Loppuraportti Tekes-tutkimushankkeesta

SSMA:
Smarter Social Media Analytics

Kuluttajatutkimuskeskus ja Tietojenkäsittelytieteen laitos,
Helsingin yliopisto
2016-2018

Avainsanat: ilmiöt, heikot signaalit, sosiaalinen media, verkkokeskustelut, dynamiikka, systeemiteoriat, kuluttajien käyttäytyminen, laskennalliset yhteiskuntatieteet

Johdanto

Sosiaalisessa mediassa vahvistetaan ja rakennetaan yrityksiin, organisaatioihin ja brändeihin liittyviä käsityksiä ja jaetaan niihin liittyviä kokemuksia. Digitaalinen mediaympäristö tarjoaa mahdollisuuden seurata ja tutkia eri toimijoihin kohdistuvia arvioita, arvosteluja, kokemuksia ja tuntemuksia laskennallisesti. SSMA-hankkeessa kehitimme isojen verkkoaineistojen avulla menetelmiä keskusteluissa syntyvien ilmiöiden ja trendien automaattiseen, reaaliaikaiseen tunnistamiseen, sekä tutkimme toimintatutkimuksen keinoin sosiaalisen median analytiikan nykytilaa ja haasteita organisaatioissa.

Käytössämme olivat satojen miljoonien viestien laajuiset sosiaalisen median aineistot: Suomi24-verkkoyhteisön koko keskusteluaineisto sekä Futusome Oy:n keräämä miljardin viestin kokoinen aineisto suomenkielistä sisältöä eri sosiaalisen median palveluista. Näiden lisäksi hyödynsimme Taloustutkimus Oy:n keräämiä edustavia kyselytutkimusaineistoja sekä S-Ryhmän hanketta varten koostamaa aineistoa eri ruokatuotteiden menekistä. Aineistoja rinnastamalla sovelsimme ja kehitimme koneoppimismenetelmiä, joiden avulla nousevia trendejä ja ilmiöitä on mahdollista tunnistaa verkkokeskusteluista. Laskennallisen data-analyysin ja sitä tukevan laadullisen analyysin ohella hankkeessa kerättiin laadullista havainnointi- ja haastatteluaineistoa toimintatutkimuksellista otetta käyttäen.

Helsingin yliopiston Kuluttajatutkimuskeskuksen ja Tietojenkäsittelytieteen laitoksen yhteistyötahoina hankkeessa oli 11 yritystä: Aller Media, Taloustutkimus, Futusome, Atria, Ilmarinen, SOK, TeliaSonera, Arvo Partners, Leiki, Sometrik ja Underhood.co.

Projektin tulokset pähkinänkuoressa:
● **Sosiaalisen median keskusteluilla on ennustevoimaa.** Vegaanituotteista käytävät keskustelut selittävät muutoksia vegaanituotteiden myynnissä. ● **Alustoilla on merkitystä.** Viraali-ilmiön kestoa ennustaa parhaiten sen leviäminen heti alussa useille eri alustoille. ● **Fiksumpi analytiikka tarvitsee ihmistä.** Laskennallinen tekstianalyysi yhdistettynä ihmistulkintaa on paras keino tunnistaa nousevia aiheita ja keskusteluavaruuksia verkkokeskusteluista.

Kolme ohjenuoraa: mitä on #smartersome?
● **Älä aliarvioi ihmistulkintaa.** Sille on varattava aikaa, jos aineistosta haluaa liiketoimintahyötyjä. ● **Vietä päivä etnografina.** Selvitä oman toimialasi kannalta oleellisimmat areenat ja tavat mitata keskustelua. ● **Älä osta mustia laatikoita.** Kysy ja selvennä, mitä menetelmät tekevät. Kysy ja varmista, kunnes ymmärrät.

Lähtökohta: Sosiaalisen median analytiikka on lapsenkengissä

Yhteiskuntatieteessä ja viestinnän tutkimuksessa on pitkä tutkimustraditio joka pyrkii ymmärtämään, miten mediasisällöt rakentavat sosiaalisia ilmiöitä ja esittävät todellisuutta (e.g., Hilgartner & Bosk, 1988). Samalla tavalla meihin vaikuttavat sosiaalisesta mediasta luetut viestit, jotka voivat olla kenen tahansa kirjoittamia. Ilmiöt eivät siis tule kansalaisille annettuina, vaan he osallistuvat itse aktiivisesti erilaisiin viestinnällisiin prosesseihin joissa näitä ilmiöitä luodaan, muokataan, vahvistetaan tai heikennetään. Osa näistä prosesseista synnyttää ilmiöitä, jotka alkavat pienestä vuorovaikutuksesta mutta kasvavat lopulta viraaleiksi hiteiksi (e.g., Watts et al., 2007), joilla voi olla merkittäviä vaikutuksia yritysten tai koko yhteiskunnan toimintaan.

Mediaseurantatyökalujen vanavedessä kehittyneet sosiaalisen median seurantatyökalut ovat tällä hetkellä keskittyneet enimmäkseen erilaisiin asiasanahakuihin sekä käyttäjän saamiin osumiin ja yksinkertaiseen analytiikkaan. Tähän on kolme syytä. Ensinnäkin, sosiaalisen median ollessa yhä kohtalaisen uusi viestinnällinen muutos asiakasyritykset eivät useinkaan osaa vaatia enempää. Toisekseen, verkkojulkisuus on monimutkainen ympäristö, eikä siellä syntyvien ilmiöiden syntymekanismeja vieläkään kunnolla ymmärretä (cf. Gulbrandsen & Just, 2011; Laaksonen et al., 2012; Pöyry et al., 2017).

