Jaakko Palvaila

VAALIENNUSTEIDEN LAADINTA PERUSTUEN TWITTER-PALVELUSTA LOUHITTUUN DATAAN

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2017

TIIVISTELMÄ

Palvaila, Jaakko
Vaaliennusteiden laadinta perustuen Twitter-palvelusta louhittuun dataan
Jyväskylä: Jyväskylän yliopisto, 2017, 35 s.
Tietojärjestelmätiede, kandidaatin tutkielma
Ohjaaja: Taipalus, Toni

Tässä tutkielmassa käsitellään Twitter-palvelusta louhitun datan hyödyntämistä vaaliennusteiden laadinnan työkaluna. Tutkielmassa esitellään ja analysoidaan tutkimuksia, joissa on louhittu ja pyritty analysoimaan Twitter-palvelun käyttäjien palveluun lähettämiä viestejä.

Tutkielman tarkoituksena on kertoa vastaus kysymykseen, voidaanko Twitteristä kerätyn datan pohjalta laatia vaaliennusteita. Aiempia tutkimuksia läpikäymällä esitellään twiittien louhinnan, seulonnan, analyysin ja ennusteiden laadinnan menetelmiä parhaan mahdollisen metodin löytämiseksi.

Tässä tutkielmassa vertaillaan louhitun datan pohjalta tehtyjen ennusteiden tuloksia ja luotettavuutta perinteisiin, kyselytutkimuksina toteutettuihin vaalitulosennusteisiin. Lisäksi pyritään määrittämään, millä menetelmillä Twitter-datan perusteella laadittujen ennusteiden tarkkuutta voidaan parantaa, ja voidaanko Suomessakin hyödyntää Twitter-käyttäjien palveluun lähettämiä viestejä vaaliennusteiden laadinnassa.

Asiasanat: Twitter, datan louhinta, tiedonlouhinta, vaaliennusteet, tunneanalyysi, sosiaalinen media

ABSTRACT

Palvaila, Jaakko
Creating election forecasts based on Twitter data mining
Jyväskylä: University of Jyväskylä, 2017, 35 p.
Information Systems, Bachelor's Thesis
Supervisor(s): Taipalus, Toni

This thesis handles using data mined from Twitter as a tool in creating election forecasts. In this thesis I shall present and analyze prior studies in which Twitter users' messages have been mined and analyzed.

The aim of this thesis is to answer the question whether data collected from Twitter can be used as a basis in the process of creating election forecasts. I will present the best models for tweet mining, filtering, analysis and forecast creation by evaluating prior studies to find the best possible methods.

In this thesis I will compare the results of election forecasts based on data mining to those achieved using more traditional polling methods. In addition to this I will attempt to define how the accuracy of forecasts based on Twitter data can be improved and whether it is possible to utilize Twitter users' messages in election forecast creation in Finland as well.

Keywords: Twitter, data mining, election polls, sentiment analysis, social media

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Aktiivisena sosiaalisen median käyttäjänä ja poliittisesti aktiivisena ihmisenä olen seurannut mielenkiinnolla eri sosiaalisten medioiden käyttöä poliittisen keskustelun välineenä. Siinä missä Facebook on poliitikoille pitkien kannanottojen sekä vaalimainonnan alusta ja Instagram julkisuuskuvan kiillottamisen väline, tarjoaa Twitter matalahierarkisen alustan poliittiseen debattiin osallistumiselle. Yksittäisten Twitter-viestin enimmäismitta on rajoitettu 140 merkkiin, joten ajatukset on onnistuttava kiteyttämään alkuperäistä viestiä olennaisesti muuttamatta. Twitterissä jokainen käyttäjä voi myös halutessaan osallistua mihin tahansa keskusteluun ja saada mielipiteensä sekä argumenttinsa kuuluviin suurelle yleisölle. Twitter on sosiaalisten medioiden joukossa joukkoviestin, jolla on omat erityispiirteensä. Jokaisella käyttäjällä on ääni, joka kuuluu kaikille.

1.1 Aihepiirin esittely

Tässä tutkielmassa käyn läpi olemassa olevia tutkimuksia koskien Twitteristä louhitun datan käyttöä poliittisten ennusteiden laadinnan työkaluna. Aiemmissa tutkimuksissa on pyritty selvittämään, voidaanko henkilövaalien lopputuloksia ennustaa Twitter-käyttäjien palveluun lähettämien viestien perusteella. Aihe on tutkimusaiheena kohtuullisen tuore: ensimmäisen kerran vaalien lopputulosta ennustettiin Twitter-viestien pohjalta vuonna 2009 Saksan liittopäivävaalien yhteydessä (Tumasjan, Sprenger, Sandner ja Welpe, 2010). Tutkimuksessa käytettiin Twitteristä kerätyn datan analysoimiseen LIWC2007-nimistä ohjelmistoa, jolla analysoitiin käyttäjien Twitter-palveluun lähettämien lyhyiden, maksimissaan 140 merkkiä sisältävien viestien emotionaalisia, kognitiivisia ja rakenteellisia ominaisuuksia. Näistä viesteistä käytetään tässä tutkielmassa jatkossa nimitystä twiitti. Sittemmin saman tyyppisiä metodeja on käytetty esimerkiksi Yhdysvaltain vuoden 2012 presidentinvaalien (Wang, Can, Kazemzadeh, Bar ja Narayanan, 2012) ja Espanjan vuoden 2015 parlamenttivaalien (Vilares, Thelwall ja

Alonso, 2015) tulosten ennustamiseen. Datan analysoimiseen käytettyjä metodeja on kehitetty eteenpäin ja tässä tutkielmassa vertailen erilaisia datan analysointityökaluja sekä niiden tuottamien tulosten tarkkuutta.

Tumasjan ym. (2010) totesivat Yhdysvaltain vuoden 2008 presidentinvaalien olleen ensimmäiset sosiaalisen median vaalit. Barack Obama käytti vaaleissa Facebookia, Twitteriä ja muita alustoja taitavasti hyödykseen aktivoimalla kannattajiaan lahjoittamaan rahaa kampanjalleen ja ylittämällä kaikki aiemmat vaalirahoitusennätykset joukkoistetulla rahoituksellaan (Williams ja Gulati, 2008). Kun Yhdysvalloissa järjestettiin presidentinvaalit 8 vuotta myöhemmin, kärkiehdokkaiden välistä vaalikamppailua käytiin puheiden ja televisiodebattien lisäksi myös Twitterissä. Varsinkin vaalit voittanut Donald J. Trump herätti niin perinteisessä kuin sosiaalisessakin mediassa vilkasta keskustelua Twitter-käyttäytymisellään, esimerkiksi toistuvilla yrityksillään kiistää aiempia lausuntojaan (ja twiittejään). Ehdokkaiden Twitter-tileillä myös hyökättiin toistuvasti vastapuolen lausuntoja ja ajatuksia vastaan.

Yhdysvaltain vuoden 2016 presidentinvaalien tiimoilta käydyissä Twitter-keskusteluissa samaa sanomaa useilta eri tileiltä tietokoneohjelmien avustuksella jakelevien twiittaajien määrä keskusteluissa oli merkittävä, mutta kuinka tunnistaa keskusteluista aidot ihmiset? Viestien louhimisen jälkeen aiemmissa tutkimuksissa on ollut merkittävä haaste seuloa oikeat viestit tietokoneohjelmien kirjoittamien viestien, eräänlaisen taustakohinan, seasta. Tietokoneohjelmia viestien kirjoittamiseen hyödyntävistä Twitter-käyttäjistä käytetään jatkossa tässä tutkielmassa nimitystä botti ja bottien kirjoittamien viestien aiheuttamasta taustakohinasta nimitystä spämmi.

Tulosennusteiden laatiminen vaaleihin on haastavaa. Yleensä tulosennusteita laaditaan kyselytutkimuksina, joissa pyritään haastattelemaan mahdollisimman useaa ihmistä mahdollisimman useasta väestöryhmästä. Tavoitteena on luoda kuva äänestäjäkunnasta mikrokoossa: kaupunkien ja maaseudun nuoria, työssäkäyviä ja työttömiä, perheellisiä ja yksinasuvia sekä kaupunkien eläkeläisiä. Jokaisen äänestäjäryhmän edustajia haastatellaan ja heiltä kysytään, ketä ja millä todennäköisyydellä he äänestäisivät tulevissa vaaleissa. Nämä ryhmien edustajien haastattelutulokset ekstrapoloidaan koskemaan koko ryhmää ja sen jälkeen ryhmien äänestysennustetta vielä korjaillaan esimerkiksi kaupungissa asuvien ja Kokoomusta kannattavien eläkeläisnaisten perinteisesti korkean äänestysaktiivisuuden perusteella. Näin saadaan laadittua koko äänestäjäkunnan äänestyskäyttäytymistä arvioiva vaalien tulosennuste, josta käytetään tässä tutkimuksessa myös nimitystä gallup.

Perinteiset gallupit ovat kuitenkin usein erehtyneet. Suomessa tuorein esimerkki on vuoden 2011 eduskuntavaalit, jolloin perussuomalaiset keräsivät vaaleissa yllättävän suuren 19,1 prosentin osuuden kaikista annetuista äänistä. Viimeinen gallup ennen vaaleja eli Yleisradion Taloustutkimukselta tilaama gallup merkittävän laajalla 3621 henkilön otoksella ennusti perussuomalaisille kuitenkin vain 15,4 prosenttiyksikön kannatusta 1,4 prosenttiyksikön virhemarginaalilla. Ennustetta laadittaessa oli mahdollisesti käytetty korjauskertoimia väärään

suuntaan perussuomalaisten kannattajien aiemman äänestyskäyttäytymisen perusteella. Samassa ennusteessa myös silloiselle pääministeripuolue Keskustalle ennustettiin 18,6 prosentin äänisaalista, mutta puolue jäi vain 15,8 prosenttiin. Molemmissa tapauksissa ennusteen ja lopullisen vaalituloksen välinen ero oli yli kaksinkertainen verrattuna Taloustutkimuksen omaan virhemarginaaliarvioon. (Yle 14.4.2011)

Twitteristä kerätyn datan pohjalta tällaista väestöryhmien ja aiemman äänestyskäyttäytymisen perusteella laadittua korjauskerrointa ei voida käyttää, sillä käyttäjien ikää, sukupuolta, työtilannetta, perhestatusta, puoluekantaa, aiempaa äänestyskäyttäytymistä tai muitakaan taustamuuttujia ei ole tiedossa ennustetta laadittaessa. Toisaalta Twitter-käyttäjien viestien pohjalta laadituissa ennusteissa otoskoko on merkittävästi laajempi kuin perinteisissä kyselytutkimuksissa. Yhteiskunnan kaikki luokat eivät kuitenkaan ole analyysin pohja-aineistossa tasaisesti edustettuina, sillä nuoremmat ikäluokat käyttävät sosiaalista mediaa vanhempia ikäluokkia todennäköisemmin eikä esimerkiksi kaupunkien eläkeläisten ääntä todennäköisesti kuulu Twitter-viesteissä.

Koska kyselytutkimusten toteuttaminen vaatii merkittävän määrän haastattelijoita ja vie aikaa, on reaaliaikaisten ennusteiden laatiminen perinteisin menetelmin mahdotonta tai ainakin hyvin kallista. Myös tämän vuoksi on lähdetty hakemaan vaihtoehtoisia tapoja tuottaa vaaliennusteita. Twitter on reaaliaikaisen luonteensa ja suuren käyttäjämääränsä vuoksi yksi mahdollinen datan lähde näiden ennusteiden luomiselle.

1.2 Tutkielman tavoite

Tämä tutkielma pyrkii vastaamaan kysymykseen ”Kuinka Twitteristä kerättyä dataa voidaan käyttää poliittisen ennusteen laatimisen työkaluna?”. Hypoteesina on, että Twitteristä kerätty data soveltuu vaaliennusteen laatimisen pohjaksi, kun otoskoko eli viestiaineiston määrä on riittävän suuri.

Seuraavissa luvuissa tulen selvittämään, kuinka Twitteristä voidaan kerätä dataa tutkimusaineiston pohjaksi ja kuinka aineistoa voidaan seuloa häiriöäänten vaimentamiseksi. Käyn läpi erilaisia aineiston analysointimenetelmiä ja arvioin, voidaanko analysoitua aineistoa käyttää tarkkojen poliittisten ennusteiden laadinnan pohjana. Lisäksi pyrin selvittämään, voidaanko Twitteristä kerättyä dataa käyttää reaaliaikaisten poliittisten ennusteiden laatimiseen.

Käytännössä tutkielmalla halutaan siis selvittää, kuinka Twitter-käyttäjien viestit saadaan louhimalla, seulomalla ja analysoimalla muutettua poliittisiksi ennusteiksi. Lisäksi tavoitteena on vertailla erilaisia analysoinnin metodeja ja niiden pohjalta laadittujen ennusteiden tarkkuutta ja löytää luotettavin aineiston analysoinnin metodi tulevia tutkimuksia ja ennusteita ajatellen.

