Mikko Koponen & Virpi Hotti

Datafikaatio

Esimerkkinä Jukolan viestiin ja Ilosaarirockiin osallistuneiden psykografiset tiedot sekä keskiarvoistettujen sanavektorien klusterointi

KARELIA-AMMATTIKORKEAKOULU

Datafikaatio

Esimerkkinä Jukolan viestiin ja Ilosaarirockiin osallistuneiden psykografiset tiedot sekä keskiarvoistettujen sanavektorien klusterointi

Tekijät: Mikko Koponen & Virpi Hotti (Itä-Suomen yliopisto, tietojenkäsittelytieteen laitos)
Sivuntaitto: Kaisa Varis
Kansikuva: Nicolas LB / Unsplash
Kustantaja: Karelia-ammattikorkeakoulu & KoDa – datan kokonaisvaltainen hallinnointi ja hyödyntäminen -hanke, 2019

ISBN:978-952-275-284-0

Karelia AMMATTIKORKEAKOULU

UNIVERSITY OF EASTERN FINLAND

BUSINESS JOENSUU

JOE WITH

Euroopan unioni
Euroopan aluekehitysrahasto
Euroopan sosiaalirahasto

Vipuvoimaa EU:lta 2014–2020

Elinkeino-, liikenne- ja ympäristökeskus

Sisällys

Esipuhe

Tietoa on saatavilla tänä päivänä enemmän kuin koskaan aiemmin. Digitaalisessa maailmassa tietoa eli dataa on aiempaa helpompi kerätä, mutta pelkkä datan kerääminen ei riitä, vaan sitä on osattava analysoida ja hyödyntää. Teknologian muuttuessa yhä edullisemmaksi ja analysointityökalujen kehittyessä, analyysiin voidaan käyttää yhä laajempaa data-aineistoa, jolloin myös tulokset ovat yhä tarkempia ja kattavampia.

Tässä selvityksessä on kuvattu, miten ihmisten kulutusmieltymyksistä saadaan tietoa heidän sosiaaliseen mediaan kirjoittamansa sisällön perusteella. Datasetit haettiin Futusomen tutkijapalvelusta ja data analysoitiin Microsoftin Power BI –sovelluksella. Data-aineistoksi valittiin Pohjois-Karjalassa pidettyihin suurtapahtumiin Ilosaarirockiin ja Jukolan viestiin liittyvät keskustelut seuraavissa sosiaalisen median kanavissa: Instagram, Twitter, Facebook, News, Forum ja Blog. Analyysin tuloksena saatua tietoa kirjoittajien kulutusmieltymyksistä voisi hyödyntää esimerkiksi tulevien tapahtumien oheistuotteiden tai -tapahtumien myynnissä ja markkinoinnissa.

KoDa – Kokonaisvaltainen datan hallinnointi ja hyödyntäminen (ESR 2017-2019)[1] on Karelia-ammattikorkeakoulun ja Itä-Suomen yliopiston yhteinen hanke, jonka tavoitteena on pienten ja keskisuurten yritysten kasvun tukeminen. KoDa tarjoaa yrityksille koulutusta datan käytöstä liiketoiminnan kehittämiseen sekä pienyrittäjille soveltuvia työkaluja datan hallinnoinnin ja hyödyntämisen tueksi. Lisäksi KoDassa on muodostettu kehittämisyhteisö, jossa datan hyödyntämisestä kiinnostuneet alueen toimijat voivat verkostoitua ja jatkaa yhteistyötä myös hankkeen päättymisen jälkeen. Hankkeen myötä Pohjois-Karjalan alueella toimivat yritykset tunnistavat datan keräämisen ja analysoinnin hyödyt, heidän tietämyksensä yrityksen omasta sekä avoimesta datasta kasvaa, ja he osaavat hyödyntää olemassa olevia työkaluja datan analysoinnissa. Tämän selvityksen toteuttivat Itä-Suomen yliopiston erityisasiantuntija Mikko Koponen ja lehtori Virpi Hotti.

Säde Lind, projektipäällikkö
Karelia-ammattikorkeakoulu

[1] https://www.eura2014.fi/rrtiepa/projekti.php?projektikoodi=S20980

1 Johdanto

Digitalisaation seuraajana pidetään datafikaatiota (datafication). Datafikaatiota pidetään jopa vallankumouksellisena mahdollisuutena tutkia ihmisen käyttäytymistä. Erilaisista digitaalisista alustoista, kuten Facebookista ja Twitteristä, kerätyn datan avulla saadaan tietoa ihmisten välisistä suhteista ja voidaan seurata ihmisten käyttäytymistä sekä asenteita. (Cukier & Mayer-Schoenberger, 2013)