Kolmanneksi, menetelmäkehitys on mahdollistanut vasta viime vuosina sen, että isoja datamassoja voidaan erilaisin tietokoneavusteisin keinoin tutkia. Laskennallisen yhteiskuntatieteen parissa on viime vuosina rakennettu tapoja käsitellä hajaantuneita, isoja tietomassoja. Tutkimuksissa on tarkasteltu esimerkiksi Twitterin ja pörssikurssien yhteyttä (Bollen et al., 2011), turvallisuuskysymyksiä ja poliittista käyttäytymistä (DiGrazia et al., 2013, Parviainen et al., 2012, Nelimarkka et al., 2015), kulttuurisia konfliktitilanteita (Leetaru, 2011), epidemian leviämistä (Barboza, et al. 2013, Yangarber et al. 2007) sekä katastrofeja ja kriisiviestintää (Bruns & Llang, 2012).

Tutkimus- ja kehityshankkeemme tarttui suoraviivaisesti yllä esitettyihin haasteisiin sosiaalisen median analytiikan nykyisessä käytössä: yhdessä yritysten kanssa kehitimme analytiikkamenetelmiä sekä suomalaisyritysten ymmärrystä analytiikan mahdollisuuksista ja riskeistä.

Aineistot ja menetelmät

Hankkeessamme käytössä on sekä vuonna 2015 tutkimuskäyttöön avattu Aller Oy:n omistama Suomi24-verkkoyhteisön koko keskusteluaineisto (n. 67 miljoonaa viestiä vuosilta 2001-2016, ks. Lagus et al., 2015), sekä Futusome Oy:n keräämä noin miljardin viestin kokoinen aineisto suomenkielistä sisältöä eri sosiaalisen median palveluista (vuosilta 2001-2018). Käytimme hankkeessa rinnakkaisaineistoina Taloustutkimus Oy:n keräämiä edustavia kyselytutkimusaineistoja sekä S-Ryhmän hanketta varten koostamaa aineistoa elintarvikkeiden myyntimenekistä vuosilta 2012-2016.

Keskityimme hankkeessa temaattisesti ruokakeskusteluihin. Ne valikoituivat kohteeksi siksi, että vastaavat aineistot löytyivät sekä Taloustutkimukselta että S-Ryhmältä. Lisäksi ruokakeskusteluja on verkkokeskusteluissa runsain määrin.

Sosiaalisen median aineistojen analyysi. Laskennallinen analyysi mahdollistaa hankkeen käytössä olevien satojen miljoonien viestien kokoisen aineiston tutkimisen makrotasolla. Hankkeen aikana testasimme ja käytimme tekstinlouhinnan keinoja yhdistettynä koneoppimismenetelmiin (ks. Hastie et al., 2009), verkostoanalyysia (Huhtamäki & Parviainen, 2013) sekä aihetunnistusta (*topic modelling*). Laskennallisen analyysin tulosten ymmärtäminen ja kontekstualisointi vaatii kuitenkin ihmissilmän suorittamaa työtä (vrt. Grimmer 2015, Laaksonen et al. 2017). Yhtäältä käytettävissä olevien aineistojen sopivasti rajattu laadullinen analyysi tuottaa tiettyä ilmiökenttää koskevan sanaston, jonka avulla ilmiön nousua ja laskua voidaan mallintaa laskennallisesti. Esimerkiksi karppauksen kaltaiseen muotidieettiin viitataan lukuisilla nimityksillä. Toisaalta laadullinen analyysi pureutuu otoksiin, jotka on rajattu laskennallisen analyysin avulla, ja luo näin syvempää ymmärrystä trendien synnystä ja elinkaaresta. Siksi käsittelimme aineistoja myös laadullisti luokittelun ja laadullisen lähiluvun avulla.