1.3 Tutkielman kokoamisessa käytetyt metodit

Tutkielma on toteutettu kirjallisuuskatsauksena, jossa on vertailtu aiheesta tehtyjä tutkimuksia ja arvioitu niiden tuloksia sekä tulosten luotettavuutta. Tutkielmassa käsiteltävä kirjallisuus on kerätty Jyväskylän yliopiston Finna-järjestelmää ja Google Scholar-hakukonetta hyödyntäen.

2 TWITTERIN KÄYTTÖ DATAN LÄHTEENÄ

Sosiaalisen median käyttäjät luovat päivittäin valtavan määrän dataa, jota voidaan hyödyntää myös tutkimusten data-aineistojen keräämisessä. Tässä luvussa esitellään sosiaalisesta mediasta kerätyn datan louhimista ja seulontaa sekä aiempien tutkimusten sovellutuksia kerätylle datalle.

2.1 Sosiaalisen median big data

Käsitettä ”Big Data” käytettiin ensimmäisen kerran vuonna 1998, kun Silicon Graphicsin John Mashey käytti käsitettä esityskalvoissaan. Silicon Graphics oli käyttänyt termiä jo aiemmin mainonnassaan, ensimmäisen kerran vuonna 1996. Diebold (2012) arvelee ajatuksen big datasta virinneen yrityksen sisällä 1990-luvun puolessavälissä. Ensimmäisen kerran käsitettä käyttivät akateemisessa tekstissä vuonna 1998 Weiss ja Indurkhya, jotka käsittelivät kirjassaan kasvavan datamäärän varastointia suuriin tietovarantoihin ja sen analysointia entistä kokonaisvaltaisemmin. (Diebold, 2012.)

Sosiaalisen median big data koostuu käyttäjien päivityksistä, käyttäjien julkaisemista kuvista ja käyttäjien välisistä viesteistä. Sosiaalisen median datan tärkeimmät lähteet ovat suhteellisen uusia (Facebook perustettiin vuonna 2004, Twitter 2006) ja kaikki palveluissa oleva tieto on syntynyt kyseisten ajankohtien jälkeen. Sosiaalisen median synty ja älylaitteiden kehittyminen on luonut uuden tiedon aikakauden ja jokaisesta sosiaalisen median käyttäjästä ja älylaitteen omistajasta on samalla tullut käveleviä tiedon generoijia. Big dataan sisältyy paljon tietoa sosiaalisen median käyttäjistä, joka pitää vain kaivaa esiin taustakohinan alta. (McAfee, Brunjolfsson, Davenport, Patil ja Barton, 2012.)

Tutkielman kohteena olevissa tutkimuksissa käsiteltävä Twitteristä kerätty data ei täytä jokaista kohtaa big datan yleisimmistä määritelmistä, esimerkiksi IBM:n neljän V:n määritelmästä big datalle (IBM, 2017). Twitterin käyttäjät luovat palvelun kautta sosiaalisen median big dataa sen kaikissa ulottuvuuksissa (volume, variety, velocity, veracity), mutta aiheena olevissa tutkimuksissa kerätty Twitter-data ei big datan kontekstissa ole tekstimuotoisena monimuotoista tai volyymiltaan suurta, vaikka joissain tutkimuksissa käsitelläänkin jopa 50 miljoonaa yksittäistä twiittiä sisältävää tutkimusaineistoa. Toisaalta Twitteristä kerätty data on nopeaa ja epävarmaa, joten osa big datan kriteereistä täyttyy myös tutkimuksissa. Käytännössä tutkimuksissa onkin kyse valikoidusta datasta, joka kerätään aihetunnisteiden avulla reaaliajassa Twitter-palvelun oman streaming-rajapinnan kautta.

Aiemmin sosiaalisesta mediasta kerätyn datan avulla on esimerkiksi kerätty joukkoistettua tietoa onnettomuuksista (Xu, Liu, Yen, Mei, Luo, Wei ja Hu, 2016) ja arvioitu sademääriä Twitter-viestien perusteella (Lampos ja Cristianini,

2012). Sosiaalisen median käyttäminen datan lähteenä ei kuitenkaan rajoitu pelkästään menneiden tapahtumien tulkitsemiseen tai meneillään olevien tapahtumien seuraamiseen: sosiaalisesta mediasta kerättyä dataa voidaan käyttää myös tulevien tapahtumien ennustamiseen. Tutkimuksissa on ennustettu esimerkiksi tautien leviämistä, vaalituloksia, makrotaloutta, elokuvien kaupallista menestymistä, tuotteiden myyntilukuja ja osakemarkkinoiden toimintaa (Kalampokis, Tambouris ja Tarabanis, 2013).

2.2 Lähdeaineiston kerääminen Twitteristä

Tässä tutkielmassa käsiteltävissä tutkimuksissa ei perehdytä Twitterin käyttäjistään keräämään dataan, vaan käytetään käyttäjien Twitter-palveluun lähettämiä viestejä eli twiittejä tutkimusaineistona. McAfee ym. (2012) esittivät, että sosiaalisen median big datan varsinainen tietosisältö tulee kaivaa taustakohinan seasta esiin ja sama pätee myös tämän tutkielman kohteena oleviin tutkimuksiin, vaikka tutkimuksissa ei varsinaisesti käsitelläkään big dataa. Enemmän dataa ei myöskään aina tarkoita, että data olisi parempaa – datan laatu riippuu vahvasti kohinan määrästä ja siitä, kuvastaako kerätty aineisto sitä joukkoa jota sen uskotaan tutkimuksessa edustavan (Fan ja Bifet, 2013). Varsinkin Twitteristä kerätyn datan yleistettävyydestä koskemaan isompia väestömassoja voidaan olla perustellusti monta eri mieltä, kuten tässä tutkielmassa myöhemmin todetaan.

Tutkimuksissa Twitteristä on kerätty ja seulottu dataa usein eri metodein. Tumasjan ym. (2010) keräsivät tutkimukseensa Saksan liittopäivävaaleista aineistoksi 104 003 twiittiä, jotka oltiin julkaistu vaaleja edeltävän viikon aikana. Aineistoon kerättiin kaikki twiitit, joissa oli mainittu liittopäivillä edustettuna oleva puolue tai näiden puolueiden poliitikko. Twiitit kerättiin käyttämällä Twitterin omaa Search-ohjelmistorajapintaa haun mahdollistajana. Tumasjan ym. eivät käyttäneet keräämänsä datan seulomiseen työkaluja, joilla mahdolliset bottitilien lähettämät twiitit oltaisiin seulottu datan joukosta pois.

Siinä missä Tumasjan ym. (2010) hakivat kaikki tutkimusaineistoon sopivat twiitit avainsanoilla yhden viikon ajalta, Gayo-Avello (2011) lisäsi ohjelmistorajapinnan avulla toteutettuun hakuunsa vielä sijaintipohjaisen suodatuksen rajoittaakseen keräämänsä twiitit ainoastaan yhdysvaltalaisten käyttäjien palveluun lähettämiin pyrkiessään analysoimaan Yhdysvaltain vuoden 2008 vaalien alla käytyä Twitter-keskustelua. Gayo-Avello keräsi lopulta tutkimusaineistoonsa 250 000 twiittiä. Myöskään Gayo-Avello ei seulonut aineistostaan bottitilien lähettämiä viestejä.

2.3 Bottien seulominen kerätystä aineistosta

Twiittien seulomiseen on kehitetty useita menetelmiä, joilla pyritään tunnistamaan palveluun viestejä lähettävien käyttäjien joukosta bottikäyttäjiä. Chu,

Gianvecchio, Wang ja Jajodia (2010) tunnistivat ihmisten ja bottien Twitter-käytöstä eroavaisuuksia: ihmisten palveluun lähettämien viestien ajankohdilla on suurempi hajonta, bottien viestit taas vaikuttavat ajastetuilta säännöllisen julkaisutahdin vuoksi ja viesteistä suuressa osassa sisältö voidaan tunnistaa roskapostiksi. Lisäksi bottien palveluun lähettämissä twiiteissä esiintyy useammin linkkejä ulkoisille sivustoille. Twiitit on usein lähetetty Twitterin REST-rajapinnan kautta rekisteröimättömän kolmannen osapuolen työkalun avulla, ihmisten suosiessa Twitterin websivustoa ja mobiilisovelluksia.

Dickerson, Kagan ja Subrahmanian (2014) käyttivät bottien tunnistamiseen viesteissä käytetyn luonnollisen kielen analysointia siihen sisältyvien tunteiden tunnistamiseksi, sillä bottien tunnistaminen kerätystä aineistosta esimerkiksi Chu ym. (2010) käyttämin metodein on vähemmän tehokasta otoskoon ollessa pienempi tai kerättäessä twiittejä ohjelmistorajapinnan avulla vain tietystä aiheesta avainsanojen avulla. Viesteissä käytetyn luonnollisen kielen analysoinnista tunteiden tunnistamiseksi käytetään tässä tutkielmassa jatkossa nimitystä tunneanalyysi. Dickerson ym. (2014) arvioivat tutkimuksessaan käyttäjiä twiittien syntaksin, semanttisen (sanojen merkityksen) analyysin, käyttäjän tilastollisen käyttäytymisen ja verkostoitumisen näkökulmista. Tutkimuksessa todettiin varsinkin semanttinen analyysi ja tunneanalyysi tehokkaaksi keinoksi tunnistaa bottikäyttäjät twiittaajien joukosta.

Chu ym. (2010) arvioivat Twitterin käyttäjäkunnasta 48,7% olevan ihmisiä, 37,5% kyborgeja (ihminen, joka käyttää bottia apunaan tai botti, jota avustaa ihminen) ja loppujen 13,8 prosentin täysin tietokoneohjelman ohjaamia botteja. Voidaan siis todeta, että merkittävä osa twiiteistä on muiden kuin ihmiskäyttäjien kirjoittamia ja bottien seulomisen pois lähdeaineistoista olevan tarpeellista mahdollisimman luotettavan poliittisen analyysin ja mahdollisen ennusteen laadinnan mahdollistamiseksi.

Käytännössä kaikissa aiemmissa tutkimuksissa tutkimusaineisto on kerätty Twitterin oman ohjelmistorajapinnan avulla hakemalla aiheeseen liittyviä twiittejä avainsanahaulla. Bottien aiheuttamaa taustakohinaa ei ole kuitenkaan seulottu pois aineistosta. Käsittelemällä lähdeaineiston Chu ym. (2010) tai Dickerson ym. (2014) metodien avulla voitaisiin tulevissa tutkimuksissa saavuttaa tarkempia lopputuloksia.

3 TWITTERISTÄ KERÄTYN DATAN ANALYSOINTI

Tutkimuksissa, joissa Twitteristä kerätyn datan avulla on laadittu ennusteita, on analysoitu kerättyä dataa erilaisin menetelmin tekstimuotoisten viestien merkitysten tunnistamiseksi. Näiden merkitysten tunnistamiseksi tutkimuksissa on toteutettu lauserakenteita ja sanojen merkitystä arvioivaa semanttista analyysia sekä käytettyjen sanojen tunneanalyysia. Tässä luvussa esitellään tutkimuksissa käytettyjä tekstien analyysimenetelmiä.

3.1 Sanamäärien laskentaan perustuva analyysi

Tumasjan ym. (2010) käyttivät tutkimuksessaan apunaan LIWC2007 (Linguistic Inquiry and Word Count) -ohjelmistoa (Pennebaker, Boyd, Jordan ja Blackburn, 2015), joka käsittelee tekstin sisältämiä tunnetiloja sekä kognitiivisia ja rakenteellisia komponentteja eli on semanttisen arvioinnin työkalu. Ohjelmisto käyttää tähän analyysiin omaa sanatietokantaansa. Ohjelmiston avulla tutkimuksessa arvioitiin, kuinka usein twiiteissä esiintyi positiivisia ja negatiivisia tunteita niissä mainittuja poliittisia puolueita ja puolueiden kärkiehdokkaita kohtaan. Ohjelmistoa on käytetty laajasti psykologian tutkimuksissa (Tausczik ja Pennebaker, 2010) arvioitaessa kohdehenkilöiden käyttämää kieltä.

Tumasjan ym. (2010) käyttivät LIWC2007-ohjelmistoa tunnistamaan 12 eri tyypin sanojen käyttöä viesteistä: tulevaisuuteen ja menneisyyteen liittyvät sanat, positiivisiin ja negatiivisiin tunteisiin liittyvät sanat, surullisuus, ahdistus, viha, epäröivyys, varmuus, työ, saavutukset ja raha. Kaikki lähdeaineisto yhdistettiin yhdeksi tekstiksi, joka käännettiin ohjelmiston oman sanaston avulla englanniksi ja analysoitiin ohjelmistolla.