Datafikaatio on sosiaalisen toiminnan muuttamista suoraan määrälliseksi tiedoksi, mikä mahdollistaa reaaliaikaisen seurannan ja ennakoivan analyysin (”the transformation of social action into online quantified data, thus allowing for real-time tracking and predictive analysis”). (Van Dijck, 2014 mukaan, Mayer-Schoenberger & Cukier, 2013)

Datafikaation laajempi määritelmä ei rajoitu pelkästään ihmisten sosiaalisen toiminnan siirtämiseen kvantitatiiviseksi dataksi, vaan sama ajatus yleistetään koskemaan mitä tahansa fyysisiä entiteettejä (entities) ja niiden tuottamaa tai niitä kuvaavaa dataa. Tämä tarkoittaa käytännössä digitaalisen teknologian hyödyntämistä fyysisistä entiteeteistä olemassa olevan tiedon erottamiseksi itse entiteeteistä – entiteettiä koskeva data kytketään tietyssä mielessä käsitteellisesti irti fyysisen entiteetin fyysisestä ilmentymästä. (Ericsson, 2014)

Tässä selvityksessä pyrimme selkiyttämään digitalisaation ja datafikaation eroavaisuuksia (Luku 2) sekä sitä, kuinka digitalisaatio on edesauttanut datafikaatiota. Datafikaatiossa psykografisilla tiedoilla on keskeinen merkitys. Psykografisista tiedoista havainnollistamme persoonallisuustyyppejä ja niiden faseteja sekä kulutusmieltymyksiä. Käytämme Futusomelta (2018) ostettuja datasettejä, kuten Ilosaarirockin datasetti (Luku 3), kun on selvitelty tapahtumakävijöiden psykografisia tietoja (Luku 4). Lisäksi selvityksessä on toteutettu analyysi, jossa sosiaalisen median käyttäjiä on pyritty erottelemaan toisistaan sen aihepiirin mukaisesti, josta he ovat kirjoittaneet. Kirjottajien erottelussa on hyödynnetty sanavektorimenetelmiä sekä klusterointia, joiden menetelmät ovat kuvattuna luvussa 5. Analyysin tulokset ovat esitettynä luvussa 6.

2 Digitalisaatiosta datafikaatioon

Tässä luvussa on käyty läpi, kuinka kolme merkittävää digitaalisen teknologian trendiä ovat myötävaikuttaneet datafikaation syntyyn. Aliluvussa 2.1 esitellään digitalisaation vaikutuksia datan määrälliseen ja laadulliseen kehitykseen. Aliluvussa 2.2 käydään läpi laskentatehon määrän ja hinnan kehitystä. Lopuksi vertaillaan digitalisaatiota ja datafikaatiota aliluvussa 2.3.

2.1 Digitalisaatio ja data

Nykyisin tietoverkot ja datakeskukset kattavat suuren osan maailmasta, mikä mahdollistaa ihmisten keskinäisten lähes välittömän viestinnän lähes mistä tahansa, minne tahansa. Digitaalisten tietoverkkojen yleistyminen on vaikuttanut suuresti siihen, kuinka ihmiset ja yritykset koordinoivat keskinäistä toimintaansa (Ericsson; Imperial College London; Sustainable Society Network, 2014).

Ajan ja erityisesti paikan merkitys kommunikaatiolle on vähentynyt. Ihmisten keskinäinen kommunikointi ei enää vaadi esimerkiksi kasvokkain näkemistä tai postin toimittamaa fyysistä paperia - mielivaltaisen välimatkan päästä tapahtuvan kommunikoinnin voisi katsoa olevan lähes ilmaista.

Digitalisaation ja tietoverkkojen yleistymisen seurauksena dataa on pyritty erottamaan fyysisestä mediasta kuten optisista levyistä, mikä edesauttaa datan saatavuutta. Saatavuus parantaa olennaisesti datan käyttöarvoa; data, johon ei pääse käsiksi, on verrannollisesti hyödyllisyydeltään samaa luokkaa kuin raakaöljy, jota ei ole vielä pumpattu öljyesiintymästä säiliöön. Internetin ulottumattomiin säilötyllä datalla on potentiaalia tulla tulevaisuudessa hyödynnetyksi, mutta tulevaisuuden arvon realisointi vaatii toimenpiteitä.

Digitaalisen teknologian ja erityisesti digitaalisten tietoverkkojen yleistyminen on lisännyt huomattavasti ihmiskunnan tuottaman datan määrää. On arvioitu, että Internetin sisältämän datan kokonaismäärä kaksinkertaistuu 18 kuukauden välein. Lisäksi 90 % maailman kaikesta saatavilla olevasta datasta on syntynyt viimeisen kahden vuoden aikana. Ennusteiden mukaan vuoteen 2030 mennessä maailmassa olisi yli triljoona sensoria. (Viitanen et al., 2017)

Mikäli oletetaan, että datan määrän lisääntyminen jatkaa kasvuaan arvioissa esitettyjen matemaattisten lainalaisuuksien mukaisesti, se tarkoittaa, että datan määrä kasvaa eksponentiaalisesti. Jos otetaan esimerkiksi 15 vuoden tarkastelujakso, johon sisältyy 10 kappaletta 18 kuukauden mittaisia jaksoja, kasvaisi Internetin saavutettavissa oleva datamäärä $1 * 2^{10} = 1024$ kertaiseksi. Toistaiseksi on epävarmaa, kuinka suuri osa tästä datasta tulee olemaan sensoreiden tuottamaa, ja kuinka suuri osa muuta dataa.