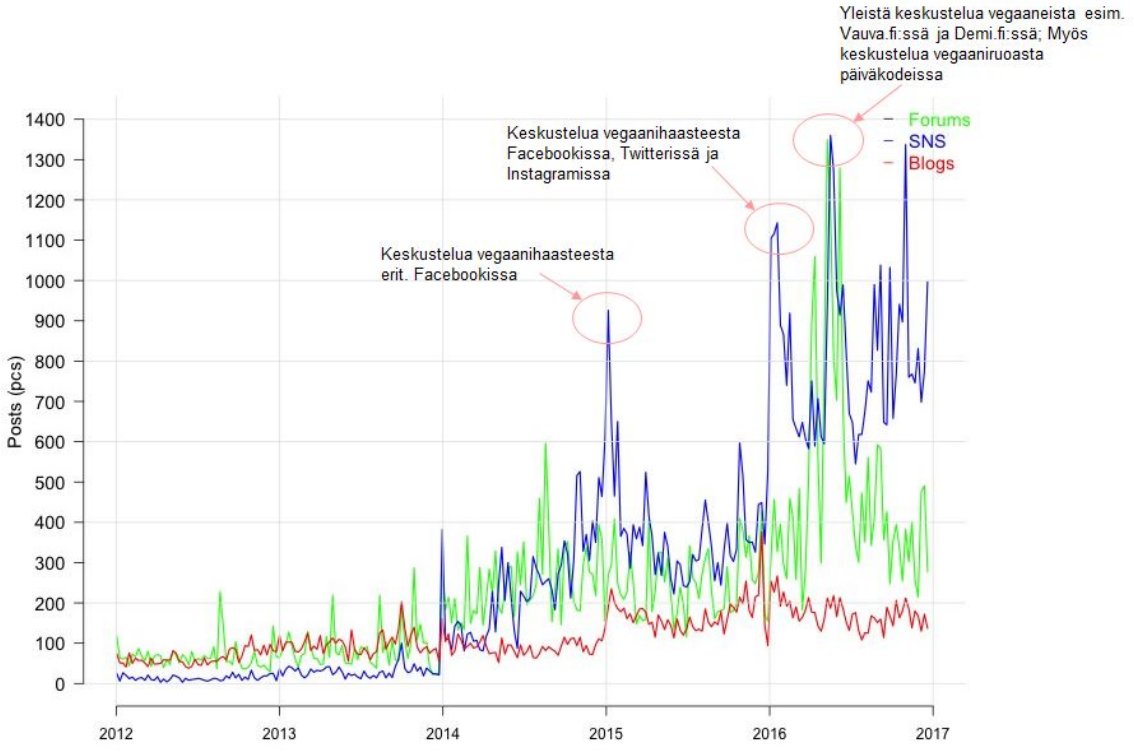
Toimintatutkimus data-analytiikkayrityksissä. Hankkeen aikana tutkijat jalkautuivat osallistuvan havainnoinnin keinoin neljän eri analytiikkayrityksen tiloihin työskentelemään ja seuraamaan yrityksissä tehtävää data-analytiikkaa. Lisäksi yritysten toimintaa ja käsityksiä data-analytiikasta ja datan arvosta tutkittiin havainnoimalla ja keräämällä tutkimusmateriaalia hankkeessa järjestettävistä työpajoista. Näin hankkeessa oli laskennallisen data-analyysin rinnalla toimintatutkimuksellinen ote. Toimintatutkimus on laadullisen tutkimuksen suuntaus, jolla pyritään analysoimaan ja kehittämään tutkimuskohdetta sen toimintatapoihin vaikuttamisen kautta (e.g., Kemmis & Wilkinson 1998). Toimintatutkimuksessa on keskeistä tutkijan osallistuminen toimintaan ja osallistuminen tutkittavan yhteisön toimintaan sen arjessa. Data-analyytikkojen käsitysten ja käytöksen tutkiminen on viime aikoina herättänyt mielenkiintoa myös kansainvälisesti (e.g., Carter & Sholler 2015; Beer, 2017), ja on välttämätöntä toimivien työvälineiden kehittämiseksi.

Lisäksi hankkeessa kerättiin analytiikan tueksi ja taustatiedoksi laadullinen haastatteluaineisto osallistuvista yrityksistä sekä muista data-analytiikka-alan suomalaisista toimijoista. Haastattelujen tarkoituksena oli kartoittaa yritysten tarpeita, toiveita ja nykytilaa sosiaalisen median analytiikan osalta, sekä analysoimaan niitä diskursseja, joita yrityksissä sosiaalisen median dataan liittyen käytetään.

Hankkeen keskeiset tulokset

1. Sosiaalisen median keskustelujen vaikutus vegaanituotteiden myyntiin

Suurin osa olemassa olevasta tutkimuksesta sosiaalisen median viestien ja kuluttajatuotteiden myynnin välisestä yhteydestä keskittyy yksittäisiin brändeihin sekä viesteihin, jotka mainitsevat kyseisen brändin. Tämä lähestymistapa ei sovi tilanteisiin, joissa halutaan tutkia nousevia kuluttajatrendejä. Kuluttajatrendejä ei voi aina tunnistaa yksittäisellä hakusanalla ja trendin vaikutus jakautuu epätasaisesti eri brändien ja tuotteiden välille. Tässä tutkimuksessa tutkitaan veganismia, nousevaa kuluttajatrendiä, heijastamalla sosiaalisen median viestejä koskien veganismi-ilmiötä, vegaanituotteita ja vegaanibrändejä vegaanituotteiden myyntiin. Analysoimalla vektoriautoregressiomenetelmällä viikoittaista myynti- ja sosiaalisen median viestidataa 2012 ja 2016 väliseltä ajalta tutkimus osoittaa, että tärkein myynnin muutoksia ennustava muuttuja on tuotetaso tarkoittaen niiden viestien määrää, jotka mainitsevat jonkin vegaanituotesanan, muttei välttämättä kyseisen tuotteeseen liittyviä brändinimiä (esim. tofu, kauramaito, vegaanipizza). Erityisesti blogit vaikuttivat olevan merkityksellinen kanava kuluttajien ostopäätösten kannalta. Se, kuinka paljon viestejä veganismista yleisesti sosiaalisessa mediassa oli (kuvassa), ei suoraan selittänyt vegaanituotteden myynnin muutoksia.



Kuva 1: Veganismiaiheiset viestit suomenkielisessä sosiaalisessa mediassa 2012-2016. Hakusanat: "vegaani", "vegaaninen", "vegaaniruoka", "veganismi".

2. Trendien semantiikka: Mitä laskennalliset menetelmät kertovat keskustelujen sisällöistä? Case karppaus

Trendien ennakoinnin lisäksi hankkeessa tutkittiin keskustelujen merkityssisällön laskennallisen mallintamisen mahdollisuuksia trendien luonteen ymmärtämiseksi. Tavoitteena oli kehittää menetelmä, jolla voitaisiin tehokkaasti tarkastella laajojen keskustelumassojen merkityssisältöjä ja tehdä tämän pohjalta päätelmiä keskustelussa eri aikoina ja alustoilla esiintyvistä merkityksistä.

