu tarkkailemalla toisiinsa merkityksiensä puolesta liittyviä sanoja tarkkailun alla olevassa sanaryöpsähdyksessä. Toisin sanoen algoritmi luo luokittelun, joka kuvainnollistaa meneillään olevan ilmiön ilmauksia ja sanoja. Tämä osaltaan helpottaa alkavan trendin käsittelyä.

Kolmannessa osassa, kun trendi on tunnistettu ja siihen liittyvät sanat ja ilmaukset on luokiteltu trendin käsittely siirtyy analyyttisemmalle tasolle. Muita havaittuun ilmiöön liittyviä sanoja indeksoidaan ja esimerkiksi twiiteissä olevia uutislinkkejä haetaan tarkoituksena tuottaa paremmin kuvaavaa analytiikkaa kyseisestä trendistä. Sovellettaessa trendien seurantaan kolmas osa antaa mahdollisuuden hyödyntää syvempää analytiikkaa, kuten erilaisia kontekstianalyysityökaluja. Käyttäjää varten Mathioudakis ja Koudas (2010) toteuttivat TwitterMonitor-mallin pohjalta käyttöliittymän, jonka avulla käyttäjät kykenivät tutkimaan trendejä eri kriteereillä sekä varustamaan näitä omilla kommenteilla. Trendeihin liittyviä twiittejä oli myös saatavilla pieni otos. TwitterMonitorin avulla kyettiin myös toteuttamaan päivittäisten trendien listaus perustuen esiintymismäärään.

Kokonaisuudessaan TwitterMonitor tarjoaa hyvin yksinkertaisen ja selkeän perusarkkitehtuurin trendien seurantaan. Sellaisenaan mallin arkkitehtuuria voisi pitää esimerkkinä kehitettäessä monenlaisiin tarpeisiin tulevia trendien löytämis- ja seurantajärjestelmiä. Malli on edelleen helposti kehitettävissä ja toiminnallisuutta voidaan laajentaa esimerkiksi lisäämällä luokittelun jälkeistä ilmiön analytiikkaa tai parantamalla seulontaa. Mallilla jo tällaisenaankin voisi kuitenkin potentiaalisesti olla monia sovelluskohteita esimerkiksi julkishallinnon kannalta. Esimerkiksi terveydenhoitoon liittyvän tekstimuotoisen datan osalta olisi mahdollista seuloa poikkeuksellisen usein ilmeneviä tietyn sairauden ilmauksia ja tarkastella näiden maantieteellistä esiintyvyyttä.

Trendien ja uutisaiheiden löytämisen kannalta voidaan soveltaa myös malleja, jotka huomioivat enemmän twiittien luonnetta. Phuvipadawat ja Murata (2010) huomioivat, että johtuen twiittien lyhyydestä on viestien sisällön samankaltaisuuden toteaminen haastavaa trendien löytämisen kannalta. Tämän vuoksi tarkkailun kohteena olevia sanoja ja sanaryhmiä voidaan painottaa merkittävyydeltään eri tavoin niiden yleisyyteen ja luotettavuuteen pohjautuen. Phuvipadawat ja Murata (2010) mukaan twiittien uutisaiheisiin liittyvä sisältö jakautuu kahteen erilaiseen kontekstiin. Näistä toinen on viestin faktuaalinen sisältö ja toinen käyttäjän uutisaihetta koskeva tunnesisältö. Perinteisen median uutiset eivät yleensä ole vastaavalla tavalla värittyneitä, joten sosiaalisen median tekstidatan osalta sentimenttisisältö tulisi mahdollisesti huomioida uutisaiheiden etsinnässä.

Phuvipadawat ja Murata (2010) esittävät ratkaisuksi keskittymistä pelkästään faktuaaliseen sisältöön. Faktuaalisen tiedon keräämiseksi on keskitytty viestien teksti-, hyperteksti- ja paikka- sekä lähdetietoihin. Toisin sanoen tunteellisesti värittyneempää kontenttia on karsittu. Haut suoritetaan Twitter API:n kautta käyttäen hyväksi hakusanoja ja hashtageja. Tämän jälkeen viestien sisältöjä indeksoidaan kontenttiin liittyen. Tässä työkaluna on käytetty Apache Lucene-työkalua. Kolmannessa vaiheessa samankaltaisiksi arvioidut viestit ryhmitellään yhteen, jotta saadaan kuvaus uutis- tai trendiaiheesta. Kun ryhmitellyt viestit muodostavat kokonaiskuvan trendistä tai uutisesta voidaan Twitterin ulkoisia uutislähteitä (kuten virallisen median uutisia) käyttää muodostamaan parempi kuva ilmiöstä. Phuvipadawatin ja Muratan (2010) lähestymistavan heikkouksia lienee sen monimutkaisuus sekä se, että väistämättä osa twiittien sisältämästä informaatiosta menetetään.

Popescun ja Pennacchiottin (2010) tutkimuksessa ”Detecting Controversial Events from Twitter” esitellään hyvin erilainen lähestymistapa. Tässä lähestymistavassa käytetään tarkkailtavina olevien ilmiöiden seurannassa ”pikakuvia” (engl. snapshot). Mallissa pikakuva muodostuu kolmiportaisesti sisältäen tiedon kohteesta eli jostain twiiteistä tunnistettavissa olevasta entiteetistä, aikayksiköstä (eli esimerkiksi yksi vuorokausi) ja tänä aikana kohdetta koskevista twiiteistä. Mallia on Popescun ja Pennacchiottin (2010) taholta käytetty kiistanalaisten tapahtumien (engl. controversial event) kartoittamiseksi. Tällä tarkoitetaan Twiitti-virrassa ilmenevää trendiä, jossa käyttäjät ovat selkeästi aktivoituneet keskusteluun esittäen vastakkaisia mielipiteitä tai epäuskoa. Huomattavan positiiviset tai negatiiviset (tai ilman reaktioita jääneet) tapahtumat sen sijaan eivät ole mielenkiinnon kohteena. Snapshotit kuvaavat joko kiistanalaisia tapahtumia tai sitten ei. Jotta vertailua pikakuvien välillä voidaan tehdä ja seurata näin ollen tapahtuman kehittymistä, täytyy olla mahdollista arvottaa viestien sisältöä jotenkin.

Tätä tarkoitusta varten pikakuvia on arvotettu erilaisten sanakirjastojen avulla. Käytössä oli viestien 7590 sanan sävyja kuvaavien sanojen kirjasto, 750 sanan ristiriitaisuutta ja kiistanalaisuutta kuvaava sanasto sekä englanninkielisten rumien sanojen sanasto. Popescun ja Pennacchiottin (2010) tutkimuksessa testataan kolmea erilaista koneoppimisen menetelmää tapahtumia kuvaavien pikakuvien mallintamiseen hyvin tuloksin. Sovellettuna trendien tarkkailuun menetelmä vaikuttaa toteutuksensa puolesta tarpeettoman raskaalta. Esimerkiksi skaalautuvuuden ja uusien trendien helpon havaittavuuden kannalta Mathioudakisin ja Koudasin (2010) sanaryöpsähdyksiä tarkkaileva malli on huomattavasti parempi.

Popescun ja Pennacchiottin (2010) malli soveltuu mahdollisesti jonkin tietyn entiteetin tilan tarkkailuun. Tarkkailtava asia tulisi siis tietää etukäteen, jolloin kyseistä asiaa koskevia twiittejä voidaan mallintaa pikakuvatekniikan avulla. Mahdollisesti sovellutuksia voisi olla esimerkiksi turvallisuusaiheita, -ilmauksia ja sanoja seuraavan järjestelmän osana. Koska menetelmä on kehitetty kiistanalaisten aiheiden tarkkailuun voisi tätä kenties myös käyttää varauksella poliittisten aiheiden seurannassa.

Cataldi, Di Caro ja Schifanella (2010) kuvaavat tutkimuksessaan trendien elämänkaareen perustuvan lähestymistavan trendien automaattiseen havainnoitsimiseen. Twiittien virrasta kerätään dataa, joka formalisoidaan analysoitavaan muotoon. Perusideana on sanaryöpsähdys-tekniikoiden tapaan tarkkailla kuinka usein termi esiintyy nyt verrattuna aiempaan. Tässä toteutuksena käytetään termeille tehtävää life cycle-mallia, joka kuvastaa termin käyttöä tietyn aikayksikön aikana. Mikäli termi on aiemmin ollut lähes käyttämättömässä tilassa ja yllättäen yleistyykin voidaan tästä indikoida alkavaa trendiä. Vastaavasti trendin hiipuessa trendin ”elinkaari” heikkenee myös. Termit voidaan järjestää esiintyvyyden suhteen järjestykseen ja näin arvottaa eri trendien vahvuutta tutkitussa aikajaksossa. Viimeisessä vaiheessa tehtäväksi jää muiden nousevaan trendiin liittyvien sanojen löytäminen ja muovaaminen ilmiötä kuvaavaksi sanavalikoimaksi.

Edellä on esitelty neljä erilaista lähestymistapaa trendien ja uutisaiheiden havaitsemiseen. Näistä Mathioudakisin ja Koudasin (2010) TwitterMonitor kuvaa perusmallin, joka on laajasti sovellettavissa yksinkertaisuutensa ja skaalautuvuutensa ansiosta. Vastaavia menetelmiä voidaan kehittää sisältämään edelleen lisää termien arvottavuutta tai huomioimaan twiittien faktuaalisen ja sentimenttikontentin (Phuvipadawat & Murata, 2010). Trendien havaitsemista voidaan lähestyä myös kuvaamalla tarkkailtavina olevien entiteettien ja asioiden pikakuvia snapshot-lähestymistavalla (Popescu & Pennacchiotti, 2010 ja Cataldi, Di Caro & Schifanella, 2010). Käsitellyistä menetelmistä Mathioudakisin ja Koudasin (2010) TwitterMonitor vaikuttaa parhaalta perusmallilta, koska tämä mahdollistaa laajan sovellettavuuden, skaalautuvuuden ja on yksinkertainen. Mallin perusrakenteen päälle on helpohkosti kehitettävissä tarkoituksenmukaista lisäfunktionaalisuutta.

4.2 Epidemioiden seuranta

Pikainen reagointi- ja varautumiskyky erilaisiin epidemioihin on tärkeää kansanterveyden kannalta ja inhimillisten kärsimysten minimoimiseksi. Nykyiset menetelmät käyttävät influenssojen raportoimisesta syntyvää dataa, joten kuvaus epidemian tilasta saadaan viiveellä. Tämän vuoksi sosiaalisen median ja hakukonepalveluiden kautta saatava tieto epidemioista voi potentiaalisesti parantaa huomattavasti terveysviranomaisten kykyä varautua ja reagoida mahdollisiin epidemioihin ja pandemioihin liittyviin uhkiin. Kansanterveyttä vaalivien instituutioiden ja organisaatioiden tehokas reagoiminen influenssoihin ja pandemioihin on sitä mahdollisempaa mitä aiemmin alkava influenssa voidaan havaita.

Achrekarin, Gandhen, Lazaruksen, Yun ja Liun (2011) mukaan perinteisillä menetelmillä syntyy noin 1-2 viikon viive taudin diagnosoimisten ja virallisten ILI-oireista (engl. influenza-like illness) raportoivien raporttien

saamisen välillä. Achrekar ym. (2011) arvioivat ja testaavat tutkimuksessaan virallista CDC-dataa (Centers for Disease Control and Prevention) sekä ilman että Twitter-datan kanssa. Tutkimuksessa osoitetaan, että Twitter-data voi parantaa mallien ennustavuuden tarkkuutta. Achrekar ym. (2011), päätyvät johtopäätökseen, jonka mukaan Twitter-data tarjoaa lähes reaaliaikaisen arvioinnin mahdollistavan työvälineen influenssan kaltaisten sairauksien aktiivisuudesta.

Epidemioiden seurannassa (engl. Epidemic Intelligence, EI) kerätään ja hyödynnetään tietoa alkavista epidemioista ja niiden leviämisestä. Tässä toiminnassa hyödynnetään sekä formaaleja että ei-formaaleja lähteitä (Quincey & Kostkova, 2010). Epidemioiden seurannan ansiosta kyetään viranomaistahojen varautumista esimerkiksi pandemioihin parantamaan. Erityisesti tautiepidemian alkamisen pikainen havaitseminen ja maantieteellisen sijainnin ripeä paikantaminen ovat tärkeitä (Lampos ja Cristianini, 2010). Tällainen seurantajärjestelmä on esimerkiksi Quinceyn ja Kostkovan (2010) mainitsema Global Public Health Intelligence Network (GPHIN). Kyseiset järjestelmät keräävät systemaattisesti tietoa julkisista lähteistä ja raporteista sekä epävirallisemmista lähteistä. Esimerkiksi Quincey ja Kostkova (2010) näkevät sosiaalisen median sivujen, kuten Facebookin ja Twitterin soveltuvan osaksi tätä ei-formaalia lähdepalettia.

Sosiaalisen median tuottama data voi olla huokea ja tehokas ratkaisu tautiepidemioiden seurantaan. Perinteisiä menetelmiä tautimäärien tilastointiin ovat muun muassa työelämän poissaolomäärien seuranta, yhteydenotot lääkäreihin ja lääkärintarkastukset ja satunnaiset kyselyt tautien kartoittamiseksi (Lampos ja Cristianini, 2010). Nämä menetelmät vaativat sekä aikaa että taloudellisia investointeja. Sosiaalisen median seurannasta saatavalla suhteellisen hyvällä korrelaatiolla voisi täten olla mahdollista ennustaa riittävissä määrin ja ennen kaikkea ripeästi tautiepidemioiden ilmaantumista ja etenemistä. Potentiaalisia etuja voisi olla saavutettavissa reaaliaikaisuuden ja kustannustehokkuuden muodossa.

Lampos ja Cristianini (2010) ovat tutkimuksessaan kehittäneet menetelmän, jolla Twitter-dataa louhitaan tarkoituksena löytää kuvauksia ja ilmaisuja sairauden oireista. Nämä maininnat muutetaan tilastollisesti yhtä aikajaksoa vastaaviksi ”flunssapisteiksi”, joilla vertailua ajallisesti voidaan tehdä. Periaatteessa flunssapisteet muodostuivat flunssan oireita kuvaavien twiittien suhteesta kaikkiin analyysiaineiston twiitteihin. Analyysin tekniikoina käytettiin koneoppimista, n-grammeja ja luokittelua. Lähdedata luokiteltiin Ison-Britannian Twitter-käyttäjistä koskemaan suurimpien asutuskeskusten käyttäjiä.

Menetelmän käyttöä testattiin Isossa-Britanniassa vuonna 2009 24 viikon ajalta kertyneellä twiittidatalla H1N1-epidemian aikoihin. Vertailtaessa menetelmän tulokset korreloivat virallisen tautidatan kanssa noin 95% tarkkuudella. Johtopäätelmänä Lampos ja Cristianini (2010) uskovat menetelmänsä olevan halpa ja nopeassa ajassa tuloksia tuottava menetelmä.

Paul ja Dredze (2011a) lähestyvät aihetta hieman erilaisesta perspektiivistä esitellen ATAM-mallinsa (Ailment Topic Aspect Model). Kyseinen malli etsii sairauden oireita, hoitokeinoja ja yleisesti sairauksiin liittyvää termistöä twiiteistä. Toisin sanoen malli ei keskity pelkästään influenssan kaltaisten oireiden etsintään vaan mallin avulla voidaan etsiä monia kansanterveydellisiä ongelmia (Paul ja Dredze (2011a) mainitsevat esimerkkeinä tulehdukset ja liikalihavuuden).

Pääasiallisena tekniikkana on ohjattu koneoppiminen, jossa oppimisaineistona on käytetty noin 1.6 miljoonan twiitin sairaus- ja oiresanastoa ja luokittelua. Lisäksi oma, reilun 5000 viestin runkosanasto luotiin sairauksien ja oireiden tunnistamista ja luokittelua varten. Oppimisessa on keskitytty etenkin sairauden oireiden etsimiseen, sillä kuten Paul ja Dredze (2011a) toteavatkin, sairauksia sellaisenaan on ongelmallista käyttää suoraviivaisessa hakusanaetsinnässä. Tuloksena voisi olla huomattava määrä harhaanjohtavia twiittauksia (kuten ”I am sick of this”). Lopputuloksena Paul ja Dredze (2011a) toteavat ATAM-mallin oppivan ryhmittelemään oireita ja hoitokeinoja sairauksien mukaan. Menetelmä kykenee siis löytämään taudin oire-kuvaukset twiittivirroista ja näitä voidaan edelleen käyttää haluttujen oireiden seurantaan. Potentiaalisesti tätä tietoa voidaan tietenkin yhdistellä käyttäjien paikkatietoihin ja twiittien aikatietoihin. Malli mahdollistaa myös ihmisten käyttämien hoitokeinojen tarkkailun löytämällä käytettyjä hoitomenetelmiä ja lääkityksiä riippuen oireista.

Paul ja Dredze (2011b) jatkoivat ATAM-mallinsa kehittämistä jatkotutkimuksessaan, jossa ATAM-mallin ja edelleen kehitellyn ATAM+-mallin soveltuvuutta kansanterveydellisiin tutkimuksiin tarkasteltiin. Lähtökohtana oli käyttäjien viestien pohjalta tehdyllä analyysillä lukuisten erilaisten asioiden tarkkailu populaatiossa. Toisin sanoen menetelmä mahdollistaa sairauksien ja käyttäytymisen riskitekijöiden tarkastelun, sairauksien lokalisaation käyttäjien paikkatietojen perusteella sekä analyysin sairauksista ja niiden hoitoon käytetyistä menetelmistä. Lopputuloksena Paul ja Dredze (2011b) esittävät Twitterin tuottamalla datalla olevan erittäin hyvä käytettävyys kansanterveydellisiä tutkimuksia ajatellen. ATAM+-mallin tuottama flunssaa koskeva data korreloi merkittävän hyvin virallisten tilastojen kanssa.

Influenssojen tarkkailu on kuitenkin vain yksi osa Twitterin luoman datan todellista potentiaalia kansanterveydellisissä tutkimuksissa. Malli mahdollistaa Paulin ja Dredzen (2011b) mukaan esimerkiksi elintapariskien (kuten tupakointi ja liikalihavuus) kartoituksen ja tutkimisen. Mallia testattaessa suurin korrelaatio löydettiinkin tupakoinnin ja syöpäoireiden välillä. Vastaavasti havaittiin negatiivinen korrelaatio aktiivisesti liikkuvien ja sairauden oireista raportoivien välillä. Merkille pantavaa on, että mallin avulla voidaan saada tietoa taudeista ja terveysongelmista, jotka eivät muuten välttämättä ikinä päätyisi julkisen terveydenhuollon piiriin. Lukuisia oireita, kuten esimerkiksi flunssa- tai unettomuuden oireita, diagnosoivat ja lääkitsevät ihmiset itse oma-aloitteisesti. Tästä ei tietoa julkiseen

terveydenhoitoon jää mutta toisaalta Twitteriin usein jää. Tämä mahdollistaa tutkimuksia siitä, miten ihmiset hoitavat oma-aloitteisesti sairauksiensa oireita ja toisaalta siitä, mihin eri oireisiin eri lääkkeitä käytetään. Tämä saattaa olla relevanttia tutkimustietoa esimerkiksi terveysviranomaisille.

Menetelmää rajoittavia tekijöitä Paul ja Dredze (2011b) mainitsevat olevan paikkatietojen liian suuri granulariteetti ja otoksien koko. Mobiililaitteiden ja geotaggauksen edistyessä paikkatietojen granulariteetti-ongelma luultavasti pienenee. Twitter-otoksien kokoon ja laatuun tilastollisessa mielessä vaikuttaa se, kuinka paljon palvelua tutkitulla alueella käytetään ja millaisten ja minkä ikäisten väestöryhmien keskuudessa.

Culottan (2010b) influenssojen alkamiseen liittyvän havaitsemismenetelmän kehitystutkimuksessa saavutettiin 95% korrelaatio virallisen tautidatan kanssa pelkästään hyödyntämällä twiittien seulonnassa pienehköä flunssiin liittyvää avainsanojen joukkoa yli 500 miljoonan twiitin aineistolla. Jo pelkällä ”flu”-sanan etsinnällä päästiin 84% korrelaatioon, mikä sinänsä on sangen yllättävää, sillä aiemmassa tutkimuksessaan Culotta (2010a) päätyi vain 78% korrelaatioon 500 tuhannen twiitin otoksella. Tulokseen tosin mahdollisesti vaikutti twiittien huonosta esivalmistelusta ja siistimisestä johtuva huono otoksen taso tekstianalytiikan hyödyntämisen kannalta. Juuri tämän tarkempiin korrelaatioihin voi olla vaikea päästäkään näin yksinkertaisella menetelmällä twiittien kohinaisen luonteen vuoksi. Kuitenkin tulokset osoittavat jo hyvin yksinkertaisen lähestymistavan antavan suuntaa-antavia tuloksia. Tämä on rohkaisevaa mikäli tautiepidemioiden tarkkailuun haluttaisiin kehittää hyvin minimalistinen muutamiin avainsanoihin perustuva ja rakenteeltaan kevyt tarkkailujärjestelmä.

Aramaki, Maskawa ja Morita (2011) esittelevät influenssojen seurantajärjestelmän, joka huomioi ja poistaa ei-relevantit influenssamaininnat twiiteistä saatavasta tekstidatasta. Tällainen voisi olla esimerkiksi kommentti ”are you going down with the flu?” twiitissä. Influenssaan liittyvä sana esiintyy, mutta sen olemassaolosta huolimatta ei voida tehdä johtopäätöstä, että jokin taho kärsisi siitä. Aramaki, Maskawa ja Morita (2011) toteuttavat influenssa-aiheisia ilmauksia etsivän järjestelmän perinteisillä ja jo esitetyillä menetelmillä mutta varustavat tämän myös työkalulla, joka jakaa twiitit positiivisiin ja negatiivisiin. Negatiiviset twiitit tässä yhteydessä tarkoittavat twiittejä, jotka sisältävät avainsanan mutteivät tarkoita sitä, että kellään twiitin osapuolilla olisi tautia. Positiiviset twiitit sen sijaan indikoivat selkeästi sairauden olemassaoloa. Arviointi on toteutettu ohjatulla oppimisella, jossa koulutusaineistona on käytetty oireita ja positiivisia ja negatiiviisia ilmauksia kuvaavaa sanastoa.

Aramaki, Maskawa ja Morita (2011) havaitsivat erilaisten Twitterissä olevien uutisaiheiden haittaavan työvälineen sovellettavuutta sekä luovan spammia tutkimusaineistoon. Tuloksissa havaittiin, että harvan uutisoinnin aikoina menetelmä saavutti huomattavasti paremman korrelaation viralliseen dataan verrattuna kuin esimerkiksi Google Flu Trends. Kuitenkin uutisoinnin ollessa runsaampaa menetelmän tulokset huononivat radikaalisti

johtuen flunssaa koskevien uutisten ja näiden jakamisen luomasta yliedustuksesta twiiteissä. Aramakin, Maskawan ja Moritan (2011) mukaan tämä ilmiö jää ongelmaksi, joka tulee ratkaista tulevaisuudessa. Mahdollisia sovelluksia suunniteltaessa tämä tulisikin ilmeisesti tiedostaa ja ottaa huomioon paremmin. Koska uutisoinnin määrä Twitterissä on kasvussa ja näitä linkkejä ulkoisiin uutislähteisiin voidaan parhaillaan jakaa hyvinkin massiivisesti voi olla, että tulevaisuudessa uutislähteiden luoma spammi vaikeuttaa tekstianalytiikan soveltamista. Tämä osaltaan voi ennakoida sitä, että jatkossa paitsi tautien mallintamisessa myös muissa tekstianalytiikan soveltamiskohteissa tulisi keskittyä entistä enemmän spammin käsittelyyn. Tässä tapauksessa duplikaattiviestien eliminoiminen, parempi viestien leviämisen ja uutisoinnin diffuusion ymmärtäminen ja etenkin ensimmäisen uutislähteen löytäminen voi korostua.

Edellä on esitelty erilaisia lähestymistapoja tautiepidemioiden seurantaan. Käsitellyistä malleista Lampoksen ja Cristianinin (2010) influenssa-mainintojen muuttaminen geolokaatiotiedoilla varustetuiksi flunssapisteiksi oli sangen suoraviivainen ja toimiva tapa kuvata epidemian tilaa. Paulin ja Dredzen (2011a & 2011b) ATAM- ja ATAM+-mallit todettiin hyvin mielenkiintoisiksi. Kyseisillä malleilla voi olla lukuisia tautiepidemoiden tarkkailun ulkopuolisia sovellutusalueita. Tällaisia sovellutuskohteita voisivat olla esimerkiksi kuntoilu- ja ruokailutottumuksien tarkkailu sekä elämäntapariskien arviointi. Culotta (2010a & 2010b) osoitti, että jo pienen tautiepidemioita koskevan avainsanojen ryhmän avulla voidaan saavuttaa suuntaa-antavia tuloksia. Aramakin, Maskawan ja Moritan (2011) esittelemä ei-relevanttien tautisanastoa sisältävien twiittien poissuodattaminen vaikuttaa monimutkaiselta lähestymistavalta. Myöskään tulokset eivät olleet kovinkaan hyviä ja twiiteissä tapahtuvan uutisoinnin todettiin vaikuttavan vahvasti ja negatiivisesti saatuihin tuloksiin. Lähestymistavan heikkous ei lienekään twiittimateriaalin seulominen relevantteihin ja ei-relevantteihin twiitteihin vaan se, että twiittejä ei saada kerättyä kohdennetusti sairauden kohteilta eli käyttäjiltä. Tämän perusteella Aramakin, Maskawan ja Moritan (2011) mallia voitaisiin parantaa keskittymällä Paulin ja Dredzen (2011a & 2011b) tapaan etenkin sairauden oireiden etsimiseen. Koska tuloksia voidaan saada yksinkertaisemmilla menetelmillä vähintään yhtä tehokkaasti on perusteltua pitää Aramakin, Maskawan ja Moritan (2011) mallia liian kompleksisena. Tautiseurannan osalta Aramakin, Maskawan ja Moritan (2011) huomio uutisoinnin mahdollisesta vaikutuksesta mallien toimimiseen on kuitenkin huomioitava. Potentiaalisesti tämä tulevaisuudessa voi olla kasvava ongelma.