Jo vuonna 2017 on esiintynyt arvioita siitä, että data olisi arvokkaampi resurssi kuin öljy (The Economist, 2017). Tällaiset arviot, olivat ne täysin tarkkoja tai eivät, kuvastavat datan jo saavuttamaa ja alati kasvavaa arvoa. Datan määrä, saatavuus ja hyödynnettävyys on kasvanut tasolle, jossa sen merkitys maailmantaloudelle on merkittävä, ellei jopa jossain määrin dominoiva – kuten öljyn tapauksessa on.

Videodatan suhteellinen ja absoluuttinen määrä Internetissä on lisääntynyt viime vuosina voimakkaasti, ja on arvioitu, että esimerkiksi vuonna 2017 videodatan osuus kaikesta verkkoliikenteestä oli 75 %. Videoliikenteen suhteellisen osuuden verkkoliikenteestä arvioidaan olevan 82 % vuonna 2022. Tässä ilmiössä ovat myötävaikuttavina tekijöinä useat trendit – etenkin kehittyvien maiden yhä yleistyvä pääsy Internetiin. Tällä hetkellä suurin kasvuvauhti IP-muotoisella liikenteellä on Lähi-Idässä ja Afrikassa. (Cisco, 2018)

Älypuhelimien käyttäjämäärän lisääntyminen tarkoittaa intuitiivisesti videodatan kokonaismäärän kasvua. Käytännössä jokaisessa älypuhelimessa on mukana kamera, jolla on mahdollista kuvata suhteellisen hyvälaatuisia videoita ja asettaa niitä saataville erilaisiin tietoverkkoihin kytkettyihin palveluihin. Lisäksi mobiiliverkkojen parantuva saatavuus ja nopeus kehittyvissä maissa luovat videodatalle kysyntää edullisten älypuhelinten mahdollistaessa kehittyvien maiden asukkaille ensimmäistä kertaa pääsyn Internetiin, ja näin ollen videoiden katsomisen lähes ilman paikkarajoitteita.

Näiden seikkojen lisäksi videodatan laatu (bittisyys ja resoluutio) paranevat jatkuvasti – ennen harvinainen ns. 4K-resoluution (3840 x 2160 pikseliä tai 4096 x 2160 pikseliä) video on nykyään Internetissä suhteellisen yleistä. 4K-resoluutioisessa videossa on esimerkiksi yleisesti käytössä olevaan Full HD -videoon (1920×1080 pikseliä) verrattuna neljä kertaa enemmän pikseleitä. Tämä tarkoittaa ymmärrettävästi sitä, että videon vaatima tilantarve kasvaa. Tulevaisuudessa tulee arvatenkin yleistymään ns. 8K-

resoluutio (7680 × 4320 pikseliä), jossa on 16-kertaisesti pikseleitä Full HD-videoon nähden.

Populaatiota, älypuhelimien määrää ja jopa internetin käyttäjämäärää nopeammin kasvaa kuitenkin Internetiin kytkettyjen laitteiden ja yhteyksien määrä. Siinä missä Internetin käyttäjien määrä lisääntyy Ciscon laatiman raportin *Cisco Visual Networking Index: Forecast and Trends, 2017–2022* mukaan 7 prosentin vuosivauhtia, lisääntyy laitteiden ja yhteyksien määrä 10 prosentin vuosivauhtia. Tämä trendi lisää laitteiden ja yhteyksien määrää asukasta kohti.

Ciscon raportissa on eritelty omaksi luokakseen laitteiden ja yhteyksien kategoria M2M (Machine to Machine) tai suomeksi laitteesta laitteeseen. M2M-luokkaan on Ciscon raportissa laskettu kuuluvaksi esimerkiksi älykkäät mittarit, videovalvonta, terveydenhuollon seurantasovellukset, kuljetusala ja pakettien sekä omistusten valvonta.

Käytännössä M2M tarkoittaa mitä tahansa kahden tai useamman laitteen välistä kommunikaatiota, internetin yli tai muutoin. Näin ollen esimerkiksi lähiverkon yli toisiinsa yhteydessä olevat älylaitteet lasketaan kuuluvaksi tähän luokkaan. M2M-luokkaan kuuluvaksi määriteltävien yhteyksien määrä lisääntyy 19 prosentin vuosivauhtia, mikä on nopeimmin kasvava yhteyksien ja laitteiden luokka Ciscon raportissa. Raportin mukaan M2M-yhteydet ja -laitteet tulevat kattamaan vuoteen 2022 mennessä 51 % kaikista laitteista ja yhteyksistä.