Tulosten perusteella laadittiin kunkin puolueen kärkipoliitikoista profiili, jolla arvioitiin kunkin kategorian sanojen esiintymistiheyttä twiiteissä, joissa kärkipoliitikko mainittiin. Tumasjan ym. (2010) eivät kuitenkaan käyttäneet näitä profiileja vaaliennusteensa laatimisessa eivätkä laatineet puolueista vastaavia profiileja, joiden pohjalta puolueiden vaalimenestystä oltaisiin voitu arvioida. Vaalitulosten ennustamiseen perehdytään tarkemmin seuraavassa luvussa.

3.2 Sanaston pohjalta tehty tunneanalyysi

Vilares, Thelwall ja Alonso (2015) analysoivat tutkimuksessaan Espanjan parlamenttivaalien tiimoilta käytyä Twitter-keskustelua käyttäen analyysissaan apuna tunneanalyysia hyödyntäen Thelwallin, Buckleyn, Paltogloun, Cain ja Kappasin (2010) kehittämää SentiStrength-ohjelmistoa. SentiStrength-ohjelmisto on kehitetty tunnistamaan ihmisten mielipiteitä lyhyistä teksteistä ja käyttää omaa sanastoaan sanojen merkitysten tunnistamiseksi (Thelwall ym., 2010).

Koska ohjelmistolla ei ennen tutkimusta ollut riittävän laajaa espanjankielistä sanastoa, loivat Vilares ym. (2015) tutkimuksessaan ohjelmistoon espanjankielisen sanaston yhteistyössä kielitieteilijöiden kanssa keräämällä satunnaisotannalla 1600 espanjankielistä twiittiä ja arvioimalla niiden sanojen merkityksen ohjelmiston tietokantaan. Alkuperäinen espanjankielinen sanasto on otettu LIWC-ohjelmistosta, jonka jälkeen Vilares ym. (2015) lisäsivät kielitieteilijöiden arvioimat sanat tietokantaan. SentiStrengthin espanjankielisen sanaston koko suureni noin 20-kertaiseksi alkuperäisestä ja sanastoon lisättiin emojien, sanontojen ja slangisanojen merkityksiä. (Vilares ym., 2015)

SentiStrength toimii hyvin eri tavalla kuin LIWC, keskittyen tietyn kategorian sanojen esiintymiskertojen laskemisen sijaan sanojen sentimentaalisiin ulottuvuuksiin. Jokaiselle sanalle on määritetty lukuarvo tunneskaalalla positiivinen-negatiivinen väliltä (1, 5) positiivisille ja väliltä (-1, -5) negatiivisille sanoille. Jokaisessa tutkimusaineiston twiitissä mainittiin poliitikko tai puolue, useamman eri puolueen mainitsevat twiitit poistettiin aineistosta. Lisäksi uudelleentwiittaukset poistettiin, sillä uudelleentwiitatut mielipiteet olisivat vääristäneet tutkimuksia. SentiStrengthia käyttäen näille viesteille määritettiin joko positiivinen tai negatiivinen lukuarvo niiden sisältämien sanojen perusteella. Mikäli viestissä ei oltu ilmaistu selkeää mielipidettä, se poistettiin aineistosta. (Vilares ym., 2015)

SentiStrengthin avulla tehdyn analyysin tuloksena Vilares ym. (2015) saivat positiivista ja negatiivista suhtautumista kuvaavat lukuarvot kullekin poliitikolle ja puolueelle. Lisäksi kerättiin jokaisen puolueen ja puolueiden puheenjohtajien mainintojen päivittäiset keskiarvot. Tuloksista selvisi esimerkiksi, että johtavien puolueiden johtajiin suhtauduttiin negatiivisemmin kuin oppositiojohtajiin. Analyysin pohjalta tehtyihin ennusteisiin palataan seuraavassa luvussa.

Vilares ym. (2015) lisäksi myös useissa myöhemmissä tutkimuksissa on käytetty sanaston avulla tehtyä tunneanalyysia määrittämään twiittien tunnesisältöä. SentiStrengthin sijaan analyysityössä on käytetty muita ohjelmistoja. Kaikkiaan sanastoon perustuva tunneanalyysimenetelmä on yleisin analyysin metodi, jota tutkimuksissa on käytetty.

3.3 Ohjattu tunneanalyysi

SASA eli Supervised aggregated sentiment analysis (vapaasti suomeksi käännettynä ohjattu koottu tunneanalyysi) on Ceronin, Curinin ja Iacusin (2015b) käyttämä termi kootun tunneanalyysin tekniikalle. Hopkins ja King (2010) esittelivät ensimmäisenä ohjatun tunneanalyysin tekniikan, joka kehitettiin ratkaisemaan kahta koneellisen tunneanalyysin ongelmaa. Ensimmäisenä ongelmana on se, että ihmiset käyttävät viestinnässään murteita ja puhekieltä standardisoidun yleiskielen sijaan. Toisaalta kielen vivahteet myös muuttuvat sen perusteella, kuka viestii. Myöskään metaforia, ironiaa tai jargonia ei voida arvioida tunneanalyysin keinoin joka yhteydessä samalla tavalla eli kieli on myös kontekstisidonnaista. Täten täysin koneellinen tunneanalyysi ei toimi toivotulla tavalla ihmisten

tuottamaa tekstiä tulkittaessa. Vilares ym. (2015) pyrkivät ratkaisemaan ongelman tekemällä ohjelmiston sanastosta mahdollisimman laajan, mutta kontekstisidonnaisuuden ohittaminen analyysissa tekee tuloksista epätarkempia.

Toinen ratkaistava ongelma on taustakohina, joka tulisi seuloa pois käsiteltävän datan seasta. Avainsanahauilla aineistoa kerättäessä mukaan tulee myös tekstiä, joka ei välttämättä liity käsiteltävään asiaan ja sekoittaa täten tunneanalyysin tuloksia. Tämän vuoksi SASA-metodia käytettäessä tunneanalyysiohjelmistoa opetetaan ihmisten arvioimalla opetusjoukolla, joka on alkuperäisestä datasta kasattu satunnaisotos. Koska ihmiset ovat konetta parempia arvioimaan tekstin liittyvyyttä käsiteltävään asiaan ja kirjoittajan tunteita kohdetta kohtaan, opetusjoukon avulla tunneanalyysin tuloksia saadaan tarkemmiksi kuin täysin koneellisilla metodeilla. (Ceron ym., 2015b)

Hopkinsin ja Kingin (2010) esittelemä ReadMe-metodi käsittelee ihmisten luokitteleman opetusjoukon avulla yksittäisten kirjoitusten mielipidesisältöä. Opetusjoukon sanojen arvottamisen jälkeen tekstin sanat riisutaan kantasanoiksi sekä lauseiden rakenteet puretaan ja jätetään analysoimatta. Tuloksena on kasa sanoja (bag of words, Wallach 2006), joiden sisältämät mielipiteet analysoidaan ohjelmiston tietokannan ja koodatun opetusjoukon avulla. Viestien sisältämien mielipiteiden arvot kootaan yhteen ja koko viestijoukolle arvioidaan kunkin mielipiteen frekvenssijakauma. Mikäli opetusjoukko on riittävän laaja (20-50 ihmisen arvioimaa twiittiä) ja satunnaisesti valittu, saavutetaan mielipidejakaumalla noin 2-3 prosenttiyksikön virhemarginaali. Koska vaalien ennustamisessa nimenomaan kokonaismielipidejakaumalla on merkitystä yksittäisten henkilöiden mielipiteiden sijaan, toimii ohjattu koottu tunneanalyysi hyvin vaalien ennustamisessa. (Ceron ym., 2015b)

4 ENNUSTEIDEN LAATIMINEN JA NIIDEN TARKKUUS

Vaaleja on pyritty ennustamaan Twitteristä kerätyn datan avulla ainakin Alankomaissa (Sanders ja van den Bosch, 2013), Espanjassa (Vilares ym., 2015), Italiassa (Caldarelli, Chessa, Pammolli, Pompa, Puliga, Riccaboni ja Riotta, 2014; Ceron ym., 2015a), Isossa-Britanniassa (Lampos, 2012), Irlannissa (Bermingham ja Smeaton, 2011), Saksassa (Tumasjan ym., 2010), Singaporessa (Choy, Cheong, Laik ja Shung, 2011; Skoric, Poor, Achananuparp, Lim ja Jiang, 2012) ja Yhdysvalloissa (Ceron ym., 2015a; Choy ym., 2012; Gayo-Avello, 2011; Jensen ja Anstead, 2013; O'Connor, Balasubramanyan, Routledge ja Smit, 2010; Shi, Agarwal, Agrawal, Garg ja Spoelstra, 2012; Washington, Thatcher, Morar ja LePrevost, 2015). Koska Twitteristä kerätyn datan pohjalta tehtyjen ennusteiden tekeminen on edullista verrattuna perinteisiin kyselytutkimuksiin ja tutkimuksissa on saavutettu rohkaisevia tuloksia ennusteiden tarkkuudesta, on tutkimuksia tehty useissa eri maissa vaalien yhteydessä.

Ennusteita on laadittu Twitteristä kerätyn ja analysoidun datan pohjalta tutkimuksissa usein eri tavoin: yksinkertaisimmat ennusteet perustuvat puolueiden ja poliitikkojen mainintamääriin Twitter-keskusteluissa ennen vaalia (Tumasjan ym., 2010), kun taas tarkempaan analyysiin perustuvissa ennusteissa tunneanalyysin tulosten perusteella on määritetty yleinen mielipidejakauma ja arvioitu sen soveltuvuutta puolueiden välisen vaalituloksen ennustamiseen (Ceron ym., 2015b). Ennusteiden laadinnassa on myös käytetty äänestysaikomuksen ilmaisuja (Gayo-Avello, 2011) ja yksittäisiä poliitikkoja kohtaan ilmaistujen positiivisten ja negatiivisten mielipiteiden jakaumaa. Seuraavaksi esittelen merkittävimmät metodit ennusteiden laadintaan ja niitä hyödyntäviä tutkimuksia.

4.1 Mainintoihin perustuvat ennusteet

4.1.1 Tumasjan ym. (2010)

Tumasjan ym. (2010) tekivät ensimmäisenä Twitteristä kerätyn datan pohjalta vaaliennusteen. Tutkimus keskittyi kerätyn Twitter-datan pohjalta tehtyyn tunneanalyysiin LIWC2007-ohjelmiston sanaston avulla, mutta tutkimuksessa esiteltiin myös metodi vaalituloksen ennustamiseen Twitteristä kerätyn datan avulla. Tunneanalyysia ei kuitenkaan käytetty ennustuksen laadinnassa, vaan ennustuksen laadinnassa käytettiin pelkästään kunkin poliittisen puolueen saamia Twitter-mainintoja äänimäärän mittarina. Puolueiden johtohahmojen saamia mainintoja ei otettu huomioon ennusteessa. Ennuste pohjautui siis puhtaasti puolueiden saamaan suosioon Twitter-keskusteluissa ja hylkäsi tunneanalyysin ennustetta laadittaessa. Ennusteessaan Tumasjan ym. (2010) saavuttavat erin-

omaiselta vaikuttavan 1,65 prosentin keskimääräisen virheen (MAE, mean absolute error) verrattuna vaalien lopullisiin tuloksiin. Perinteisten gallupien esittämä pienin MAE oli 0,80% ja suurin MAE 1,48%, eli Twitterin pohjalta laadittu ennuste oli kuitenkin perinteisiä tulosennusteita epätarkempi vaalituloksen ennustaja.

Tumasjan ym. (2010) tutkimus on sittemmin saanut runsasta kritiikkiä esimerkiksi Jungherrin, Jürgensin ja Schoenin (2012) arviossa tutkimustuloksen toistettavuudesta ja ennustusmetodin satunnaisuudesta, twiittien keruuvälin ajankohdan valinnan näennäisestä satunnaisuudesta ja piraattipuolueen jättämisestä ennusteen ulkopuolelle. Varsinkin tutkimuksen ajatus siitä, että puolueiden Twitter-maininnoista voitaisiin luotettavasti päätellä puolueen vaalimenestys, saa kritiikkiä Jungherrilta ym. (2012).