Ohjaamaton koneoppiminen on viime aikoina ollut suosittu menetelmä yhteiskuntatieteellisessä tekstiaineistojen tutkimuksessa (ks. esim. Grimmer & Stewart 2013). Tutkimuksessamme sovelsimme aineiston rajaamiseen ohjattua koneoppimista ja tekstien mallintamiseen ohjaamatonta, sanavektorirepresentaatioita hyödyntävää Word2vec -menetelmää (Mikolov et al. 2013). Word2vec on neuroverkkoihin perustuva uusi tekstinmallinnusmenetelmä, joka mallintaa aineistossa esiintyvien sanojen merkitystä perustuen sanojen esiintymiskontekstiin. Menetelmä luo kullekin aineiston sanalle numeerisen vektorirepresentaation, joihin perustuva malli parhaiten ennustaa aineistossa sanojen vieressä esiintyviä sanoja. Näin menetelmällä pystytään mallintamaan sanojen merkitystä hienosyisemmin kuin esimerkiksi aihemallinnuksella (Blei 2012), jossa sanojen esiintymistä tarkastellaan kokonaisten viestien tasolla, kiinnittämättä huomiota esiintymiskontekstiin.

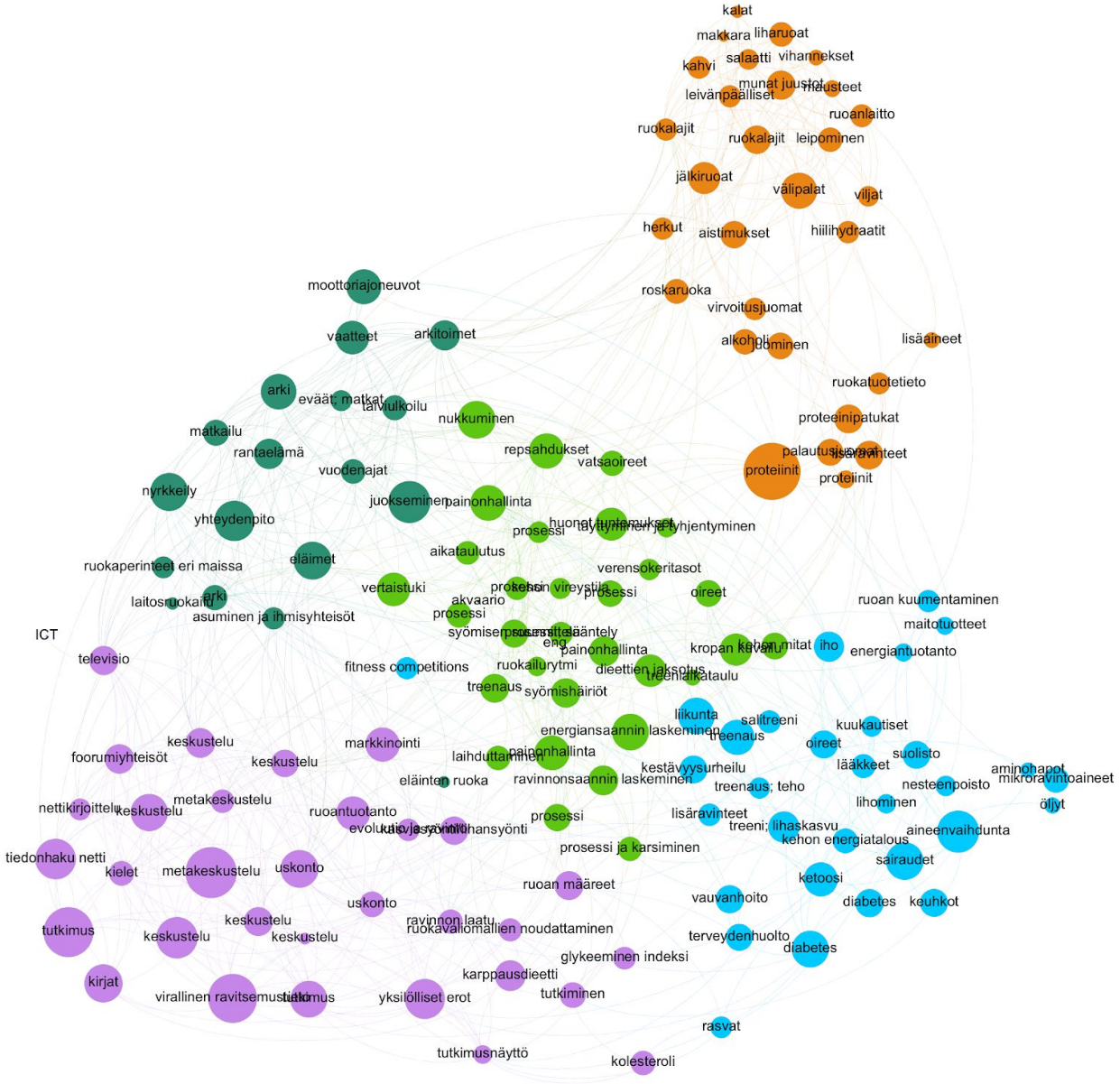
Tarkastelimme tapauksena vähähiilihydraattista ruokavaliota ja “karppausta” koskevaa keskustelua suomalaisessa sosiaalisessa mediassa, käyttäen aineistona Futusomen rajapinnasta ladattuja karppaukseen ja hiilihydraatteihin liittyviä viestejä. Lopullisessa aineistossamme oli yhteensä 428,838 uniikkia viestiä eri sosiaalisen median alustoilta (esim. Facebook, Twitter, erilaiset keskustelufoorumit ja blogit). Karppauskeskustelut olivat tarkoitukseemme sopiva tapaus, sillä karppauksen tiedetään olleen ilmiö Suomessa 2010-luvulla (Huovila 2014; Jauho 2016). Lisäksi ilmiö näkyi myös väestötasolla, mikä auttaa ilmiön huipun ajoittamisessa jotakuinkin vuosiin 2012-2013.