Laitteiden välisten yhteyksien kasvuvauhdin suhteesta laitteiden ja yhteyksien kokonaismäärään voidaan päätellä, että erilaiset älylaitteet ovat yhä yleisemmin yhteydessä toisiinsa. Yhtenä arkipäiväisenä tämän trendin ilmentymänä voitaisiin pitää esimerkiksi Amazonin Echo- ja Googlen Home-älykaiuttimia. Nämä älykaiuttimet kykenevät vastaanottamaan puhuttuja komentoja ja suorittamaan niiden perusteella erilaisia tehtäviä, kuten esimerkiksi hakemaan Internetistä säätiedotuksen ja kertomaan sen syntetisoidulla äänellä käyttäjälle. M2M-trendin kannalta oleellinen ominaisuus näissä älykaiuttimissa on, että äänikomennoilla voi myös ohjata kodin älylaitteita, mikäli älylaitteessa on ominaisuutena yhteensopivuus älykaiuttimen kanssa. Esimerkiksi Philips Hue -älyvalaistusta on mahdollista käyttää puhutuin komennoin Google Homen tai Amazon Echon kautta – eli esimerkiksi valojen päällä

oloa ja kirkkautta voidaan säätää äänikomennoin. Tällainen toiminnallisuus vaatii ymmärrettävästi, että älykaiutin ja älyvalot keskustelevat jollakin tavalla keskenään.

Tietoverkkojen ja etenkin älylaitteiden, sensoreiden ja muiden vastaavien laitteiden välisten yhteyksien yleistymisen osuminen suunnilleen samaan aikajaksoon kuin datan määrän voimakas kasvu on datan hyödynnettävyyden kannalta positiivinen ilmiö. Yhä suurempi osuus datasta on luotu sellaisena aikakautena, jona tietoverkot ovat olleet jo hyvinkin käyttökelpoisia datan saattamiseksi laajemman käyttäjäkunnan saataville. Lähiaikoina tuotettu data on tällä perusteella tietyllä tavalla arvokkaampaa – sitä on helpompi hyödyntää, se on tyypillisesti reaaliaikaista ja sen on digitaalisen teknologian kehittymisen myötä mahdollista olla entistä hienojakoisempaa.

2.2 Laskentatehon kustannukset

Ihmiskunnan käytettävissä oleva laskentateho on kasvanut merkittävästi viimeisten vuosikymmenten aikana. Intelin vuonna 1972 julkaisemaan ensimmäiseen x86-käskykantaa tukeva suoritin, mallimerkinnältään Intel 8008, pystyi suorittamaan noin 45 000 – 100 000 käskyä sekunnissa (CPU-World, 2018). Tällä hetkellä suhteellisen tavanomainen, samaa käskykantaa tukeva Intelin pöytäkoneisiin suunnattu suoritin Intel Core i7-8700K kykenee 7zip-suorituskykytestin perusteella suorittamaan 38 794 miljoonaa käskyä sekunnissa (Proclockers, 2018).

Intelin 8008-suoritin maksoi markkinoille tulonsa aikoihin 120 $ (ComputerWorld, 2018), joka on inflaatiokorjattuna noin 740 $. Vuonna 2008 Intel Core i7-8700K maksoi 369,99 $ (Newegg, 2018). Näiden tietojen perusteella on mahdollista laskea molemmille suorittimille karkea arvio yksittäisen laskutoimituksen hinnalle sekuntia kohti – vaikkakin muitakin parannuksia laskutoimitusten suorittamisessa on näiden suorittimien välillä tietenkin tapahtunut, joita tämä metriikka ei kunnolla tuo esille.

Tämän laskutoimituksen perusteella saadaan tulos, joka kertoo samalla rahalla saatavan tänä päivänä noin 200 000 -kertaisesti suoritintehoa verrattuna vuoteen 1972. Jotakin tiettyä rajattua tehtävää – kuten koneoppimista – varten suunniteltujen mikropiirien laskentateho on toki vielä useita kertaluokkia suurempi niissä laskutoimituksissa, joihin ne ovat suunniteltu.