Kokonaisuudessaan Paulin ja Dredzen (2011a & 2011b) ATAM- ja ATAM+-mallit vaikuttavat parhaimmilta ja monikäyttöisimmiltä – tautiepidemioihin liittyvän datan lisäksi nämä mahdollistavat myös monenlaisen muun informaation tarkkailun. Myöskin Lampoksen ja Cristianinin (2010) esittämää yksinkertaista flunssapisteisiin perustuvaa järjestelmää voidaan pitää tuloksiltaan vahvana ja helposti sovellettavana menetelmänä terveysviranomaisten käyttöön. Culottan (2010a & 2010b)

tulokset antavat myös tukea sille otaksumalle, että tautitarkkailua voitaisiin tehdä hyvin yksinkertaisella menetelmällä ja keskittymällä suppeaan ryhmään taudin keskeisiä avainsanoja.

4.3 Sävyanalyysiin perustuvia tutkimuksia

Sävyanalyysilla (engl. sentiment analysis) tarkoitetaan erilaisten tekstianalytiikan menetelmien soveltamista analysoitavana olevan tekstin tunteellisen kontentin värittyneisyyden selvittämiseksi. Tyypillisesti nämä menetelmät pyrkivät arvioimaan tekstin negatiivista ja positiivista sävyä sekä arvioimaan minkä tyyppisistä tunteista on kyse. Ytimekkäästi ilmaistuna sävyanalyysi on positiivisten ja negatiivisten tunteiden, mielipiteiden ja arvioiden tunnistamista (Wilson, Wiebe & Hoffman, 2005). Menetelmät ovat esimerkiksi erilaisia luonnollisen kielen käsittelyn menetelmiä tai ontologioita erilaisista tunteita kuvaavista sanoista (esimerkiksi EMOTIVE-projekti, 2013b). Samoin ohjattua oppimista voidaan soveltaa sävyanalyysiin esimerkiksi käyttämällä hymiöillä valikoituja twiittejä opetusmateriaalina (Go, Bhayani & Huang, 2009). Alla on esitelty erilaisia sävyanalyysiin pohjautuvia tutkimuksia sekä selvitetty näissä käytettyjä menetelmiä.

Yleisesti ottaen twiittien tarkoituksellinen sisältö jakautuu kahteen eri kategoriaan. Näitä ovat twiitit, joissa käyttäjät twiittaavat itsestään ja toisaalta twiitit, joissa jaetaan muuta tietoa. Bollen, Pepe ja Mao (2011) toteavat molemmissa tapauksissa olevan mahdollista, että twiitit sisältävät jotain, joka paljastaa lähettäjän mielialan tai mielen tilan. Mikäli riittävän laaja määrä lähdeaineistoa kerätään analyysia varten, pitäisi siis olla mahdollista analysoida tunnetiloja laajemmassa mittakaavassa. Tämä mahdollistaa esimerkiksi kansakunnan vallitsevan mielialan mallintamisen. Bollen, Pepe ja Mao (2011) käsittelevät tutkimuksessaan Twitter-viestien emotionaalisen sisällön suhdetta makroskooppisiin sosioekonomisiin tapahtumiin eli käytännössä todellisessa maailmassa tapahtuviin muutoksiin. Tällaisia ovat esimerkiksi vaalien tulokset.

Tutkimuksessa käytettiin psykologista emotionaalisia aspekteja sanojen perusteella luokittelevaa kategorisointimallia, POMS:ia (POMS, Profile of Mood States). Lähdeaineisto (vajaa 10 miljoonaa twiittiä loppuvuodesta 2008) luokiteltiin mielentilan tai emotionaalisen latauksen perusteella POMS:in mukaisesti kuuteen eri kategoriaan. Näitä olivat jännitys, masennus, viha, väsymys, sekavuus ja tarmokkuus. Tämän perusteella luotiin päivittäiseksi kollektiivisen tunnetilan esitykseksi kuusiulotteinen mielialamalli. Tämän jälkeen oli mahdollista suorittaa vertailuja päivittäisten mallinnosten ja esimerkiksi kyseisen päivän uutisointien välillä.

Lopputuloksena Bollen, Pepe ja Mao (2011) havaitsivat, että tapahtumat sosiaalisessa, poliittisessa, kulttuurisessa ja taloudellisessa todellisuudessa vaikuttivat merkittävästi ja välittömästi kyseisen mielialamallin eri osiin. Tämän perusteella tutkimuksen tekijät uskovat olevan mahdollista

käyttää Twitter-dataa suuressa mittakaavassa kollektiivisten tunnetilojen ennustamiseen tarkastellussa populaatiossa. Bollen, Pepe ja Mao (2011) toteavat myös, että tämänkaltaisessa tutkimuksessa koneoppiminen ei ole tarpeellista vaan verrattain yksinkertainen mielialojen luokitteluun perustuva menetelmä tuottaa riittävän suurilla datamäärillä tyydyttävän tuloksen.

Onnellisuuden tutkimusta harjoitetaan enenevissä määrin. Soveltavia tutkimuksen aloja, jotka hyötyisivät populaation onnellisuuden arvioinnista (ja niihin vaikuttavista tekijöistä) on monia. Tällaisia voisivat olla esimerkiksi yhteiskunnallinen ja psykologinen tutkimus sekä onnellisuuden taloustiede (engl. happiness economics). Sosiaalisen median tarjoamalla datalla on teoriassa valtava potentiaali lähdeaineistoksi erilaisiin koettua onnellisuutta koskeviin tutkimuksiin. Mitchell, Harris, Frank, Dodds ja Danforth (2013) tutkivat Twitter-datalla onnellisuuden jakautumista maantieteellisesti. Tavoitteena oli tutkia Yhdysvaltojen alueella erilaisten maantieteellisten ja demografisten tekijöiden yhteyttä koettuun onnellisuuteen ja terveyteen.

Analyysiaineistona oli yli 80 miljoonaa sanaa sisältävä ja paikkatiedoilla varustettu twiittikokoelma sekä vertailuaineistona vuosittain kootut tiedot osavaltiokohtaisesti. Työkaluna käytettiin yksittäisiä sanoja onnellisuus-asteikolla pisteyttävää työkalua ja edelleen nämä pisteytetyt sanat yhdistettiin paikkatietoihin. Tutkittaessa onnellisuuden tasojen vaihtelua eri paikoissa tuli ottaa huomioon myös eri tavoin käytetyt sanat eri puolilla Yhdysvaltoja. Lopulta tätä arvotettua dataa voitiin vertailla osavaltiokohtaisesti koottuun vertailuaineistoon.

Aineiston perusteella havaittiin esimerkiksi onnellisuuden korreloivan Yhdysvalloissa vahvasti varallisuuteen. Korkein korrelaatio havaittiin tulotasossa. Vastaavasti vahva negatiivinen korrelaatio havaittiin onnellisuuden ja liikalihavuuden välillä. Lopputuloksena Mitchell, Harris, Frank, Dodds ja Danforth (2013) esittävät, että sosiaalisen median tuottamaa dataa voitaisiin potentiaalisesti käyttää reaaliaikaisiin arviointeihin todellisen maailman tapahtumista ja populaatiossa tapahtuvista muutoksista. Tällä voisi edelleen olla julkishallinnon kannalta sovellettavuutta vaikkapa yhteiskunnallisen muutoksen seurauksien arvioinnissa.

Twitterin tuottaman datan tekstianalyysin hyödyntämistä on myös tutkittu osakekurssien ennakoinnissa. Yleisellä tasolla on havaittavissa ainakin kolme potentiaalista tekijää, jotka tekevät Twitter-datasta varteenotettavan tutkimuskohteen osakekurssien ennakointia ajatellen. Nämä ovat potentiaali nopeaan uusien uutisten löytämiseen, käyttäjien kuluttajatottumuksien tarkkailu sekä Twitter-käyttäjäpopulaation yleisen mielialan huomiointi. Tässä tutkimuksessa perehdytään vain sävyanalyysin osalta relevanttiin yleisen mielialan huomiointiin. Koska taloudellisiin päätöksiin vaikuttaa myös kuluttajien tunnetila, on osakekurssien muutoksien ennakoinnin kannalta perusteltua tutkia Twitterissä ilmenevän kollektiivisen mielialan suhdetta osakekurssien käyttäytymiseen.

Bollen, Mao ja Zeng (2011) ovat tutkimuksessaan keskittyneet Twitter-datasta saatavan datan analyysiin arvottamalla kohdepopulaation

kollektiivisen tunnetilan suhdetta pörssikurssien liikkeisiin. Tutkimuksessa kerättiin lähdeaineistoksi kokoelma (noin 10 miljoonaa) twiittejä vuoden 2008 helmi- ja joulukuun väliseltä ajalta. Aineisto käsiteltiin siten, että välimerkit sekä stop word-sanat poistettiin näytteestä. Tämän jälkeen lähdeaineistosta louhittiin esiin eksakteja tunnetilaa kuvaavia ilmauksia, kuten ”I am feeling” ja ”I feel”.

Jäljellä olevan aineiston arvottamiseen käytettiin kahta eri työkalua. Nämä olivat OpinionFinder ja GPOMS. OpinionFinder on työkalu, jolla voidaan arvottaa yksittäisiä sanoja tunneskaalalla positiiviseksi ja negatiiviseksi. Tämän avulla oli mahdollista luoda päivittäinen kollektiivisen tunnetilan aikasarja positiivisen ja negatiivisen latauksen esityksenä. Vastaavasti GPOMS oli työkaluna edelleen kehitetty versio psykologiassa käytetystä kyselyyn perustuvasta mittaus-välineestä, POMS:ista (Profile of Mood States). POMS-kyselyn termistöä laajennettiin automaattista tiedonhakua varten siten, että alkuperäisistä 72 mielenkiinnon kohteena olevasta mitattavasta termistä muokattiin näihin liittyvien termien avulla 962 termin sanakirja. Tekstianalytiikan tarpeita ajatellen tämä lienee järkevää, jotta saavutetaan enemmän hakuosumia lähdeaineistosta. Tätä sanakirjaa vastaavia termejä etsittiin lähdeaineistosta ja luotiin kuusi-ulotteinen esitys kollektiivisesta tunnetilasta päivittäisenä aikasarjana. Tunnetilan ulottuvuuksia tässä olivat tyyneys (engl. calm), valppaus (engl. alert), varmuus (engl. sure), elinvoimaisuus (engl. vital), kiltteys (engl. kind) ja onnellisuus (engl. happy). Yhdistämällä OpinionFinderin ja GPOMSin luomat aikasarjat saatiin näin 7-ulotteinen päivittäinen aikasarja tutkittavalta ajanjaksolta. Näitä tuloksia verrattiin kyseisen ajanjakson DJIA-arvoihin (Dow Jones Industrial Average).

Lopputuloksena havaittiin, että tietyt tunnetilojen muodot ennustavat hyvin osakemarkkinoiden muutoksia, mutta eivät kuitenkaan kaikki. Esimerkiksi GPOMSin kategoriat tyyneys ja onnellisuus korreloivat tarkasteltavan datan kanssa jopa ennakoivasti mutta sen sijaan OpinionFinderin avulla luotu aikasarja yleisestä onnellisuustasosta ei korreloinut. Bollen, Mao ja Zeng (2011) pitävätkin mielenkiintoisena, että erityisesti GPOMsin kategorioista tyyneys (yhdistettynä GPOMS-onnellisuuteen) vaikutti olevan yhteydessä tuleviin muutoksiin DJIA-indeksissä. Muutokset näillä GPOMSin kategorioilla vaikuttivat ilmenevän 3-4 päivän viiveellä DJIA-indeksissä. Toisin sanoen tämä tulos antaa syytä olettaa, että Twitter-datasta esitetyllä tavalla saatava kuvaus kollektiivisesta tunnetilasta tai ilmapiiristä voi paitsi korreloida myös olla luonteeltaan ennustavaa osakekurssien kehityksen kannalta.

Vastaavia tuloksia saatiin Zhangin, Fuehresin ja Gloorin (2011) tutkimuksessa.

Zhang ym. (2011) mittasivat tutkimuksessaan kuuden kuukauden aikana kerätyllä twiittien otoksella kollektiivista toiveikkuuden ja pelon tuntemuksien määrää päivittäisellä tasolla. Tarkoituksena oli tutkia, voidaanko kyseisten ominaisuuksien aktiivisuuden tasosta havaita korrelaatioita osakemarkkinoiden tunnuslukuihin ja voidaanko näiden liikkeistä ennakoida osakemarkkinoiden liikkeitä.

Twiitit kerättiin vuoden 2009 maalis- ja syyskuun välisenä aikana sattumanvaraisena otoksena (satunnaisotos oli noin sadasosa päivittäisistä twiiteistä). Tästä otoksesta etsittiin pelkoa, epävarmuutta ja huolta kuvaavia sanoja. Tämän jälkeen Zhang, Fuehres ja Gloor (2011) tutkivat korrelaatiota eri osakemarkkinoiden tunnuslukuihin (Dow Jones, Nasdaq, S&P 500 ja VIX). Tutkimuksen perusteella taloudellisesti epävarmoina aikoina ihmiset alkavat käyttää enenevissä määrin tunteita kuvaavia sanoja kuten toivo, pelko ja huoli. Näiden käyttö korreloi vahvasti pörssikurssien muutoksien kanssa riippumatta siitä oliko sanojen käyttöyhteys negatiivinen tai positiivinen kontekstiltaan. Yksinkertaistaen Zhang, Fuehres ja Gloor (2011) väittävät tuloksiinsa nojaten Dow Jonesin laskevan mikäli edellisenä päivänä ihmiset ovat ilmaisseet paljon toivoon, pelkoon ja huoleen viittaavia tunteita. Vastaavasti näiden tunteiden ilmaisemisien vähentyessä Dow Jones nousee. Johtopäätöksenä Zhang, Fuehres ja Gloor (2011) toteavat Twitterissä esiintulevien voimakkaiden tunteellisten viestien määrän kasvun tai laskun antavan ennakoivaa tietoa siitä, kuinka pörssikurssit seuraavana päivänä reagoivat.

Julkisen mielialan arviointi ja mittaus on tärkeää monista eri tutkimusnäkökulmista katsottuna. Tämänkaltainen informaatio voi toimia myös merkittävänä tekijänä päätöksenteossa monilla eri aloilla. Perinteisin metodein tämänkaltaisen tiedon saaminen edellyttäisi usein kyselytutkimusten toteuttamista, joten Twitter-dataan perustuva tutkimusmenetelmä lienee tervetullut vaihtoehto.

Bertrand ym., (2013) mukaan sosiaalisen median tuottama data parantaa tämän kaltaisten tutkimusten tarkkuutta ajallisesti ja paikallisesti. Twitterin osalta geotaggays ja mobiilikäyttäjät sekä jokaisen twiitin varustaminen täsmällisellä aikatiedolla johtaa siihen, että hyvin tarkkoja mallinnuksia voidaan tehdä. Bertrand, Bialik, Virpee, Gros ja Bar-Yam (2013) mallintavat tutkimuksessaan New Yorkin alueella lähetettyjen twiittien avulla alueellisesti vallitsevaa mielialaa luoden ”sentimentti”-kartan New Yorkin alueesta kahden viikon ajalta. Aineistoksi kerättiin noin 600 tuhatta eksaktein geolokaatiotiedoin varustettua twiittiä New Yorkin alueelta kahden viikon ajalta huhtikuussa 2012. Aineiston analyysiin kehitettiin ohjatulla oppimisella luokittelutyökalu, joka huomioi twiittien sentimentin ja luokittelee nämä negatiiviseen ja positiiviseen sekä varustaa twiitit sävyn vahvuutta kuvaavalla arvolla. Luokittelutyökalun koulutusmateriaalina käytettiin hymiöitä sisältäviä twiittejä. Ennen analyysityötä twiittiaineisto esikäsiteltiin ja siistittiin poistamalla linkitykset ja käyttäjänimet. Tämän jälkeen twiitit arvotettiin sentimenttityökalulla. Sentimenttiarvoilla varustetut twiitit voitiin edelleen yhdistää geolokaatio- ja aikatietoihin, jolloin lopputuloksena voitiin luoda yleisen mielialan karttaesitys paikkatietojen osalta sekä tarkastella mielialan aaltoilua ajallisesti.

Tuloksissa havaittiin, että lähestymistavalla voitiin havaita yleisiä suuntauksia mielialoissa (Bertrand ym., 2013). Sentimenttien arvotuksen mukaan muokatulla kartalla kyettiin havaitsemaan hyvin selkeästi alueet, joilta vahvoja sentimenttiarvoja lähetettyjä twiittejä lähetettiin. Positiivisimmat twiitit

olivat pääosin puistoalueilta Central Parkin ollessa vahvinta ”onnellisuuden” aluetta. Negatiiviset sävyt liittyivät usein liikenteeseen ja negatiivisen sentimentin twiittien lähetyspaikkoja olivat esimerkiksi bussi- ja juna-asemat, sillat ja tunnelit. Vahvoja tunnelatauksia sisältäviä twiittejä lähetettiin myös hautuumailta, sairaaloista, vankiloista ja saastumisesta kärsiviltä alueilta.

Twiittien sävyanalyysissä havaittiin myös temporaalisia kaavamaisuuksia (Bertrand, 2013). Viikonpäivien aikana toistui sama kaavamaisuus, joka oli samankaltainen arkisin kuin myös viikonloppuna (viikonloppuisin vahvempi). Tuloksien perusteella ihmiset ovat keskimäärin onnellisempia kun eivät työskentele ja viikonloppuisin mielialat ovat positiivisempia kuin viikolla. Yleisesti ottaen havaittiin myös, että päivittäin sama mieliala-käyttäytyminen toistui; aamuisin sävyt olivat alavireisiä ja iltaisin korkeammalla.

Bertrand, Bialik, Virpee, Gros ja Bar-Yam (2013) arvioivat lähestymistavassaan olevan useita vahvuuksia. Verrattuna perinteisiin kartoittaviin kyselyihin geolokaatio-data on todella tarkkaa kuten myös aikatiedot. Twiittejä on myös saatavilla massiiiviset määrät ja niiden prosessointi voi olla huomattavasti nopeampaa ja kustannustehokkaampaa kuin perinteiset menetelmät. Tähän osaltaan vaikuttaa twiiteille ominainen lyhyys. Runsaasti asutulla ja paljon Twitteriä hyödyntävällä alueella saavutetaan myös varsin tasainen twiittien saatavuus analyysiaineistoksi. Tämä on eritoten hyvä mielialan ajallisten vaihteluiden tarkasteluun. Lisäksi Bertrand ym., (2013) toteavat menetelmän hyväksi puoleksi kieliriippumattomuuden, koska luokittelutyökalu on koulutettu hymiöiden eikä esimerkiksi ulkoisten sanastojen avulla.

Tämän kaltaisen datan entistä nopeampi, kustannustehokkaampi ja laajempi saatavuus varustettuna entistä tarkemmilla aika- ja paikkatiedoilla on omiaan avaamaan sovellusaloja erilaisiin tutkimuksiin, joilla voidaan tukea päätöksentekoa monilla eri alueilla. Suoria sovellutusalueita voisivat olla esimerkiksi kaupunkisuunnittelu ja ihmisten sosiaalisen käyttäytymisen ja elämänrytmien tutkiminen. Kenties hieman soveltamalla on mahdollista hyödyntää tämän tyyppistä dataa esimerkiksi liikennesuunnittelussa tarkkailemalla liikenteestä lähetettyjen twiittien aika- ja paikkatietoja sekä sentimenttiä.

Edellä on käsitelty sävyanalyysiä (engl. sentiment analysis) hyödyntäviä tutkimuksia. Bollenin, Pepen ja Maon (2011) mielialatutkimuksessa käyttämä psykologian alalta lainattu POMS (engl. Profile of Mood States) vaikuttaa erinomaiselta tavalta ryhmitellä twiittien sävyjä sanojen ja ilmauksien avulla. Samaa työvälinettä on käytetty myös monissa tutkimuksissa ja esimerkiksi Bollen, Mao ja Zeng (2011) hyödyntävät tämän laajennettua versiota GPOMS:ia osakemarkkinoita koskevassa tutkimuksessaan. Vaikuttaisi, että GPOMS soveltuu erinomaisesti Twitter-datan sentimenttianalyysiin.

Mitchell ym. (2013) esittämä onnellisuuspisteisiin perustuva malli voisi olla hyödynnettävissä esimerkiksi reaaliaikaisissa

arvioinneissa populaatiossa tapahtuvista muutoksista. Julkishallinnon kannalta sovellusalueena voisi olla yhteiskunnallisen muutoksen ja koetun onnellisuuden tilan tarkkailu Twitter-datasta. Lisäksi Bertrand ym. (2013) esittämä geolokaatiotietoja ja sentimenttianalyysia yhdistävä tutkimus lienee laajasti sovellettavissa erilaisiin käyttökohteisiin. Kollektiivisen mielialan ja osakekurssien välistä korrelaatiota tutkivat artikkelit (Bollen, Mao & Zeng, 2011 ja Zhang, Fuehres & Gloor, 2011) osoittavat tuloksiensa puolesta, että Twitter-datan pohjilta kyetään tekemään hyvin mielenkiintoisia havaintoja, mikäli oikeita riippuvuussuhteita ja indikaattoreita osataan tarkastella.

4.4 Poliittinen tutkimus

Twitter-dataa voidaan hyödyntää poliittisessa tutkimuksessa monella tapaa. Esimerkiksi poliittisia rakenteita on tutkittu tekstin kontenttia ja verkostoitumista tutkimalla (Conover ym. (2011a). Myös esimerkiksi poliittista polarisaatiota on tutkittu poliittis-sävytteisiä hashtageja hyödyntämällä (Conover, Ratkiewicz, Francisco, Goncalves, Flammini & Menczer, 2011b). Tekstianalytiikan soveltamisessa poliittiseen tutkimukseen on kuitenkin esteitä (Gayo-Avello, 2012) eikä esimerkiksi sävyanalyysi vaikuta olevan välttämättä kovinkaan käyttökelpoinen menetelmä poliittisissa aiheissa.

Tumasjan, Sprenger, Sandner ja Welpe (2010) ovat tutkineet Twitterin tuottaman datan analyysiä poliittisessa kontekstissa ja vaalien tuloksien ennakoinnissa. Tutkimuksessa tutkittiin Saksan liittotasavaltavaalien aikana Twitter-viestien sisältöä. Tarkoituksena oli havaita käytettiinkö Twitteriä poliittisten ilmauksien välittämisen välineenä ja voitiinko tästä datasta tehdä johtopäätöksiä koskien vallitsevaa poliittista asennetta. Lisäksi mielenkiinnon kohteena oli voidaanko vaalien lopputulosta ennakoida Twitter-datan analytiikalla.

Tutkimuksessa käytettiin tekstianalyysiin tarkoitettua ohjelmistoa, jolla sisällön analyysi lähdeaineistosta suoritettiin. Lähdeaineistona toimi kyseiseltä ajalta kerätty reilun sadan tuhannen twiitin kokoelma, joka oli kerätty hakusanojen perusteella. Näitä hakusanoja olivat viittaukset tiettyihin puolueisiin ja puolueiden edustajiin. Aineistoa käsiteltiin mielipidelouhinnan (engl. sentiment analysis) välineellä, jolla analysoitiin erilaisia sävyjä poliittisesti värittyneestä aineistosta. Näitä olivat esimerkiksi positiiviset ja negatiiviset tunteet, tulevaisuus- tai menneisyys-orientoituneisuus, viha, suru, levottomuus sekä rahaan, saavutuksiin ja työhön liittyvät tunteet.

Tuloksien pohjalta Tumasjan, Sprenger, Sandner ja Welpe (2010) kommentoivat, että jo pelkkä twiittauksien lukumäärä eri puolueisiin heijasteli vaalien lopputulosta ja on loppujen lopuksi lähellä virallisia ennusteita vaalien lopputuloksesta. Tämä johtopäätös on saanut osakseen paljon kritiikkiä. Esimerkiksi Skoricin, Poorin, Achananuparpin, Limin ja Jiangin (2012) mukaan korrelaatiota twiittien lukumäärän ja lopputulosten välillä on havaittavissa mutta ei kuitenkaan siinä määrin, että ennakointia vaalien lopputuloksista

voitaisiin tehdä. Tumasjan ym. (2010) myös toteavat, että useampia puoluemainintoja sisältäneet viestit heijastelivat todellisessa maailmassa olevaa poliittista ”lähekkäisyyttä” kyseisien puolueiden välillä. Analyysin perusteella he myös päätyivät lopputulokseen, että Twitterissä ilmaistut poliittiset mielipiteet heijastelivat hyvin todellisessa maailmassa vallitsevaa poliittista ilmapiiriä. Tutkimuksen mukaan ihmiset käyttävät Twitteriä poliittisen keskustelun ylläpitämiseen, joskin tässä erottuu selkeästi toisistaan huomattavan aktiiviset ”keskustelun ylläpitäjät” ja passiivisempi muu käyttäjäkunta. Tutkimuksen perusteella tekijät kokevat, että Twitteriä voitaisiin käyttää muiden vaalien ennustamiseen tarkoitettujen menetelmien ohella täydentävänä tiedonlähteenä.