Tutkimuskysymyksemme liittyivät eri keskusteluissa esiintyvien merkitysten kartoittamiseen sekä Futusomen aineistossa havaitsemaamme mielenkiintoiseen ilmiöön. Kun karppausta ja hiilihydraatteja koskevan keskusteluaineiston määrä suhteutettiin koko Futusomen aineiston määrään eri vuosina, niin karppaus- ja hiilihydraattikeskustelussa näytti olleen toinen huippu jo ennen vuotta 2012, noin vuosina 2004-2005. Miksi karppaus oli trendi vuonna 2012, mutta ei 2005? Keskustelun merkityksiä erottelemalla etsimme vihjeitä ilmiön luonteesta eri vuosina, joilla trendin käyttäytymistä voisi ymmärtää.

Teimme Futusomesta ladatulle aineistolle Word2vec-mallinnuksen kahdella eri aikavälillä, jotka sisälsivät kaksi keskusteluhuippua (2003-2007, 2011-2013). Tämä tuotti molempien aikavälien aineistojen sanoille 200-ulotteiset numeeriset vektorirepresentaatiot, joita käyttäen voimme vertailla sanojen merkityksiä aikavälien sisällä. Klusteroimme kunkin aikavälin sanojen vektorit sanaryhmiksi käyttäen k-means -klusterointialgoritmia

(R-paketti *cclust*) ja annoimme kullekin sanaryhmälle tulkinnan siinä esiintyvien sanojen perusteella.

Alla on esitetty verkostorepresentaatio ensimmäisen aikavälin sanaryhmien suhteista. Kuvassa nimetyt solmut kuvaavat sanaryhmiä, jotka ovat verkostossa lähellä toisiaan, jos niiden merkitys on samanlainen. Solmujen koko kuvaa sanaryhmien “tiiviyttä”. Jos solmu on pieni, niin ryhmän sanojen merkitykset ovat Word2vec-mallin vektoriavaruudessa lähellä toisiaan. Suurissa solmuissa taas sanojen merkitykset ovat hajaantuneita. Solmujen värit vastaavat verkosta Gephi-ohjelmalla tunnistettuja solmujoukkoja. Samanväristen joukkojen merkitykset ovat keskenään samanlaisia. Alla kuvatussa verkossa on selkeyden vuoksi esitetty vain sanaryhmät, joiden merkitykset olivat vahvasti samanlaisia keskenään.



Verkostokuva havainnollistaa Word2vec-mallinnuksen mahdollisuuksia suurten tekstiaineistojen merkityksen tutkimisessa. Menetelmä tuottaa helposti tulkittavia ja

semanttisesti koherentteja sanaryhmiä, joiden merkityksiä voi vertailla toisiinsa esimerkiksi verkostorepresentaatioita käyttäen. Menetelmä soveltuu suurten tekstiaineistojen käsittelyyn ja mahdollistaa siten ilmiöihin liittyvien sosiaalisen median keskustelujen merkityssisältöjen pitkän aikavälin vertailevan tarkastelun. Työmääräällisesti rajoittavin tekijä on käsityö, jota sanojen merkityksiä tulkitessa tutkija joutuu tekemään. Sama rajoite tosin pätee muihinkin tekstiaineistojen ohjaamattomiin mallinnusmenetelmiin. Esimerkiksi aihemallinnukseen verrattuna Word2vec vaikuttaa tuottavan helposti tulkittavia tuloksia, joita voi lisäksi käyttää yksittäisten sanojen merkityksen tarkasteluun ja vertailuun aineistojen sisällä.

Tavoitteenamme oli vertailla kahden aikavälin aineistojen malleja toisiinsa ja näin etsiä eroja karppausta ja hiilihydraattejaa koskevalle keskustelulle eri aikoina. Vertasimme yllä esiteltyä ensimmäisen aikavälin aineiston verkostokuvausta toisen aikavälin aineiston verkkoon, mutta näillä kuvauksilla keskustelusta ei vielä löytynyt merkittäviä eroja. Seuraavina mahdollisina askelina tutkimuksessa on eritasoisten verkostokuvausten vertailu toisiinsa ja sanaryhmien merkitysten validoiminen suhteessa eri aikavälin keskustelun viesteihin.

3. Trendien tunnistaminen: Ruokatrendit 2017 sosiaalisessa mediassa

Trendeihin liittyvän keskustelun semantiikan tarkastelun lisäksi hankkeessa tutkittiin laskennallisen tekstianalyysin mahdollisuuksia vasta syntymässä olevien trendien tunnistamiseksi. Lähtökohtamme tässä tutkimuksessa oli hypoteesi, että trendiin liittyvien sanojen yleistyminen yli esiintymiskeskiarvon ennakoisi ko. trendin nousua. Näin tarkastelemalla sosiaalisen median keskusteluissa esiintyvien sanojen yleisyyttä voitaisiin kehittää automaattinen menetelmä trendin synnyn ja tyypin tunnistamiselle.

Pohjasimme kehittämämme menetelmän aikaisemmassa tutkimuskirjallisuudessa esiteltyihin sanojen yleisyyttä kuvaaviin mittareihin (Shamma et al. 2011). Näistä käytimme niin kutsuttua *normalisoitua termifrekvenssiä* (*ntf)*, joka kuvaa termin esiintymisfrekvenssiä jonkin aikavälin viesteissä suhteutettuna termin esiintymisfrekvenssiin koko tarkastellussa aineistossa. Kehitimme tätä mittaria ottamalla lisäksi huomioon kausivaihtelut eri termien käytössä (esim. “pääsiäinen”, “joulu”). Tavoitteenamme oli tunnistaa termejä, joiden aineistossa jollakin aikavälillä esiintyvä korkea ntf-arvo ei selittyisi kausittaisella vaihtelulla. Tämän kuvaamiseksi kehitimme *kausisuhteutetun normalisoidun termifrekvenssin (scntf, seasonality-corrected normalized term frequency)*, joka suhteuttaa termin ntf-arvot aikaisemman aineiston vastaavien aikavälien ntf-arvojen keskiarvoon. Lisäksi tarkastelimme termien käytön kehityksen kulmakerrointa.