Yleisesti tunnettu laki liittyen tietokoneiden laskentatehoon on ns. Mooren laki. Mooren lain esittämän säännönmukaisuuden havaitsi Gordon Moore julkaisussaan *Cramming more components onto integrated circuits* vuonna 1965. Mooren laki ennustaa, että transistoreiden määrä tiheissä mikropiireissä kaksinkertaistuu noin kahden vuoden välein. Tällainen tietyin väliajoin kaksinkertaistuva transistorien määrän tai laskentatehon kasvu on luonteeltaan eksponentiaalista. (Moore, 1965)

On myös esitetty arvioita siitä, että laskentatehon kasvu olisi luonteeltaan jopa ylieksponentiaalista. Mikäli näin on, laskentatehon kasvu ei näytä edes logaritmisella asteikolla mitattuna lineaariselta, vaan kasvunopeuden eksponentti kasvaa jatkuvasti. (Kurzweil, 2018)

Tämä viime vuosina saataville tullut laskentateho on lisännyt datafikaation mielekkyyttä paradigmana huomattavasti, sillä kuten aliluvussa 2.3 käy ilmi, datafikaation menetelmien hyödyntäminen vaatii runsaasti laskentatehoa. Tulevaisuudessa laskentatehon vaikutukset datafikaatioon ilmiönä ja paradigmana tulevat luultavasti jatkamaan nykyistä trendiä, mutta lopulta myös laadullisia muutoksia datafikaatioon ja datan käsittelyyn yleisesti oletettavasti ilmenee – hieman kuten siirtyessä puhtaasta digitalisaatiosta datafikaatioon kävi. Tällöin joutunemme keksimään esiin nousseelle, laadullisesti datafikaatiosta hieman eroavalle ilmiölle jälleen uuden nimen.

2.3 Digitalisaatio verrattuna datafikaatioon

Vaikkakin datafikaatiolla ja digitalisaatiolla on monia yhteisiä piirteitä (Taulukko 1), ne ovat kuitenkin erillisiä käsitteitä. Digitalisaationa tunnettu prosessi, joka alkoi 1950-luvulla puolijohdeteollisuuden synnyttyä, keskittyy osaksi luomaan fyysisen maailman ilmiöille digitaalisia vastineita. Tästä esimerkkinä voitaisiin pitää esimerkiksi tekstin muuttamista HTML-sivuiksi sekä musiikin muuttamista digitaalisiksi mp3-tiedostoiksi. Näin ollen digitalisaatiota voisi kuvata prosessiksi, joka mahdollistaa ideoiden ja käsitteiden muuttamisen digitaaliseen muotoon lähetystä, uudelleenkäyttöä ja muokkausta varten. (Ericsson; Imperial College London; Sustainable Society Network, 2014)

Taulukko 1. Digitalisaation ja datafikaation ominaispiirteet (mukaillen Ericsson, 2014)

DIGITALISAATIO	DATAFIKAATIO
Alustataloudellinen liiketoiminta	Kehitystyön alla oleva tuotteet ja alustat
Prosessiautomaatio, arvoketju yritysten hallinnassa	Massakustomointi, loppukäyttäjän mahdollisuus vaikuttaa arvoketjuun
Täydentävät tuotteet (kolmannen osapuolen sovellukset)	Tekijäkulttuuri ja tehostettu tuotanto
Vain digitaalinen	Digitaalinen sekä vuorovaikutus fyysisen maailman kanssa
Otoksiin pohjautuva data-analytiikka	Kvantifiointiin perustuva data-analytiikka

Datasta muodostuu eräänlainen heijastuma digitaalisiin tietoverkkoihin, joihin tutkittava ilmiö on tavalla tai toisella yhteydessä. Tätä datasta koostuvaa heijastumaa koneellisesti analysoimalla on mahdollista tutkia itse ilmiötä – usein reaaliajassa – sekä tehdä mahdollisesti ennusteita sen käyttäytymisestä tulevaisuudessa.

Datafikaation viitekehys käyttää osittain digitalisaation luomia työkaluja, mutta sitä pidetään kuitenkin digitalisaatiosta erillisenä ilmiönä ja paradigmana. Digitalisaation menettelytapoihin kuului yleisesti datan näytteistäminen (sampling), eli pyrkimys saada ilmiöön liittyvästä datasta edustava otos kaiken ilmiöön liittyvän datan käsittelyn välttämiseksi. Datafikaatiossa taas pyritään käsittelemään kaikkea oleelliseksi katsottua ilmiöön liittyvää dataa kokonaisina datasetteinä. (Ericsson, 2014)

Tämä kaiken saatavilla olevan datan paradigma vaatii paitsi suuria analysoitavia datamassoja analysoitavaksi, myös suurta laskentakapasiteettia, jotta kaikki haluttu data saadaan käsiteltyä. Luultavasti tulevaisuudessa yhä kiihtyvä laskentatehon halpeneminen tulee tuomaan saataville mahdollisuuden löytää monenlaisia korrelaatioita, asiayhteyksiä ja löydöksiä sellaisista paikoista, joista niitä ei ole aikaisemmin laskentatehon kalleuden vuoksi ollut mielekästä etsiä. Mikäli jatketaan aikaisemmin käytettyä vertausta öljyyn, tai tässä tapauksessa kaivostoimintaan, tehokkaammat malmin rikastusmenetelmät sallivat pitoisuudeltaan heikomman malmin käsittelyn siten, että voittoa syntyy malmin heikosta pitoisuudesta huolimatta.