Poliittisten suuntauksien analysoinnissa Twitter-datan soveltuvuutta luonnollisesti rajoittaa jossain määrin tilastollisessa mielessä riittämätön otos esimerkiksi alueilla, joilla Twitterin käyttö ei ole kovin suosittua. Myös käyttäjät sinällään eivät ole poikkileikkaus koko äänestävästä populaatiosta. Tumasjan, Sprenger, Sandner ja Welpe (2010) toteavat myös tutkimuksessa käytetyn tekstianalyysityökalun olleen räätälöimätön Twitter-viestien lyhyyteen.

Jungherr, Jürgens ja Schoen (2011) ovat osoittaneet kritiikkiä Tumasjanin ym. (2010) tutkimusta kohtaan. Ensinnäkin tutkimuksessa kerätty tutkimusaineisto oli huonosti valittu ja kerätty. Lisäksi aineiston kokoamista ei ole kuvattu riittävän tarkasti, jotta tutkimus voitaisiin toistaa. Näin ollen Jungherr, Jürgens ja Schoen (2011) eivät pidä johtopäätöksiä valideina mikä johtaa näiden käyttökelvottomuuteen tulevia vaaleja arvioitaessa. Tutkimusaineisto koski lisäksi pelkästään jälkikäteen valittuna puolueita, jotka pääsivät hallitukseen. Mikäli saksalainen Pirate Party-puolue olisi otettu tarkasteluun mukaan, olisi se twiittiaineiston perusteella ollut vaalien suurin voittaja, vaikkei nyt päässyt hallitukseen. Johtopäätöksenä tästä Jungherr, Jürgens ja Schoen (2011) päätyvät väittämään, että twiiteissä esiintyvien poliittisten puolueiden mainintojen lukumäärä itsessään ei ole pätevä indikaattori todellisen maailman poliittisen kentän kuvaamiseen tai vaalien lopputuloksen ennustamiseen. Myös Tjong, Sang ja Bos (2012) päätyivät tutkimuksessaan siihen tulokseen, että pelkkä twiittien lukumääräinen laskeminen ja arvioiminen ei ennakoi vaalien lopputuloksia. Toisaalta Skoricin ym. (2012) mukaan valtiollisella tasolla on mahdollista saavuttaa suuntaa-antava yleiskuva poliittisista vahvuussuhteista mutta tämä ei päde paikallisella tasolla.

Gayo-Avello (2012) kritisoi artikkelissaan Twitter-datan käytettävyyttä poliittisessa ennustamisessa. Hänen mukaansa yhdessäkään tutkimuksessa ei ole vielä ennustettu vaalien lopputuloksia vaan tuloksia on käsitelty pelkästään takautuvasti. Twiittien tekstianalytiikkaan pohjautuvaa poliittista tutkimusta koskevina ongelmina Gayo-Avello (2012) näkee muun muassa seuraavat tekijät:

- Ei ole vielä järjellistä tapaa arvioida twiittejä toteutuvina ääninä vaaleissa
- Twiittien ei voida olettaa olevan luotettavia. Huhut, propaganda ja potentiaalisesti väärä tieto täytyy huomioida
- Demografisia tekijöitä ei juuri oteta huomioon
- Data-aineisto eli twiitit tulevat vain poliittisesti aktiivisilta Twitterin käyttäjiltä

Johtöpäätöksenä Gayo-Avello (2012) toteaa, että Twitterin ennustavaa voimaa politiikan suhteen on suuresti liioiteltu ja että vaikeita ongelmia on edelleen jäljellä ratkaistavaksi, jotta poliittista tutkimusta voidaan toteuttaa. Esimerkiksi Metaxas, Mustafaraj ja Gayo-Avello (2011) tutkimuksessa sosiaalisen median datan perusteella ei saavutettu juurikaan sattumanvaraista tulosta parempaa ennustavuutta vaalien osalta. On mahdollista, että poliittinen pelikenttä osoittautuu vaikeaksi sovellusalueeksi Twitterin tekstianalytiikalle.

Ottaen huomioon esitetyn kritiikin on tarpeen kehittää uusia lähestymistapoja, jotka kykenevät parantamaan poliittisen ennustamisen menetelmiä. Poliittisesti aktiivisten Twitterin käyttäjien aiheuttamaa tutkimusaineiston vääristymää voidaan tasoittaa esimerkiksi karsimalla näiden käyttäjien twiittejä aineistosta. Näin onkin tehty esimerkiksi Tjong ym. (2012) tutkimuksessa. Tumasjanin ym. (2010) puoluemainintojen lukumääräisen lähestymistavan sijaan sävyanalyysiä yhdistelemällä syvälliseen geolokaatio- ja demografiseen tietoon olisi mahdollista parantaa tuloksia. Esimerkiksi Metaxas ym. (2011) mukaan vaalien ennakoimista varten käytetyn menetelmän tulisi kyetä ottamaan huomioon demografiset erot Twitter-käyttäjien ja todellisen populaation välillä. Mikäli voidaan luoda alueellinen kuvaus poliittisista sävyistä – tässä tapauksessa siis negatiivisista ja positiivisista tuntemuksista eri puolueita kohtaan – ja verrata tätä alueella asuvan populaation määrän ja aiemman vaalidatan perusteella saatuun äänestysprosenttiin voitaneen luoda tarkempi kuvaus poliittisesta kentästä Twitterin avustuksella. Toisaalta jo sävyanalyysin suorittaminen poliittisista aiheista voi osoittautua vaativaksi haasteeksi. Tjong ym. (2012) tutkimuksen mukaan sävyanalyysi parantaa poliittisten tutkimuksien tuloksia jossain määrin mutta esimerkiksi Metaxas ym. (2011) mukaan tarvitaan jatkotutkimusta siitä, missä määrin sävyanalyysiä ylipäänsä voidaan hyödyntää poliittisissa aiheissa.

Edellä on lyhyesti käsitelty poliittisen tutkimuksen tekemistä Twitter-dataan perustuvan tekstianalytiikan avulla. Keskeisenä huomiona voidaan havaita, ettei twiittien lukumäärää voida pitää ennakoivana indikaattorina vaalien lopputuloksesta (Jungherr ym., 2011). Lisäksi havaitaan poliittisten aiheiden vaikea tulkittavuus tekstianalytiikan avulla.

4.5 Joukkoistettu aistinta ja seuranta

Tässä alaluvussa käsitellään Twitter-datan tekstianalytiikkaa joukkoistetun aistimisen näkökulmasta sekä käytännön seurantajärjestelmänä. Joukkoistamisen osalta esimerkkitutkimuksina perehdytään kahteen Twitter-pohjaiseen maanjäristysten havaitsemisjärjestelmään. Lopuksi käsitellään edistyksellistä toteutusta automatisoidun seurantajärjestelmän malliksi.

Joukkoistamisella (engl. crowdsourcing) tarkoitetaan toimeksiantajan tietyn tehtävän hajautettua suorittamista yhteisön taholta. Twitterin osalta tämä tarkoittaa jonkin tietyn ongelman tai tehtävän suorittamista käyttäjien avustuksella. Käyttäjät voivat osallistua joidenkin tehtävien suorittamiseen vapaaehtoisesti tai tahattomasti yksinkertaisesti twiittaamalla ja tiedottamalla ympärillä tapahtuvista asioista. Tällä tavoin he toimivat sosiaalisen sensorin roolissa (Sakaki, Okazaki & Matsuo, 2010). Teoriassa mobiilikäyttäjien osalta olisi myös mahdollista varustaa käyttäjien suostumuksella mobiililaitteita erilaisilla sensoreilla, jotka käyttäisivät Twitteriä informaation jakamisen alustana (Demirbas, Bayir, Akcora, Yilmaz & Ferhatosmanoglu, 2010). Tämä tarjoaisi kokonaan uusia mahdollisuuksia esimerkiksi kansanterveydellisille tutkimuksille.

Steelen (2011) mukaan lisääntyvä sensorien integrointi ja mobiilien laitteiden yleistyminen sosiaalisessa mediassa mahdollistavatkin uusia tapoja terveysviranomaisille terveysdatan keräämiseen. Steele (2011) ei kuitenkaan pidä todennäköisenä, että näitä voitaisiin ottaa välittömästi käyttöön ilman vakavia pohdintoja yksityisyyden suojasta ja eettisistä kysymyksistä.

Maanjäristykset ovat tarjonneet hyvän tutkimuskohteen Twitterin pohjalta toteutetuille joukkoistettuun aistintaan perustuville järjestelmille. Twitteriin perustuvat nopeat joukkoistetut jäljitysjärjestelmät ovat mahdollisia, koska ihmiset alkavat lähettää twiittejä kymmenien sekuntien viiveellä järistyksen tuntemisesta ja havainto merkittävästä maanjäristyksestä voidaan tehdä noin kahden minuutin jälkeen (Earle, Bowden ja Guy, 2012). Toisin sanoen taustalla on yksinkertainen olettamus siitä, että missä maanjäristyksiä tapahtuu ja ihmisillä on Twitter-tilejä, tulevat nämä myös niistä twiittaamaan. Maanjäristyksistä raportoivat twiittaajat muodostavat näin joukkoistetun sensoriverkon, jonka aika- ja paikkatietoja analysoimalla voidaan muodostaa esimerkiksi kohtuullisen reaaliaikainen kuva käynnissä olevan maanjäristyksen sijainnista. Käytännön sovellutuksena tämän yhteyteen voidaan toteuttaa esimerkiksi varoitusjärjestelmä (Sakaki, Okazaki & Matsuo, 2010).

Sakaki, Okazaki ja Matsuo (2010) ovat tutkineet Twitterin luoman datan käyttöä maanjäristysten seurannassa ja paikallistamisessa. Tutkimuksessa käytettiin maanjäristyksiin liittyvien twiittien osalta tunnistusalgoritmia, joka teki hakusanaetsintöjä sekä analysoi twiittien pituutta ja kontekstia ja tämän perusteella luokitteli twiitit maanjäristystä koskeviksi tai merkityksettömiksi. Tämän luokittelijan toteuttamisessa käytettiin ohjattua koneoppimista ja harjoitusdataa, sillä pelkkä hakusanaetsintä ei ole riittävä. Kunkin

maanjäristystä käsittelevän twiitin taustalla oleva käyttäjä ymmärretään sensoriksi, joka ilmoittaa tapahtuvasta maanjäristyksestä. Paikannuksen arvioinnissa käytettiin näiden ”sosiaalisten sensorien” tuottaman datan osalta erilaisia filtteröintitekniikoita, joiden tuloksena maanjäristyksien keskuksia voitiin määrittää.

Tämän järjestelmän yhteyteen kehitettiin maanjäristyksistä varoittava raportointijärjestelmä, jota testattiin Japanissa. Japani oli luonnollisesti testaamiseen oiva maa lukuisten ja säännöllisten maanjäristysten sekä kohtuullisen korkean Twitterin käyttäjämäärän vuoksi. Mahdollisesti osittain tästä johtuen tuloksena oli 96 % korrelaatio Japanin meteorologisen viraston virallisten maanjäristysrapottien kanssa. Lisäksi varoitusjärjestelmän kautta rekisteröityneet käyttäjät saivat sähköpostitse varoitusviestin meneillään olevasta järistyksestä nopeammin kuin kyseiseltä virastolta. Vastaavia menetelmiä on jo jossain määrin otettukin käyttöön. Earlen, Bowdenin ja Guyn (2012) mukaan Yhdysvaltain maanjäristyksiä tutkiva organisaatio U.S. Geological Survey (USGS) tutkii miten Twitteriä voidaan käyttää osana maanjäristys-reagointijärjestelmiä. Tällä hetkellä Twitteriä käytetään jo sangen yleisesti tiedostusväylänä seismisesti virallisiksi todetuille maanjäristysvaroituksille esimerkiksi Euroopassa, Kanadassa ja Indonesiassa (Earle, Bowden ja Guy, 2012).

Twiittaajien käyttäminen eräänlaisina sosiaalisina sensoreina on mahdollisesti tämän kaltaisten tekniikoiden kannalta ongelmallista johtuen sensoreiden ihmismäisestä luonteesta. Sakaki, Ozakaki ja Matsuo (2010) huomauttavatkin, että näillä sensoreilla on valtavia keskinäisiä eroja ja ne ovat hyvin erilaisia. Esimerkiksi jotkut sensorit (käyttäjät) ovat hyvin aktiivisia ja toiset eivät, kun taas jotkin sensorit saattavat olla poissa päältä (nukkuvat). Lisäksi käyttäjät puhuvat paljon muustakin kuin maanjäristyksistä. Tämä johtaa siihen, että Twitter-käyttäjien luoma sosiaalisten sensorien verkko on paljon epävarmempi kuin fyysisten sensorien luoma verkko.

Earle, Bowden ja Guy (2011) ovat tutkineet myös Twiitteihin perustuvan tekstianalytiikan käyttöä maanjäristysten havaitsemisessa ja tutkimuksessa esitellään pelkästään twiitteihin perustuva maanjäristysten havaitsemisjärjestelmä. Earlen, Bowdenin ja Guyn (2012) mukaan twiiteissä esiintyvien maanjäristykseen viittavien sanojen eri kieliset versiot selvästi korreloivat maanjäristysten tapahtumisen ajankohtien kanssa. Tutkimuksessa keskityttiin näiden sanojen esiintymistiheyden ripeän kasvamisen löytämiseen ”short-term over long-term”-algoritmin avulla. Algoritmi siis vertasi sanojen suhteellista esiintymistä menneen esiintymisen kanssa ja päätteli tätä kautta milloin järistyksistä kertovien twiittien runsaus ylitti kriittisen rajan. Tutkimusaineistona käytettiin viiden kuukauden aikana vuonna 2009 globaalista twiittidatasta kerättyä twiittiä, josta valikoitiin esiin maanjäristystä kuvaavia sanoja sisältävät twiitit. Twiiteistä kerättiin aika-, paikka- ja tekstitiedot. Mikäli twiitit eivät sisältäneet GPS-pohjaista geolokaatiotietoa, arvioivat tutkijat twiitin sijainnin käyttämällä käyttäjän ilmoittaman paikkakunnan koordinaatteja, jotka saatiin Google Maps ohjelmointirajapinnan

kautta. Menetelmällä löydettiin 48 maanjäristystä, joista 2 oli vääriä hälytyksiä. Vastaavasti virallisten seismologisten tilastojen mukaan kyseisellä aikavälillä tapahtui 5175 maanjäristystä. On siis selvää, että ihmismäisten sensorien havaitsemat maanjäristykset ovat pelkästään kokoluokassa, joka voidaan selkeästi havaita (Earle, Bowden ja Guy, 2011). Toisena heikkoutena oli se, että tarkkaa lokaatiotietoa ei saatu. Toisaalta menetelmän vahvuuksia ovat Earlen, Bowdenin ja Guyn (2012) mukaan se, että isot maanjäristykset havaitaan hyvin nopeasti. Nämä ovat merkittävimpiä ja vaativat vasteajaltaan ripeän hälytysjärjestelmän, jotta mahdollisiin toimenpiteisiin voidaan varautua optimaalisesti. Menetelmän avulla nämä isot järistykset havaittiin 2 minuutin sisällä. Twiittien tekstianalytiikan avulla voidaan saada myös tilanteen kartoittamisen kannalta oleellista tietoa, jota voidaan käyttää päätöksenteossa alueella tapahtuvista pelastustoimenpiteistä. Toisin sanoen twiitit voivat kätkeä sisäänsä informaatiota alueella tapahtuneista potentiaalisista vahingoista (Earle, Bowden & Guy, 2012).

Tutkimusta voidaan kritisoida geolokaatiotietojen epäluotettavasta käsittelystä. Luotettavampien paikkatietojen saamiseksi olisi tutkimusaineistona voinut käyttää pelkästään eksaktin geolokaatiotiedon (mobiilikäyttäjät ja GPS-paikkatiedon sisältävät twiitit) sisältäviä twiittejä. Nyt käytetty menetelmä, jossa epävarman paikkatiedon sisältävät twiitit huomioitiin ei ole luotettava. Käyttäjän Twitteriin rekisteröitymisen yhteydessä ilmoittaman paikkatiedon käyttäminen Google Maps API:ssa antaa toki koordinaatit, mutta tämä geolokaatiotieto tuskin juuri koskaan on täsmällinen. Tältä osin menetelmää voidaan helposti parantaa. Tulevaisuudessa mobiilikäytön ja täten myös eksaktien geolokaatiotietojen lisääntyessä lieneekin perustellumpaa karsia ei-täsmälliset paikkatiedot kokonaan pois havaintoaineistosta. Earle, Bowden ja Guy (2012) huomioivat tämän myös itse tyytyen toteamaan epätarkkojen geolokaatiotietojen olevan vakava ongelma tarjoamatta kuitenkaan parannusehdotusta. Tutkimuksen lopputuloksena Earle, Bowden ja Guy (2012) toteavat osoittaneensa mahdolliseksi rakentaa twiitteihin perustuvan maanjäristysten havaitsemisjärjestelmän, jonka virhemarginaali on pieni. Järjestelmä ei kuitenkaan kykene havaitsemaan pienen mittakaavan järistyksiä, koska ihmiset eivät näitä havaitse eivätkä kaikki sensorit (Twitter-käyttäjät) välttämättä näistä raportoi. Täten Earle, Bowden ja Guy (2012) tulevat johtopäätökseen, ettei Twitter-pohjainen järjestelmä voi olla korvaava seismologisille järjestelmille. Kuitenkin hyötynäkökulmasta Twitter-pohjainen järjestelmä on globaali, lähes reaaliaikainen ja kustannustehokas (Earle, Bowden & Guy, 2012). Lisäksi twiittien tekstisosuuksista voidaan saada tilannetietoisuuden kannalta merkityksellistä ajantasaista tietoa.

Esitellyt maanjäristyksien havaitsemismenetelmät nostavat esiin kysymyksen soveltuvuudesta muihin käyttötarkoituksiin. Menetelmät voisivat olla eri toten hyödynnettävissä erilaisten kriisien havaitsemiseen ja päätöksentekoon vaikuttavan informaation saamiseen paikan päältä. Edellä käsiteltyjen menetelmien pohjalta edelleen kehitellyt trend detection-järjestelmät yhdistettynä esimerkiksi trendien endo- ja eksogeeniseen

tunnistukseen (Naaman, Becker ja Gravano (2011) voisivat olla käytettävissä myös monitarkoitukselliseen, universaaliin hälytysjärjestelmään, jossa havaittavaa ilmiötä ei etukäteen tiedetä. Esimerkiksi ripeästi nousevien trendien tunnistusmenetelmä yhdistettynä riittävän tarkasti esitettyyn taksonomiaan erilaisesta kriisisanastosta voisi olla toteutettavissa. Tällainen järjestelmä voisi havaita (ja tiedottaa viranomaistahoja) minkälaisen kriisin tahansa puhjetessa. Käyttösovellutukset voisivat vaihdella havaintojen tekemisestä mellakoiden ja ampumavälikohtauksien osalta aina infrastruktuurisiin ongelmiin ja luonnonkatastrofeihin.

Aggarwal ja Abdelzaher (2011) ovat käsitelleet sensorien ja sosiaalisten verkkojen hyödyntämisen haasteita joukkoistamisen näkökulmasta. Koska sosiaalisen median palveluissa hyödynnetään nykyisin monenlaisia sensoreita (kenties tärkeimpänä paikkatiedon antavat sensorit) nousee yksityisyys ja sen suojeleminen tärkeäksi seikaksi joukkoistettujen järjestelmien kannalta. Koska tämänkaltainen tieto on käyttäjän yksityistä dataa, ei Aggarwalin ja Abdelzaherin (2011) mukaan pelkkä tiedon anonymiteetti takaa yksityisyyden säilyttämistä. On esimerkiksi selkeää näyttöä, että twiittien perusteella voidaan päätellä käyttäjän sijainti sangen tarkasti (Hecht, Hong, Suh & Chi, 2011). Myös Steele (2011) pitää yksityisyyden suojaa ja eettisiä kysymyksiä esteinä esimerkiksi kansanterveydellisten tutkimuksien toteuttamiselle sensoripohjaisesti sosiaalisen median välityksellä. Aggarwalin ja Abdelzaherin (2011) mukaan joukkoistettuun sensorien hyötykäyttöön suunniteltujen järjestelmien arkkitehtuuri tulisi suunnitella siten, että kyettäisiin vastaamaan kolmeen keskeiseen haasteeseen. Näitä ovat heidän mukaansa yksityisyyttä varjelevien tekniikoiden kehittäminen, suurien datamäärien käsittely ja reaaliaikaista päätöksentekoa tukevien järjestelmien toteuttaminen.

Twiitti-virtaa seuraavien joukkoistettujen järjestelmien kannalta on mahdollista kehittää monenlaista toiminnallisuutta. Esimerkiksi Lee ja Sumiya (2010) ovat esittäneet menetelmän, jolla twiittien tekstianalytiikalla voidaan määrittää joukkokäyttäytymistä yhdistäen tämä geolokaatiotietoihin. Tämä periaatteessa mahdollistaa alueellisten tapahtumien, esimerkiksi musiikkifestivaalien, havainnoinnin pelkästään twiittien perusteella. Varsinaista sovellutusta tälle esimerkiksi terrorismin ehkäisyssä ei liene mutta joukkomellakoinnin paikallistamisessa ja ehkäisemisessä vastaaville menetelmille lienee käyttöä (EMOTIVE-projekti, 2013a & 2013b).

Rikosten ennakoimisella pyritään hahmottamaan kaavamaisuuksia, joilla voidaan arvioida missä ja milloin rikoksia tehdään todennäköisimmin. Tyypillisesti vertailemalla historiallista rikosdataa ja seuraamalla spatiaalis-temporaalisia trendejä voidaan saada käsitys siitä, mitkä alueet milloinkin ovat todennäköisimpiä rikosten tapahtumispaikkoja. Tämä helpottaa esimerkiksi optimaalisessa poliisivoimien resurssien sijoittelussa ja käytössä. Esimerkiksi Wang, Brown ja Gerber (2012) ovat tässä yhteydessä tutkivat spatiaalis-temporaalisen mallin yhdistämistä Twitter-datan tekstianalytiikkaan pyrkien parantamaan rikollisuuden ennakointiin käytettävän datan tarkkuutta hyvin

tuloksin. Potentiaalisten rikosten, mellakoiden ja muiden ei-toivottujen tapahtumien toteutumista voidaan myös ennakoida tarkkailemalla populaation mielialaa ja vihaa kuvaavia tunteellisia twiittejä alueellisesti. Seuraavassa käsitellään EMOTIVE-projektia (engl. Extracting the Meaning Of Terse Information in a Visualisation of Emotion), missä toteutetaan tällainen automatisoitu seurantajärjestelmä.

Onnistuneesti toteutetussa seurantajärjestelmässä populaation reaktioiden tarkkailun on oltava automatisoitua ja ennen kaikkea skaalautuvaa. Viranomaisten kannalta erilaisen valvonnan suorittaminen pelkästään ihmisen tekemän asiantuntijatarkkailun avulla tulee hyvin nopeasti tiensä päähän yksinkertaisesti datan määrän vuoksi. EMOTIVE-projektissa (2013a & 2013b) esitellään erityisen hienojakoinen ja tarkoitukseensa räätälöity ontologia sekä tätä hyödyntävä valvontajärjestelmä tunneskaalojen visualisoituun tarkkailuun maantieteellisesti. Kyseinen ontologia mallintaa twiiteistä havaittavia tunteisiin liittyviä ilmauksia ja käsitteitä sekä niiden välisiä suhteita. EMOTIVE-projektissa (2013a & 2013b) kiinnitetään huomiota siihen, että aiemmissa sävyjen luokittelua koskevissa tutkimuksissa ollaan usein yksioikoisesti keskitytty jakamaan tunteita positiivisiin tai negatiivisiin taikka muutoin tyydytty sangen karkean tason jaotteluun. EMOTIVE-projektissa (2013a & 2013b) on todettu, ettei tämä kuitenkaan ole optimaalista kuvastamaan reaalimaailmaa vastaavaa tilaa. EMOTIVE-projektissa (2013a & 2013b) esittellään kahdeksan tunteen tulkintaan perustuva ontologia. Näitä tunteita ovat viha, hämmennys, inho, pelko, onnellisuus, surullisuus, häpeä ja yllättyneisyys. Luokittelu perustuu nopeaan luonnollisen kielen käsittelyn (Natural language processing, NLP) työvälineeseen, joka kykenee luokittelemaan noin 1500 twiittiä sekunnissa normaalilla tietokonelaitteistolla. Ontologia perustuu yli 300 tunteisiin liittyvään termiin. Lisäksi edustettuna ovat slangisanat, hymiöt, tunteiden vahvuutta lisäävät ja vähentävät sanat sekä negaatio-ilmaukset (kuten: ”en ole vihainen”). Tunteisiin liittyvät sanat on ontologiassa painotettu ilmaistun tunteen vahvuuden mukaan. Testattaessa ontologian ja luonnollisen kielen työkalun tuottamia arvioita twiittien tunteellisesta kontentista saavutettiin 98% korrelaatio kahden ihmisen suorittamien tulkintojen kanssa. Toisin sanoen twiittien tunnelatausta kyettiin arvioimaan hyvin eksaktisti automatisoidulla ja nopealla järjestelmällä. Yhdistettynä geolokaatio- ja aikatietoihin sekä visualisointitekniikoihin tämä mahdollistaa automaattisen seurantajärjestelmän luomisen. Potentiaalisesti tämä seurantajärjestelmä voitaisiin varustaa myös automaattisilla hälytysjärjestelmän ominaisuuksilla esimerkiksi soveltamalla kriittisiksi miellettäviä sanoja ja ilmauksia yksinkertaisella trendin tunnistusmenetelmällä. Näiltä osin esimerkiksi Mathioudakisin ja Koudasin (2010) sanaryöpsähdysten tunnistamiseen liittyvä menetelmä voisi toimia järjestelmän rinnalla ja laukaista automaattisen hälytyksen erilaisten kriisi-, rikos-, terrorismi- tai katastrofisanojen poikkeavan lisääntymisen kohdalla.