Käytimme mittareiden kehityksessä aineistona Futusomesta ladattua ruokaa koskevaa keskustelua vuosilta 2014-2017. Laskimme kuukausittaisen ntf-arvon ja kulmakertoimen vuoden 2017 aineiston (yht. 942,603 uniikkia viestiä) kullekin sanalle. Lisäksi laskimme

kuukausittaisen ntf-arvon vuosien 2014–2016 aineiston (3,531,778 viestiä) sanoille, jota käytimme laskeaksemme vuoden 2017 sanojen scntf-arvon.

Näin saimme vuoden 2017 ruokakeskustelussa käytetyille sanoille kuukausittaiset scntf-arvot ja sanojen käytön kulmakertoimet. Tällä tarkastelulla voidaan erottaa toisistaan eri tavalla käyttäytyviä trendejä. Esimerkiksi keskustelussa terävinä huippuina näkyvillä “kohutermeillä” on huippukuukautena suuri scntf-arvo, mutta pieni kulmakerroin. Erityisen kiinnostavia sanoja trendien tunnistamisen näkökulmasta olisivat sanat, joiden käyttö on kasvavaa ja jatkuvasti suurempaa kuin aikaisemmassa aineistossa. Näillä sanoilla olisi huippukuukautena suuri scntf-arvo, sekä lisäksi suuri kulmakerroin. Tarkastelimme sanojen esiintymistä vuoden 2017 ruokakeskusteluissa, mutta selkeitä esimerkkejä tämänkaltaisista trendisanoista ei löytynyt. Tämä voi johtua aineiston kokoon liittyvistä rajoituksista tai tarkastelussa käytetystä aikaikkunasta (kuukausi). Tulevaisuudessa olisikin kiinnostavaa kokeilla ja kehittää mittaria edelleen käyttäen eri aineistoja ja lisäten vaihtelevasti mittariin lisätekijöitä. Kehittämämme menetelmä toimii automaattisesti ja on lisäksi suurillakin aineistoilla tehokas. Näin ollen sanojen esiintymisen tarkastelu vaikuttaa lupaavalta tavalta tunnistaa keskusteluaineistosta syntyviä trendejä.

4. Etnografin työpöydältä: Mitä on fiksumpi sosiaalisen median analytiikka?

Osana projektia tutkimme myös data-analytiikan todellisuutta organisaatioissa etnografisen havainnoinnin keinoin, sekä haastattelemalla mukana olevia yrityksiä. Seuraavat neljä kohtaa tiivistävät laadullisen tutkimuksen keskeisimmät havainnot. Niiden tausta-aineistona on myös käytetty hankkeen työpajoissa tehtyjä ryhmätehtäviä ja kerättyä materiaalia.

1. Fiksumpi sosiaalisen median analytiikka on ihmisen ja koneen yhteistyötä

Sosiaalisen median analytiikkaan — ja tekoälykeskusteluun laajemminkin — liittyy vahvasti laskennallisuuden rationalisointi ja ns. big data -myytti (Desrosières, 2001; Couldry, 2014; Beer, 2017): mikä tahansa numeroiksi muunnettava tieto, jota voidaan käsitellä algoritmisesti, on automaattisesti luotettavaa ja totta. Näin on varsinkin, jos taustalla on isoja aineistoja eli kaikkien himoitsemaa big dataa. Todellisuudessa automaattinen analytiikka vaatii yleensä algoritmin opettamista ja yhteistyötä ihmisen kanssa. Opettaminen tapahtuu esimerkiksi luokittelemalla useita satoja tai tuhansia esimerkkiviestejä halutun kysymyksen mukaisesti. Siksi opetusaineisto vaikuttaa suuresti siihen, minkälainen koneoppijasta tulee. Matemaatikko ja data scientist Cathy O’Neil (2016) varoittaakin algoritmien vinoutumisesta: algoritmit automatisoivat status quo -tilaa, sillä ne rakentuvat aina historiallisen datan ja sen rakenteen päälle. Maailma ei ole täydellinen, ja sen epätäydellisyys heijastuu myös koneoppimiseen ja tekoälyyn. Siksi rinnalle tarvitaan ihmisajattelua arvioimaan algoritmien oikeellisuutta ja vaikutuksia.

2. Fiksumpi someanalytiikka vaatii mietittyä datan esikäsittelyä

Automaattiseen tekstianalytiikkaan piiloutuu paljon valintoja. Niiden tekeminen alkaa aineiston rajauksesta: harvoin on laskentaresursseja tutkia kaikkea saatavilla olevaa dataa, joten se pitää ensimmäiseksi rajata hakusanoilla. Hakusanojen kehittely vaatii usein sekin ihmisasiantuntijaa. Analyysialgoritmit vaativat usein myös aineiston esikäsittelyä. Suomen kielen kohdalla se tarkoittaa esimerkiksi aineiston perusmuotoistamista, joka vie aikaa ja resursseja. Lisäksi tekstimassasta poistetaan tyypillisesti yleisimmät, merkityksettömät sanat eli ns. stopwordit. Niiden poistaminen on kuitenkin samalla myös valinta siitä, mikä on merkityksellistä ja mikä ei. Kiveen hakattuja ohjeita tai yleisesti hyväksyttyä listaa ei kuitenkaan ole olemassa, vaan ratkaisuja tehdään tapauskohtaisesti. Tiedossa on, että poistettujen sanojen lista vaikuttaa lopulliseen analyysiin, mutta on epäselvää millä tavoin.