Tumasjan ym. (2010) valitsivat aineistokseen twiitit viiden viikon ajalta väliltä torstai 13. elokuuta 2009 - lauantai 19. syyskuuta 2009, vaalien ollessa 8 päivää twiittien keräämisen päättymisen jälkeen. Ajankohdan valinta herätti kritiikkiä, sillä twiittien keräämisen päättäminen viikkoa ennen varsinaista vaalipäivää herätti kummastusta. Kun Jungherr ym. (2012) keräsivät omaan tutkimukseensa mukaan twiitit myös noilta kahdeksalta päivältä ja sisällyttivät myös piraattipuolueen tunnisteet hakuunsa, kasvoi twiittien määrä 70 prosentilla alkuperäisestä tutkimuksesta. Jungherrin ym. (2012) verratessa keräämäänsä aineistoa lopulliseen vaalitulokseen, todettiin keskimääräisen virheen kasvaneen 2,13 prosenttiyksikköön. Myös muut mittausvälin vaihtamiset Jungherrin ym. (2012) tutkimuksessa muuttivat virhemarginaalia, kuten taulukosta 1 ilmenee.

Myös piraattipuolueen ohittaminen tutkimuksessa herätti kritiikkiä Jungherrilta ym. (2012), sillä piraattien kampanja internet-sensuuria vastaan herätti suurta keskustelua saksankielisessä Twitterissä ennen vaaleja kääntymättä kuitenkaan vaalimenestykseksi. Puolue mainittiin 34,8 prosentissa Jungherrin ym. (2012) keräämistä viesteistä, mutta sai vaaleissa vain 2,1 prosenttia äänistä. Tumasjan ym. (2012) vastasivat kuitenkin kritiikkiin halunneensa säilyttää tutkimuksensa ennusteessa vertailukelpoisuuden saksalaisten mittauslaitosten ennusteisiin, joissa piraattipuoluetta ei oltu huomioitu.

Tumasjan ym. (2012) toteavat vastineessaan myös, että jopa Jungherrin ym. (2012) saavuttama 3,34 prosenttiyksikön keskimääräinen virheprosentti (Taulukko 1) vain kahden vuorokauden twiittien keräämisellä on rohkaiseva merkki menetelmän toimivuudesta, varsinkin ottaen huomioon ennustuksen laatimi-

Taulukko 1: Ennusteiden virheprosentti, Tumasjan ym. (2010) (taulukossa TSSW) verrattuna Jungherr ym. (2012). Kuva Jungherr ym. (2012)

	13.8–19.9 (TSSW)	13.8–27.9	13.8–19.9	20.8–19.9	27.8–19.9	3.9–19.9	10.9–19.9	17.9–19.9
CDU	1.0	1.95	0.39	0.58	1.42	1.62	2.65	2.60
CSU	1.3	2.22	2.23	2.28	2.3	1.75	2.03	3.00
SPD	2.2	2.21	1.9	1.99	1.75	2.33	1.82	4.43
FDP	1.7	3.04	1.67	2.01	2.22	2.83	2.59	3.14
Linke	0.3	0.03	0.04	0.03	0.31	0.40	0.53	0.39
Green	3.3	3.31	2.81	2.81	2.93	2.47	3.38	6.51
MAE	1.6	2.13	1.51	1.62	1.82	1.90	2.17	3.34

seen käytetyn yksinkertaisen mainintojen laskemisen metodin ja tarkastelujakson pidentyessä pienenevän keskimääräisen virheprosentin. Toisaalta vastineessa myös todetaan, että Twitterin pohjalta käytetyillä metodeilla laadittu ennuste sopii paremmin perinteisen gallupin lisäksi kuin korvaamaan niitä.

4.1.2 Muut mainintoihin perustuvat ennusteet

Myös muissa tutkimuksissa on laadittu ennusteita Twitter-mainintojen perusteella. Sanders ja van den Bosch (2013) ennustivat Alankomaiden parlamenttivaaleja keräämällä 170 tuhatta twiittiä vaaleja edeltäviltä 10 päivältä. Lopulta tarkimmat ennusteet saatiin vertaamalla vaalipäivää edeltävien viiden päivän mainintoja vaalitulokseen, tuloksena 1,9 prosentin keskimääräinen virhe perinteisten gallupien yltäessä 1,1 prosenttiin. Twiittien keruuaikaväliä vaihtelemalla MAE vaihteli välillä 1,9 – 2,4. Vaikka tutkijat eivät saaneet selville parasta mahdollista aikaväliä twiittien keräämiselle, tutkimuksessa todetaan kirjoittajien uskovan vaalien ennustamiseen Twitter-mainintojen pohjalta.

Vilares ym. (2015) keräsivät puolueista ja puolueiden kärkipolitiikoista positiivisia ja negatiivisia tuntemuksia, onnistumatta hyödyntämään näin keräämiään suosiolukuja vaalituloksen ennustamiseen. Vaikka mainintojen määrä Twitterissä oli poliitikkojen suosiota parempi vaalituloksen ennuste, eivät maininnatkaan vastanneet vaalitulosta. Tähän annettiin tutkimuksessa monia syitä, esimerkiksi vasemmiston kannattajien korkea Twitter-aktiivisuus ja analysoitujen twiittien rajoittuminen ainoastaan espanjankielisiin. Mikäli Espanjan valtakielistä esimerkiksi baski ja katalaani olisivat luvuissa mukana, tulokset olisivat kattavampia. (Vilares ym., 2015)

Caldarelli ym. (2014) ennustivat Italian parlamenttivaaleja vertaamalla kerätyn Twitter-aineiston sisältämiä mainintoja lopulliseen vaalitulokseen. Maininnat seuraavat useiden puolueiden osalta varsinaista vaalitulosta, mutta pienpuolueet ja pääministeripuolue SC ovat Twitter-aineistossa yliedustettuina suhteessa vaalitulokseen. Tarkastelujakson muuttaminen aiheuttaa myös suurta heilahtelua mainintojen prosenttiosuuksissa eikä tarkastelujakson pidentäminen paranna ennusteen tarkkuutta. Tutkimuksessa saatiin tuloksia puolueiden voimasuhteista, mutta gallupien korvaamiseen riittävää tarkkuutta ei saavutettu.

Bermingham ja Smeaton (2011) pyrkivät ennustamaan Irlannin parlamenttivaaleja 2011 vertaamalla mainintoja ja koneellisen tunneanalyysin tuloksia gallup-lukuihin ja vaalitulokseen. Tarkimmat tulokset tutkimuksessa saatiin mainintojen määrällä, mutta MAE gallup-lukuihin verrattaessa oli 3,67% ja vaalitulokseen pienimmillään 5,85%, gallupien yltäessä 1,61% keskimääräiseen virheeseen. Sanastoon vertaamalla tehdyn tunneanalyysin avulla ennustaminen oli vielä epävarmempaa, puolueen osuuden positiiviseksi tunnistetuista twiiteistä ollessa toiseksi paras vaalituloksen ennakoija mainintojen jälkeen.

Skoric ym. (2012) ennustivat Twitteristä kerätyn datan pohjalta Singaporen vuoden 2011 parlamenttivaaleja vertaamalla mainintojen määrää vaalitulokseen. Tutkimuksessa keskimääräinen virhe (MAE) oli 5,23 prosenttia, yhden puolueen kohdalla ennusteen ollessa 17,34 prosenttiyksikköä liian matala.

Jensen ja Anstead (2013) käyttivät keräämäänsä Twitter-dataa ennustamaan republikaanien esivaaliehdokkaiden suoriutumista ja gallup-lukujen tulevaa kehitystä vuoden 2012 presidentinvaalien esivaaleissa. Mainintoihin perustuvalla ennusteella tutkimuksessa saavutettiin Iowan esivaalissa ennusteen MAE 3,1% gallupien MAE:n ollessa 2,2%. Muita esivaaleja ei tutkimuksessa ennustettu datan perusteella.

Shi ym. (2012) ennustivat Yhdysvaltain 2012 republikaanien esivaaleja Twitter-datan pohjalta sekä mainintojen että semanttisen analyysin pohjalta. Mainintojen määrä korreloi osan ehdokkaista galluplukujen kanssa vahvasti, mutta toisten ehdokkaiden, esimerkiksi esivaalien kärkiehdokkaiden, osalta ei soveltunut lainkaan ennustamaan gallupeissa suoriutumista. Myöskään semanttisen analyysin pohjalta ei saatu kaikkien ehdokkaiden osalta luotettavia tuloksia.

Washington ym. (2015) käsittelivät tutkimuksessaan Yhdysvaltojen vuoden 2012 presidentinvaalien tulosta ja äänten kokonaisjakaumaa, sivuuttaen osavaltiokohtaiset tulokset. Käytännössä tutkimuksessa verrattiin lopullista vaalitulosta vaalipäivän aikaisiin Twitter-mainintoihin ja Radian6-ohjelmiston avulla tehtyyn tunneanalyysiin. Mainintojen ei todettu korreloivan vaalin tulokseen. Tutkimuksen aiemmassa versiossa oli myös tarkempia Twitterin pohjalta tehtyjä laskelmia, mutta kirjoittajien todettua laskelmansa vääriksi on virheelliset laskelmat sisältävä osio sittemmin poistettu tutkimuksesta. (Washington ym., 2015.)

4.1.3 Mainintoihin perustuvien ennusteiden luotettavuus

Edellisessä kappaleessa käytiin läpi 10 eri tutkimusta, joissa oltiin ennustettu vaalien lopputulosta Twitteristä kerätyn datan perusteella vertaamalla lopullista vaalitulosta ja/tai gallup-tuloksia ehdokkaiden tai puolueiden Twitterissä saamien mainintojen määrään. Taulukossa 2 on käyty läpi tutkimusten ennusteiden keskimääräistä virhettä mittaavia MAE-lukuja.

Kuten taulukosta 2 ilmenee, on kahdessa tutkimuksessa (Sanders ja van den Bosch, 2013; Tumasjan ym., 2010) saavutettu alle kahden prosenttiyksikön keskimääräinen virheprosentti varsinaiseen vaalitulokseen verrattaessa. Tumasjan ym. (2010) ennustamissa Saksan parlamenttivaaleissa perinteisten gallupien MAE-prosentit olivat välillä 0,8 – 1,48 ja Alankomaiden parlamenttivaaleissa (Sanders ja van den Bosch, 2013) perinteisten gallupien MAE oli 1,1 prosenttiyksikköä. Parhaissakin tapauksissa mainintoihin perustuvat ennusteet ovat siis suoriutuneet perinteisiä kyselytutkimuksia heikommin.

Useissa mainintojen laskemiseen perustuvissa tutkimuksissa ei myöskään löydetty kunnollista korrelaatiota mainintojen ja vaalituloksen välille, joten keskimääräistä virheprosenttia ei ole laskettu kaikissa tutkimuksissa. Lisäksi esimerkiksi Bermingham ja Smeaton (2011) sekä Skoric ym. (2012) saavuttivat yli 5 prosenttiyksikön keskimääräiset virheet, johtaen käytännössä hyvin epäluotettaviin ennusteisiin.

Mainintoihin perustuvia ennusteita ei voida siis pitää uskottavina perinteisten vaalitulosennusteiden korvaajina, vaan niitä on parhaimmillaankin pidettävä

Taulukko 2: Mainintoihin perustuvien ennusteiden keskimääräiset virheprosentit verrattuna vaalitulokseen

Tutkimuksen nimi ja vuosi	**Ennustettavat vaalit**	**Korrelaatio mainintojen määrä ja vaalitulos**
Sanders ja van den Bosch, 2013	Alankomaiden parlamenttivaalit 2012	MAE 1,9%
Vilares ym. (2015)	Espanjan parlamenttivaalit 2015	heikko
Caldarelli ym. (2014)	Italian parlamenttivaalit 2013	olemassa, yksi puolue yliedustettuna
Bermingham ja Smeaton (2011)	Irlannin parlamenttivaalit 2011	MAE 5,85%
Tumasjan ym. (2010)	Saksan parlamenttivaalit 2009	MAE 1,65%
Jungherr ym. (2012)	Saksan parlamenttivaalit 2009	MAE 2,13%
Skoric ym. (2012)	Singaporen parlamenttivaalit 2011	MAE 5,23%
Jensen ja Anstead (2013)	Republikaanien esivaalit 2012	MAE 3,1%
Shi ym. (2012)	Republikaanien esivaalit 2012	heikko
Washington (2015)	Yhdysvaltain presidentinvaalit 2012	heikko

perinteisten ennusteiden lisänä. Parhaita mainintoja laskemalla saavutettuja tutkimustuloksiakin on pidettävä lähinnä onnekkaina sattumina, sillä menetelmä ei huomioi esimerkiksi missä sävyssä mistäkin ehdokkaasta tai puolueesta on sosiaalisessa mediassa keskusteltu.

4.2 Tunneanalyysiin perustuvat ennusteet

4.2.1 Lampos (2012)

Lampos (2012) käsitteli väitöskirjatutkimuksessaan Twitter-datan perusteella Ison-Britannian parlamenttivaaleja 2010. Tutkimuksen tavoitteena oli ennustaa twiittien pohjalta parlamentin kolmen suurimman puolueen (konservatiivit, työväenpuolue Labour ja liberaalit) äänijakaumat vaaleissa. Tutkimusta varten kerättiin 50 miljoonaa twiittiä, mutta näistä twiiteistä vain 300 000 käytettiin tutkimuksen ennusteen laatimisessa.