Itse järjestelmä käyttää hyväkseen ontologiaa ja geolokaatiotietoja luoden visualisoidun tunnekartan karttapohjalle (EMOTIVE-projekti, 2013a &

2013b). Tämä mahdollistaa monipuolisen seurannan, kuten sen miten paikalliset reaalimaailman tapahtuvat heijastuvat Twitterissä. Järjestelmää voidaan käyttää esimerkiksi mellakoiden ennakoimiseen etsimällä. Indikaattoreita tulevasta voidaan havaita negatiivisten joukkotunteiden heräämisenä paikallisesti ja syvemmän analyysin osalta voidaan twiiteistä paikallistaa sovittuja tapaamisia. Yksinkertaistettuna järjestelmän hyöty on se, että sen käyttö auttaa analysoimaan tapahtumassa tai tapahtumaisillaan olevia asioita. Näin ollen viranomaistahojen toimintavalmius paranee.

Järjestelmän toiminnan kannalta EMOTIVE-projektin (2013a & 2013b) kehittäjät mainitsevat useita edistyksellisiä asioita. Tekstinkäsittelyn osalta järjestelmä huomioi spam-viesteiksi luokiteltavat twiitit sekä twiittien kontentin osalta järjestelmä löytää kuvalinkatut twiitit ja lähteet. Tunneilmausten osalta järjestelmällä voidaan rakentaa myös käyttäjien psykologisia profiileja perustuen käyttäjän aiempien twiittien tunnekontenttiin. Tietyllä tavalla radikaali muutos twiittien sävyssä riskikäyttäjäksi tunnistetun käyttäjän kohdalla mahdollistaisi teoriassa jossain määrin automatisoidun tarkkailun tämän tapaisissa tilanteissa. Luonnollisesti näiltä osin yksittäisen käyttäjän henkilökohtaiseen seuraamiseen liittyy lainsäädännöllisiä ja eettisiä ongelmia yksityisyyden kannalta. EMOTIVE-projektin (2013a & 2013b) kehittäjien mukaan järjestelmän toiminnallisia osia voidaan laajalti soveltaa esimerkiksi käsittelemään Facebook- ja sähköpostiperäistä tekstimuotoista dataa.

Kokonaisuutena EMOTIVE-projektissa (2013a & 2013b) esitelty ontologia ja sitä hyödyntävä järjestelmä ovat perusmalli käyttökelpoiselle automaattiselle valvontajärjestelmälle. Menetelmän eräänä sovelluksena on nähty Lontoon 2011 mellakoiden kaltaisten tapahtumien ennakoiminen Twitter-dataan pohjautuvien indikaattorien havaitsemisen avulla. Tekstianalytiikan osalta tässä yhdistyvät aika- ja paikkatietojen sekä sävyanalyysin tuottamien tietojen hyödyntäminen yhdessä visualisointitekniikoiden kanssa. Järjestelmän toteutustapaa voitaneen pitää hyvänä pohjana alettaessa kehittää automaattisia seurantajärjestelmiä. Huomio kiinnittyy kuitenkin jo välittömästi siihen, mitä muuta analytiikkaa järjestelmään voitaisiin sisällyttää. Eräs mahdollisuus on varustaa seurantajärjestelmä Mathioudakisin ja Koudasin (2010) esittelemän yksinkertaisen trendin havaitsemisjärjestelmän kaltaisella toiminnallisuudella. Tämä mahdollistaisi automaattisten hälytysten tekemisen kriisejä koskevien twiittien lisääntyessä radikaalisti. Samoin monella tapaa mallin analyyttistä puolta voitaisiin jatkokehittää. Mielekästä olisi myös kyetä käyttämään samaa järjestelmää kriisinajan toiminnallisessa ohjaamisessa informaation saamiseen. Käytännössä tämä tarkoittaisi kriisialueella paikallisesti tapahtuvan twiittauksen seuraamista reaaliaikaisen informaation saamiseksi.

Edellä on käsitelty Twitter-dataan perustuvan tekstianalytiikan hyödyntämistä joukkoistettuna. Esimerkkeinä on käsitelty kahta maanjäristysten havaitsemisjärjestelmää (Sakaki ym., 2010 ja Earle ym., 2011). Lopputuloksena voidaan todeta järjestelmien toimineen erittäin hyvin riittävän

isojen maanjäristysten kohdalla. Sovellutuskohteita vastaaville tekniikoille on potentiaalisesti lukuisia. Lopulta on esitelty sangen edistyksellinen malli reaaliaikaisen seurantajärjestelmän toteuttamiseksi (EMOTIVE-projekti, 2013a & 2013b).

4.6 Kirjallisuuskatsauksen tulokset ja yhteenveto

Tässä alaluvussa on käsitelty kirjallisuuskatsauksen eli lukujen kolme ja neljä tuottamia tuloksia. Tulokset on esitelty lyhyesti viitaten keskeisiin havaintoihin. Samalla esitetään aiheeseen liittyvää pohdintaa ja yhteenvetoa sekä vastataan kirjallisuuskatsauksesta saadun informaation valossa ensimmäiseen tutkimuskysymykseen. Alaluvun päättää kirjallisuuskatsauksen pohjalta luotujen hypoteesien esittely empiiristä vaihetta varten. Arvioitaessa käytettyjä menetelmiä ja niiden soveltuvuutta julkishallinnon kohdeorganisaatioiden kannalta on pyritty pitäytymään löyhähkösti Hevner ym. (2004) arviointiperusteissa. Toisin sanoen tutkija on muodostanut oman käsityksensä esimerkkitapauksissa käytettyjen IT-artifaktien käyttökelpoisuudesta. Samoin on pyritty arvioimaan menetelmien kompleksisuutta sekä muita menetelmien kannalta olennaisia ominaisuuksia. Lopulta tutkielman tekijä on pyrkinyt luomaan alustavia arvioita potentiaalisista käyttökohteista ja –skenaarioista julkishallinnon kannalta.

Kolmannessa luvussa on käsitelty Twitteriä, Twitterin käytön tyypillisiä piirteitä ja sosiaalisen median sekä Twitterin tuottaman datan hyödyntämistä tekstianalytiikalla. Lisäksi luvussa käsiteltiin Twitter-datan analysoitavuutta.

Twitterin osalta todettiin, että ihmiset käyttävät Twitteriä keskustellakseen ja jakaakseen uutisia sekä linkkejä (Java ym., 2007). Käyttäjät ilmaisevat twiiteissään tyypillisesti mielipiteitä, tunteita ja informaatiota (Phuvipadawat & Murata, 2010), joiden kartoittaminen olisi moninaisten tutkimuskohteiden kannalta tavoiteltavaa. Koska Twitter-data on julkista tekstidataa ja varustetaan aika- ja paikkatiedoilla, havaittiin tämä otolliseksi kohteeksi tekstianalytiikan soveltamiselle. Havaittiin myös, että tämänkaltaista tutkimusta oltiin jo tehty sangen laajasti.

Laajemmassa kontekstissa tarkasteltiin sosiaalisen median asettamia haasteita tekstianalytiikan soveltamiselle (Aggarwal & Zhai, 2012). Todettiin, että keskeisiä haasteita ovat tyypillisesti sosiaalisen median lyhytmuotoinen tekstidata sekä tämän tekstidatan kohinaisuus (Barbier & Liu, 2011). Haasteita ovat myös skaalautuvuus, dynaamisuus ja reaaliaikaisuus (Aggarwal & Zhai, 2012). Lisäksi ihmisten käyttämä kieli, huumori ja sarkasmi aiheuttavat tulkinnallisia ongelmia (esimerkiksi sarkasmin osalta González-Ibáñez ym., 2011). Lisääntyvissä määrin oleva multimediasisältö voi osaltaan aiheuttaa muutospaineita tekstianalytiikan kehittämisessä (Hu & Liu, 2012).

Tarkastelun tarkentuessa koskemaan nimenomaan Twitter-dataa huomattiin, että yleisemmin sosiaalisen median dataa koskevat

tekstianalytiikan ongelmat pätivät myös pitkälle Twitter-datan osalta. Twitter-datan analysoitavuuden osalta havaittiin tiedon julkisuuden sekä aika- ja paikkatietojen olevan positiivisia puolia. Lisäksi todettiin twiittien ilmaisevan vahvasti tunteita ja mielipiteitä, mikä tekee Twitter-datasta sävyanalyysin soveltamisen kannalta otollista. Havaittiin twiittien lyhyyden olevan toisaalta positiivinen (Zhang, Fuehres & Gloor, 2011) ja toisaalta negatiivinen tekijä (Hu & Liu, 2012) analysoitavuuden kannalta. Samalla todettiin myös käyttäjien harrastaman hashtagien käytön olevan hyvä asia datan analysoitavuuden kannalta (Hu & Liu, 2012). Analysoitavuuden kannalta tehtiin huomio sarkasmin hyvin ongelmallisesta tunnistamisesta (González-Ibáñez ym., 2011). González-Ibáñez ym. (2011) tutkimuksen tuloksien pohjalta on myös syytä olettaa, että sarkasmi jää ongelmalliseksi tulkinta-alueeksi tekstianalytiikan kannalta. Geolokaatiotietojen osalta havaittiin mobiilikäyttäjiltä saatavat eksaktit paikkatiedot analysoitavuuden kannalta erinomaiseksi seikaksi. Sen sijaan perinteisten käyttäjien itse ilmoittamat paikkatiedot havaittiin jossain määrin ongelmallisiksi (Hecht ym., 2011) vaikkakin lähestymistapoja estimoida sijainteja on kehitetty (Hecht ym., 2011 ja Cheng ym., 2010). Analysoitavuuden kannalta havaittiin myös tyypillisiä demografisia piirteitä, jotka tulee tekstianalytiikan soveltamisessa kenties ottaa huomioon. Näitä olivat yliedustus runsaasti asutuilla alueilla ja miesten yliedustus varhaisessa Twitterin omaksumisvaiheessa (Mislove ym., 2011). Lisäksi joissain tutkimuksissa koettiin käyttäjien ikäryhmiin jakautuminen ongelmalliseksi (Paul & Dredze, 2011b).

Neljännessä luvussa pureuduttiin Twitter-datan tekstianalytiikan käyttötapauksiin. Esimerkkitapauksia pyrittiin käsittelemään siten, että tutkielman puitteissa kohtuullisen laaja-alainen esitys erilaisista sovellusalueista saatiin katettua.

Trendien ja uutisaiheiden havaitsemisen osalta käsiteltiin useampia eri malleja. Näistä lupaavimmalta trendihavaitsemisen perusmalliksi vaikutti Mathioudakisin ja Koudasin (2010) esittelemä TwitterMonitor. Tämä vastaa toimintaperiaatteeltaan ja yksinkertaisuudeltaan hyvin esimerkiksi skaalautuvuuden vaatimukseen (Aggarwal & Zhai, 2012). Lisäksi hyvin yksinkertaisen mallin päälle on helppo rakentaa syvempää analytiikkaa ja voi täten olla laajalti sovellettavissa. Sanaryöpsähdysten havaitsemiseen perustuva tekniikka on tehokas. Tässä yhteydessä käytiin läpi myös muita malleja trendien havaitsemiseen, jotka olivat rakenteeltaan monimutkaisempia (Popescu & Pennacchiotti, 2010 ja Phuvipadawat & Murata, 2010 sekä Cataldi ym., 2010). Lisäksi tutkittiin Naamanin ym. (2010) kehittämää trendien taksonomiaa, jonka avulla voidaan analytiikkaa ohjata tunnistamaan onko alkava trendi Twitterin sisäinen vai ulkoinen.

Tautiseurannan osalta tutkittiin useita erilaisia lähestymistapoja. Twiiteissä mainittujen flunssaoireiden perusteella laskettujen flunssapisteiden esittäminen paikkatietojen kanssa karttapohjalla oli sangen yksinkertainen tekniikka (Lampos & Cristianini, 2010). Tämä oli myös yllättävän tehokas lähestymistapa. Toteutustapa vastaa sangen suoraviivaisesti Google Flu

Trends:in käyttämää lähestymistapaa eli flunssa-mainintoja suhteutetaan ja esitetään maantieteellisesti. Lampos ja Cristianini (2010) tosin toteuttavat tämän koneoppimisen ja luokittelun avulla. Verrattain selkeän toiminnallisuutensa puolesta tämä olisi käyttökelpoinen ja reaaliaikainen sovellutus tautiseurantaan. Paulin ja Dredzen (2011a & 2011b) kehittämät ATAM- ja ATAM+-mallit todettiin hyvin potentiaalisiksi työvälineiksi tautiseurantaan. Mallit hyödyntävät koneoppimista keskittyen oireita kuvaavien ilmauksien etsintään. ATAM-mallin kehittyneempi versio mahdollistaa paljon muutakin kuin pelkän tautiepidemioiden seurannan. Paulin ja Dredzen (2011b) mukaan malli tekee mahdolliseksi esimerkiksi elintapariskien kartoituksen ja tutkimisen. Tämän kaltaiselle mallille potentiaalisia käyttökohteita kansanterveydellisissä tutkimuksissa voisi olettaa olevan runsaasti. Tämän lisäksi mallien soveltava käyttö voi tuoda myös monenlaista muuta tutkimustulosta, joka on tällä hetkellä julkisen terveydenhoidon palveluiden ulottumattomissa. Tällaista voisi olla esimerkiksi tieto siitä, millä ihmiset hoitavat omatoimisesti sairausoireitaan.Työvälineenä ATAM-mallit ovat edistyksellisiä ja niillä voisi siis olla käytettävyyttä muidenkin tutkimusalojen käytössä.

Tautiseurannan osalta havaittiin myös, että jo hyvin yksinkertainen pieneen avainsanojen joukkoon perustuva lähestymistapa toi merkittävän hyviä tuloksia vallitsevasta epidemiatilanteesta (Culotta, 2010a & 2010b). Lampoksen ja Cristianinin (2010) ja Culottan (2010a & 2010b) tutkimusten perusteella vaikuttaa siis siltä, että Twitter-dataan perustuva lähes reaaliaikainen tautiepidemioiden seuranta olisi mahdollista toteuttaa hyvin yksinkertaisestikin seuraamalla tautia ja taudin oireita koskevien keskeisten sanojen esiintymistä. Lisäksi käsiteltiin Aramaki ym. (2011) tutkimusta, jossa havaittiin uutisoinnilla olevan negatiivinen vaikutus tutkimustuloksiin. Tämä todennäköisesti johtuu kyseisessä tutkimuksessa olleesta menetelmästä, joka pyrki keskittymään ei-relevantin kontentin suodattamiseen ja pitäytymiseen twiittien informatiivisessa osassa. Tämä todennäköisesti johti otoksen kannalta yliedustukseen uutisoinnin kaltaisten twiittien osalta. Tästä johtuen huomattavasti parempana lähestymistapana voidaan kirjallisuuskatsauksen osalta pitää Paulin & Dredzen (2011a & 2011b) ATAM-malleja sekä yksinkertaisuutensa puolesta Lampoksen ja Cristianinin (2010) sekä Culottan (2010a & 2010b) lähestymistapoja. Kokonaisuudessaan tautiepidemioiden seurannan osalta johtopäätös on, että näitä voidaan käyttää vähintään osana ei-formaalia seurantaa.

Sävyanalyysin sovellettavuuden osalta perehdyttiin tutkimuksiin, joissa sävyanalyysiä oli käytetty kansallisten mielialojen mallintamisessa (Bollen, Pepe & Mao, 2011), onnellisuuden tutkimuksessa (Mitchell ym., 2013), pörssikurssien ennakoinnissa (mm. Bollen, Mao & Zeng, 2011) sekä alueellisen temporaalis-spatiaalisen mielialakartan luomisessa (Bertrand ym., 2013). Bollenin, Pepen ja Maon (2011) tutkimuksen perusteella havaittiin, ettei koneoppiminen ole välttämätöntä mielialojen luokittelun kannalta vaan pelkkä mielialojen luokitteluun perustuva menetelmä on suurilla datamäärillä riittävä. Erityisen käyttökelpoiseksi työvälineeksi havaittiin psykologian alalta lainatun

ja modifioidun POMS:in (engl. Profile of Mood States) käyttö. Tämän mallin soveltavalla käyttämisellä on mahdollista luoda hyvin kuvaavia mallinnoksia erilaisista koetuista tunteista. Mikäli POMS:in käyttöön sovelletaan laajennettua tunteita kuvaavien sanojen kokoelmaa voidaan työvälineen tehokkuutta twiittien tunteellisen sisällön kategorisoinnissa mahdollisesti edelleen tehostaa. Tämän tulisi mahdollistaa varsinkin tutkimustarkoitukseen räätälöitynä erinomaisen tavan hahmottaa kollektiivisesti koettuja mielialoja ja tunteita. On syytä olettaa, että jonkinlainen POMS:ista edelleen kehitetty versio tulee olemaan tulevaisuuden standardiratkaisu tehtäessä sosiaalisen median tekstidataan perustuvia mielialoihin liittyviä tutkimuksia. Merkille pantavaa on tässä yhteydessä myös ohjatun oppimisen menetelmin sävyjen tunnistaminen. Täten on toimittu esim. Bertran ym. (2013) tutkimuksessa, jossa koulutusdatana käytettiin hymiöitä sisältäviä twiittejä. Hymiöihin perustuvan koulutusdatan hyväksi puoleksi havaittiin se, että näin mielialoja luokitteleva luokittelija on kieliriippumaton. Sävyanalyysipohjaisten tutkimusten osalta päädyttiin lopputulokseen, että riittävillä otosmäärillä populaatiosta voidaan tehdä hyvinkin seikkaperäistä kuvausta mielialoista ja koetusta onnellisuudesta (Bollen, Pepe & Mao, 2011 ja Mitchell ym., 2013).

Tämänkaltaisella tutkimuksella ja seurannalla voi olla hyvin laaja-alaisia ja mielenkiintoisia sovellutusalueita. Tästä esimerkkinä osakekursseihin liittyvät mielialatutkimukset (esimerkiksi Bollen, Mao & Zeng, 2011). On todennäköistä, että kollektiivinen mieliala on tutkinnallisesta ja julkishallinnollisesta näkökulmasta katsottuna mielenkiintoinen kohde laajemminkin esimerkiksi sosiaalisen hyvinvoinnin, yhteiskunnan ja talouden tutkimuksen kannalta. Sävyanalyysin kannalta tärkeimmiksi työvälineiksi havaittiin siis POMS:in soveltava käyttö sekä ohjatun oppimisen hyödyntäminen twiittien semanttisessa arvottamisessa.

Lopulta tarkasteltiin erilaisia tekstianalytiikan soveltamisen alueita, joissa erilaisia lähestymistapoja ja tekniikoita käytettiin monin tavoin. Joukkoistetun aistinnan ja yhteistyön osalta käsiteltiin kahta tutkimusta, joissa molemmissa tarkasteltiin Twitter-dataan perustuvaa maanjäristyksien havaitsemista. Tulokset olivat lupaavia etenkin Sakaki ym. (2010) tutkimuksessa, jossa hyvin nopean maanjäristysten havaitsemisjärjestelmän pohjalta lopulta toteutettiin hälytysjärjestelmä. Tämä hälytysjärjestelmä oli nopeampi kuin tuolloin käytössä oleva Japanin virallinen järjestelmä. Toteutuksessa käytettiin ohjatun oppimisen avulla luotua luokittelijaa, jonka avulla tunnistettiin relevantit maanjäristysaiheita koskevat twiitit. Näiden twiittausten lisääntyessä radikaalisti voitiin olettaa maanjäristyksen olevan käynnissä. Toisin sanoen menetelmässä hyödynnettiin sanaryöpsähdysten havaitsemisen kaltaista tekniikkaa. Earlen, Bowdenin ja Guyn (2012) lähestymistavassa taasen on pyritty löytämään eri kielisiä ilmauksia maanjäristyksistä ja myös havainnoimaan näiden tiettyjen sanojen radikaalia kasvua twiittien joukossa. Earle ym. (2012) tutkimuksen osalta voitaneen kritiikkiä antaa estottomalle käyttäjän ilmoittamien geolokaatiotietojen käytölle.

Tässä yhteydessä nämä paikkatiedot olisi voinut sangen hyvin karsia otoksesta pois keskittyen pelkästään käyttämään eksakteja paikkatietoja.

Tutkimusten perusteella on selvää, että joukkoistettua aistintaa voidaan tehdä Twitter-datan pohjalta yhdistämällä ohjattua oppimista, sanaryöpsähdysten tai trendien havaitsemistekniikkaa ja tarkkailtavan ilmiön asiasanaston tarkkailua. Todennäköisesti sovellusalueita tällaiselle sosiaalisten sensorien tekemälle joukkoistetulle havainnoinnille olisi muitakin. Tutkimustulosten perusteella korostuukin enemmän se, missä määrin ihmisiin sosiaalisina sensoreina voidaan luottaa tiedon oikeellisuuden ja sensorin toiminnan jatkuvuuden kannalta. Mielenkiintoinen kysymys on myös se, että voidaanko jonkinlaista joukkoistettua yhteistyötä toteuttaa Twitteriä alustana käyttäen siten, että käyttäjät kysyttäessä raportoivat älypuhelimiin sijoitettavien sensorien tuloksia (Demirbas ym., 2010).

Kirjallisuuskatsauksen pohjalta poliittinen tutkimus on Twitter-datan tekstianalytiikalle kenties vaikein osa-alue. Tumasjan ym. (2010) esittivät sangen rohkeita tuloksia poliittisen tutkimuksen osalta mutta nämä ovat saaneet osakseen suurta kritiikkiä. Muun muassa Jungherr ym. (2011) osoittavat tutkimusmenetelmissä ja –otoksessa selkeitä aukkoja. Lisäksi Gayo-Avello (2012) on kritisoinut voimakkaasti tähän mennessä suoritettuja poliittiseen ennakointiin pyrkiviä tekstianalytiikkaan pohjautuvia tutkimuksia. Yleisesti ottaen vaikuttaa, että poliittisesti aktiiviset twiittaajat dominoivat otoksissa siten, että ennustuksia esimerkiksi vaalien lopputuloksesta on vaikea tehdä. Tältä osin olen ehdottanut uudeksi lähestymistavaksi sävyanalyysin yhdistämistä tarkkailtavana olevan alueen demografisiin ja edeltävissä vaaleissa saavutettuihin tuloksiin. Tätä kautta otosta voitaisiin pyrkiä normalisoimaan. On syytä olettaa, että tällä lähestymistavalla voitaisiin saavuttaa parempia tuloksia mutta menetelmänä se ei ole yksinkertainen. Vaikuttaa, että automatisoitu tekstianalytiikkaan pohjautuva poliittinen tutkimus voi jatkossakin olla haastavaa ja esimerkiksi kykenemättömyys sarkasmin automaattiseen tunnistamiseen eittämättä hankaloittaa tätä poliittisissa aiheissa.

Yleisesti ottaen tutkimuksissa käytetyimmät tekniikat olivat luokitteluita, koneoppimisen erilaisia tekniikoita (pääosin ohjattua oppimista) sekä eri tavoin toteutettua sävyanalyysia. Tämä on ymmärrettävää, sillä tutkimuksissa joudutaan yleensä huomioimaan twiittien sävy tavalla tai toisella. Käytännössä tämä vaatii semanttista analyysiä tai erilaisia koneoppimisen muotoja yhdistettynä aihepiiriä kuvaavaan sanahakemistoon. Erityistä huomiota kiinnittämällä tutkittavan aihepiirin ontologian kehittämiseen voidaan myös saavuttaa merkittävän hyviä tuloksia (esimerkiksi EMOTIVE-projekti, 2013a & 2013b).

Kaikissa tutkimuksissa lukuunottamatta poliittisia tutkimuksia voidaan todeta Twitter-datan analyysin olevan vähintäänkin hyödyllistä ei-formaalina tiedonlähteenä virallisten tutkimuslähteiden rinnalla. Lisäksi suurimmassa osassa tutkimuksista havaittiin korkea korrelaatio verrattuna vertailudataan, kuten esimerkiksi virallisiin tautiepidemia-tilastoihin. Tässä

Twitter-datan louhinnan etu on luonnollisesti sen nopeus sekä potentiaalisesti kustannustehokkuus. Nopeasti tuotettavissa oleva suuntaa-antava estimaatio eri tutkimuksen kohteista on monilla aloilla hyvin tärkeässä roolissa verrattuna hitaasti tuotettaviin ja marginaalisesti parempiin kuvauksiin ilmiöistä. Äärimmäisenä esimerkkinä tästä voisi olla rikollisen toiminnan tai terrorismin jäljitys ja torjunta. On todennäköistä, että Twitter-dataa analysoivien mallien kehittyessä on löydettävissä runsaasti lisää käyttökohteita.