Oleellista on myös ymmärtää käytössä olevan datan mahdolliset rajoitukset ja niiden vaikutukset analyysiin. Esimerkiksi tutkimuskäyttöön luovutettu Suomi24-aineisto on periaatteessa koko aineisto, mutta tietokantavirheen vuoksi aineistosta puuttuu paljon viestejä vuosilta 2004-2005. Tällainen kuoppa näkyy jokaisessa aineistosta piirrettävässä aikajanassa, ja sitä tuijottaessaan tutkija tulee helposti tehneeksi virheellisiä tulkintoja keskusteluaiheen katoamisesta ellei aineiston koostumus ole tiedossa.

3. Fiksumpi sosiaalisen median analytiikka tarvitsee ymmärrystä alustoista ja niiden kulttuureista

Laskemisen ja big datan huumassa on helppoa unohtaa laadullisen analyysin ja kulttuurisen ymmärryksen merkitys. Sosiaalisen median keskusteludata on hyvin kontekstuaalista dataa, jonka syntymiseen vaikuttaa paitsi yhteiskunta ympärillä, myös alustan teknologia ja kyseiselle alustalle muodostunut alakulttuuri. Palstoille voi esimerkiksi syntyä oma slangi ja hyvinkin erikoistunutta sanastoa. Suomen kielen käsittelijä ei välttämättä tunnista verkossa syntyviä uudissanoja saatika tuttujen sanojen erikoisia käyttötapoja. Esimerkiksi keppihevonen tarkoittaa toisaalla oikeasti keppihevosta, mutta toisaalla tietynlaista poliittista diskurssia.

Lisäksi automaattisen tekstianalytiikan on osoitettu olevan hyvin kontekstiriippuvaista. Erot tulevat ilmi varsin pienissäkin muutoksissa: Yhdysvalloissa senaatin ylähuoneen puheesta koostuvalla aineistolla koulutettu luokittelualgoritmi ei enää toimikaan alahuoneen puhetta analysoitaessa (Yu et al., 2008). Vuoden 2005 ruokapuhetta käsittelevä algoritmi tuskin pärjää tarpeeksi hyvin vuoden 2015 uuden kielen ja sanaston kanssa.

Myös monet teknologian tuottamat artefaktit muodostuvat hankalaksi automaattiselle analytiikalle. Esimerkiksi monella keskustelufoorumilla viestit lähetetään anonyymisti, jolloin kirjoittajana näkyy “Vierailija”. Kuin vierailija vastaa näihin vierailijan viesteihin lainaamalla niitä, syntyy ketjuja, joissa on hämmentävän monta kertaa mainittu sana vierailija. Lopputuloksena esimerkiksi ohjaamaton aihemallinnus erottaa datasta aiheen,

jossa puhutaan kovasti vierailijoista. Sen todellinen olemus ei avaudu kuin esimerkkiviestejä lukemalla.

4. Fiksumpi sosiaalisen median analytiikka on vähemmän mustia laatikoita

Viimeinen ja ehkä tärkein fiksumman sosiaalisen median analytiikan väittämä liittyy analytiikan tekemiseen ja palveluiden ostamiseen. Ala rakentuu tällä hetkellä hämmentävän vahvasti erilaisten mustien laatikoiden ympärille; käytössä on teknologioita ja algoritmeja, jotka on hienosti paketoitu tekoälyksi, mutta todellisuudessa niiden takana ovat samat kontekstiin, kieleen ja validiteettiin riippuvat ongelmat kuin yllä mainituissa esimerkeissä. Monet organisaatiot mittaavat esimerkiksi Facebookista suoraan saatavaa engagement-lukua ymmärtämättä täysin, mistä siinä oikeastaan on kysymys. Analytiikkayrityksen kauppaama keskustelun sentimenttiä kuvaava hieno piirakkadiagrammi ostetaan tyytyväisenä kyseenalaistamatta analyysissa käytettyä algoritmia.

Tämä ei tarkoita, että kaikki tehty automaattinen analytiikka olisi automaattisesti virheellistä. Mutta se tarkoittaa sitä, että analytiikan tekijöiltä vaaditaan lisää avoimuutta käytettyjen menetelmien sekä niiden heikkouksien suhteen sekä sitä, että analytiikan ostajat osaavat kysyä tarkentavia kysymyksiä mustan laatikon sisuksista. Kysymys on lopulta kielenkäytöstä: samalla tavalla kuin lääkärin on osattava selventää diagnoosi potilaalle, on datatieteilijän ja analytiikkayrittäjän osattava selittää analyysin kulku kansankielellä asiakkaalleen. Siksi hankkeen keskeisimpänä oppina peräänkuulutamme lisää kriittisyyttä sosiaalisen median analytiikkaan; sekä sen tekijöiltä että ostajilta.