Lampos käytti tutkimuksessaan kuutta eri tunneanalyysin variaatiota parhaan mahdollisen tekniikan löytämiseksi. SentiWordNet-ohjelmistoa käytettiin positiivisten ja negatiivisten tunteiden asteiden löytämiseksi twiiteistä, eli sanat arvotettiin niiden tunnepitoisuuden perusteella yksinkertaisen positiivinen/negatiivinen-merkinnän sijaan. Ensimmäisessä metodissa (SnPOS) twiittien sanat arvioitiin kantasanoiksi riisuttuina SentiWordNetin avulla ja kullekin twiitille laskettiin yhteen sen sisältämien sanojen positiiviset ja negatiiviset arvot. Mikäli SentiWordNetin sanastosta ei löytynyt sanalle merkitystä, sen tunnearvoksi merkittiin 0.

Toisessa metodissa (SPOS) Lampos käsitteli twiittien sisältämiä sanoja ja niiden välisiä suhteita part-of-speech (POS)-merkinnän avulla. Sanat merkittiin

sanatyypeittäin (esimerkiksi verbit, adjektiivit) ja mikäli samaan kantasanaan perustuvia saman sanatyypin sanoja esiintyi twiitissä useampi, laskettiin näille sanoille tunnearvojen keskiarvo. Koko twiitin tunnearvo laskettiin jälleen laskemalla yhteen twiitin sanojen positiiviset ja negatiiviset lukuarvot.

Kolmannessa metodissa (SPOSW) POS-merkintää laajennettiin lisäämällä mukaan analyysiin WordNet-ohjelmiston 5000 sanan synonyymisanasto. Mikäli twiitissä esiintyi WordNetin sanastosta löytyvä sana, lisättiin twiittiin myös sanan synonyymit. Täten voitiin tehostaa tunneanalyysia varmistamalla, että SentiWordNetistä löytyisi todennäköisemmin lukuarvo jollekin synonyymisanoista. Lisäksi pyrittiin vahvistamaan twiitissä ilmaistuja tunteita analyysin helpottamiseksi, sillä Twitterin 140 merkin viestin enimmäispituus saattaa heikentää ilmaisua. Twiitille laskettiin kokonaistunnearvo laskemalla yhteen siinä ilmaistut positiiviset ja negatiiviset tunteet.

Sanojen tunneilmaisun analysoimisen jälkeen saadut tunnearvot laskettiin yhteen kahdella eri tavalla. Ensimmäinen metodi, MTS (Mean Thresholded Sentiment) karsi ensin semanttisen analyysin kannalta epävarmimmat eli lähimpänä nollaa positiivisen ja negatiivisen tunnearvon erotukseltaan olevat twiitit. Tämän jälkeen laskettiin kunkin mainitun puolueen osalta loppujen twiittien positiivisten ja negatiivisten tunnekeskiarvojen erotus. Regressioanalyysia käyttäen määritettiin tiedossa olevan äänestysaikomusprosentin (tässä tapauksessa kyselytutkimuksella määritetty gallupin ennuste) ja laskemalla määritetyn tunnekeskiarvon välistä suhdetta kuvaava painoarvo kullekin puolueelle. Lopuksi laskettiin äänestystodennäköisyys kullekin puolueelle kertomalla tunneanalyysin keskiarvo puolueen painoarvolla ja normalisoimalla kaikkien puolueiden summien yhteenlasketuksi arvoksi 1.

Toinen metodi, DSC (Dominant Sentiment Class) on muuten samanlainen kuin MTS, mutta käyttää tunnekeskiarvon määrittämiseen twiittien lukumäärää. Positiivisiksi tunnistettujen twiittien ja negatiivisiksi tunnistettujen twiittien lukumäärän erotus lasketaan ja jaetaan twiittien määrällä, jonka jälkeen määritetään puolueille äänestystodennäköisyydet MTS-metodin tapaan.

Kun kolme tunneanalyysin metodia ja kaksi äänestystodennäköisyyden laskennan metodia kerrotaan, saadaan yhteensä kuusi erilaista tapaa kunkin puolueen äänimäärän ennustamiseen. Metodeita vertailtaessa huomataan (Taulukko 3), että tarkimmat tulokset saavutetaan käytettäessä POS-merkittyjen sanojen yhteydessä WordNetin synonyymisanastoja. MTS ja DCS olivat käytännössä SPOSW-tunneanalyysimetodin kanssa yhtä tarkkoja äänestystodennäköisyyden määrittäjiä, mutta muiden tunneanalyysimetodien kanssa DCS:n avulla tehdyt ennusteet olivat tarkempia.

Lamposin (2012) tutkimuksessa päästiin ennusteita laadittaessa parhaassa tapauksessa 3,49 prosentin keskimääräiseen virheprosenttiin (MAE), kun aineistosta poistettiin ensin tunneanalyysin kannalta epäselvimmät twiitit. Muut tunneanalyysin metodit eivät olleet tehokkaita WordNetin synonyymisanastoa käyttäneeseen SPOSW-metodiin verrattuna, joten synonyymien lisäämistä twiitteihin tunneanalyysin tulosten tehostamiseksi voidaan pitää onnistuneena.

Taulukko 3: Lampos (2012), ennusteiden laadinnan metodien vertailu.

	δ	CON	LAB	LIBDEM	All Parties	MRE	p-value
SnPOS	0	12.97 ± 10.89	10.56 ± 10.05	6.66 ± 7.08	10.06 ± 9.72	0.4038	0.222
SnPOS	0.025	12.98 ± 10.92	10.51 ± 10.03	6.7 ± 7.19	10.06 ± 9.74	0.4038	0.259
SPOS	0	35.03 ± 6.96	21.31 ± 9.4	12.71 ± 7.16	23.02 ± 12.11	0.9231	0.531
SPOS	0.0072	35.14 ± 6.82	21.33 ± 9.41	12.77 ± 7.13	23.08 ± 12.1	0.9231	0.527
SPOSW	0	4.44 ± 3.18	2.66 ± 1.85	3.74 ± 2.86	3.61 ± 2.76	0.1731	0.019
SPOSW	0.0238	4.44 ± 3.31	2.65 ± 1.8	3.84 ± 2.71	3.64 ± 2.75	0.1346	0.006

Table IV. : MAE ± MAE's standard deviation and MRE for Dominant Class Sentiment (**DCS**).

	δ	CON	LAB	LIBDEM	All Parties	MRE	p-value
SnPOS	0	10.56 ± 6.75	9.82 ± 9.18	7.69 ± 9.88	9.36 ± 8.68	0.4038	0.467
SnPOS	0.025	9.66 ± 6.89	9.46 ± 9.05	7.25 ± 9.04	8.79 ± 8.35	0.3654	0.529
SPOS	0	10.63 ± 8.94	8.09 ± 6.37	6.12 ± 5.12	8.28 ± 7.14	0.3846	0.238
SPOS	0.0072	10.51 ± 9.14	7.95 ± 6.18	6.08 ± 5.5	8.18 ± 7.26	0.4038	0.149
SPOSW	0	4.51 ± 3.45	2.87 ± 2.06	3.53 ± 3.29	3.64 ± 3.04	0.1154	0
SPOSW	0.0238	4.49 ± 3.49	2.46 ± 1.81	3.51 ± 3.14	3.49 ± 2.98	0.0962	0

4.2.2 Muut tunneanalyysiin perustuvat ennusteet

Vilares ym. (2015) käyttivät SentiStrength-ohjelmiston avulla tehtyä tunneanalyysia Espanjan parlamenttivaalien 2015 puolueiden ja niiden kärkipoliitikkojen vaalimenestyksen ennustamiseen. Tutkimuksessa todettiin tunneanalyysin tulokset heikommaksi vaalimenestyksen mittariksi kuin esimerkiksi mainintojen määrän. Analyysin tulosten poliitikkojen suosiosta todettiin kuitenkin seuraavan vahvasti poliitikkojen mielipidekyselyissä saamia suosiolukuja.

Bermingham ja Smeaton (2011) käyttivät mainintojen lisäksi tunneanalyysia Irlannin parlamenttivaalien ennustamiseen usean eri metriikan avulla. Tehokkain tunneanalyysia hyödyntävä metodi eli puolueen osuus positiivisista twiiteistä on kuitenkin tutkimuksen tulosten perusteella heikompi vaalituloksen ennustaja kuin gallupit tai mainintoihin perustuva ennuste.

Choy ym. (2011) ennustivat Singaporen presidentinvaaleja käyttämällä ennusteessaan apuna edellisten vaalien tuloksia, viimeisimmän väestönlaskennan ikäjakaumatilastoja sekä tilastoa kansalaisten internetin käytöstä ikäryhmittäin. Tutkimuksessa luotiin ennustemalli, jossa Twitterissä ilmaistut mielipiteet kuvastivat internetiä käyttävän kansanosan mielipidettä ja internetiä käyttämättömien oletettiin äänestävän 2011 parlamenttivaalien puolueiden välisen äänijakauman mukaisesti. Tutkimuksessa päädyttiin 6,07 prosentin keskivirheeseen, eli ennustusmalli ei toiminut vaalituloksen ennustamisessa merkittävän hyvin.

Choy ym. (2012) pyrkivät myös ennustamaan Yhdysvaltain vuoden 2012 presidentinvaaleja samankaltaisella internetin käyttöön eri ikäryhmissä perustuvalla ennustemallilla kuin edellisessä tutkimuksessa. Ennusteen tarkkuus oli parempi kuin Singaporen vaaleissa, osavaltiokohtaisen ennusteen keskimääräisen virheprosentin ollessa 2,60 prosenttiyksikköä. Toisaalta voidaan myös todeta, että ennustemallin ottaessa huomioon Obaman voittamat 2008 presidentinvaalit

ja Obaman ollessa ehdolla toistamiseen 2012 vaaleissa tämä aiheuttanee vääristymää varsinaisen tunneanalyysin tulosten luotettavuudesta.

Gayo-Avello (2011) pyrki ennustamaan Yhdysvaltain presidentinvaaleja 2008 useilla eri metodeilla. Ennusteissa otettiin huomioon ehdokkaiden saamat maininnat, positiivisten ja negatiivisten twiittien määrä, sanaston avulla arvioitu ilmaistujen tunteiden vahvuuden analysointi ja äänestysaikomuksen ilmaisseiden määrä ehdokasta kohden. Tutkimuksessa keskityttiin vaa'ankieliosavaltioiden tulosten ennustamiseen ja aineistoa kerätessä käytettiin geofiltteröintiä, joka on saattanut johtaa vääristymiin dataa kerättäessä. Millään neljästä metodista ei saavutettu selkeää korrelaatiota ehdokkaiden saaman äänimäärän kanssa.

O'Connor ym. (2012) ennustivat Yhdysvaltain vuoden 2008 presidentinvaaleja tunneanalyysin keinoin. Ennusteen laadinnassa käytettiin yksinkertaista erottelua positiivisiin ja negatiivisiin twiitteihin ilman tunteiden vahvuuden analysointia. Tutkimuksessa ei löydetty vahvaa korrelaatiota positiivisten twiittien ja ehdokkaiden galluplukujen välillä.

Washington ym. (2015) käyttivät asiakastyytyväisyyden analysointiin kehitettyä Radian6-ohjelmistoa Yhdysvaltain vuoden 2012 presidentinvaalien ennustamiseen. Koska kyseessä on kaupallinen suljetun lähdekoodin ohjelmisto, ei tunneanalyysin metodeista ole tutkimuksessa tarkempaa analyysia. Satunnaisotannalla valitut miljoona toisen ehdokkaan mainitsevaa twiittiä ajettiin ohjelmiston analyysin läpi ja molemmille ehdokkaille laskettiin heidät mainitsevat positiiviset twiitit. Menetelmällä saavutettiin 1,8 prosentin MAE.

4.2.3 Tunneanalyysiin perustuvien ennusteiden luotettavuus

Sanaston avulla tehtyyn koneelliseen tunneanalyysiin on perehdytty edellisissä luvuissa ja yllä on listattuna kahdeksan eri tutkimusta, joissa vaalitulosta on pyritty ennustamaan koneellisen tunneanalyysin tulosten pohjalta. Kuten tutkimusten tuloksista voidaan päätellä, vaikuttaa tunneanalyysin pohjalta tehtyjen ennusteiden keskimääräinen virheprosentti olevan vielä mainintojen laskemiseen perustuvien ennusteiden vastaavia keskimääräisiä virheitä suurempi.