Julkishallinnon tekstimuotoisten datavarantojen ominaisuuksista analysoitavuuden kannalta voitaneen tehdä muutamia oletuksia. On syytä olettaa, että julkishallinnon tekstidata antaa heterogeenisemman poikkileikkauksen populaatiosta kuin sosiaalisen median tekstidata. Lisäksi oletusarvoisesti tekstidatan sisällössä tuskin esiintyy juurikaan sarkasmia taikka tarkoituksellisesti valheellista tietoa. Nämä tekijät tekisivät tekstimuotoisen datan analytiikasta tietyllä tavalla mutkattomampaa. Toisaalta on oletettavaa, että julkishallinnon hallussaan pitämä tekstimuotoinen data ei ole vastaavalla tavalla sentimenttien kannalta värittynyttä kuin sosiaalisen median data. Twitter-datan osalta ihmiset heijastavat twiiteissään runsaasti erilaista tunne- ja mielipidekontenttia. Tämän tyyppinen itsensä ilmaiseminen lienee julkishallinnon omistamassa tekstimuotoisessa datassa vähäisempää. Twitterin ja sosiaalisen median osalta käyttäjiä nimenomaan kannustetaan kertomaan mielipiteistään ja tunteistaan kun taas julkishallinnon eri instanssien kanssa asioidessaan kansalaiset yleensä täyttävät lomakkeiden tekstiosuuksia hyvin virallisella ja varovaisella kirjoitustyylillä. Tämän olettamuksen pohjin on perusteltua olettaa, että julkishallinnon hallussaan pitämä tekstimuotoinen data ei ole kovinkaan otollinen varsinkaan sävyanalyysin kohteeksi. Sen sijaan on oletettavaa, että esimerkiksi trendien havaitsemistekniikoiden soveltamisella voitaisiin julkishallinnon tekstidatasta etsiä esiin nousevia, kenties yllättäviäkin tutkimuskohteita. Tällainen voisi esimerkiksi olla vaikkapa tietyn sairauden oireiden poikkeuksellisen suuri esiintymismäärä tietyltä maantieteelliseltä alueelta saadussa tekstimuotoisessa terveyttä koskevassa datassa. Myös esimerkiksi takautuvien visualisointien tekeminen esimerkiksi liikenneonnettomuuksista ajallisena ja paikallisena esityksenä onnettomuusraporttien perusteella on toteutettavissa. Kirjallisuuskatsauksessa käsiteltyjen tekniikoiden osalta hyvin monenlainen soveltaminen julkishallinnossa on mahdollista edellyttäen, että tekstimuotoista dataa on saatavilla.

Vastauksena ensimmäiseen tutkimuskysymykseen on edellä esitelty laaja-alaisesti kirjallisuuskatsauksesssa esiintulleita Twitter-datan tekstianalytiikan käyttökonteksteja. Samoin on esitelty työkaluja, menetelmiä ja lähestymistapoja. Näiden toimivuutta on arvioitu eri konteksteissa ja parannusehdotuksia tietyissä tapauksissa on esitetty. Kokonaisuudessaan on havaittu tekstianalytiikkaan pohjautuvien tutkimuksien tuottavan lupaavan laadukasta tietoa. Tämä on tutkimuksissa ilmennyt esimerkiksi voimakkaana korrelaationa vertailudataan. Poikkeuksena tästä on poliittinen tutkimus, jonka

osalta on havaittu tekstianalytiikan soveltaminen ja saadut tulokset heikohkoiksi.

Kirjallisuuskatsauksen pohjalta voidaan siis esittää seuraavaa.

- Toistaiseksi ennakoiva tutkimus politiikasta tekstianalytiikalla on ollut vaikeaa
- Menetelmät ovat soveltuneet hyvin trendien ja uutisaiheiden havaitsemiseen, tautien seurantaan, kollektiivisen mielialan arviointiin ja joukkoistettuun havaitsemiseen sekä seurannan suorittamiseen
- Menetelmät soveltuvat eritoten mobiilikäyttäjien tutkimiseen, koska eksaktit paikkatiedot ovat saatavissa
- Sovellettavien perusmenetelmien tulisi olla jokseenkin yksinkertaisia koska skaalautuvuus, nopeus, tehokkuus ja muokattavuus korostuvat sovitettaessa malleja julkishallinnon käyttöön

Alla on esitetty kirjallisuuskatsauksen pohjalta tehdyt hypoteesit empiiristä tutkimusta varten. Nämä hypoteesit koskevat toista tutkimuskysymystä, joten empiirisen vaiheen haastattelukysymyksiä muotoiltiin myös hypoteesien testaamisen kannalta. Hypoteesien toteutumista arvioidaan empiirisen osuuden johtopäätöksiä käsiteltäessä.

Ensimmäinen hypoteesi

Paul & Dredze (2011a & 2011b) ATAM- ja ATAM+-malli on mahdollisesti suoraan käytettävissä tai sovellettavissa kohdeorganisaatioiden omaan tekstimuotoiseen dataan.

Toinen hypoteesi

Trendien havaitsemis-tekniikoiden käyttö kohdeorganisaatioiden omaan tekstimuotoiseen dataan on mahdollista ja kannattavaa.

Kolmas hypoteesi

Twitter-datan tekstianalytiikan tekniikoiden käyttäminen sosiaalisen median tuottaman tekstidatan seurantaan tuottaa julkishallinnon kannalta lisäarvoa.

Neljäs hypoteesi

Twitter-dataan perustuvia tautimallinnustekniikoita voidaan käyttää täydentävänä informaatiolähteenä käytössä olevien menetelmien rinnalla.

Viides hypoteesi

Kohdeorganisaatioiden tekstidata ei ole hyvää sävyanalyysin kannalta.

5 Empiirinen tutkimus

Tämän luvun ensimmäisessä alaluvussa on käsitelty empiirisen osuuden haastattelutilanteiden kulkua ja tuloksia analysoimalla haastattelututkimuksessa saatu materiaali kysymyskohtaisesti. Toisessa alaluvussa on esitelty tämän perusteella tehtävät johtopäätökset.

5.1 Empiirisen tutkimuksen tulokset

Koska haastattelu oli luonteeltaan puolistrukturoitu teemahaastattelu oli haastattelutilanteet keskustelunomaisia ja ideoivia. Haastattelija esitti kummassakin tilaisuudessa etukäteen suunnitellut 12 kysymystä. Ensimmäinen kysymys koski nykyistä tilannetta. Tämän jälkeen kysymykset oli jaettu teemoittain ja koskivat käytettyjä menetelmiä ja niiden sovellettavuutta kyseisiin organisaatioihin. Viimeinen kysymys koski tulevaisuuden näkymiä. THL:n osalta kysymyksien esittämisessä tehtiin kaksi pientä muutosta. Koska Kela-haastattelussa havaittiin kahdeksas kysymys huonosti muotoilluksi, esitettiin tämä nyt ilman tiukkaa IT-arkkitehtuurikontekstia. Toisin sanoen kysymys laajennettiin koskemaan soveltuvuutta THL:n koko organisaatiolle. Kysymys numero 12 muokattiin koskemaan sekä tarkemmin THL:ää että julkishallintoa yleisellä tasolla. Tällä tavoiteltiin kattavampaa informaatiota haastateltavilta. Lisäksi haastattelija ohjasi kummankin haastattelun kulkua tarjoamalla tarkentavia kysymyksiä ja pyrkimällä pitämään tarpeettoman keskustelun rönsyilyn aihepiirin puitteissa minimissä.

Seuraavassa käsitellään tulokset kysymyksittäin ja aihealueittain siten, että molempien organisaatioiden edustajien tarjoamat vastaukset ovat kunkin kysymyskohdan alla käsiteltyinä rinnasteisesti. Kysymykset etenevät aihealueittain. Nämä aihealueet oli jaettu nykyisen tilanteen kartoitukseen (kysymys 1), Hevner ym. (2004) design science- arviointiperusteisen lähestymistavan osalta observatiivisiin (kysymykset 2-4), analyyttisiin (kysymykset 5, 6 ja 7 staattisten ominaisuuksien arvioinnin kannalta ja

kysymykset 8 ja 9 arkkitehtuurin huomioiden) ja deskriptiivisiin (kysymykset 10 ja 11) kysymyksiin sekä tulevaisuuden näkymien kartoitukseen (kysymys 12).

Vastausten analysoinnissa huomioidaan IT-artifaktien osalta Hevnerin ym. (2004) design science arviontiperusteista jo mainitut observatiiviset, analyyttiset ja deskriptiiviset lähtökohdat sekä vapaamuotoisesti muu aihepiirin kannalta relevantti informaatio, jota haastatteluista oli saatavissa. Lisäksi haastatteluita analysoitaessa kiinnitetään huomiota tulevaisuuden visioihin sekä eritoten erilaisiin käyttöskenaarioihin, joissa malleja jatkossa voitaisiin hyödyntää. Haastatteluiden vastaukset kummankin organisaation kohdalta on tässä luvussa esitetty rinnasteisesti ja jossain määrin alkuperäistä keskustelunomaista rakennetta säilyttäen. Näin on tehty, jotta mielikuva kummankin tahon näkemyksistä on helposti nähtävissä, vertailtavissa ja esitettävissä.

KELA:n taholta esityksessä läsnä olivat ja haastatteluun osallistuivat IT-johtaja Markku Suominen ja tietohallintojohtaja Veikko Hytönen. THL:n osalta esitykseen ja haastatteluun osallistuivat kehittämispäälliköt Hilkka Miettinen (Sosiaali- ja terveydenhuollon tietohallinnon operatiivisen ohjauksen yksikkö), Eija Hukka (Elintavat ja osallisuus –osasto) ja Sari Atkins (Julkaisu ja verkkopalvelut-yksikkö).

1.Kysymys

Ensimmäisessä kysymyksessä pyydettiin haastateltavia kuvaamaan tekstianalytiikan käyttöä, käyttötarkoituksia ja tekstimuotoisen datan määrää organisaatiossa tällä hetkellä. Tavoitteena oli luoda näkemys nykytilanteesta ja siitä, missä yhteyksissä ja miten tekstianalytiikkaa käytetään sekä kartoittaa kuinka paljon tekstimuotoista dataa organisaatiolla on sekä millaista se on.

KELA

Tutkimuksessa kuvatun kaltaista tekstianalytiikkaa vähäisesti käytössä tällä hetkellä. Tekstimuotoista dataa on organisaatiossa jonkin verran, mutta pääsääntöisesti data on käytössä rakenteisessa muodossa. Kelalla on pitkä perinne käyttää rakenteista tietoa ja tieto käsitellään kenttämuotoisena. Eräänä tekstimuotoisen datan lähteenä huomioidaan sähköpostipalaute-kanava sekä asiakaspalvelujärjestelmä. Ongelmana tekstianalytiikan soveltamisessa näiden kohdalla huomioidaan datan sangen pieni määrä. Lisäksi tekstidata on luonteeltaan sellaista, että arvioidaan jonkun ihmisen täytyvän aina nämä läpi käydä. Toisin sanoen automatisoidulle tekstianalytiikalle nämä eivät ole otollisia kohteita. Lisäksi esimerkiksi sävyanalyysin kannalta palautekanavien tekstidata ei ole hyödynnettävissä, sillä toimihenkilöt läpikäyvät nämä viestit ja kirjoittavat puhtaaksi, täten alkuperäinen palautetekstidata muuttuu. Tekstimuotoisen datan olemassaolosta huomioidaan lisäksi, että lääkärilausuntoja on tekstimuotoisena. Kuitenkin sama tieto on jo käytettävissä

rakenteellisena datana henkilötunnuksien ja koodituksien avulla ja esimerkiksi reseptitietojen perusteella. Täten tarvetta analyyttisemmalle tekstianalytiikan käytölle oman tekstimuotoisen datan suhteen ei ole.

THL

THL:n osalta arvioidaan ensin tekstimuotoista dataa olevan kyllä paljon mutta haastateltavilla ei ole tietoa, onko tekstianalytiikan menetelmiä käytetty missä määrin. Arvio on, että systemaattisesti tekstianalytiikkaa ei käytetä tällä hetkellä. Kuitenkin on mahdollista, että yksittäisenä työkaluna erilaisissa tutkimuksissa on ollut käytössä. Vaikka tekstidataa arvioidaan olevan sangen runsaasti, huomioidaan myös tutkimusaineiston osalta vahva strukturoidun datan määrä. Tämä tekee tekstianalytiikan käytöstä organisaation oman datan kannalta merkityksettömämpää. Tekstimuotoisen datan lähteiksi havainnoidaan verkkosisällöt, vapaasanainen palautekanava (verkkosivut, mailit, Facebook). Tämä tieto ei kuitenkaan ole tallessa missään tietokannoissa vaan esimerkiksi palautekanavien datan osalta joku aina lukee ja tekee omat johtopäätökset, vastaa ja raportoi. Tietokannoissa olevan datan osalta tieto on strukturoitua. Mahdolliseksi tulevaksi tekstimuotoisen datan lähteeksi huomioidaan myös terveydenhuollon palveluiden arvioimista varten oleva palautekanava, joka tosin vielä ei paljoa dataa tuota. Joka tapauksessa on epätodennäköistä, että THL:n oma datamäärä olisi siinä määrin merkittävä, että tekstianalytiikkaa tarvittaisiin.

2.kysymys

Toisessa kysymyksessä viitattiin haastattelijan esityksessä esittelemiin tekstianalytiikan tapoihin hyödyntää Twitteristä saatavaa tekstimuotoista dataa. Kysymyksessä pyydettiin haastateltavia arvioimaan esiteltyjen menetelmien mielekkyyttä ja toimivuutta esitellyissä asiayhteyksissä. Tältä osin toisen kysymyksen vastaukset antavat materiaalia Hevner ym. (2004) arvioimisperusteiden (Observational, field study) kannalta.

KELA

Haastateltavat kokevat esitellyt esimerkit toimiviksi sekä mielekkäiksi. He huomioivat, että julkishallinnon osalta käyttöä olisi varmuudella esimerkiksi eräänlaisen ”nettipoliisi”-toiminnan osalta, koska automaattinen tarkkailu mahdollistaisi tätä tehokkaasti. Lisäksi esitellyt menetelmät koetaan erinomaisesti soveltuviksi monille muille viranomaistahoille. Esimerkkeinä mainitaan erilaiset onnettomuus- ja kriisitilanteet, uutisten saaminen ripeää toimintaa vaativissa tilanteissa sekä mahdolliset kartoitustarpeet siitä, ”mitä mieltä ollaan missäkin päin suomea". Teoriapohjalta esitellyt menetelmät koetaan mielekkäiksi. Lisäksi huomioidaan laaja hyödynnettävyys kaupallisten yritysten kannalta, joille nähdään potentiaalia menetelmien käyttämisessä johtamisen, tuotekehityksen sekä

kohdennetun markkinoinnin saroilla. Lisäksi tekstianalytiikan käyttö ”puskaradiomaisen” kuluttaja-palautedatan seurannassa pohdituttaa ja ilmoille jää idea siitä, että kaupallisten yritysten kannattaisi seurata jonkin tapaisia signaaleja siitä, mitä heidän tuotteistaan sosiaalisessa mediassa puhutaan.

THL

Haastateltavat kokevat esitellyt esimerkit toimiviksi sekä mielekkäiksi. Esiteltyjen menetelmien osalta ei niinkään nouse epäilyksiä mutta sen sijaan Suomen kohdalla näiden soveltaminen herättää epäilystä. Tämä johtuu lähinnä vähäisestä Twitterin käytöstä Suomessa. Tautiseurannan osalta kokemusta organisaatiosta löytyy jo vastaavista menetelmistä (Google Flu Trends). Lisäksi mainitaan jossain vaiheessa harkitun vaihtoehtoa, jossa vastaavaa suomalaista sovellusta olisi käytetty tautiseurannan tukena. Tästä ideasta kuitenkin luovuttiin, sillä tautiseurannan osalta koetaan nykyiset menetelmät hyvin toimivina. Tautiseurannan osalta kuitenkin mahdollinen kokeileminen kiinnostaa siten, että tavoitteena olisi selvittää miten hyvin nämä toimisivat verrattuna nykyiseen käytäntöön. Todetaan myös, että menetelmän osalta ei ole epäilyksiä mutta Suomen osalta tautidatan pitäisi olla peräisin muualta kuin Twitteristä johtuen suomalaisten tämän hetkisestä vähäisestä Twitter-käyttäjämäärästä. Tautiseurannan osalta myös huomioidaan se, että hyvin merkittävä korrelaatio (viralliseen tautidataan) on periaatteessa saavutettavissa lähes reaaliaikaisesti, joten tämä menetelmä eritoten kiinnostaa.

Onnellisuus- ja mielialatutkimuksien osalta ei myöskään esitetä epäilyksiä menetelmien toimivuuden tai mielekkyyden osalta. Lähinnä vastaukset siirtyvät pohtimaan erilaisia sovellutuskohteita. Eräänä tällaisena mainitaan sosioekonomisen statuksen ja siihen liittyvän muutoksen seuraamisen sosiaalisen median kautta. Tämä voisi esimerkiksi vaikeina taloudellisina aikoina heijastella yhteiskunnassa koettua muutosta. Teknisesti tämä olisi mahdollisesti rinnastettavissa esityksessä esiteltyyn onnellisuustutkimukseen ja vastaavia menetelmiä voitaisiin soveltaa. Lisäksi huomioidaan mahdolliset sovellutuskohteet onnellisuus- ja mielialatutkimuksille sosiaalihuollon osalta.

Esiin nousee myös idea aktiivisesta kritiikin ja palautteen etsinnästä tekstianalytiikan avulla eri lähteistä. Esimerkkeinä mainitaan keskustelupalstojen keskustelut. Näistä on saatavilla kärkästä kritiikkiä mitä muista lähteistä ei ole saatavilla. Menetelmien osalta nähdään, että soveltamalla esiteltyjä työvälineitä erilaisiin sosiaalisen median lähteisiin ja keskusteluforumeille palautetta voitaisiin kerätä esimerkiksi rokotuksista.

3.kysymys

Kolmannessa kysymyksessä haastateltavia pyydettiin arvioimaan millainen mielikuva Twitter-datan tekstianalytiikan keinoista jää sekä arvioimaan kyetäänkö tämän kaltaiseen analytiikkaan perustuvaa dataa

käyttämään päätöksenteon tukena. Vastauksista välittyy missä määrin haastateltavat kokevat esimerkkitapauksissa esiteltyjen tekniikoiden olevan päteviä tuottamaan päätöksentekoon sopivaa dataa.

KELA

Haastateltavat toteavat, että ilman muuta voidaan käyttää päätöksenteossa olettaen, että aihetta tarkastellaan laajemmin. Juuri tällä hetkellä ei koeta pelkästään Kelan näkökulmasta olevan juurikaan hyödynnettävyyttä. Haastateltavat arvioivat merkityksen jatkossa olevan kasvava. Menetelmien arvioidaan olevan tulevaisuudessa niitä harvoja tapoja heijastaa tätä hetkeä nopeasti, jotka mahdollistavat nopean reagoinnin esimerkiksi kriisitilanteissa. Tämä koetaan päätöksenteon kannalta paitsi hyödylliseksi myös kriittiseksi. Sosiaalisen median tekstianalytiikka nähdään myös työkaluna, jolla kyetään hankkimaan etukäteisinformaatiota päätöksentekoa varten tutkimalla populaation reaktioita esimerkiksi kaavailtuun muutokseen. Arvioidaan, että ”päätöksenteko tulisikin ehkä harkita siten, että heittää sen pallon ensin ulos ja kuuntelee ennen kuin menee tekemään päätöksiä”.

Menetelmien käytettävyydestä päätöksenteossa kommentoidaan myös, että tämä riippuu paljon siitä mihin käyttää. Huomioidaan, että kaikissa päätöksissä ei tarvitse olla nopea ja voidaan odottaa eksaktimpaa tutkimustietoa. Eritoten menetelmät havaitaan edullisiksi päätöksenteon kannalta tilanteissa, jotka vaativat nopeaa reagointia ja joissa tekstianalytiikan tuottaman datan ei tarvitse olla eksaktisti oikeaa vaan suuntaviivoja antavaa. Tällaisia tapauksia voisivat olla esimerkiksi luonnonmullistukset, kriisit ja pandemiat – ajantasaisen tiedon välityksen nopeus korostuu päätöksenteon kannalta eikä esimerkiksi tautitartuntojen eksakti määrä ole juuri sillä hetkellä relevanteinta. Koetaan, että tiettyä tiedon epävarmuutta voidaan sietää, mikäli riskit ovat hyvin isoja.

Johtopäätöksenä haastateltavat saapuvat lopputulokseen, että menetelmiä voidaan käyttää päätöksenteossa normaaliin tapaan mutta on oltava jokin virallinen data, jota vasten menetelmien tuloksia voidaan myös tarkistaa. Päätöksenteon osalta huomioidaan vielä terrori-iskujen ja kouluammuskeluiden torjunta. Menetelmät koetaan hyvin soveltuviksi tällaisten iskujen ennalta ehkäisemiseen ja tällaisiin tapauksiin viittaavien indikaattorien etsintään.

THL

Eräs haastateltavista epäilee menetelmien sovellettavuutta päätöksenteossa ja arvioi, että jotta menetelmiä voitaisiin käyttää niitä tulisi testata ja nähdä koetuloksia. Toiset huomioivat, että soveltuvuus päätöksentekoon riippuu käyttökohteesta. Itse menetelmien tekniseen puoleen ei oteta kantaa. Erityisesti nähdään viestinnän ja organisaatioviestinnän päätöksenteon kannalta mahdolliseksi hyödyntää esiteltyjä menetelmiä.

Käyttötarkoitukseksi koetaan myös mahdollisuus antaa indikaatioita siitä, mihin suuntaan palveluja tulisi kehittää. Itse menetelmiä ei juurikaan arvioida kriittisesti mutta huomioidaan, että Suomen kohdalla Twitter-dataa ei juurikaan kerry joten sovellettavuus tällä hetkellä ei liene hyvä. Toisaalta ajatellaan, että ollaan tulossa siihen aikakauteen, jolloin tällaiseen siirrytään. Päätöksenteon osalta havainnoidaan mahdollisuus hyödyntää globaalia dataa, koska koetaan että kansainväliset trendit jollakin aikataululla siirtyvät Suomeen. Esimerkiksi globaalien trendien vertaaminen Suomen tilanteeseen voi auttaa arvioimaan rokotusmyönteisyyttä eri tuotteiden kohdalla tai uusien trendeinä etenevien ravintotottumusten terveysvaikutuksia. Lisäksi nähdään, että erilaisissa tutkimustarkoituksissa tuloksille voidaan hakea vahvistusta ja verrainnollistamista Twitter-datan kanssa – esimerkkinä tästä annetaan nuorisokulttuurin raittius-suuntaus ja sen tutkiminen näkyykö tämä jotenkin sosiaalisessa mediassa.

4.kysymys

Neljäs kysymys koski esiteltyjen menetelmien, työkalujen ja lähestymistapojen sovellettavuutta joko suoraan tai muokattuna kohdeorganisaatioiden oman tekstimuotoisen datan tekstianalytiikkaan.

KELA

Haastateltavat huomioivat edelleen vähäisen tekstimuotoisen datan määrän organisaatiossa. Lisäksi tiedonkäsittelymenetelmät on alusta asti rakennettu rakenteisiksi, joten käyttötarkoituksia tekstianalytiikalle ei juuri ole. Ainoana käyttökohteena suoralle sovellettavuudelle mainitaan palautekanava sävyanalyysin menetelmien osalta, mutta varsinaista käyttötarkoitusta tällaiselle analysiille ei ole. Haastateltavat kokevat kuitenkin, että tulevaisuudessa varsinkin vuorovaikutteisuuden lisääntyessä viestinnässä asiakasnäkökulman huomioon ottamisessa on mahdollista olla sovellutuskohteita.

THL

Haastateltavat huomioivat edelleen vähäisen tekstimuotoisen datan määrän organisaatiossa toteamalla, että periaatteessa tekstiaineistoa olisi, mutta tämä ei ole käytettävissä. Tällä viitataan sairaaloiden rekistereissä olevaan tekstimuotoiseen dataan. Tämän hyödyntämisen esteeksi nähdään erilaiset yksityisyyteen liittyvät ongelmat sekä lakisääteiset ja toisaalta eettiset esteet. Nämä ovat haastateltavien mielestä myös sangen merkittäviä esteitä vaikka dataa lähestyttäisiin siten, että otos anonymisoitaisiin ja sattumanvaraistettaisiin. Huomioidaan kuitenkin, että kansallisen potilasarkiston laajamittaisen käyttöönoton jälkeen voi tulevaisuudessa tulla

mahdolliseksi tietyin edellytyksin. Nämä edellytykset olisivat, että dataa on riittävästi ja toiminnan tulisi olla aina tutkimuskohtaisesti luvanvaraista.

5.kysymys

Viidennessä kysymyksessä pyydettiin haastateltavia kertomaan arvionsa esiteltyjen mallien ja toimintapojen rakenteesta, logiikasta, toimivuudesta ja sovellettavuudesta. Mielenkiinnon kohteena oli myös, että nähdäänkö joidenkin menetelmien olevan jo siinä määrin kehittyneitä, että niitä voitaisiin pitää standardoituina ratkaisuina.

KELA

Molemmat haastateltavat pitävät esiteltyjä tekniikoita loogisina, toimivina ja sovellettavissa olevina. Koetaan myös, että esimerkiksi trendi- ja uutisaiheiden etsinnän osalta tiettyjen parhaaksi koettujen menetelmien stardardoiminen on järkevää.

THL

Haastateltavat eivät suoranaisesti arvioi esiteltyjen tekniikoiden loogisuutta tai monimutkaisuutta mutta toteavat, että tärkeämpää on toimivuus tulosten osalta. Periaatetasolla lähestymistavat koetaan loogisiksi ja toimiviksi ja myös sovellettavissa oleviksi. Haastateltavat kokevat, että parhaat käytännöt voidaan standardisoida ja otaksuvat myös kaupallisella puolella näin tehdynkin.

6.kysymys

Kuudes kysymys koski analyyttistä arviota Twitter-datan tekstianalytiikan työkalujen kompleksisuudesta.