Tutkimusryhmä

- Hankkeen tieteellsenä johtajana toimi Helsingin yliopiston (HY) Kuluttajatutkimuskeskuksen ylopistotutkija Mikko Jauho ja CS/HIIT-rinnakkaisprojektin johtajana dosentti, FT **Antti Salovaara**.
- Hankkeen koordinaattori: **Salla-Maaria Laaksonen** (HY)
- Tutkijat: Veikko Isotalo, Piia Jallinoja, Arto Kekkonen, Matti Nelimarkka, Juho Pääkkönen, Essi Pöyry

Hankkeen johtoryhmä:

- Puheenjohtaja: **Kristina Hännikäinen** (Aller Media).
- Jäsenet: **Ville Henttonen** (Sonera), **Aleksi Kallio** (CSC – Tieteen Tietotekniikan Keskus), **Camilla Magnusson** (Underhood), **Juho Muhonen** (Futusome), **Tomi Näres** (Arvo Partners), **Mika Pantzar** (KTK), **Olli Parviainen** (Sometrik), **Mira Pakarinen** (Ilmarinen), **Petrus Pennanen** (Leiki), **Kari Roose** (Taloustutkimus), **Minna Ruckenstein** (KTK), **Jukka Saarenpää** (Atria), **Maaret Valli** (SOK).

Lisätietoja projektista ja tulevista julkaisuista myös hankkeen päättymisen jälkeen: https://blogs.helsinki.fi/smartersocialmedia/julkaisut

Kirjallisuus:

Barboza P, Vaillant L, Mawudeku A, Nelson NP, Hartley DM, Madoff LC, Linge JP, Collier N, Brownstein JS, Yangarber R, Astagneau P. (2013). Evaluation of epidemic intelligence systems integrated in the early alerting and reporting project for the detection of A/H5N1 influenza events. PLoS One Journal, 8(3).

Beer, D. (2017). Envisioning the power of data analytics. Information, Communication & Society, 21(3), 1–15.

Blei, D. (2012). Probabilistic Topic Models. *Communications of the ACM,* 55(4).

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

Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. First Monday, 17(4).

Carter, Daniel, and Dan Sholler. (2015). Data Science on the Ground: Hype, Criticism, and Everyday Work. *Journal of the Association for Information Science and Technology*.

Couldry, N. (2014). The Myth of Big Data. Teoksessa: Schäfer, M. T., & Van Es, K. (Eds.). *The datafied society : studying culture through data*. Amsterdam: Amsterdam University Press. Retrieved from http://oapen.org/search?identifier=624771

Desrosières, A. (2001). How Real Are Statistics? Four Possible Attitudes. *Social Research*, 68(2), 339–355.

DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. Foot (Eds.), Media Technologies: Essays on Communication, Materiality, and Society (pp. 167–194).

Grimmer, J. & Stewart, B.M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis* 21(3).

Gulbrandsen, I. T., & Just, S. N. (2011). The collaborative paradigm: towards an invitational and participatory concept of online communication. Media, Culture & Society, 33(7), 1095–1108.

Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The Elements of Statistical Learning*. *Elements*. Springer Series in Statistics. New York, NY: Springer New York.

Hilgartner, S., & Bosk, C. L. (1988). The rise and fall of social problems: A public arenas model. American journal of Sociology, 53-78.

Huhtamäki, J. & Parviainen, O. (2013) Verkostoanalyysi sosiaalisen median tutkimuksessa. Teoksessa Salla-Maaria Laaksonen, Janne Matikainen & Minttu Tikka (toim.) *Otteita verkosta - Verkon ja sosiaalisen median tutkimusmenetelmät*. Tampere: Vastapaino, s. 245–273.

Huovila, J. (2014). Kansa ei enää tottele: Karppaus individualistisen ja universalistisen ravitsemuspuheen ristiaallokossa Helsingin Sanomissa vuosina 2010–2012 [The low-carbohydrate diet movement in the Helsingin Sanomat newspaper 2010–2012]. Sosiaalilääketieteellinen Aikakauslehti, 51(1), 18-31. https://journal.fi/sla/article/view/41365

Jallinoja, P, Jauho, M, Mäkelä, J (2016, published ahead of print) Newspaper debates on milk fats and vegetable oils in Finland, 1978 – 2013: An analysis of conflicts over risks, expertise, evidence and pleasure. *Appetite*. doi:10.1016/j.appet.2016.05.035

Jauho, M. (2016). The social construction of competence: Conceptions of science and expertise among proponents of the LCHF-diet in Finland. *Public Understanding of Science* 25(3), 332–345.

Kemmis, S., & Wilkinson, M. (1998). Participatory action research and the study of practice. Teoksessa Atweh, B., Kemmis, S., & Weeks, P. (Eds.). *Action research in practice: Partnership for social justice in education*. Routledge.

Laaksonen, S-M.; Falco, A.; Salminen, M.; Aula, P.; Ravaja, N.; Ainamo, A., & Neiglick, S. (2012). *Digital Reputation. Characterizing and measuring reputation, reputation risk, and emotional responses to reputation in digital publicity*. Final Report. Media and Communication Studies Research Reports 2/2012. Communication Research Centre CRC: Helsinki.

Lagus, K; Pantzar, M. & Ruckenstein, M. (2015). Keskustelun tunneaallot – Suomi24-hanke. *Tieteessä tapahtuu* 33(6).

Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. First Monday, 16(9).

O'Neill, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Allen Lane.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781

Parviainen, O., Poutanen, P., Laaksonen, S-M., & Rekola, M. (2012). Measuring the effect of social connections on political activity on Facebook. Paper presented at OII Internet, Politics, Policy:'Big Data, Big Challenges, September 2012.

Shamma, D., Kennedy, L. & Churchill, E. (2011). Peaks and Persistence: Modeling the Shape of Microblog Conversations. *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*.

Van Dijck, J., & Poell, T. (2013). Understanding Social Media Logic. Media and Communication, 1(1), 2. doi:10.17645/mac.v1i1.70

Watts, D. J., Peretti, J., & Frumin, M. (2007). Viral marketing for the real world. Harvard Business School Pub. Chicago.

Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, 5(1), 33–48.