Parhaat tulokset tutkimuksissaan saavuttivat Washington ym. (2015) ennustaessaan Radian6-ohjelmiston avulla Yhdysvaltain presidentinvaaleja 2012 1,8 prosentin virheellä verrattuna lopulliseen vaalitulokseen. Lisäksi Choy ym. (2012) ennustivat Yhdysvaltain presidentinvaaleja 2012 2,60 prosentin osavaltiokohtaisella keskimääräisellä virheprosentilla. Choyn ym. (2012) käyttämässä ennustemallissa otettiin kuitenkin tunneanalyysin tulosten lisäksi huomioon myös Yhdysvaltain vuoden 2008 presidentinvaalien tulokset, joten tutkimuksen muihin tutkimuksiin verrattaessa pieni keskimääräinen virheprosentti selittyy osittain edellisiä vaaleja seuranneilla äänestystuloksilla. Tutkimus käytti tunteiden ilmaisun arviointiin Twitteriä varten kehitettyä AFINN-ohjelmistoa (Nielsen, 2011), joka arvioi kullekin sanalle sen sisältämän tunnearvon oman sanastonsa perusteella.

Tutkimuksissa, joissa ei arvioitu tunteiden ilmaisun astetta, vaan jaoteltiin twiitit yksinkertaisesti positiivisiin ja negatiivisiin ei saavutettu merkittävää korrelaatiota tunneanalyysin tulosten ja varsinaisen vaalituloksen välille. Washington ym. (2015) ja Choy ym. (2012) saavuttivat kuitenkin hyviä tuloksia ennustaessaan Yhdysvaltain vuoden 2012 presidentinvaaleja tunteiden ilmaisun vahvuutta mittaavien ohjelmistojen avulla. Myös Lamposin (2012) WordNet-sanaston avulla saavuttamia tuloksia voidaan pitää rohkaisevana osoituksena tunneanalyysin toimivuudesta vaalien ennustamisessa.

Voidaankin todeta, ettei twiittien yksinkertainen jaottelu vaikuta saavuttavan edes mainintojen määrää mittaavien ennusteiden tarkkuutta. Tunteiden ilmaisun mittaaminen vaikuttaa sen sijaan toimivan vaalituloksen ennustamisessa. Toisaalta myös ilmaistujen tunteiden vahvuutta mittaavat tutkimukset ovat epäonnistuneet vaalitulosten ja tunneanalyysin korrelaation osoittamisessa (Vilares ym., 2015).

4.3 Ohjattuun tunneanalyysiin perustuvat ennusteet

4.3.1 Ceron ym. (2014, 2015a)

Ceron ym. (2015a) pyrkivät ennustamaan Yhdysvaltain vuoden 2012 presidentinvaaleja ja Italian vaalikoalitio IBC:n puheenjohtajavaalia 2012 keräämänsä ja analysoimansa Twitter-datan pohjalta. Analyysissaan he käyttivät Hopkinsin ja Kingin (2010) kehittämää ReadMe-metodia (Hopkins, King, Knowles ja Melendez, 2010) twiittien tunnesisällön arvioimiseen. Analyysimenetelmää on käyty tarkemmin läpi luvussa 3.3.

Hopkinsin ja Kingin (2010) metodilla osa twiiteistä valitaan ensin satunnaisotannalla ihmisten analysoitaviksi, jolloin twiiteille arvioidaan niiden tunnesisältöä kuvaava positiivinen tai negatiivinen lukuarvo. Tämän jälkeen tätä opetusjoukkoa käytetään automaattisen analyysin kalibroimiseen ja analysoidaan loput twiiteistä. Metodi soveltuu hyvin Twitter-viestien analysoimiseen, sillä twiittien 140 merkin maksimipituus saattaa rajoittaa ilmaisun vahvuutta viesteissä ja johtaa esimerkiksi sanaston pohjalta tehtyä analyysia hankaloittavaan lyhennettyjen sanojen käyttöön. Lisäksi valvotulla analyysilla saadaan parempia tuloksia esimerkiksi ironiaa ja sarkasmia sisältävien viestien analyysissa, sillä ihmiset tunnistavat sarkasmin konetta paremmin.

Ceron ym. (2015a) ennustemalli keskittyy kokonaismielipidejakauman määrittämiseen kerättyjen twiittien kootun tunneanalyysin pohjalta. Analysointimetodia kutsutaan ohjatuksi kootuksi tunneanalyysiksi, sillä sen tunneanalyysi on opetusjoukolla ohjattua pelkkään sanastoon vertaamisen sijaan ja se laskee koko mielipidejoukon tunnejakauman yksittäisten mielipiteiden jakauman sijaan. Käytännössä kaikki twiitit riisutaan yksittäisiksi sanoiksi, jotka kootaan yhteen ja joiden mielipidejakauma lasketaan manuaalisesti arvioituun opetusjoukkoon sekä olemassa olevaan sanastoon vertaamalla.

Yhdysvaltain 2012 presidentinvaaleissa Ceron ym. (2015a) pyrkivät ennustamaan vaalien kokonaistulosta ja 11 vaa'ankieliosavaltion äänestystulosta. Twiitit jaettiin osavaltiotasolle geofiltteröinnin perusteella. Tutkimuksessa verrattiin ennustetta mielipidetutkimusten keskiarvoon. Twitterin pohjalta laadittu ennuste pääsi 2,93 prosentin keskivirheeseen ja ennusti koko maan äänijakauman 0,4 prosentin tarkkuudella sekä oikean voittajan yhdeksässä 11 arvioidusta osavaltiosta. Gallupien keskivirheprosentti oli 2,84 ja virheprosentti koko maan äänijakaumaa arvioitaessa 3,2%. Tulosennusteet erehtyivät osavaltion voittajasta kuitenkin vain yhdessä osavaltiossa. Twitter-ennuste oli gallupeita lähempänä oikeaa tulosta kokonaisäänimäärän lisäksi myös seitsemässä osavaltiossa, kahdessa ennusteet päätyivät samaan ja gallupit ennustivat kaksi osavaltiota Twitter-ennustetta paremmin.

Ceron ym. (2015a) ennustivat myös Italian IBC-koalition puheenjohtajavaalia ReadMe-metodin pohjalta. Tutkimusaineisto oli merkittävästi pienempi kuin Yhdysvaltain vaaleissa, noin 50 000 twiittiä Yhdysvaltain vaalien 50 miljoonaan twiittiin verrattuna. Verrattaessa perinteisiin tulosennusteisiin tutkimuksessa saavutettiin 1,96 prosentin keskimääräinen virhe puheenjohtajavaalin ensimmäisen kierroksen äänijakaumassa, gallupien päätyessä 1,06 ja 2,48 prosentin välille. Toisella kierroksella Twitter-ennuste suoriutui toiseksi parhaiten kahdeksaan perinteiseen tulosennusteeseen verrattaessa.

Ceron, Curini, Iacus ja Porro (2014) käyttivät Hopkinsin ja Kingin (2010) ReadMe-metodia jo aiemmin ennustaessaan Ranskan parlamenttivaaleja twiittien pohjalta. Laaditun ennusteen MAE oli tutkimuksessa 2,38% gallupien päästessä 0,69% - 1,93% keskimääräiseen virheeseen. Tutkimuksessa perehdyttiin myös ennusteen virheiden mahdollisiin syihin paikallistasolla ja todettiin ennusteen luotettavuuden paranevan sen mukaan, mitä suurempi twiittimäärä ennustetta laadittaessa on käytettävissä. Lisäksi huomattiin, että kuten perinteisissä gallupeissa, myös Twitter-ennusteiden tarkkuus kärsii ihmisten äänestämättä jättämisestä. 10% vaihtelu äänestysaktiivisuudessa vaikutti keskivirheeseen jopa 1,2 prosenttiyksikön verran.

4.3.2 Ohjattuun tunneanalyysiin perustuvien ennusteiden luotettavuus

Ceron ym. (2014, 2015a) ovat saaneet tutkimuksissaan kohtuullisen hyviä tuloksia käyttämällä tunneanalyysissaan apuna Hopkinsin ja Kingin (2010) ohjatun kootun tunneanalyysin metodia. Varsinkin 2012 Yhdysvaltain presidentinvaalien osalta Twitter-ennustetta voidaan pitää onnistuneena paitsi kokonaisäänimäärän osalta myös osavaltiotasolla. Ceron ym. (2015b) listasivat ReadMe-metodilla saavutettujen tulosten lisäksi myös metodiin kehittämänsä laajennuksen, iSA-analyysin avulla saavuttamiaan tuloksia. iSA-metodin kerrotaan tutkimuksessa olevan ReadMe-metodia tehokkaampi, mahdollistaen jopa reaaliaikaisen analyysin, mutta tarkempia iSA-metodia käyttäviä tutkimuksia ei olla vielä julkaistu.

Verrattaessa yksinkertaisiin mainintoihin perustuviin ennusteisiin voidaan ohjattua tunneanalyysia pitää tehokkaana vaaliennusteiden laadinnan työkaluna.

Vaikka ennusteissa on saavutettu suhteellisen matalat keskimääräiset virheprosentit ja joissain tapauksissa ennuste suoriutuu jopa tulosennusteita paremmin, ei myöskään Hopkinsin ja Kingin (2010) metodeilla tehtyyn tunneanalyysiin perustuvat ennusteet ole vielä valmiita korvaamaan perinteisiä kyselytutkimuksena toteutettuja mielipidetutkimuksia. Ceronin ym. (2015b) saavuttamien rohkaisevien tulosten pohjalta voidaan kuitenkin todeta, että suhteellisen edullisesti toteutettavissa olevilla tunneanalyysiin perustuvilla tutkimuksilla on mahdollisuus toimia perinteisten tulosennusteiden lisänä vaaleja ennustettaessa.

4.4 Ennusteiden saama kritiikki

Twitteristä kerätyn datan pohjalta laaditut ennusteet ovat kohdanneet tutkijayhteisössä myös kritiikkiä (Gayo-Avello, Metaxas ja Mustafaraj, 2011; Metaxas, Mustafaraj ja Gayo-Avello, 2011; Gayo-Avello, 2011, 2013).

Tutkimuksissa on esitetty kritiikkiä etenkin siitä, ettei yhdessäkään Twitterin pohjalta vaaleja ennustaneessa tutkimuksessa ennustetta ole tehty ennen vaaleja. Koska ennusteet on laadittu vasta vaalien jälkeen, ovat ne käytännössä ”datan prosessointia jälkikäteen, jotta sen voitaisiin väittää mahdollisesti ennustaneen vaalituloksen” (Metaxas ym., 2011).

Artikkeleissa kritisoidaan myös yksinkertaisia, mainintoihin perustuvia ennustemalleja niiden tulosten heikosta toistettavuudesta. Gayo-Avello ym. (2011) esimerkiksi kritisoivat Tumasjanin ym. (2010) ja O’Connorin ym. (2010) saavuttamia tuloksia, yrittäen ennustaa Yhdysvaltain senaattorivaaleja 2010 samoin metodein (maininnat ja positiivinen/negatiivinen tunneanalyysi) siinä onnistumatta. Tulosten satunnaisuuden ja mahdollisten eri ihmisryhmien erilaisista sosiaalisen median käyttötavoista johtuvien tilastovääristymien lisäksi esiin nostetaan myös huoli tarkasteluajanjaksojen muuttamisesta vääristyviin tuloksiin. Esimerkiksi Tumasjanin ym. (2010) viikkoa ennen vaaleja päättyvää tarkastelujaksoa kritisoidaan artikkelissa.

Lisäksi artikkeleissa kritisoitiin automatisoidun tunneanalyysin metodeja. Gayo-Avello ym. (2011) esimerkiksi nostavat esiin huolen automatisoidun tunneanalyysiin kyvystä tunnistaa twiitteihin sisältyvää sarkasmia ja ironiaa sekä virheellistä informaatiota sisältäviä tekstejä. Myös spämmitilien aiheuttamat mahdolliset tilastovääristymät nostetaan esiin artikkeleissa.

Gayo-Avello (2013) totesi, että sekä mainintoihin että koneellisesti sanastoon vertaamalla suoritettuun tunneanalyysiin sisältyy merkittäviä ongelmia ennusteiden luotettavuuden kannalta. Mainintoihin perustuvat ennusteet ovat liian satunnaisia, riippuvaisia twiittien keruuajankohdasta ja alttiita esimerkiksi spämmille, ettei niihin voida täysin luottaa. Lisäksi Gayo-Avello (2013) kritisoi mainintojen laskemiseen sisältyvää ”kaikki julkisuus on hyvää julkisuutta”-ajatusta. Tukea ajatuksilleen hän hakee Gayo-Avellon (2011) tuloksista, joissa todettiin, ettei Tumasjanin ym. (2010) käyttämillä metodeilla pystytty ennustamaan Yhdysvaltain vuoden 2008 presidentinvaaleja.