KELA

Haastateltavat eivät näe käytetyissä tekniikoissa ongelmia siinä mielessä, että malleja ei koeta kompleksisiksi. Asiayhteyksissä esitellyt tekniikat vaikuttavat haastateltavien mielestä selkeiltä ja arvio on, että työvälineet eivät ole liian konstikkaita tarkoitusperiään varten. Haastateltavat kokevat, että mikäli työvälineet tuottavat todennetusti oikean suuntaista tietoa tutkittavasta kohteesta, ei sen jälkeen enää teknisellä osaamisella juurikaan ole merkitystä työkalun käyttäjän taholta. Haastateltavien mielestä tässä vaiheessa korostuu tutkimuskohdetta koskeva spesifimpi osaaminen.

THL

Haastateltavat eivät koe esiteltyjä tekniikoita kovin kompleksisina. Sen sijaan huomio kiinnittyy enemmän siihen, että tekniikat toimivat eli tuottavat tarkoitusperiin soveltuvaa dataa.

7.kysymys

Seitsemännessä kysymyksessä kysyttiin esiteltyjen menetelmien osalta hyödyntämisen kannattavuutta kohdeorganisaatioissa. Mielenkiinnon kohteina oli, nähdäänkö menetelmät sellaisina, että niitä voitaisiin käyttöönottaa sellaisenaan tai sovellettuna nyt tai tulevaisuudessa.

KELA

Haastateltavat kokevat, että johtuen organisaation omasta vähäisestä tekstimuotoisen datan määrästä ja siitä, että kaikki hoidetaan jo strukturoidusti, ei tällä hetkellä ole tarvetta käyttää esiteltyjä menetelmiä sellaisenaan eikä sovelletusti. Kuitenkin koetaan, että mahdollisesti tulevaisuudessa voitaisiin soveltaa mutta näissä tapauksissa mahdollisia käyttökohteita ei vielä ole tiedossa.

THL

Haastateltavien mielestä menetelmiä voitaisiin mahdollisesti jossain määrin hyödyntää tutkimuksien osalta sillä varauksella, että työkalut olisivat helppoja ja luotettavia eivätkä vaatisi kovin paljoa resursseja. Tutkimuksien tekemisessä pitkän aikavälin tutkimuksien osalta menetelmät koetaan tarpeettomiksi. Kuitenkin tietyissä tapauksissa nähdään sovellettavuutta tutkimuksissa, joissa nopeus ja tarkka geolokaatiotieto olisivat oleellisia. Tällaisia ei kuitenkaan juurikaan ole ja täten mahdollinen hyöty koetaan jossain määrin rajalliseksi. Viestinnän osalta menetelmien hyödynnettävyys nähdään huomattavasti relevantimmaksi esimerkiksi palautejärjestelmien ja viranomaistahojen viestinnän välisenä apukeinona.

8.kysymys

Kahdeksannen kysymyksen tarkoituksena oli saada tietoa siitä, miten hyvin menetelmät olisivat istutettavissa osaksi organisaatioiden IT-arkkitehtuuria. Tämä tarkoitusperä kysymykselle tuli Hevnerin ym. (2004) arviontiperusteista mutta soveltui sangen huonosti kysymyksen muotoiluun tässä tapauksessa. Toisessa haastattelussa THL:n kohdalla kysymystä muokattiinkin laajentamalla kysymyksen käsittelemään koko organisaatiota.

KELA

Haastateltavat kokevat, että esitellyillä menetelmillä ei ole juurikaan IT-arkkitehtuurin kanssa tekemistä. Menetelmät nähdään irrallisten

välineiden asemaan, jotka eivät muuta mitään itse IT-arkkitehtuurissa. Toisin sanoen tämänkaltaisia menetelmiä ei integroida olemassa oleviin järjestelmiin vaan käytetään lähinnä apuvälineenä tutkimuksessa. Yleisellä tasolla julkishallinnon käytössä nähdään käyttömahdollisuuksia tukevana ja lisätietoa antavana työvälineenä mutta osana järjestelmiä tai pysyvänä osana automaattista päätöksentekoa työvälineitä ei nähdä.

THL

Haastateltavat näkevät, että jossain vaiheessa erityyppiset sosiaalisen median signaalien kuuntelemiset ja tarkkailut tulevat kuulumaan viestinnän ihmisten työtehtäviin. Koetaan myös, että erilaisissa turvallisuuteen liittyvissä konteksteissa menetelmillä voisi olla oleellista merkitystä organisaatiorakenteen kannalta, oletettavasti esimerkiksi jonkinlaisen varoitusjärjestelmän roolissa. Toisaalta nähdään myös THL:n näkökulmasta tarvetta erilaiselle palautteen ja viestinnän seurannalle ja tarkkailulle. Tällaista voisi olla esimerkiksi palautteen saaminen THL:n kannanottoihin. Tautiseuranta koetaan myös keskeiseksi osaksi toimintaa ja siinä tietyissä rajoissa sovellettavuutta nähdään. Loppujen lopuksi haastateltavat tiivistävät, että käytettävyyttä olisi ”viestinnän näkökulmasta erityistilanteissa tai palautteen saamisessa – mutta se käytetäänkö Twitteriä THL:ssä tutkimusaineistona niin siihen me ei varmaan voida vastata vaan tulevaisuus näyttää tämän".

9.kysymys

Yhdeksännessä kysymyksessä kysyttiin millaisia ajatuksia julkishallinto ja tämän rooli Twitter-datan aktiivisessa seurannassa herättää. Tavoitteena oli kartoittaa, nähdäänkö että populaatiosta voidaan kerätä Twitteriä valvomalla relevanttia seurantadataa ja toisaalta miten turvallisuuden edistäminen seurannalla ja yksityisyyden vaaliminen nähdään.

KELA

Haastateltavat arvioivat, että niiden viranomaistahojen, jotka tarvitsevat tämän tyyppistä seurantaa ja siihen liittyvää dataa, tulisi tehdä seurantaa. Esimerkkeinä annetaan poliisi ja THL. Yleisesti siis julkishallinnon osalta organisaatiot, jotka kaipaavat tietoa siitä, mitä missäkin päin voi lähitulevaisuudessa tapahtua. Selkeästi koetaan siis erilaiset turvallisuuteen liittyvät näkökulmat tärkeiksi esimerkiksi terrorin tai pandemioiden varalta varautumisessa. Kuitenkin koetaan tärkeäksi se, että seuranta ei värity julkisuudessa negatiiviseksi ja korostetaan, että esimerkiksi Yhdysvalloissa NSA:n tekemän seurannan kaltaista julkisuutta ei haluta.

THL

Haastateltavat kokevat tällä hetkellä Twitterissä olevien suomalaisten edustuksen liian pieneksi, että seurantaa voitaisiin järkeviä tuloksia odottaen tehdä. Seurannan toteuttamista ei juuri tällä hetkellä pidetä ajankohtaisena mutta huomioidaan, että tulevaisuudessa tällaista luultavasti tehdään sillä olettamuksella, että Twitterin käyttö yleistyy Suomessa. Laajemmin ajateltuna kokonaisuudessaan sosiaalista mediaa ja hakukonepalveluiden tekstimuotoista dataa nähdään käyttöpotentiaali seurannassa huomattavasti hedelmällisemmäksi, koska suomalaisten edustavuuskin otoksessa on tällöin huomattavasti parempi. Käytännössä potentiaali nähdään palautteen saamisessa ja viestinnässä THL:n osalta. Haastateltavat eivät näe THL:llä tässä mielessä roolia, johon sisältyisi tulevien asioiden ennakoiminen sosiaalisen median tekstidataa seuraamalla. Kokonaisuudessaan pelkän Twitter-dataan perustuvan otoksen edustavuus nähdään tällä hetkellä ongelmana Suomen osalta.

10.kysymys

Kymmenennessä kysymyksessä pyydettiin haastateltavia kuvaamaan erilaisia tilanteita ja skenaarioita, joissa esiteltyjen menetelmien hyödyntäminen nähdään järkevänä julkishallinnon kannalta. Osaltaan tämä vastaa Hevner ym. (2004) arviointiperusteluissa deskriptiivistä skenaarioiden kuvausta, joissa tutkittavat mallit ja menetelmät voisivat toimia ja olla käytettävissä.

KELA

Yleisellä tasolla haastateltavat näkevät kaksi eri näkökulmaa skenaarioiden kannalta. Näistä toinen on sen ennakoiminen mitä tulee tapahtumaan ja toinen on vastaavasti taaksepäin katsova lähestymistapa. Toisin sanoen käyttötapausskenaarioiden osalta mielipide on se, että tulevaisuutta voidaan tietyin indikaattorein ennakoida eri tavoin tekstimuotoisesta datasta. Esimerkkinä mainitaan erilaisten turvallisuusriskitilanteiden ennakointi seuraamalla Twitter-dataa. Toisaalta huomioidaan myös taaksepäin katsova tutkimus, jonka tarkoituksena on oppia miten päätöksentekoa voidaan mukauttaa ja mitä voidaan oppia tutkimalla tekstianalytiikalla takautuvasti sosiaalisen median dataa.

Twitteriä pidetään myös ajantasaistiedon ja uutisten saamisen väylänä. Tämä voi esimerkiksi kriisialueilta tulevassa viestinnässä olla tärkeää. Edelleen eräänä käyttöskenaariona nähdään erilainen testaaminen. Toisin sanoen esimerkiksi suuria rakennepoliittisia tai etuusjärjestelmiä koskevia muutoksia tehtäessä voitaisiin muutosten vastaanottoa etukäteen arvioida ennen lopullisen päätöksen sosiaalisen median tekstianalytiikalla. Erääksi käyttöskenaarioksi koetaan myös palautteen saaminen – joko suoraan tai sitten epäsuorasti koottuna tekstianalytiikan avulla (esimerkiksi toteutuksena voisi olla kollektiivinen mielipide Twitter-datasta koskien jotain asiaa).

THL

Haastateltavat pitävät selkeimpänä sovellutusskenaariona erilaisia kriisitilanteita ja näiden seurantaa sekä viestintää kriisien aikana. Pohtivaa asennoitumista otetaan palvelujärjestelmän palautejärjestelmän toteuttamiseen tekstianalytiikan keinoin samoin kuin mahdollisuuteen arvioida poliittista pelikenttää. Varsinaisesti kantaa näihin ei oteta vaan skenaariot tulevat esiin lähinnä mahdollisena sovellutuskohteena. Haastateltavat näkevät myös jonkinlaisen reaaliaikaisen palvelujärjestelmän toteuttamisen Twitter-pohjaisesti mahdollisena. Esimerkiksi terveydenhuollon peruutusaikojen viestintä kansalaisille olisi eräs tällainen mahdollisuus. Epäselväksi jää kuitenkin tekstianalytiikan osuus tässä yhteydessä.

Yhtenä käyttöskenaariona nähdään varoitusjärjestelmän toteuttaminen. Tässä varoitusjärjestelmässä ensimmäiset merkit puhkeavasta vaarasta havaitaan Twitterin välityksellä ja viranomaisvarmistuksen jälkeen varoitusviestit voidaan jakaa Twitterissä suoraan mobiilikäyttäjille. Tästä onkin jo sovellutuksia mutta kokonaan uusi idea on tämän soveltaminen käänteisesti siten, että kriisialueelta ajantasaista tietoa saataessa voidaan järjestelmää käyttää myös lähellä olevien viranomaisten informoimiseen asiasta. Tämä voisi osaltaan parantaa viranomaistahojen reagoimis- ja varautumisaikaa. Tämän kaltaisen järjestelmän toteuttaminen koettiin jo hyvinkin mahdolliseksi tiedotus- ja viestintätarkoituksiin Suomessa.

11.kysymys

Yhdennessätoista kysymyksessä pyydettiin haastateltavia olettamaan, että julkishallinnon taholta sovellettaisiin aktiivisesti tekstianalytiikkaa sosiaalisen median tuottaman tekstimuotoisen datan seurantaan. Tämän jälkeen haastateltavia pyydettiin arvioimaan tuottaisivatko nämä menetelmät lisäarvoa vanhojen tutkimus- ja seurantamenetelmien rinnalle. Osittain tämä kysymys vastaa myös Hevner ym. (2004) arvosteluperusteisiin deskriptiivisistä käyttöskenaarioista.

KELA

Haastateltavat näkevät, että tällaisessa tapauksessa menetelmät tuottaisivat lisäarvoa. Sen sijaan haastateltavia pohditutti se, onko Twitter-datasta esiinsaatavat tiedot relevantteja siinä mielessä, että Twitter-käyttäjät olisivat kuvaava otos populaatiosta. Kumpikin näkivät kuitenkin tämänkaltaisen seurannan tuottavan lisäarvoa mutta skeptisyyttä ilmeni otoksien tuottamien vääristymien kannalta. Esimerkiksi ikäluokkien oletettiin olevan eri tavalla edustettuina datassa.

Haastateltavat pohtivat, olisiko Twitter-populaatiota tutkimuksissa mahdollista joillain menetelmillä normalisoida. Tämä tekisi tutkimustuloksena saatavasta datasta populaation kannalta huomattavasti relevantimpaa. Tämä on

erittäin hyvä huomio, eikä tämän tutkimuksen aikana tehdyn kirjallisuuskatsauksen aikana aiemmissa tutkimuksissa esiintynyt tällaista lähestymistapaa. Sinällään idea on siis uusi ja erittäin mielenkiintoinen. Kuitenkin toteutuksen osalta tähän saattaisi liitttyä pientä problematiikkaa. Normalisoinnin toteuttamiseksi Twitter-datasta pitäisi käyttäjäkohtaisesti kyetä seulomaan tietynlaiset käyttäjät, arvioimaan heitä toista datavarantoa vasten ja tämän jälkeen koostamaan haluttu otos, joka vastaa demografisia keskiarvoja. Ottaen huomioon, että pyrkimyksenä on saada aikaan tekniikoita, jotka ovat nopeita, keveitä, kustannustehokkaita ja luonteeltaan suuntaa-antavia tämä voi osoittautua liian monimutkaiseksi lähestymistavaksi. Toisaalta esimerkiksi pitkittäistutkimuksiin sovellettuna tämä voisi olla järkevintä. Tietynlaisen informaation nopea jakaminen koettiin myös eräänä selkeänä käyttökohdeskenaariona sen kummemin erittelemättä mikä tekstianalytiikan rooli tässä olisi.

THL

Haastateltavat näkevät tämänkaltaisen seurannan tuottavan lisäarvoa varsinkin palautteen saamisessa. Haastateltavien mukaan ihmisten into vastailla tutkimus- ja kyselylomakkeisiin on hiipunut, joten vaihtoehto tälle on tervetullut. Twitter-datan ja laajemmin sosiaalisen median tekstidatan tekstianalytiikka voisi tarjota tämän vaihtoehdon. Samalla nähdään mahdollisia kustannussäästöjä vastaavan mielipidedatan tuottamisessa. Lisäksi informaatiota populaation muustakin käyttäytymisestä voitaisiin seurannalla saada ja esimerkiksi ravitsemustutkimukset nähdään tällaisina. Yleisesti ottaen tutkimuksen ja tiedonkeruun kannalta koetaan, että lisäarvoa tulee olemaan tulevaisuudessa mutta ei ehkä vielä tällä hetkellä. Joka tapauksessa Twitter-datan (ja laajennettuna muun sosiaalisen median ja hakukonepalveluiden) tekstianalytiikan rooli nähdään perustoimintaa tukevana työvälineenä. Haastateltavien näkemyksen mukaan nämä olisivat täydentäviä vaihtoehtomenetelmiä, jotka olisivat luonteeltaan perinteisiä menetelmiä tukevia ja tuottaisivat vastaavaa dataa hieman toisesta näkökulmasta katsoen.

12.kysymys

Kahdestoista kysymys koski tulevaisuuden visiota ja odotuksia siitä millaisena tekstianalytiikan rooli tulevaisuudessa nähdään julkishallinnon kannalta. Tältä osin mielenkiinnon kohteena oli sekä organisaation oman tekstimuotoisen datan tekstianalytiikka että julkishallinnon mahdollisesti harjoittama sosiaalisen (Twitter-datan) median tekstianalytiikka. THL:n kohdalla suoritetussa haastattelussa kahdettatoista kysymystä muokattiin siten, että korostettiin myös THL:n eikä pelkästään julkishallinnon roolia tekstianalytiikan käyttäjänä.

KELA

Haastateltavat huomioivat tekstimuotoisen datan olevan periaatteessa lähes ikuisesti tallennettavissa kokonsa puolesta. Tämä mahdollistaisi monenlaisia takautuvasti tehtäviä tutkimuksia, joiden tarkoitusta vielä ei voida edes arvioida. Toisin sanoen, koetaan mielekkäänä ylläpitää tekstimuotoisen datan varantoja, vaikka ei vielä olekaan selvillä mitä sieltä haluttaisiin tutkia. Haastateltavat huomioivat kuitenkin sosiaalisenkin median muuttuvan ja koska enenevissä määrin data sisältää muuta mediaa, kuten videoita, tulee tämäkin huomioida tallenuskapasiteetissä sekä analytiikan menetelmissä.

Tekstianalytiikalla nähdään olevan osansa ja merkityksensä varoitusjärjestelmässä, joka olisi viranomaistahon valvoma. Esimerkiksi tekstianalytiikan keinoin havaitut yllättävät tapahtumat vaatisivat aina viranomaistahon varmistuksen ennen kuin varoitusviestit lähetettäisiin. Tekstianalytiikalla ehostettu Twitterpohjainen kommunikointiväylä organisaatioiden ja yritysten väliseen (kriisi-)kommunikointiin nähdään myös merkityksellisenä. Syynä tähän on se, että Twitter nähdään infrastruktuurina olevan ulkopuolella kaikkia virallisia järjestelmiä, jolloin tätä voitaisiin käyttää viestintäväylänä mikäli viralliset väylät ja yhteydet ovat käyttökelvottomia. Tekstianalytiikan soveltamisen osalta jää tosin epäselväksi näiden tekniikoiden rooli. Samoin järjestelmien häiriötiedottaminen voisi kulkea tätä kautta. Tämä herättää myös esiin mahdollisen uuden tutkimuskysymyksen koskien Twitterin soveltuvuutta viranomaiskäyttöön esimerkiksi vanhojen menetelmien rinnalle ja kriisinajan viestintään ja tiedottamiseen kansalaisille.

Haastateltavat näkivät myös tekstianalytiikan ja neuroverkkoteknologian yhdistämisen tuomat mahdollisuudet potentiaalisesti mielenkiintoisiksi tulevaisuudessa. Tämä voisi omalta osaltaan luoda täysin uusia, tällä hetkellä vaikeasti kuviteltavissa olevia tapoja hyötyä Twitter-datan tekstianalytiikasta.

THL

Haastateltavat kokevat sosiaalisen median tekstianalytiikan menetelmien merkityksen julkishallinnon kannalta olevan kasvamassa. Viestintänäkökulmasta katsottuna haastateltavat korostavat jatkossa tarvittavan lisää vuorovaikutteisuutta siten, että tekstianalytiikan keinoin saatavaan palautteeseen reagoitaisiin. THL:n osalta tutkimuksissa, joissa THL:llä on velvollisuus toimia kansalaisten kanssa koetaan olevan sosiaalisen median tekstianalytiikan menetelmillä merkitystä.

Haastateltavat pohtivat myös sosiaalisen median viestintäkanavien keskittämistä siten, että tekstianalytiikkaa voitaisiin soveltaa siellä, missä ihmiset kritiikkiä esittävät. Toisin sanoen eri sosiaalisen median palveluissa ja keskusteluforumeilla. Tämä lisäisi myös läpinäkyvyyttä ja organisaation lähestyttävyyttä. Toisin sanoen ihmisille vastattaisiin sinne, missä he asioita kysyvät – ja vastattaisiin vaikka eivät kysyisikään vaan esittäisivät pelkästään kritiikkiä tai huolta jostain asiantilasta. Tämä koetaan mahdolliseksi mikäli

valtaosa suomalaisista siirtyy helposti seurattavissa olevan sosiaalisen median piiriin, kuten esimerkiksi Twitterin käyttäjiksi. Haastateltavat pitävät todennäköisenä, että tällöin myös erilaisia seurantajärjestelmiä otettaisiin käyttöön julkishallinnossa. Seurantajärjestelmien osalta potentiaalia nähdään tulevaisuutta ajatellen esimerkiksi levottomuuksien ennakoinnin kannalta. Seurantajärjestelmiä ei kuitenkaan nähdä pelkästään yhden datalähteen varaan rakennetuilta vaan ennemminkin kombinaationa eri sosiaalisen median ja hakukonepalveluiden tekstidataa ja tekstianalytiikkaa.

Haastateltavat huomioivat myös, että monenlaiset terveyttä ja kuntoilua seuraavat sensoriset laitteet ovat nykyisin suosittuja. Näiden pohjalta voisi olla mahdollista tehdä jotain aiheeseen liittyvää seurantaa tai tutkimusta. Pohditaan myös mahdollisuutta kuvastaa tällä reaaliaikaista kansakunnan terveydentilaa. On kuitenkin sangen selvä, että tällaisessa datassa tilastollinen vääristymä olisi potentiaalisesti massiivinen, koska terveysdataa saataisiin pääosin hyväkuntoisilta ja kuntoilua harrastavilta populaation jäseniltä. Lisäksi pohditaan erilaisten terveydenedistämiskampanjoiden kehittämistä Twitter-pohjaisesti. Esimerkkinä haastateltavat mainitsevat vanhuksille toimitettavat varoitukset liukkaista keleistä tai ohjeet kotijumpasta. Mielenkiintoista tässä on, että voisivatko nämä ohjeistukset olla tekstianalytiikan tai jonkin sensorisen datan vapaaehtoisen luovuttamisen pohjalta personalisoituja tai lokalisoituja.

Twitterin ja yleisemmin sosiaalisen median tuottamaa tekstimuotoista dataa nähdään myös käytettävän reaaliaikaisen tiedon tuottamisessa hitaammin saatavilla olevan tutkimustiedon täydentämiseksi. Esimerkkinä tästä annetaan rokotuskattavuuden tutkimus tekstianalytiikalla terveyskeskusdatan lisäksi.

Yleisellä tasolla merkittävimmiksi sovellusaloiksi koetaan alat, joissa tulee kyetä nopeasti reagoimaan muutoksiin. Tältä osin haastateltavat huomaavat potentiaalisen käyttökohteen viranomaisyhteistyössä esimerkiksi terveysviranomaisten, poliisin ja tullin välillä. Esimerkiksi mikäli sosiaalisen median ja hakukonepalveluiden tekstidatan analytiikasta ja terveyskeskuspotilaiden oireistosta voidaan päätellä, että jokin tietty huume on nousemassa muotiin, voidaan tulliviranomaisia asiasta informoida. Tämä viestintä voi kulkea myös toiseen suuntaan, jolloin terveysviranomaiset osaisivat varautua tulevaan.

5.2 Empiirisen tutkimuksen johtopäätökset

Tässä alaluvussa käsitellään tutkimuksen empiirisen osuuden tuottamia tuloksia ja näistä tehtäviä johtopäätöksiä.

Nykytilanteessa kummallakaan tutkimuksen kohteena olleella organisaatiolla ei ole juurikaan tekstimuotoista dataa. Toisaalta kummankaan osalta ei ilmene nykyisessä käytössä myöskään merkittäviä syitä tekstianalytiikan käyttöön, sillä molemmilla organisaatioilla on perinteisesti käsitelty omaa dataa rakenteisessa muodossa. THL:n osalta tekstianalytiikkaa

on mahdollisesti saatettu käyttää osana tutkimusta itsenäisenä apuvälineenä. Kummankin organisaation osalta omaa tekstidataa syntyy lähinnä palautekanavien kautta ja tätäkään kautta ei kummassakaan tapauksessa runsaasti. Lisäksi molempien kohdalla koetaan tärkeäksi analysoida tämä palautedata inhimillisesti jonkin toimihenkilön taholta. Tältä pohjin tekstianalytiikalle ei ole juurikaan käyttöä KELA:n ja THL:n oman datan kannalta. Koska kohdeorganisaatioilla ei ole omaa tekstimuotoista dataa tai tarvetta tekstianalytiikalle oman datan suhteen, on Hevner ym. (2004) design science-arviointiperusteita mahdoton käyttää menetelmien arvioinnissa organisaatioiden oman datan suhteen. Tämän vuoksi huomio on merkittävästi keskittynyt koskemaan kohdeorganisaatioiden kuvailemia käyttöskenaarioita sekä mahdollista käyttöä sosiaalisen median tuottaman datan analytiikassa.

Tekstianalytiikan soveltaminen sosiaalisen median tuottamaan dataan kiinnosti haastateltavia merkittävästi. Molemmat organisaatiot, etenkin THL ja yleisemmällä tasolla muut julkishallinnon tahot voisivat selkeästi hyötyä sosiaalisen median tuottaman tekstidatan tekstianalytiikasta. Kummassakin haastattelussa tuli ilmi, että menetelmiin oli suurta mielenkiintoa tiettyjen aihepiirien kohdalla. Tällaisia olivat esimerkiksi erilaiset turvallisuuteen liittyvät seurantajärjestelmät ja kriisitilanteiden aikainen tiedon louhiminen twiiittivirrasta.

Molemmissa haastatteluissa kävi ilmi, että haastateltavat pitivät menetelmiä toimivina, loogisina, mielekkäinä, luotettavina ja sovellettavissa olevina erilaisille aloille. Lisäksi käytettyjen tekniikoiden rakenteen puolesta menetelmien ei koettu olevan liian kompleksisia hyödynnettäväksi julkishallinnon taholta. Enemmänkin arvioitiin muiden toimialojen osaamisen korostuvan datan tuloksia analysoitaessa.