Paradigmana datafikaation lähestymistapa ei ole täysin ongelmaton – on olemassa vaara, että analysoitava datasetti on tavalla tai toisella suodattunut tai vinoutunut. Esimerkiksi Twitterissä, joka on suhteellisen yleisesti käytetty sosiaalisen median

datalähde ihmisten välisten sosiaalisten verkostojen, mielipiteiden, sentimenttien ja muiden vastaavien ilmiöiden analysoinnissa, oli vuonna 2012 rekisteröityneenä vain 15 % amerikkalaisista Internetin käyttäjistä. Twitteriin liittyminen ylipäänsä saattaa toimia suodattavana tekijänä, joka vinouttaa Twitter-keskustelun edustavuutta suhteessa koko populaatioon. (van Dijck, 2014)

Edustavuuden lisäksi ongelmana analyysin osuvuutta haittaavana tekijänä sosiaalisilla media-alustoilla on yleisenä tapana algoritmisesti heikentää tai voimistaa jonkin tietyn sisällön näkyvyyttä omien tavoitteidensa – yhtiömuotoisilla organisaatioilla yleensä voiton tavoittelu – mukaisesti, mikä on omiaan vinouttamaan kohteena olevasta aiheesta käytävää keskustelua. Myöskään mielipidevaikuttajien mahdollista vaikutusta tutkittavasta ilmiöstä käytyyn keskusteluun ei tule jättää huomioimatta. (van Dijck, 2014)

Yllä kuvatut rajoitteet tulee ottaa sosiaalista mediaa tietolähteenä käyttäessä huomioon – sosiaalisen median viestimassoista eristettyjä tuloksia ei voi ainakaan perusoletuksena pitää koko populaatiota edustavana otoksena. Vinoutumia ja epäedustavuutta on mahdollista jossain määrin korjata, mutta se vaatii ulkopuolisen datan käyttöä.

Datafikaatioon kuuluu aiemmin esitetyn laajennetun määritelmän mukaisesti muitakin ilmiöitä kuin ihmisten sosiaalinen toiminta. Esimerkiksi erilaisten verkkoon kytkettyjen sensoreiden tuottaman datan voitaisiin katsoa olevan laadultaan sosiaalisen median viestejä objektiivisempaa – joskin lopullisen päätöksen jonkin tietyn sensorin kytkemisestä verkkoon, sijoituspaikasta, ja muista vastaavista seikoista, on kuitenkin ainakin toistaiseksi lähes poikkeuksetta tehnyt ihminen, mikä aiheuttaa riskin vinoumasta.

3 Ilosaarirock-datasetti

Tätä selvitystä varten on ladattu datasetti Futusomen tutkijapalvelusta, josta on mahdollista ladata haluttujen hakuehtojen mukaan rajattuja koosteita sosiaalisen median viestisisällöstä. Käytetty datasetti on saatu rajaamalla haku sellaisiin viesteihin, joissa esiintyy sana ”Ilosaarirock”. Ilosaarirock on vuosittain Joensuussa heinäkuun toisena viikonloppuna järjestettävä rockfestivaali. Datasetin sisältämät viestit ovat aikaväliltä 1.1.2015 – 25.8.2018.

Ladattu datasetti on taulukko, joka koostuu 29 159 rivistä, pois lukien ensimmäinen rivi, johon on tallennettu muuttujien nimet. Yksittäinen rivi kuvastaa yksittäistä sosiaalisen median viestiä. Viestirivi koostuu muuttujista, joista kukin on tallennettu omaan sarakkeeseensa: *kirjoittaja*: Viestin kirjoittajan käyttämä nimimerkki; *julkaisuaika*: Viestin julkaisuaika kuten “2018-08-25 08:52:24 +0000”; *tyyppi*: Viestin tyyppi, josta löytyy sekä viestikanava että tarkenne, joka mahdollisesti kertoo tarkemmin, millaisesta viestistä on kyse (Taulukko 2); *linkki alkuperäiseen*. URL-osoite, josta viesti on alun perin noudettu; *tekstisisältö*: Viestin sisältö.

Taulukko 2. Esimerkkejä viestityypeistä.