Sanastoon vertaavaan tunneanalyysiin Gayo-Avello (2013) suhtautuu kriittisesti todeten sen olevan kykenemätön tunnistamaan poliittisen kielen piirteitä, esimerkiksi sarkasmia. Lisäksi artikkelissa pidetään olemassa olevien tutkimusten tuloksia heikkoina ja kritisoidaan tulosten tasapainottomuutta, vieden pohjaa väittämältä, että datamäärän kasvattaminen poistaisi tutkimusten virheet.

4.5 Reaaliaikaisen ennustamisen mahdollisuudet

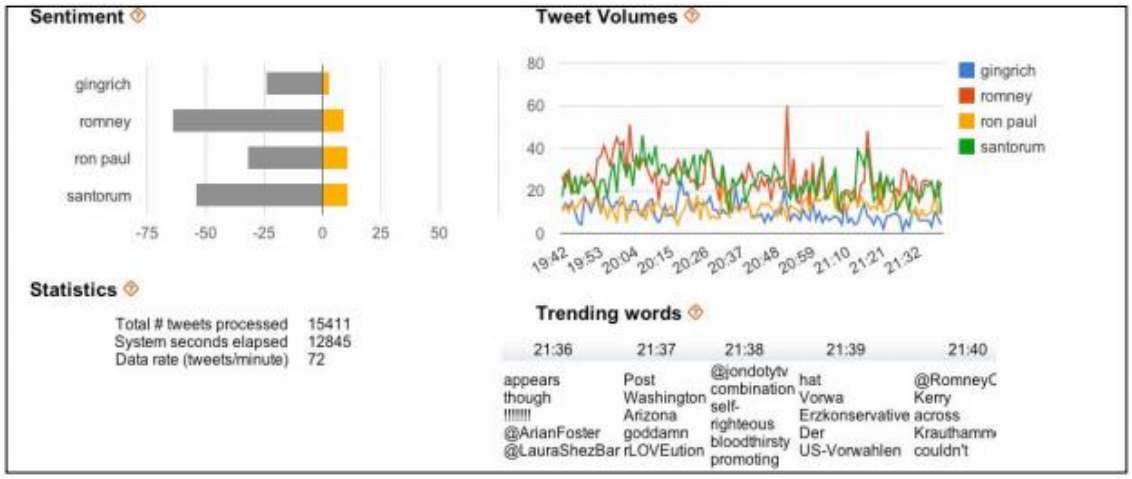
Wang ym. (2012) kävivät tutkimuksessaan läpi Twitteristä kerätyn datan mahdollista käyttöä reaaliaikaisen vaaliennusteen laadinnassa. Tutkimuksessa pyrittiin ennustamaan Yhdysvaltain vuoden 2012 presidentinvaalien republikaanien esivaaleja.

Tutkimuksessa käytettiin twiittien analysointiin IBM:n InfoSphere Streams-alustaa, joka on kehitetty datan reaaliaikaiseen analyysiin ja joka mahdollistaa skaalautuvuuden Twitter-liikenteen mukaan. Analyysia varten kukin twiitti saneistettiin tekstin osasten erottelemiseksi toisistaan. Tunneanalyysimalli kehitettiin analyysia varten Amazon Mechanical Turkin käyttäjien avulla, käyttäjien arvioidessa kunkin twiitin sarkasmisisältöä, ilmaistua tunnetta (positiivinen, negatiivinen, neutraali) sekä twiitin kirjoittajan oletettua poliittista suuntautumista välillä konservatiivi-liberaali. Näin luodussa opetusjoukossa oli lähes 17000 twiittiä.

Käyttäjien luomia luokituksia käytettiin twiittejä analysoitaessa opetusjoukkona ja analyysin pohjalta kunkin ehdokkaan mainitsevat twiitit jaettiin positiivisiin, negatiivisiin ja neutraaleihin. Tämän jälkeen viimeisen viiden minuutin twiittien analyysin tulokset visualisoitiin analyysin ohjausnäkymässä (kuvio 1).

Koska tutkimuksessa ei verrattu reaaliaikaisen analyysin tuloksia lopulliseen vaalitulokseen eikä laadittu ennustetta, ei Wangin ym. (2012) reaaliaikaisen

Kuvio 1: Wang ym. (2012) analyysin tulokset visualisoituna ohjausnäkymässä.



analyysin tulosten luotettavuuden arviointi ole mahdollista. Toisaalta tutkimuksessa myös todetaan ”järjestelmän kyvyn tuottaa reaaliaikaista analyysia vaalien alla tapahtuvista mielipiteiden muutoksista olevan erityisen merkityksellistä”. Wang ym. (2012) analyysi tarjoaakin lisäarvoa perinteisille mielipidekyselyille reaaliaikaisella asenneilmapiirin analyysillaan.

Ceronin ym. (2015b) väitteet iSA-metodin soveltuvuudesta reaaliaikaisen Twitter-analyysin suorittamiseen ovat vielä toistaiseksi vailla tutkittua faktapohjaa, mutta väite itsessään on mielenkiintoinen. Sekä iSA-metodi että Wangin ym. (2012) käyttämä analyysimetodi perustuvat opetusjoukon käyttöön analyysissa ja ohjattuun tunneanalyysiin pelkkään sanastoon vertaamisen sijaan. Tulevaisuudessa ohjattuun tunneanalyysiin perustuville metodeille on entistä enemmän käyttömahdollisuuksia käytettävissä olevan laskentatehon kasvaessa ja menetelmien kehittyessä.

Ceronin ym. (2015a, 2015b) saavuttamien tulosten perusteella ohjattu tunneanalyysi vaikuttaa kohtalaisen luotettavalta vaalien ennustamisen metodilta, joten mikäli metodia voidaan hyödyntää myös reaaliaikaisessa analyysissa Wangin ym. (2012) tavoin voidaan reaaliaikaiselta analyysilta odottaa merkittävää lisäarvoa perinteisten gallupien rinnalle. Aiheesta tarvitaan kuitenkin enemmän tutkimusta, jotta johtopäätöksiä reaaliaikaisten analyysimetodien toimivuudesta vaalien ennustamiseen voidaan tehdä.

5 YHTEENVETO

Tässä tutkimuksessa on käyty läpi lukuisia tutkimuksia, joissa on pyritty ennustamaan vaalien lopputulosta Twitteristä kerätyn datan perusteella. Tutkimuksissa on käytetty lukuisia eri metodeita niin tekstien analyysiin kuin ennusteiden laadintaankin. Tämän tutkielman tavoitteena oli selvittää vastaus kysymykseen, voidaanko Twitteristä kerättyä dataa käyttää vaaliennusteiden laadinnan työkaluna.

Kuten Gayo-Avello (2013) artikkelissaan nostaa esiin, ei yhdessäkään tutkimuksessa ole onnistuttu varsinaisesti ennustamaan vaalien tulosta, vaan Twitteristä kerätyn datan avulla on pyritty löytämään yhtäläisyyksiä toteutuneeseen vaalitulokseen. Toisaalta myös epäselvyys parhaista ennustamisen metodeista on vaikeuttanut ennusteiden laatimista ennakkoon, sillä asioiden syy-seuraussuhteita voidaan analysoida riittävän tarkasti vasta jälkikäteen. Täten vaalien jälkeen tehdyillä ennusteilla on myös arvoa parhaiden ennustemetodien löytämisessä.

5.1 Kuinka vaaleja voidaan ennustaa Twitter-datan pohjalta?

5.1.1 Tutkimuksissa käytetyt metodit ja niiden luotettavuus

Suurin osa tässä tutkielmassa käsitellyistä tutkimuksista on hyödyntänyt ennusteen laadinnassa yksinkertaista ehdokkaan tai puolueen mainintojen laskemista. Joissain tapauksissa, kuten Tumasjan ym. (2010), Jungherr ym. (2011) sekä Sanders ja van den Bosch (2013) on pelkkiä mainintoja laskemalla saavutettu lähes gallupien tarkkuuteen yltäviä tuloksia vaalien ennustamisessa. Toisaalta useissa tutkimuksissa, kuten esimerkiksi Vilares ym. (2015) ja Washington ym. (2015) ei ole löydetty minkäänlaista vahvaa korrelaatiota mainintojen ja toteutuneen vaalituloksen välillä. Pelkkiin mainintoihin perustuvissa ennusteissa satunnaisuus vaikuttaa olevan liian vahvasti läsnä ja joissain tapauksissa tuloksissa olevan liian suuria vääristymiä, jotta menetelmää voitaisiin pitää uskottavana vaalituloksen ennustamisen välineenä. Metodi on myös hyvin altis spämmin vääristävälle vaikutukselle, sillä ilman tehokasta spämmiviestien seulontaa jokainen botin lähettämä viesti vääristää mainintojen pohjalta laadittua oletettua mielipidejakaumaa.

Useissa tutkielmassa käsitellyissä tutkimuksissa oltiin myös käytetty yksinkertaista, koneellista tunneanalyysia twiittien analyysissa. Lampos (2012) esimerkiksi vertasi useita sanaston avulla tehdyn tunneanalyysin metodeja tutkimuksessaan, todeten ennusteen tarkkuuden paranevan synonyymisanaston käytön avulla. Sanaston avulla toteutetussa analyysissa on twiittien tulkitsemisessa merkittäviä haasteita, kuten ironian ja sarkasmin käyttö, jota koneellinen analyysi ei useinkaan tunnista. Washington ym. (2015) saavuttivat kuitenkin hyviä

tuloksia sanaston avulla toteutetulla tunneanalyysilla käyttämällä kaupallista Radian6-ohjelmistoa twiittien tunnesisällön arvioimiseen. Koska ohjelmiston toiminnasta ei ole saatavilla tarkkaa tietoa, on mahdotonta arvioida, miksi juuri kyseisellä ohjelmistolla on saavutettu hyviä tuloksia twiittien avulla ennustamisessa. Toisaalta kyse voi myös olla sattumasta.

Osa tunneanalyysia hyödyntävistä tutkimuksista, kuten Vilares ym. (2015) sekä Bermingham ja Smeaton (2011) ei löytänyt vahvaa korrelaatiota tunneanalyysin tulosten ja toteutuneen vaalituloksen välillä. Koska käytännössä jokainen tunneanalyysia hyödyntänyt tutkimus on toteuttanut tunneanalyysin eri tavalla, on tutkimusten tulosten ja metodien vertaaminen keskenään haastavaa. Toisaalta Lamposin (2012) ja Washingtonin ym. (2015) saavuttamia tuloksia voidaan pitää rohkaisevana merkkinä siitä, että tunneanalyysilla voidaan saavuttaa luotettavia tuloksia vaalien ennustamisessa. Varsinkin Washington ym. (2015) onnistuivat tutkimuksessaan osoittamaan, että tunneanalyysin tulokset olivat luotettava vaalituloksen ennustaja samalla, kun mainintojen määrä oli painottunut vahvasti toiseen ehdokkaaseen.

Yksinkertaisen, täysin koneellisen tunneanalyysin lisäksi on tutkimuksissa käytetty myös ohjattua tunneanalyysia. Wang ym. (2012) hyödynsivät reaaliaikaisessa analyysissaan Amazon Mechanical Turkin käyttäjien arvioimista twiiteistä laadittua opetusjoukkoa, mutta eivät laatineet analyysinsa pohjalta varsinaista vaaliennustetta. Varsinaisia ennusteita ohjattua tunneanalyysia käyttäen ovat laatineet ainoastaan Ceron ym. (2014, 2015a, 2015b). Tutkimuksissaan Ceron ym. ovat ennustaneet esimerkiksi Yhdysvaltain presidentinvaaleja, Italian puoluejohtajan valintaa ja Ranskan parlamenttivaaleja kohtuullisen hyvällä tarkkuudella. Pelkkään sanastoon vertaamiseen verrattuna ohjatulla analyysilla ja opetusjoukkoa käyttämällä saavutetaan parempia tuloksia esimerkiksi sarkastisten twiittien tunnistamisessa ja analysoinnissa, jolloin tunneanalyysin tulosten luotettavuus nousee.

Tunneanalyysin tulosten pohjalta laadituissa ennusteissa on käytetty erilaisia metodeja analyysin tulosten muuttamiseksi ennusteiksi. Ceron ym. (2015a) käyttävät koottua tunneanalyysia, jossa koko analysoitavan viestijoukon koottu tunnejakauma lasketaan määrittämään kaikkien joukon twiittien keskimääräistä tunnesisältöä. Toisaalta muissa tutkimuksissa on useimmiten käytetty yksittäisten twiittien tunnesisällön määrittämistä ja laskettu sen jälkeen koko joukon tunnesisältö yksittäisistä twiiteistä. Yksittäisten twiittien analyysissa mahdollisesti tapahtuneet luokitteluvirheet saattavat kertautua yhteen laskettaessa suuren hajonnan seurauksena (Ceron ym., 2015b). Toisaalta yksittäisten twiittien sisältämien virheiden pitäisi joissain määrin kumota toisensa ja analyysin tulosten hajonnan tulisi pienentyä suurempia joukkoja yhteen laskettaessa. Ennusteen laskennan parhaasta metodista ei siis voida olla täysin varmoja, mutta Ceronin ym. (2015a) käyttämän Hopkinsin ja Kingin metodin ja kootun tunneanalyysin tuloksia voidaan pitää lupaavina.