Kummankin organisaation edustajat pitivät menetelmien tuottamaa tietoa soveliaana päätöksentekoon tietyin reunaehdoin. Näitä olivat käyttötarkoituksen konteksti ja se, että aina tarvittaessa menetelmän tuottamaa dataa voidaan tarkastella jälkikäteen ns. ”kovaa dataa” eli virallista tutkimusdataa vasten. Lisäksi kummankin organisaation edustajien taholta nähtiin menetelmien tuottavan lisäarvoa monenlaisissa konteksteissa esimerkiksi reaaliaikaisuuden ja kustannustehokkuuden ansiosta sekä tarjoamalla vaihtoehtoisen ja uuden kanavan tiedon saamiselle entisten menetelmien rinnalle.

On epäselvää, missä määrin tekstianalytiikkaa voitaisiin hyödyntää muissa julkishallinnon osissa. On kuitenkin luultavaa, että muillakin julkishallinnon instansseilla on perinteiset tietojenkäsittelytapansa rakenteisen datan pohjilta. Tämä antaisi syytä olettaa, ettei tekstimuotoista dataa ole siinä määrin, että esitellyt tekstianalytiikan menetelmät toisivat suurta lisäarvoa julkishallinnon oman tekstidatan käsittelyyn. Sen sijaan julkishallinnon taholta katse voitaisiin siirtää enenevissä määrin kohti sosiaalisen median tekstidataa ja tekstianalytiikan soveltamista tämän seurannassa ja analysoimisessa.

Tulevaisuuden osalta arviot tekstianalytiikan käytöstä nimenomaan sosiaalisen median tekstimuotoisen datan käsittelyyn olivat

optimistisia ja haastateltavat näkivät paljon erilaisia sovellutusaloja tekniikoille. Näitä olivat muun muassa palautteensaamis- ja varoitusjärjestelmät, tekstianalytiikan avulla tehostetut viranomaisviestinnän järjestelmät sekä yleisemmällä tasolla Twitterin käyttäminen viestinnässä. Lisäksi nähtiin täsmällisten geolokaatiotietojen mahdollistavan eri tavoin esimerkiksi terveydenhuollon palveluiden tehostamista. Sovellutuksia tästä voisivat olla esimerkiksi geolokaatiotiedolla painotettu hoitopaikkajärjestelmä ja populaatiota koskevan terveysdatan kerääminen Twitterin avulla.

Haastatteluissa nähtiin, että tulevaisuudessa tämän tyyppisen datan merkitys päätöksenteossa voi kasvaa. Twitter-datan ja laajemmin sosiaalisen median tuottaman tekstidatan tekstianalytiikka nähtiin keinoksi kuvastaa sitä, mitä reaalimaailmassa tapahtuu. Lisäksi arvioitiin, että esimerkiksi yhteiskunnallisia rakennemuutoksia tehtäessä voitaisiin etukäteen ennakoida kansakunnan mielipiteitä tarkkailemalla sosiaalisen median indikaattoreita.

Yleisellä tasolla koettiin, että sosiaalisen median tarkkailu tekstianalytiikan avulla olisi hyödyllisintä päätöksenteon aloilla, joissa on kyettävä toimimaan muutenkin erittäin nopeasti ja usein epävarman tiedon perusteella. Tällaisia tilanteita ovat esimerkiksi kriisitilanteet ja luonnonmullistukset sekä varautuminen potentiaalisiin uhkiin, kuten mellakoihin ja tauteihin. Koettiin myös, että mallit soveltuisivat globaalien trendien tarkkailuun siten, että esimerkiksi eri rokotuskampanjoiden saamaa vastaanottoa globaalisti voitaisiin käyttää arviona siitä, minkälaisen vastaanoton kampanja Suomessa saa. Käyttöarvoa nähtiin myös virallisen tutkimusdatan testaamisen ja vahventamisen näkökulmasta.

KELA:n ja THL:n omiin organisaatioihin tekstianalytiikan menetelmät soveltuvat sikäli huonosti, että käyttökohteita ei ole juurikaan. Kuitenkin THL:n osalta arvioitiin että mahdollisesti tulevaisuudessa luvanvaraisesti tutkimuskäytössä voitaisiin käyttää sairaaloiden hallussa pitämiä terveysarkistoja esimerkiksi lääkärinlausunnoista. Kuitenkin tätä koskien lainsäädännölliset, eettiset ja yksityisyyttä koskevat ongelmat ovat ilmeisiä.

Johtuen aktiivisten Twitterin käyttäjien verrattain pienestä määrästä Suomessa arvioivat haastateltavat, että mikäli sosiaalisen median tekstianalytiikkaa jatkossa tehdään tulisi sen kohdistua laajennettuna useampiin sosiaalisen median lähteisiin, keskustelufoorumeihin ja hakukonepalveluiden tekstidataan. Twitterin osalta pidettiin nykyistä suomalaisten käyttäjien määrää liian pienenä, jotta kunnollista tutkimusta tai seurantaa voitaisiin suorittaa. Samoin haastateltavat huomioivat, että tilastollisena otoksena suomalaiset twiittaajat eivät välttämättä ole demografisesti kovinkaan heterogeeninen. Vaihtoehtoina nähtiin muiden sosiaalisen median tietolähteiden lisäämisen tai esimerkiksi tautiseurannan osalta kansainvälisen datan käyttämisen.

Demografisten tekijöiden osalta haastattelussa tuli myös esiin idea, jota ei oltu käytetty kirjallisuuskatsauksen esimerkkitutkimuksissa. Tämä oli

Twitter-datan normalisointi populaation tiedossa olevien demografisten tekijöiden avulla. Vaikkakin ongelmallista toteuttaa, tämä voisi mahdollisesti parantaa Twitter-datasta saatavien tulosten tilastollista kuvaavuutta koko kansakunnasta.

Tekstianalytiikan työvälineillä ei todettu olevan merkitystä IT-arkkitehtuurin kannalta vaan menetelmät nähtiin erillisinä, tiettyyn spesifiin tarkoitukseen tarkoitettuina välineinä. Toisaalta THL:n yhteydessä pidettiin todennäköisenä, että tämänkaltaisten työvälineiden hyödyntäminen viestinnän osalta korostunee jatkossa.

Julkishallinnon aktiiviseen rooliin Twitter-datan seuraamisessa suhtauduttiin positiivisesti. Haastateltavat pitivät sangen selkeästi menetelmiä käyttökelpoisina ja järkevinä niille julkishallinnon osa-alueille, jotka vastaavat esimerkiksi turvallisuudesta. Tällaisia instansseja voisivat olla esimerkiksi poliisi ja tulli. Onkin mahdollista, että tämän tutkielman osalta erilaista tutkimusdataa oltaisiin saavutettu mikäli kyseiset tahot olisivat olleet mukana tarkastelussa.

Käyttöskenaarioita koskevissa näkemyksissä koettiin, että Twitter-datan seurannasta voitaisiin saada hyötyä monin tavoin. Huomioitiin, että mikäli tekstimuotoista dataa on pysyvästi tallennettuna, voidaan tehdä takautuvasti tutkimuksia ja etsiä korrelaatioita eri indikaattorien ja tapahtumien välillä. Samoin voidaan havainnoida sitä, mitä on tapahtumassa ”juuri nyt” esimerkiksi trendien ja nousevien uutisaiheiden havaitsemistekniikoilla sekä joukkoistetun aistinnan menetelmillä kriisitilanteissa. Lisäksi huomioitiin mahdollinen tulevaisuuden ennakoiminen eräänä käyttösovellutuksena erityisesti turvallisuuteen liittyvissä asioissa. THL:n osalta kiinnostusta herätti myös tutkimukselliselta kannalta erilaisten ruokavalio- ja terveysaiheisten trendien ennakointi. Koettiin, että Suomessa nämä trendit syttyvät viiveellä ja näin esimerkiksi kansainvälisestä Twitter-datasta olisi mahdollista havaita, mikä tulee olemaan seuraava ruokavaliotrendi. Yleisellä tasolla myös palvelu- ja viestintäjärjestelmien kehittäminen tekstianalytiikkaan pohjautuvan palautteen muodossa koettiin mahdolliseksi.

Koettiin, että tekstianalytiikkaa tulee tehdä sieltä, missä ihmiset ovat. Toisin sanoen tulevaisuudessa tarkastelussa ei välttämättä tulisi olla pelkästään Twitter vaan menetelmiä tulisi käyttää soveltaen eri sosiaalisen median palveluiden, hakukonepalveluiden ja keskustelufoorumien tekstidataan. Tämä osaltaan korostaa vaatimusta käyttöön valittavien mallien yksinkertaisuudesta, skaalautuvuudesta ja muokattavuudesta.

Haastattelun rakenne oli suunniteltu siten, että arvioita kyettiin tekemään myös Hevner ym. (2004, s.86) design science-arviointiperusteiden perusteella. Haastattelujen perusteella observatiivisen arvioinnin osalta Twitter-dataan pohjautuvat tekstianalytiikan menetelmät havaittiin tarkoitusperiinsä sopiviksi ja toimiviksi. Lisäksi haastateltavat kokivat, että menetelmät tuottavat päätöksentekoon soveltuvaa informaatiota. Vastaukset ovat tältä osin linjassa tutkimuksissa saavutettujen sangen hyvien tulosten kanssa.

Haastateltavat arvioivat Twitter-datan tekstianalytiikan menetelmiä myös staattisten ominaisuuksien ja organisaatioon soveltumisen kannalta. Tätä kautta saatiin osaltaan vastauksia menetelmien soveltuvuudesta Hevner ym. (2004, s.86) analyyttisestä arviointinäkökulmasta. Erityisesti painotettiin kysymystä menetelmien kompleksisuudesta. Tämän taustalla oli pyrkimys todeta, koetaanko jotkut tekniikat liian monimutkaisiksi sovellettavuutensa kannalta. Haastateltavat eivät kokeneet esiteltyjä menetelmiä kompleksisiksi vaan pitivät menetelmiä tarkoituksenmukaisina ja rakenteeltaan loogisina. IT-arkkitehtuurin kannalta kysymys menetelmien soveltuvuudesta osoittautui huonoksi, sillä kummassakaan tapauksessa ei koettu, että tekstianalytiikalla olisi mitään konkreettista roolia IT-arkkitehtuurin kannalta. Tekstianalytiikan rooli tältä osin nähtiin lähinnä tutkimuksellisena työvälineenä ja apukeinona.

Deskriptiivisen arvioinnin osalta haastateltavia pyydettiin kuvailemaan esiteltyjen menetelmien sovellettavuutta sellaisenaan tai modifioituna organisaation käyttöön. Haastateltavia pyydettiin myös arvioimaan tulevaisuuden käyttöskenaarioita. Haastattelun aikana esiin tulleita käyttöskenaario-ideoita olikin runsaasti. Tällaisia olivat muun muassa seuraavat.

- Erilaiset valvontajärjestelmät
- Tiedonkeruu- ja tiedotusjärjestelmät onnettomuus- ja kriisitilanteissa
- Uutisaiheiden ja trendien havaitsemisjärjestelmät
- Kansalaisten mielipiteiden huomioimisjärjestelmä
    - Etenkin ennakoivassa merkityksessä mahdollistaen päätösten esivalmistelun pohjautuen sosiaalisesta mediasta havaittavaan mielialareaktioon
- Palautekanava
    - Epäsuoran palautteen kerääminen tekstianalytiikalla eri sosiaalisen median lähteistä, hakukonedatasta ja keskustelufoorumeilta
- Tautiseurantajärjestelmä
    - Aineiston osalta täytyy olla laajempi kuin pelkät Suomen Twitter-käyttäjät. Tällöinkin rooli olisi virallista kanavaa tukeva.
- Työväline yhteiskunnalliselle ja sosioekonomista statusta tutkivalle tutkimukselle
- Sävyanalyysi ja kollektiivisen mielialan tutkimukset
- Viranomaisten tiedotusväylänä, jossa tekstianalytiikalla on oma lisäinformaatiota tuottava roolinsa

Kirjallisuuskatsauksen perusteella luotuihin hypoteeseihin voidaan empiirisen osuuden perusteella esittää seuraavat vastaukset.

Ensimmäinen hypoteesi:

Paul & Dredze (2011a & 2011b) ATAM- ja ATAM+-malli on mahdollisesti suoraan käytettävissä tai sovellettavissa kohdeorganisaatioiden omaan tekstimuotoiseen dataan.

Hypoteesi ei pidä paikkaansa, koska kumpikaan kohdeorganisaatioista ei omista tekstimuotoista dataa siinä määrin, että mallilla olisi mitään käyttökohteita. Teoriassa mallia voisi soveltaa lääkärinlausuntojen muodostamaan dataan esimerkiksi oireiden etsinnässä ja näiden visualisoinnissa karttapohjalle. Kuitenkin on epäselvää, missä määrin tätä tietoa on tekstimuotoisena. Lisäksi tämä data on jo joka tapauksessa rakenteisessa muodossa, joten tarvetta menetelmälle ei ole. Sen sijaan mallia voitaisiin hyvin käyttää Twitterin ja yleisemmin sosiaalisen median tuottaman tekstimuotoisen datan osalta.

Toinen hypoteesi:

Trendien havaitsemis-tekniikoiden käyttö kohdeorganisaatioiden omaan tekstimuotoiseen dataan on mahdollista ja kannattavaa.

Hypoteesi ei pidä paikkaansa, koska kumpikaan kohdeorganisaatioista ei omista tekstimuotoista dataa siinä määrin, että trendien havaitsemis-tekniikoiden käyttö olisi järkevää. Jälleen havaitaan, että sovellettuna sosiaalisen median tarkkailuun menetelmillä olisi laajaltikin käyttöä tiettyjen julkishallinnon instanssien taholta.

Kolmas hypoteesi:

Twitter-datan tekstianalytiikan tekniikoiden käyttäminen sosiaalisen median tuottaman tekstidatan seurantaan tuottaa julkishallinnon kannalta lisäarvoa.

Hypoteesi pitää paikkansa. Monilla julkishallinnon tahoilla olisi käyttöä reaaliaikaiselle ja kustannustehokkaalle tavalle tarkkailla sosiaalisen median tapahtumia ja sitä kautta tarkkailla ja ennakoida reaalimaailman tilaa. Haastateltavat pitivät esiteltyjen menetelmien tuottamaa tietoa soveltuvana päätöksentekoon etenkin tilanteissa, joissa virallista reittiä saatavaa dataa ei ole vielä käytössä. Tämä korostaa etenkin mielenkiintoa nopeasti saatavaan informaatioon. Tyypillinen sovelluskohde voisi olla esimerkiksi THL:n kannalta epäsuoran palautteen kerääminen rokotuskampanjoista. Suurin käyttöpotentiaali olisi todennäköisesti kuitenkin saatavilla erilaisista

tutkimuskohteista ja mahdollisista seurantajärjestelmistä. Tällaisia voisivat olla esimerkiksi tautiseurannat sekä tutkimukset, joissa pyrittäisiin hyötymään sosiaalisen median datan mielipiteitä ja tunteita kuvaavasta kontentista.

Neljäs hypoteesi:

Twitter-dataan perustuvia tautimallinnustekniikoita voidaan käyttää täydentävänä informaatiolähteenä käytössä olevien menetelmien rinnalla.

Hypoteesi pitää paikkansa. On tosin huomioitava Suomen kohdalla vähäinen Twitter-käyttäjien määrä. Täten on syytä harkita joko kansainvälisellä tasolla menetelmän hyödyntämistä tai mahdollisesti laajempaa tarkkailudataa. Tällainen voisi olla esimerkiksi viranomaistahojen osalta useiden eri sosiaalisen median ja hakukonepalveluiden tuottama tekstimuotoinen data yhdistettyinä erilaisten forumien tekstidataan.

Viides hypoteesi:

Kohdeorganisaatioiden tekstidata ei ole hyvää sävyanalyysin kannalta.

Hypoteesi pitää paikkansa. Hypoteesia luotaessa oletettiin, että KELAn ja THLn kaltaisilla organisaatioilla tekstimuotoinen data on luonteeltaan ja sävyltään hyvin virallista. Tämä vaikuttaisi sävyanalyysin käyttömahdollisuuksiin negatiivisesti. Sen lisäksi, että hypoteesi todettiin oikeaksi todettiin myös tekstimuotoisen datan olematon määrä kohdeorganisaatioissa. Toisaalta palautekanavien kautta olisi saatavilla hyvinkin värittynyttä tekstimuotoista dataa, mutta molemmat organisaatiot käsittelevät nämä henkilökohtaisesti jonkin työntekijän taholta.

Empiirisen osuuden analysoinnin jälkeen on mahdollista muodostaa vastaus toiseen tutkimuskysymykseen. Toinen tutkimuskysymys koski sitä, voidaanko Twitter-datan tekstianalytiikan menetelmiä, työkaluja ja malleja soveltaa julkishallinnon tarpeisiin. Kysymys oli sikäli kaksijakoinen, että tarkastelussa oli näiden mallien ja menetelmien soveltaminen sekä organisaation omaan että sosiaalisesta mediasta saatavaan tekstimuotoiseen dataan.

Esiteltyjen tekstianalytiikan menetelmien soveltaminen KELA:n ja THL:n osalta omaan tekstimuotoiseen dataan ei yksiselitteisesti ole mahdollista, koska tekstimuotoista dataa ei ole eikä tarvetta tekstianalytiikalle oman datan suhteen ole olemassa. Koska toinen tutkimuskysymys koski kuitenkin tekstianalytiikan käyttöä julkishallinnon osalta on huomioitava, että mahdollisesti muilla julkishallinnon instansseilla omaa tekstidataa on. Näiltä osin sovellettavuus oman tekstimuotoisen datan osalta jää auki. Soveltamismahdollisuuksia julkishallinnon omiin tekstivarantoihin ei kuitenkaan voida arvioida KELA:n ja THL:n osalta, sillä tekstidataa ei ole. Tältä osin tutkimuskysymys jää avoimeksi.

Sen sijaan tekstianalytiikan menetelmiä voidaan laajastikin nähdä käytettävän julkishallinnon taholta sosiaalisen median tuottamaan dataan. Näiltä osin vastaus toiseen tutkimuskysymykseen on, että Twitter-datan tekstianalytiikan menetelmiä voidaan soveltaa hyvin laajasti ja monissa eri yhteyksissä sosiaalisen median tuottaman tekstimuotoisen datan analytiikkaan, keräämiseen ja seurantaan.

6 TUTKIELMAN TULOKSET JA JOHTOPÄÄTÖKSET

Tässä luvussa käsitellään tutkielman aikana saatuja tuloksia. Tässä yhteydessä on käsitelty tutkimuskysymyksiä ja niihin tutkielman perusteella saatavia vastauksia sekä esitetty näiden perusteella johtopäätöksiä. Samoin esitetään lyhyesti vastaukset kirjallisuuskatsauksen jälkeen luotuihin hypoteeseihin. Tämän jälkeen käsitellään tutkielman tekemisen aikana ilmitulleet parannusehdotukset, jatkotutkimuskysymykset sekä arviot käytetyistä tutkimusmenetelmistä.

Ensimmäinen tutkimuskysymys käsitteli miten Twitteristä saatavaa tekstimuotoista dataa voidaan käyttää tekstianalytiikan avulla. Näiltä osin tutkimuskysymys oli jaettu kolmeen ala-aiheeseen, joita olivat käyttökontekstit, käytetyt menetelmät ja menetelmien perusteella saadut tulokset. Tutkimuskysymykseen vastaukset saatiin pääosin kirjallisuuskatsauksessa tutustumalla aihetta käsitteleviin tutkimuksiin. Näiden pohjalta Twitter-dataa voidaan tekstianalytiikan keinoin hyödyntää hyvin monissa eri käyttökonteksteissa. Tutkimuksen aikana on tullut esiin tai sivuttu ainakin seuraavia käyttökohteita.

- Trendien ja uutisaiheiden havaitseminen
- Epidemioiden havaitseminen ja seuranta
- Sävyanalytiikkaa soveltavat tutkimukset
    - Mielialatutkimukset
    - Onnellisuus-tutkimukset
    - Osakekurssien ennakointi
- Elintapariski-tutkimukset
- Ennakoivat tutkimukset
    - Rikollisuuden ennakoiminen
    - Mellakoiden ennakoiminen
- Joukkoistetun aistinnan hyödyntäminen
    - Maanjäristyksien havaitseminen
    - Kriisitilanteet
    - Sensorisen terveysdatan saaminen

- Poliittinen tutkimus
- Seurantajärjestelmät

Käytettyjä menetelmiä ja työkaluja ovat tyypillisimmin olleet erilaiset ohjatun koneoppimisen menetelmät, luokittelut ja sävyanalyysimenetelmät. Haasteina näille menetelmille on nähty Twitter-datan kohinaisuus ja lyhyt mitta (Barbier & Liu, 2011), datan määrä (Aggarwal & Zhai, 2012), käytetyn kielen sisältämä sarkasmi ja huumori (González-Ibáñez ym., 2011) sekä tietyissä tapauksissa epävarmat geolokaatiotiedot (Hecht ym., 2011).

Jatkossa tekstianalytiikan menetelmien haasteet tulevat keskittymään enenevästi esimerkiksi skaalautuvuuden ja multimediakontentin huomioimiseen (Hu & Liu, 2012). Tutkimuksen perusteella vaikutti, että menetelmistä psykologian alalta lainattu tunteiden mallintamistapa POMS (engl. Profile of Mood States) on vakiinnuttamassa erilaisine sovellutuksineen paikkaansa Twitter-datan tekstianalytiikassa sävyanalyysin osalta. Kaikkien käyttökontekstien tutkimuksissa lukuunottamatta poliittisia aiheita on saatu vähintään rohkaisevia ja paikoin erinomaisia tuloksia tutkielmassa esitellyillä menetelmillä. Poliittisen tutkimuksen osalta tekstianalytiikan menetelmien soveltamisella vaikuttaisi olevan suuria haasteita edessään. Esimerkiksi sävyanalyysin soveltaminen poliittisiin aiheisiin sekä poliittis-demografisten taustatekijöiden huomioiminen on ongelmallista.

Toisessa tutkimuskysymyksessä kysyttiin ovatko Twitter-datan tekstianalytiikan menetelmät, työkalut ja mallit sovellettavissa julkishallinnon tarpeisiin. Tutkimuskysymyksen kannalta tarkasteltiin soveltamista sekä julkishallinnon oman tekstidatan kannalta että sosiaalisen median tuottaman tekstidatan kannalta. Tekstianalytiikkaa ei voida soveltaa kohdeorganisaatioiden eli THL:n ja KELA:n oman tekstimuotoisen datan suhteen. Tämä johtuu siitä, että kummallakaan organisaatiolla ei ole tekstimuotoista dataa merkittäviä määriä eikä täten varsinaista tarvettakaan tekstianalytiikan soveltamiseen oman datan kannalta. Julkishallintoa yleisemmin ajatellen ei voida siis arvioida tekstianalytiikan sovellettavuutta omaan tekstimuotoiseen dataan. On kuitenkin todennäköistä, että mikäli tekstimuotoista dataa eri julkishallinnon instansseilla on, on löydettävissä laajoja sovellutusmahdollisuuksia ja käyttökohteita. Sen sijaan kävi ilmi, että koskien sosiaalisen median tuottamia tekstidatavarantoja kohdeorganisaatioiden edustajilla oli suurta mielenkiintoa näiden potentiaaliselle hyödyntämiselle tekstianalytiikan keinoin. THL:n osalta havaittiin mahdollisia käyttöskenaarioita olevan epidemioiden seurannassa, palautteen saamisessa, viestinnässä sekä erilaisissa tutkimuksissa.

Yleisesti ottaen todettiin julkishallinnon hyötyvän eniten erilaisista seurantaan tai turvallisuuden parantamiseen liittyvistä järjestelmistä. Lisäksi todettiin sosiaalisen median tekstianalytiikan menetelmien tärkeimmän piirteen julkishallinnon kannalta liittyvän näiden tuottamaan nopeaan tietoon. Toisin sanoen julkishallinnon potentiaalisimmat mahdollisuudet sosiaalisen median tekstidatan hyödyntämisessä olisivat aloilla, joissa ripeä tiedonsaaminen

tilannetietoisuuden kannalta on tärkeää. Tämä viittaisi siihen, että eniten julkishallinnon taholta voisivat hyötyä poliisin ja tullin kaltaiset osapuolet. Menetelmiä näissä tapauksissa voitaisiin soveltaa alkuperäiseen käyttötarkoitukseensa eli Twitter-datan analytiikkaan suoraan. Näiden aspektien perusteella voidaan toiseen tutkimuskysymykseen vastata, että menetelmät ovat sovellettavissa julkishallinnon taholta etenkin sosiaalisen median tekstidataan ja varauksella julkishallinnon eri instanssien omaan tekstidataan, mikäli tätä on.

Kirjallisuuskatsauksen jälkeen luotujen hypoteesien osalta todettiin, etteivät Paulin ja Dredzen (2011a & 2011b) ATAM- ja ATAM+-mallit ole sovellettavissa kohdeorganisaatioiden oman datan analysoimiseen. Tämä johtuu siitä, että kohdeorganisaatioilla ei tällaista dataa ole. Samasta syystä myöskään trendien havaitsemistekniikoiden soveltaminen organisaatioiden omaan dataan ei ole mahdollista. Kolmas hypoteesi todettiin oikeaksi, sillä empiirisen osuuden perusteella Twitter-datan tekstianalytiikan soveltaminen sosiaalisen median tuottamaan dataan nähdään lisäarvoa tuottavaksi julkishallinnon kannalta.