5.1.2 Twitteristä kerätyn datan ongelmat

Twitteristä kerätyssä datassa on merkittäviä haasteita verrattuna perinteisiin kyselytutkimuksiin. Taustamelun eli käsiteltävään aiheeseen liittymättömien, mutta silti aiheeseen liittyvän tunnisteen sisältävien twiittien määrä vaikeuttaa kaikkien ennustemetodien käyttöä. Muiden kuin pelkkien ihmiskäyttäjien määrä Twitterissä on yli 50%, kuten Chu ym. (2010) tutkimuksessaan totesivat. Lisäksi spämmi ja virheellistä informaatiota sisältävät viestit vaikeuttavat ennusteiden laadintaa. Käytettävä tutkimusaineisto on syytä seuloa hyvin ennen tutkimusta. Varsinkin mainintoihin ja täysin koneelliseen tunneanalyysiin perustuvien ennusteiden tulosten luotettavuuden kannalta on äärimmäisen tärkeää, että tutkimuksessa analysoitava data on mahdollisimman puhdasta.

Twitteristä kerätyn datan ongelmana on myös sen yleistettävyys koskemaan koko äänestäjäkuntaa. Choy ym. (2011, 2012) ratkaisivat ongelman arvioimalla datan pohjalta ainoastaan internetiä käyttävän kansanosan äänestyskäyttäytymistä ja tuomalla ennusteeseen mukaan myös edellisten vaalien tulokset. Toisaalta edellisten vaalien tulosten hyödyntäminen myös vääristää tutkimuksen tuloksia, luoden varjon Choyn ym. (2012) tulosten luotettavuuden ylle. Mitä suurempi äänestävän väestön Twitter-käytön prosentti on, sitä luotettavampana pelkän Twitter-viestidatan pohjalta laadittuja ennusteita voidaan pitää. Käytettävissä olevan datan määrän pitäisi siis korreloida myös tutkimusten tulosten luotettavuuteen. Käytettävissä olevan datamäärän pieneneminen taas vastaavasti korostaa analyysimetodien luotettavuuden merkitystä, nostaen esimerkiksi ohjatun tunneanalyysin arvoa ennusteiden laadinnassa parantuneen analyysitarkkuuden ansiosta.

5.2 Voidaanko vaaleja ennustaa Twitteristä kerätyn datan pohjalta?

Vastauksena tutkimuskysymykseen voidaan todeta, että vaalien ennustaminen Twitteristä kerätyn datan pohjalta on mahdollista. Twitterin pohjalta laadittujen ennusteiden ei kuitenkaan voida odottaa syrjäyttävän perinteisiä kyselytutkimuksia, mutta ne voivat tarjota lisäarvoa gallupien ohelle. Parhaan metodin tunnistaminen vaatii kuitenkin edelleen lisää tutkimuksia, mutta ohjattua tunneanalyysia käyttämällä saavutettujen tulosten valossa on todennäköistä, että ohjattu koottu tunneanalyysi osoittautuu tehokkaimmaksi tavaksi ennustaa vaaleja twiittien pohjalta.

Tulevissa tutkimuksissa olisi hyödyllistä hyödyntää ohjattua tunneanalyysia vaaliennusteiden reaaliaikaiseen laadintaan. Sekä Washingtonin ym. (2015), Wangin ym. (2012) että Ceronin ym. (2015b) tulosten perusteella suurten datamäärien käsittelyyn soveltuvia alustoja voitaisiin hyödyntää tunneanalyysin tekemiseen ja ennusteiden laatimiseen lähes reaaliajassa. Metodeilla olisi mahdollista laatia ennusteita vaalien lopputuloksesta esimerkiksi edellisen vuorokauden tai viikon twiittien perusteella ja seurata Wangin ym. (2012) tapaan myös

erilaisten kampanjatapahtumien vaikutusta Twitter-käyttäjien ehdokkaista ilmaisemiin mielipiteisiin. Vaaliennusteiden lisäksi tällaisilla työkaluilla olisi lisäarvoa myös vaalikampanjoille itselleen, sillä niiden avulla voitaisiin seurata sosiaalisen median reaktioita kampanjoiden toimintaan. Samalla sosiaalista mediaa voitaisiin hyödyntää edullisten vaaliennusteiden laadintaan ja tarjota lähes reaaliaikainen vaihtoehto verrattain hitaille ja kalliille kyselytutkimuksille.

5.3 Voidaanko Suomessa ennustaa vaaleja Twitterin pohjalta?

Suomessa on noin 471 000 Twitter-tiliä ja aktiivisia Twitter-käyttäjiä on enimmillään ollut yhden viikon aikana 58 000 (Suomi-Twitter, 2017). Aktiivisten käyttäjien määrä on kuitenkin laskenut viimeisen kahden vuoden aikana, pudoten 35 tuhanteen aktiiviseen twiittaajaan vuoden 2016 lopulla. Suomessa on siis aktiivisia twiittaajia noin prosentti väestöstä ja enimmilläänkin alle 10 prosentilla väestöstä on Twitter-tili. Tämä aiheuttaa merkittäviä ongelmia Twitterin pohjalta laadittaville ennusteille, sillä Twitteristä kerätyn datan yleistäminen koko äänestäjäkuntaa koskevaksi saattaa johtaa merkittäviin vinoutumiin aineiston alhaisen edustavuuden takia. Eduskuntavaaleissa 2015 annettiin lähes kolme miljoonaa ääntä, joten aktiivisten Twitter-käyttäjien osuus äänestäjistä on noin 1,5 prosenttia. Mikäli esimerkiksi vasemmistoliiton 211 000 äänestäjästä on Twitterissä aktiivisia 10%, on puolue edustettuna Twitter-aineistossa jo yli 35% osuudella verrattuna todelliseen 7 prosentin kannatukseen vaaleissa. Twitter-datan pohjalta laadittavissa ennusteissa on siis merkittävä tilastoharhojen mahdollisuus.

Toisaalta suomalaisten Twitter-käyttäjien puoluekannoista ei ole olemassa tutkimustietoa emmekä näin ollen voi olla varmoja, edustavatko suomalaiset Twitterin käyttäjät läpileikkausta suomalaisesta äänestäjäkunnasta. On siis mahdollista, että myös Suomessa Twitter-viestien tunneanalyysilla voidaan ennustaa ihmisten äänestyskäyttäytymistä perustuen heidän tuntemuksiinsa puolueita ja puoluejohtajia kohtaan. Vahvasti henkilökeskeisissä vaaleissa, kuten esimerkiksi presidentinvaaleissa 2018, on mahdollisuus kokeilla tunneanalyysin toimivuutta vaalien ennustamisen työkaluna myös Suomessa.

Tutkimuksessa olisi mahdollista hyödyntää SentiStrengthin olemassa olevaa suomenkielistä sanastoa analyysin työkaluna, mutta toisaalta aiempien tutkimustulosten perusteella ohjattu tunneanalyysi voisi olla toimiva metodi tunneanalyysin suorittamiseksi luotettavasti aineiston pieni koko huomioon ottaen.

Mikäli varsinainen vaaliennuste ei osoittautuisi luotettavaksi, tuottaisi reaaliaikainen tunneanalyysi kuitenkin kuvan suomalaisten Twitter-käyttäjien suhtautumisesta ehdokkaisiin ja samalla arvokasta tietoa heidän kampanjoilleen. Analyysin tulokset tuottaisivat siis ainakin lisäarvoa hitaasti reagoiville mielipidekyselyille.

LÄHDELUETTELO

Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results. Teoksessa Proceeding of IJCNLP conference, Chiang Mai, Thailand.

Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., & Riotta, G. (2014). A multi-level geographical study of Italian political elections from Twitter data. PloS one, 9(5), e95809.

Ceron, A., Curini, L., Iacus, S. M., & Porro, G. (2014). Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media & Society, 16(2), 340-358.

Ceron, A., Curini, L., & Iacus, S. (2015a). Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy. Social Science Computer Review, 33(1), 3-20.

Ceron, A., Curini, L., & Iacus, S. (2015b). Using social media to forecast electoral results: A review of state-of-the-art. Italian Journal of Applied Statatistics, 25(3), 237-259.

Choy, M., Cheong, M. L., Laik, M. N., & Shung, K. P. (2011). A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. Research Collection School Of Information Systems, Singapore Management University

Choy, M., Cheong, M., Laik, M. N., & Shung, K. P. (2012). US presidential election 2012 prediction using census corrected Twitter model. Research Collection School Of Information Systems, Singapore Management University

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2010, December). Who is tweeting on Twitter: human, bot, or cyborg?. In Proceedings of the 26th annual computer security applications conference (pp. 21-30). ACM.

Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. (2014, August). Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?. In Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on (pp. 620-627). IEEE.

Diebold, F. X. (2012). A Personal Perspective on the Origin (s) and Development of'Big Data': The Phenomenon, the Term, and the Discipline, Second Version. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. ACM sIGKDD Explorations Newsletter, 14(2), 1-5.

Gayo-Avello, D., Metaxas, P. T., & Mustafaraj, E. (2011). Limits of electoral predictions using twitter. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. Association for the Advancement of Artificial Intelligence.

Gayo-Avello, D. (2011). Don't turn social media into another 'Literary Digest' poll. Communications of the ACM, 54(10), 121-128.

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. Social Science Computer Review, 31(6), 649-679.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229-247.

Hopkins, D., King, G., Knowles, M., & Melendez, S. (2010). ReadMe: Software for automated content analysis. Institute for Quantitative Social Science.

IBM: What is big data? - Bringing big data to the enterprise. (17.3.2017). Haettu osoitteesta http://www-01.ibm.com/software/data/bigdata/

Jensen, M. J., & Anstead, N. (2013). Psephological investigations: Tweets, votes, and unknown unknowns in the republican nomination process. Policy & Internet, 5(2), 161-182.

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, 30(2), 229-234.

Kalampokis, E., Tambouris, E., & Tarabanis, K. (2013). Understanding the predictive power of social media. Internet Research, 23(5), 544-559.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4), 72.

Lampos, V. (2012). Detecting Events and Patterns in Large-Scale User Generated Textual Streams with Statistical Learning Methods. University of Bristol. Väitöskirja

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data. The management revolution. Harvard Bus Rev, 90(10), 61-67.

Metaxas, P. T., Mustafaraj, E., & Gayo-Avello, D. (2011, October). How (not) to predict elections. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 165-171). IEEE.

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. Proceedings of the ESWC 2011 Workshop on "Making Sense of Microposts": Big Things Come in Small Packages.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. ICWSM, 11(122-129), 1-2.

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.

Sanders, E., & Van Den Bosch, A. (2013). Relating Political Party Mentions on Twitter with Polls and Election Results. In DIR (pp. 68-71).

Shi, L., Agarwal, N., Agrawal, A., Garg, R., & Spoelstra, J. (2012). Predicting US primary elections with Twitter. URL: http://snap.stanford.edu/social2012/papers/shi.pdf.

Skoric, M., Poor, N., Achananuparp, P., Lim, E. P., & Jiang, J. (2012, January). Tweets and votes: A study of the 2011 singapore general election. In System Science (HICSS), 2012 45th Hawaii International Conference on (pp. 2583-2591). IEEE.

Suomi-Twitter. (17.3.2017). Haettu osoitteesta http://www.toninummela.com/suomi-twitter/

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544-2558.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2012). Where there is a sea there are pirates: Response to Jungherr, Jürgens, and Schoen. Social Science Computer Review, 30(2), 235-239.

Vilares, D., Thelwall, M., & Alonso, M. A. (2015). The megaphone of the people? Spanish SentiStrength for real-time analysis of political tweets. Journal of Information Science, 41(6), 799-813.

Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning (pp. 977-984). ACM.

Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

Washington, A. L., Thatcher, J. B., Morar, D., & LePrevost, K. (2015). What Is the Correlation between Twitter, Polls and the Popular Vote in the 2012 Presidential Election?(Correction). GMU School of Public Policy Research Paper No. 15-13

Weiss, S. M., & Indurkhya, N. (1998). Predictive data mining: a practical guide. Morgan Kaufmann.

Williams, C., & Gulati, G. (2008). What is a social network worth? Facebook and vote share in the 2008 presidential primaries. American Political Science Association.

Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., & Hu, C. (2016). Crowdsourcing based description of urban emergency events using social media big data. IEEE Transactions on Cloud Computing.

Yle Uutiset. (14.4.2011). Kokoomus paalupaikalta vaaleihin. Haettu osoitteesta http://yle.fi/uutiset/3-5343236