Neljäs hypoteesi todettiin myös paikkansa pitäväksi, sillä Twitter-dataan perustuvia tautimallinnustekniikoita voidaan käyttää täydentävänä informaatiolähteenä käytössä olevien menetelmien rinnalla. Tälle asettaa kuitenkin Suomen osalta vähäinen Twitter-käyttäjien määrä omat rajoituksensa. Täten toteutuksessa tulisi vaihtoehtoisesti käyttää kansainvälistä tautidataa tai käyttää menetelmiä myös muiden sosiaalisen median palveluiden ja esimerkiksi hakukonepalveluiden tekstidataan. Viidennen hypoteesin osalta kohdeorganisaatioiden tekstidata todettiin sävyanalyysin kannalta sangen huonoksi kohteeksi.

Tutkielman perusteella on syytä olettaa sarkasmin olevan jatkossakin vaikea ongelma automaattisen Twitter-datan tekstianalytiikan kannalta (González-Ibáñez ym., 2011). Tekstianalytiikan tulkintaan vaikuttavia demografisia piirteitä havaittiin olevan runsaasti asuttujen alueiden yliedustus ja miesten yliedustus varhaisessa Twitterin omaksumisvaiheessa (Mislove ym., 2011). Lisäksi joissain tutkimuksissa koettiin käyttäjien ikäryhmiin jakautuminen ongelmalliseksi (Paul & Dredze, 2011b). Epätäsmälliseen geolokaatiotietoon perustuvat ongelmat lienevät väistymässä mobiilikäyttäjien määrän ja saatavilla olevan täsmällisen geolokaatiotiedon lisääntyessä. Käytetyiden menetelmien ja tekniikoiden osalta lupaavimmalta trendihavaitsemisen menetelmältä vaikutti Mathioudakisin ja Koudasin (2010) esittelemä TwitterMonitorin arkkitehtuuri. Tautiseurannan osalta havaittiin lukuisia tekniikoita, joiden soveltaminen olisi epäilemättä tehokasta (esimerkiksi Lampos & Cristianini, 2010 ja Culotta, 2010b sekä Paul & Dredze, 2011a & 2011b). Sävyanalyysin osalta voidaan yleisesti ottaen suositella ohjattuun oppimiseen perustuvaa luokittelijaa ja POMS-mallintamisen (Profile of Mood States) soveltavaa käyttöä käyttökontekstista riippuen.

Tutkielman tekemisen aikana tuli eteen lukuisia erilaisia tekniikoita ja menetelmiä sekä näitä kohtaan esitettyä kritiikkiä. Näiltä osin on

tutkimuksessa pyritty esittämään joitakin parannusehdotuksia. Hankalaksi Twitter-datan tekstianalytiikan sovellutusalueeksi todetun poliittisen tutkimuksen osalta olen esittänyt parannetun lähestymistavan, jolla demografisia taustatekijöitä voidaan ottaa enemmän huomioon. Lisäksi empiirisessä vaiheessa haastattelutilaisuudessa esiintullut ajatus Twitter-datan twiittien normalisoinnista siten, että vastaavuus normaaliin populaatioon on tilastollisesti pätevämpi, oli uusi.

Tutkielman perusteella voidaan luoda useita jatkotutkimuskysymyksiä ja -aiheita. Eräs näistä on muiden julkishallinnon instanssien tekstianalytiikan käytön tutkiminen. Toisaalta jo tämän tutkielman perusteella on syytä olettaa, että julkishallinnon taholta tekstianalytiikkaa voidaan hyödyntää vähintäänkin sosiaalisen median tuottaman tekstimuotoisen datan osalta. Näin ollen jatkotutkimuksessa hedelmällisempi tutkimuskohde voisi olla käytännön sovellutuksen toteuttaminen. Tällainen voisi olla esimerkiksi simplifikoidun epidemioiden havaitsemisjärjestelmän kehittäminen tai seuranta- ja hälytysjärjestelmän luominen julkishallinnon käyttöön.

Tutkielman perusteella voidaan THL:n osalta suositella toteutukseltaan yksinkertaisia epidemioiden havainnointimenetelmiä. Eritoten Lampos ja Cristianinin (2010) ja Culottan (2010b) lähestymistavat tulee tässä kyseeseen. Tutkimuksellisessa käytössä Paulin ja Dredzen (2011a & 2011b) ATAM- ja ATAM+-mallit voisivat olla käytettävissä esimerkiksi elintapariskien kartoitukseen. Yleisemmin julkishallinnon osalta voidaan suositella pyrkimystä sosiaalisen median tuottaman datan seurantajärjestelmän luomiseen. Tämän kaltaista työvälinettä voitaisiin hyödyntää trendien ja uutisaiheiden keräämiseen, haluttujen indikaattoreiden seurantaan, automaattisena havaitsemis- ja hälytysjärjestelmänä sekä tutkimuksellisessa käytössä.

Tutkimusaiheen ja menetelmien osalta kritiikkiä voidaan osoittaa valittuihin design science-arviointiperusteisiin, joiden soveltaminen olisi kenties ollut perustellumpaa mikäli tutkimuksen aikana olisi toteutettu jokin käytännön tekstianalytiikan työväline kohdeorganisaatioille. Koska tutkielmassa kyseessä ei ollut mikään tietty IT-artifakti vaan kokonainen ryhmä erilaisia ja hyvin vaihtelevia tekstianalytiikan menetelmiä oli arviointiperusteiden käyttäminen jossain määrin haastavaa. Tutkimusaihe oli sangen laaja-alainen ja tämän vuoksi kirjallisuuskatsauksen osalta on pyritty esittelemään aihepiiriä laajasti. Tämän osalta kritiikkiä voitaneen esittää sikäli, että tiukkarajaisempi keskittyminen tiettyyn tekstianalytiikan sovellutusalueeseen olisi voinut tuoda tutkimukseen lisää syvyyttä. Toisaalta tällöin myöskään nykyisen kaltaista kartoittavaa lähestymistä menetelmiin ei olisi voitu suorittaa. Tutkielman tutkimuskysymyksien kannalta olisi myös voinut olla harkittavissa vaihtoehtoisia tutkimusmenetelmiä. Koska julkishallinnon tekstianalytiikan käyttömahdollisuudet oli keskeinen asia, olisi voinut olla mahdollista toteuttaa laajempi kyselyhaastattelu useiden julkishallinnon eri organisaatioiden kanssa. Toisaalta tällöin syvällisempää

keskustelua teemahaastatteluiden puitteissa kohdeorganisaatioiden asiantuntijatahojen kanssa ei oltaisi voitu toteuttaa.

7 YHTEENVETO

Tutkielmassa on kirjallisuuskatsauksen osalta käsitelty laaja-alaisesti Twitter-datan tekstianalytiikan käyttöä. Tässä yhteydessä on havaittu Twitter-dataan perustuvaa tekstianalytiikkaa voitavan käyttää monilla eri sovellutusalueilla. Erityisen tehokkaiksi menetelmät ovat osoittautuneet epidemioiden seurannassa, sävyanalyysissä ja trendien sekä uutisaiheiden havaitsemisessa. Lupaavia tuloksia on saatu myös maanjäristyksien havaitsemisjärjestelmien toteutuksina joukkoistetun aistinnan avulla. Erilaisten seuranta- ja ennakointijärjestelmien osalta tekstianalytiikkaa on myös sovellettu onnistuneesti. Sen sijaan poliittisten tutkimuksien tekeminen vaikuttaa Twitter-dataan perustuvan tekstianalytiikan avulla olevan haastavaa. Tyypillisimpinä tekstianalytiikan menetelminä ja lähestymistapoina on käytetty ohjattua koneoppimista, erilaisia luokitteluita ja sävyanalyysiä. Tulokset ovat tekstianalytiikan soveltamisen osalta olleet vähintään lupaavia lukuunottamatta poliittista tutkimusta. Twitter-datan tekstianalytiikan haasteiksi on havaittu skaalautuvien ja multimedia-kontentin sekä käyttökontekstin paremmin huomioivien menetelmien kehittämisen. Lisäksi on todettu Twitter-datalle tyypillisten ominaisuuksien olevan haasteita tekstianalytiikan soveltamiselle. Tällaisia ominaisuuksia ovat tekstiosuuksien lyhyys ja kohinainen luonne sekä sarkasmin käyttö. Sen sijaan geolokaatiotiedon eksaktius on vähenevissä määrin ongelma. Kaikilla Twitter-datan tekstianalytiikan sovellusalueilla lukuun ottamatta poliittista tutkimusta todettiin Twitter-datan analyysin olevan vähintäänkin hyödyllistä ei-formaalina tiedonlähteenä virallisten tutkimuslähteiden rinnalla.

Tekstianalytiikan menetelmät tarjoavat julkishallinnon eri organisaatioille potentiaalisesti tehokkaita työvälineitä sosiaalisen median tekstimuotoisen datan osalta. Tältä osin tutkielmassa esitellyt menetelmät ovat sovellettavissa laajasti. Käyttökohteita ovat trendien ja uutisaiheiden havaitseminen, epidemioiden seuranta, kollektiivisen mielialan, onnellisuuden ja elintapariskien kartoittaminen, erilainen ennakoiva tutkimus (esimerkiksi rikollisuuden spatiaalis-temporaalisen ennakoinnin parantaminen),

joukkoistetun aistinnan hyödyntäminen (esimerkiksi sensorisen terveysdatan saamiseksi) ja seurantajärjestelmien kehittäminen.

Julkishallinnon omaan tekstimuotoiseen dataan Twitter-datan tekstianalytiikan menetelmien sovellettavuutta ei voida tämän tutkimuksen osalta arvioida. Tämä johtuu siitä, että empiirisessä osuudessa mukana olleilla kohdeorganisaatoilla ei ole hallussaan vastaavaa tekstimuotoista dataa eikä täten varsinaisesti tarvetta tekstianalytiikan käytölle. Toisin sanoen tältä osin tarkastelua kirjallisuuskatsauksessa olleiden menetelmien soveltamisesta julkishallinnon organisaatioiden tekstimuotoiseen dataan ei voitu tehdä. On kuitenkin oletettavaa, että menetelmiä kyettäisiin laaja-alaisesti soveltamaan myös julkishallinnon tekstidataan, mikäli tätä vain on. Tutkielman pohjalta käy kuitenkin selväksi, että Twitter-dataan pohjautuvien tekstianalytiikan tekniikoiden käyttö on julkishallinnon kannalta hyödyllisintä alkuperäisessä kontekstissaan. Tässä yhteydessä julkishallinnon tarpeita ajatellen havaittiin kuitenkin tarve käyttää lukuisia eri datavarantoja tiedonlähteinä, sillä Suomen osalta Twitterin vähäinen käyttö todettiin sekä tilastolliselta että demografisten ominaisuuksien kannalta ongelmalliseksi seikaksi. Tulevaisuudessa sosiaalisen median muut palvelut, keskustelufoorumit sekä hakukonepalveluiden tuottama tekstimuotoinen data tulisi sisällyttää julkishallinnon tekstianalytiikan kohteiksi.

Tutkielman tuloksien perusteella voidaan THL:n osalta suositella alustavaan tutkimukselliseen koekäyttöön simplifikoituja epidemioiden seurannan menetelmiä sekä yleisemmin julkishallinnon organisaatioille pyrkimystä sosiaalisen median seurantaan perustuvien järjestelmien kehittämiseen.

LÄHTEET

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S-H. & Liu, B. (2011). Predicting flu trends using Twitter data. *2011 IEEE Conference on Communications Workshops* (s. 702-707).

Aggarwal, C. C. & Abdelzaher, T. (2011). Integrating sensors and social networks. Teoksessa C. C. Aggarwal (toim.), *Social Network Data Analytics* (s. 379-412). Boston: Springer.

Aggarwal, C. C. & Wang, H. (2011). Text mining in social networks. Teoksessa C. C. Aggarwal (toim.), *Social Network Data Analytics* (s. 353-378). Boston: Springer.

Aggarwal, C. C. & Zhai, C. (2012). An introduction to text mining. Teoksessa C. C. Aggarwal (toim.), *Mining Text Data* (s. 1-10). Boston: Springer.

Aramaki, E., Maskawa, S. & Morita, M. (2011). Twitter catches the flu: Detecting influenza epidemics using Twitter. Teoksessa *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (s. 1568-1576).

Asur, S. & Huberman, B.A. (2010). Predicting the future with social media. Teoksessa *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Vol 1* (s. 492-499).

Barbier, G. & Liu, H. (2011). Data mining in social media. Teoksessa C. C. Aggarwal (toim.), *Social Network Data Analytics* (s. 327-352). Boston: Springer.

Barbosa, L. & Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. Teoksessa *Proceedings of the 23rd International Conference on Computational Linguistics* (36-44).

Bertrand, K.Z., Bialik, M., Virpee, K., Gros, A. & Bar-Yam, Y. (2013). *Sentiment in New York City: A high resolution spatial and temporal view.* arXiv:1308.5010.

Bollen, J., Mao, H. & Zeng, X-J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2 (2011),* 1-8.

Bollen, J., Pepe, A. & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Teoksessa *Proceedings of the Fifth Inter-national AAAI Conference on Weblogs and Social Media* (s. 450-453).

Castillo, C., Mendoza, M. & Poblete, B. (2011). Information credibility on Twitter. Teoksessa *Proceedings of the 20th International Conference on World Wide Web* (s. 675-684).

Cataldi, M., Di Caro, L. & Schifanella, C. (2010). Emerging topic detection on Twitter based on temporal and social terms evaluation. Teoksessa *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (Article No. 4).

Cheng, Z., Caverlee, J. & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. Teoksessa X. J. Huang, G. Jones, N. Koudas, X. Wu & K. Collins-Thompson (toim.), *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (s. 759-768).

Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A. & Menczer, F. (2011a). Predicting the political alignment of Twitter users. Teoksessa *IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing* (s. 192-199).

Conover, M.D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A. & Menczer, F. (2011b). Political polarization on Twitter. Teoksessa *Proceedings of the 5th International Conference on Weblogs and Social Media* (s. 89–96).

Culotta, A. (2010a). Towards detecting influenza epidemics by analyzing Twitter messages. Teoksessa *Proceedings of the First Workshop on Social Media Analytics* (s. 115-122).

Culotta, A. (2010b). *Detecting influenza outbreaks by analyzing Twitter messages*. arXiv:1007.4748.

Demirbas, M., Bayir, M.A., Akcora, C.G., Yilmaz, Y.S. & Ferhatosmanoglu, H. (2010). Crowd-sourced sensing and collaboration using Twitter. Teoksessa *Proceedings of the 2010 IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks* (s. 1-9).

EMOTIVE-projekti. (2013a). Extracting the Meaning of Terse Information in a Geo-Visualisation of Emotion. Haettu 29.10.2013 osoitteesta http://emotive.lboro.ac.uk/

EMOTIVE-projekti. (2013b). Resources: Sykora M., Jackson, T.W., O'Brien, A. & Elayan, S. - Presentation slides. Haettu 29.10.2013 osoitteesta http://emotive.lboro.ac.uk/?page_id=24

Earle, P.S., Bowden, D.C. & Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics* 54(6), 708–715.

Gayo-Avello, D. (2012). *I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper – A Balanced Survey on Election Prediction using Twitter Data.* arXiv:1204.6441v1.

Go, A., Bhayani, R. & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford Digital Library Technologies Project.

González-Ibáñez, R., Muresan, S. & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. Teoksessa *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2* (s. 581-586).

Hart, C. (1998). *Doing a Literature Review: Releasing the Social Science Research Imagination*. London: Sage.

Hart, C. (2001). *Doing a Literature Search: A Guide for the Social Sciences*. London: Sage.

Hawn, C. (2009). Take two Aspirin and tweet me in the morning: How Twitter, Facebook and other social media are reshaping health care. *Health Affairs* 28(2) (s. 361–368).

Hecht, B., Hong, L., Suh, B. & Chi, E.D. (2011). Tweets from Justin Bieber's Heart: The dynamics of the "location" field in user profiles. Teoksessa *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (s. 237-246).

Hevner, A.R., March, S.T., Park, J. & Ram, S. (2004). Design science in information systems research. *MIS Quarterly Volume 28 Issue 1*, 75-105.

Hirsjärvi, S. & Hurme, H. (2000). *Tutkimushaastattelu: Teemahaastattelun teoria ja käytäntö*. Helsinki: Yliopistopaino.

Hirsjärvi, S.,Remes, P. & Sajavaara, P. (2010). *Tutki ja kirjoita. (16. uud. painos)*. Helsinki: Tammi.

Hu, X. & Liu, H. (2012). Text analytics in social media. Teoksessa C. C. Aggarwal (toim.), *Mining Text Data* (s. 385–414). Boston: Springer.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. Teoksessa *Proceedings of the 6th International ISCRAM Conference*, Gothenburg, Sweden.

Jansen, B. J., Zhang, M., Sobel, K. & Chowdury, A. (2009). Twitter Power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology archive Volume 60 Issue 11*, 2169-2188.

Java, A., Song, X., Finin, T. & Tseng, B. (2007). Why we Twitter: Understanding microblogging usage and communities. Teoksessa *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis* (s. 56-65).

Ji, X., Chun, S.A. & Geller, J. (2012). Epidemic Outbreak and Spread Detection System Based on Twitter Data. Teoksessa J. He, X. Liu, E. Krupinski & G. Xu (toim.), *Health Information Science : First International Conference, Beijing, China, April 8-10* (s. 152-163).

Jungherr, A., Jürgens, P. & Schoen, H. (2012). Why the Pirate Party won the German election of 2009 or the trouble with predictions: A response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. "Predicting elections with Twitter: what 140 characters reveal about political sentiment". *Social Science Computer Review Volume 30 Issue 2*, 229-234.

Järvinen, P., Järvinen, A. (2011). *Tutkimustyön metodeista*. Tampere: Opinpajan kirja.

Kaufmann, J. & Kalita, J. (2011). Syntactic normalization of Twitter messages. Teoksessa *International Conference on Natural Language Processing (ICON 2011)* (s, 149–158).

Lampos, V. & Cristianini, N. (2010). Tracking the flu pandemic by monitoring the social web. Teoksessa *IAPR 2nd Workshop on Cognitive Information Processing* (411-416).

Lee, R. & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. Teoksessa *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (s. 1-10).

Lovejoy, K., Waters, R. & Saxton, G.D. (2011). Engaging stakeholders through Twitter: How nonprofit organizations are getting more out of 140 characters or less. *Public Relations Review 38(2)*, 313-318.

Mathioudakis, M. & Koudas, N. (2010). TwitterMonitor: Trend detection over the Twitter stream. Teoksessa *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (s. 1155-1158).

Metaxas, P.T., Mustafaraj, E. & Gayo-Avello, D. (2011). How (not) to predict elections. Teoksessa *Proceedings in IEEE International Conference on Privacy, Secu-*

rity, Risk and Trust and IEEE International Conference on Social Computing (s. 165-171).

Mills, A., Chen, R., Lee, J. & Rao, H.R. (2009). Web 2.0 emergency applications: How useful can Twitter be for emergency response? *Journal of Information Privacy & Security vol. 5*, 3-26.

Mislove, A., Lehmann, S., Ahn, Y-Y., Onnela, J-P. & Rosenquist, J.N. (2011). Understanding the demographics of Twitter users. Teoksessa *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (s. 554-558).

Mitchell, L., Harris, K. D., Frank, M. R., Dodds, P. S. & Danforth, C. M. (2013). *The Geography of Happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place.* arXiv:1302.3299.

Moturu, S. (2009). *Quantifying the Trustworthiness of User-Generated Social Media Content.* PhD thesis, Arizona State University.

Naaman, M., Becker, H. & Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology, Vol. 62 Issue 5*, 902-918.

Nagarajan, M., Gomadam, K., Sheth, A.P., Ranabahu, A., Mutharaju, R. & Jadhav, A. (2009). Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. *InWeb Information Systems Engineering (WISE 2009) Lecture Notes in Computer Science, 5802*, 539–553.

Paul, M. & Dredze, M. (2011a). *A model for mining public health topics from twitter.* Technical report, Johns Hopkins University.

Paul, M.J. & Dredze, M. (2011b). You Are What You Tweet: Analyzing Twitter for Public Health. Teoksessa *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media* (265-272).

Peffers, K., Tuunanen, T., Rothenberger, M.A. & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems 24 (3)*, 45-77.

Petrovic, S., Osborne, M. & Lavrenko, V. (2011). RT to win! Predicting message propagation in Twitter. Teoksessa *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (s. 586-589).

Phuvipadawat, S. & Murata, T. (2010). Breaking news detection and tracking in Twitter. Teoksessa *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Vol. 03* (s. 120-123).

Popescu, A-M. & Pennacchiotti, M. (2010). Detecting controversial events from Twitter. Teoksessa *Proceedings of the 19th ACM international conference on Information and knowledge management* (s. 1873-1876).

Quincey, E. & Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: The potential of twitter. Teoksessa *Electronic Healthcare* (s. 21-24). Springer: Berlin.

Sakaki, T., Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Teoksessa *Proceedings of the 19th international conference on World wide web* (s. 851-860).

Skoric, M., Poor, N., Achananuparp, P., Lim, E-P. & Jiang, J. (2012). Tweets and Votes: A Study of the 2011 Singapore General Election. Teoksessa *Proceedings of the 45th Hawaii International Conference on System Sciences* (s. 2583 - 2591).

Steele, R. (2011). Social Media, Mobile Devices and Sensors: Categorizing New Techniques for Health Communication. Teoksessa *Proceedings of the 5th International Conference on Sensing Technology* (s. 187 - 192).

Terpstra, T., de Vries, A., Stronkman, R. & Paradies, G.L. (2012). Towards a real-time Twitter analysis during crises for operational crisis management. Teoksessa L. Rothkrantz, J. Ristvej, Z. Franco (toim.) *Proceedings of the 9th International ISCRAM Conference*, Vancouver, Canada.

Terveyden ja hyvinvoinnin laitos – Organisaatio. Thl.fin www-sivusto.
Haettu 20.10.2013 osoitteesta
http://www.thl.fi/fi_FI/web/fi/organisaatio

Tjong, E., Sang, K. & Bos, J. (2012). Predicting the 2011 Dutch senate election results with Twitter. Teoksessa *Proceedings of the Workshop on Semantic Analysis in Social Media Pages* (s. 53-60).

Tumasjan, A., Sprenger, T. O., Sandner, P. G. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Teoksessa *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (s. 178-185).

Twittercensus. (19.2.2013). Twittercensus - Counting every Finnish speaking tweep. Presentation Finnish Twitter. Haettu 29.10.2013 osoitteesta
http://www.finnishtwitter.com/2013/02/19/presentation-finnish-twitter/

Wang, X., Brown, D.E. & Gerber, M.S. (2012). Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. Teoksessa D. Zeng, L. Zhou, B. Cukic, G. A. Wang, & C. C. Yang

(toim.), *IEEE International Conference on Intelligence and Security Informatics: Cyberspace, Border, and Immigration Securities* (s. 36-41).

Webster, J. & Watson, R.T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly 26 (2),* pxiii-xxiii.

Wilson, T., Wiebe, J. & Hoffman, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Teoksessa *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (s. 347-354).

Zhang, X., Fuehres, H. & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as i fear". *Procedia Social and Behavioral Sciences 26*, 55–62.

Zhao, X., Jiang, J., Weng, J., He, J. & Lim, E.P. (2011). Comparing Twitter and traditional media using topic models. Teoksessa P. Clough, C. Foley, C. Gurrin, H. Lee & G. J. F. Jones (toim.), *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR)* (s. 338–349).

LIITE 1 HAASTATTELUKYSYMYKSET

1. Kuvatkaa tekstianalytiikan käyttöä, käyttötarkoituksia ja tekstimuotoisen datan määrää organisaatiossa.

2. Esityksessä on käyty läpi erilaisia tapoja hyödyntää Twitter-datan tekstianalytiikkaa. Vaikuttavatko esitellyt menetelmät mielekkäiltä ja toimivilta esitellyissä asiayhteyksissä?

3. Millainen mielikuva Twitter-datan tekstianalytiikan keinoista jää - voidaanko saatua dataa käyttää esimerkiksi päätöksenteon tukena?

4. Esityksessä on käyty läpi erilaisia tapoja hyödyntää Twitter-datan tekstianalytiikkaa. Näettekö, että jotain näistä työkaluista, menetelmistä, malleista tai lähestymistavoista olisi mahdollista käyttää joko suoraan tai muokattuna organisaationne tekstimuotoisen datan tekstianalytiikassa?

5. Mitä mieltä olette esiteltyjen mallien ja toimintatapojen rakenteesta?

6. Ovatko esitellyt Twitter-datan tekstianalytiikan työkalut kompleksisia?

7. Onko edellä käsiteltyjen menetelmien hyödyntäminen kannattavaa ylipäänsä organisaatiossanne?

8. Näettekö esiteltyjen menetelmien / työkalujen / lähestymistapojen soveltuvan yleisellä tasolla hyvin julkishallinnon organisaatioiden IT-arkkitehtuuriin?

9. Millaisia ajatuksia julkishallinto ja Twitter-datan seuranta herättää?

10. Minkälaisissa tilanteissa ja skenaarioissa näkisitte esiteltyjen kaltaisten menetelmien olevan hyödynnettävissä julkishallinnossa?

11. Oletetaan, että tekstianalytiikkaa sovellettaisiin aktiivisesti julkishallinnon taholta
sosiaalisen median tuottaman datan seurantaan sekä olemassa olevaan

julkishallinnon dataan. Näettekö, että menetelmät tuottaisivat lisäarvoa vanhojen tutkimus- ja seurantamenetelmien rinnalle?

12. Tiedon määrän kasvaessa valtavasti varsinkin sosiaalisen median osalta, millaisena näette tekstianalytiikan merkityksen tulevaisuudessa julkishallinnon kannalta?