Tyyppi	Tyypin selite	Lukumäärä
blog_answer	Vastaus blogiviestiin	481
blog_comment	Kommentti blogiviestiin	127
blog_post	Blogiviesti	1 047
facebook_comment	Facebook-kommentti	3 782
facebook_event	Facebook-tapahtuma	70
facebook_link	Facebook-linkki	2 052
facebook_photo	Facebook-valokuva	2 064
facebook_post	Facebook-viesti	749
facebook_status	Facebook-tila	370
facebook_video	Facebook-video	615
forum_post	Keskustelupalstan viesti	2 613
googleplus_post	Google+ -viesti	60
instagram_image	Instagram-kuva	4 847
instagram_image_comment	Kommentti Instagram-kuvaan	428
news_comment	Kommentti uutisessa	1 042
pinterest_pin	Pinterestin pin, eli mediaviesti	5
twitter_retweet	Twitter-viestin uudelleenlähetys	2 320
twitter_tweet	Twitterin viesti	6 077
youtube_video	Youtube-video	289
youtube_video_comment	Kommentti Youtube-videoon	121
Yhteensä		**29 159**

Microsoftin Power BI:tä (Microsoft, 2018) käytettiin datasetin ajalliseen analysointiin. Microsoft Power BI on liiketoiminnan analytiikan sovellus, joka mahdollistaa datan visualisoinnin ja tulosten jakamisen. Tuettuna on myös upotus ohjelmistoihin (app) tai web-sivustoihin. Power BI -analyysin tukena käytettiin Power BI:n tukemia DAX-lausekkeita. DAX-dokumentaatio kuvaa DAX-kieltä seuraavasti (Microsoft, 2018): ”DAX on kokoelma funktioita, operaattoreita ja vakioita, joita voidaan käyttää kaavoissa tai lausekkeissa yhden tai useamman arvon laskemiseen ja palauttamiseen”. Instanssien esiintyvyydet kuukausitasolle sekä Pohjois-Karjalan maininnat ja Joensuu maininnat saatiin seuraavilla DAX-lausekkeilla:

- *Kuukausi = DATE(YEAR([julkaisuaika]); Month([julkaisuaika]); 1)*
- *Pohjois-Karjala maininnat = COUNTAX(FILTER('Sheet1';[Pohjois-Karjala]=true);[Pohjois-Karjala])*
- *Joensuu maininnat = COUNTAX(FILTER('Sheet1';[Joensuu]=true);[Joensuu])*

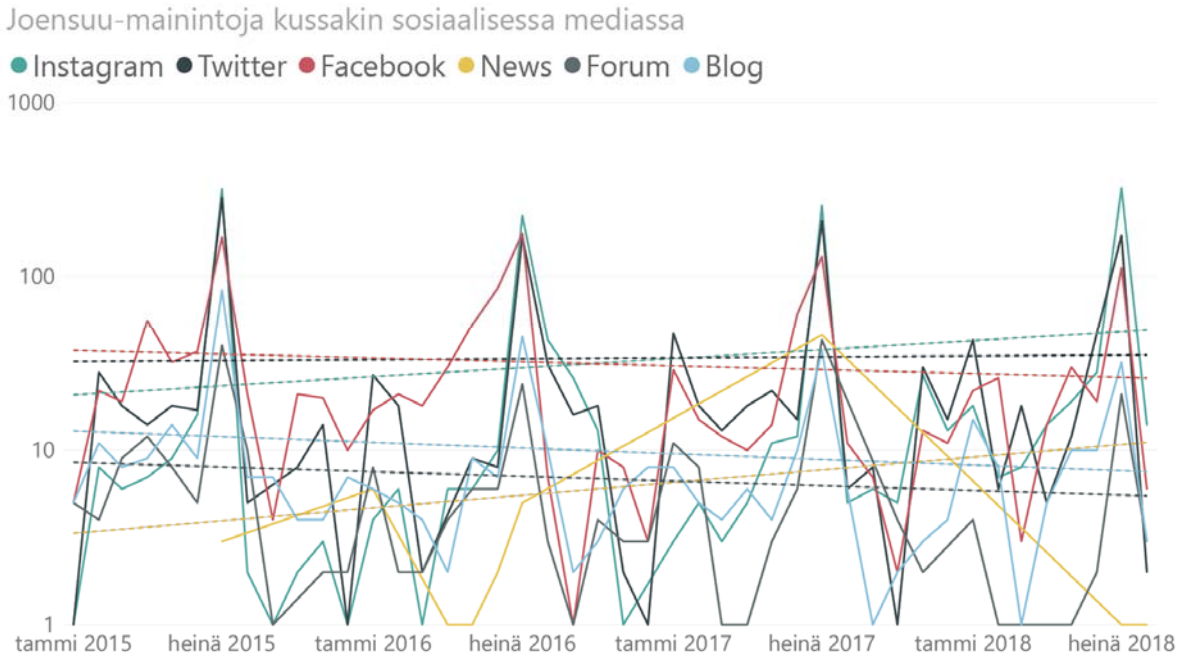
Vertailtavaksi valittiin kuusi eri sosiaalisen median tyyppiä: Instagram, Twitter, Facebook, News (uutiset), Forum (keskustelupalsta) ja Blog (blogi). Kullekin tyypille laadittiin oma DAX-lauseke oikeiden rivien valitsemiseksi rivin ”tyyppi”-muuttujan perusteella. Alla ovat kuvattuna käytetyt lausekkeet muuttujakohtaisesti:

- *Instagram = COUNTAX(FILTER('Sheet1';FIND("instagram"; Sheet1[tyyppi];true;false));[tyyppi])*
- *Twitter = COUNTAX(FILTER('Sheet1';FIND("twitter"; Sheet1[tyyppi];true;false));[tyyppi])*
- *Facebook = COUNTAX(FILTER('Sheet1';FIND("facebook"; Sheet1[tyyppi];true;false));[tyyppi])*
- *News = COUNTAX(FILTER('Sheet1';FIND("news"; Sheet1[tyyppi];true;false));[tyyppi])*
- *Forum = COUNTAX(FILTER('Sheet1';FIND("forum"; Sheet1[tyyppi];true;false));[tyyppi])*
- *Blog = COUNTAX(FILTER('Sheet1';FIND("blog"; Sheet1[tyyppi];true;false));[tyyppi])*

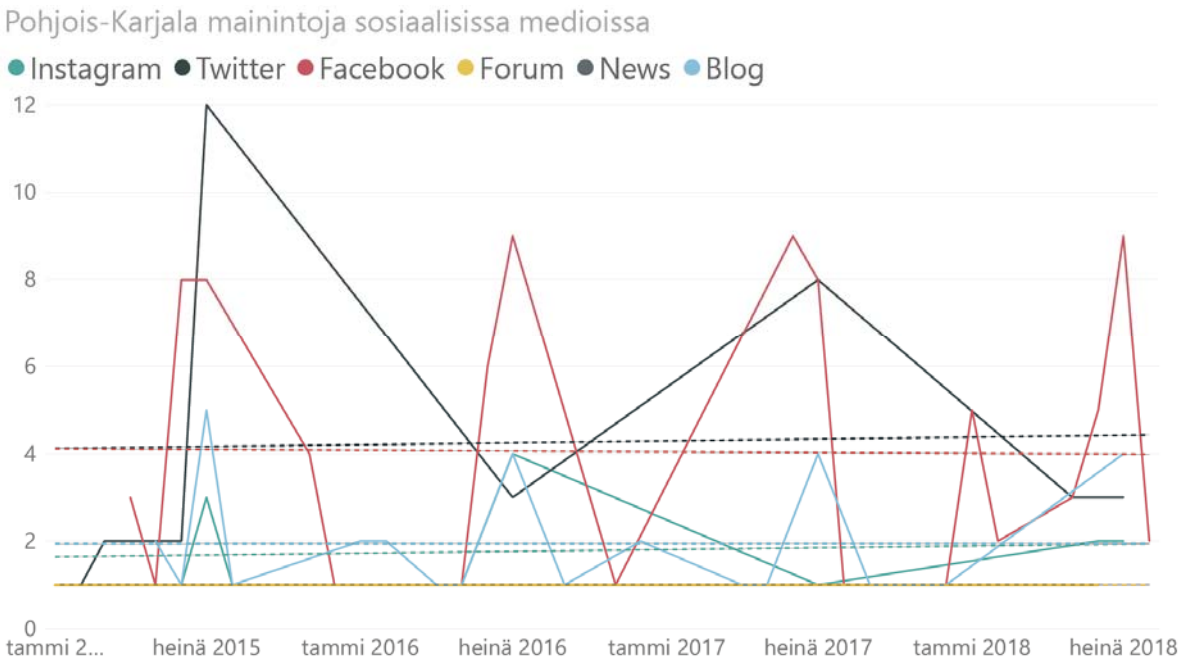
Näiden kuvaajien osalta käytettiin instanssien rajaamiseen kaaviokohtaisesti Power BI:n tarjoamaa `Page level filters` -ominaisuutta, johon asetettiin Joensuun tapauksessa ”Joensuu -> is true” ja Pohjois-Karjalan tapauksessa ”Pohjois-Karjala -> is true”. Kuvaajiin on lisätty trendiviivat mukaan otettujen muuttujien kehityksen trendin seuraamiseksi. Trendiviivat on toteutettu Power BI:n oman sisäänrakennetun toiminnallisuuden avulla.

Kuva 1 ja Kuva 2 havainnollistavat mainintojen painottumista heinäkuulle. Eri sosiaalisten medioiden painotus ja trendiviivat voidaan tulkita seuraavasti:

- Joensuun kuvaajan trendiviivoissa on nähtävissä Facebookin erittäin lievä johtoasema vielä vuonna 2015, mutta trendi on Facebookin osalta selvästi alaspäin. Tästä johtuen Twitter sekä Instagram ovat molemmat ohittaneet mainintojen määrässä Facebookin noin vuoteen 2017 mennessä. Instagramilla on trendiviivan mukaan voimakkaampi kasvutrendi, ja se onkin viimeisimmän datan mukaan alusta, jossa on esiintynyt eniten Joensuu-mainintoja.
- Joensuun kuvaajassa on myös havaittavissa mainintojen huippumäärien itse tapahtuman aikana järjestäytyvän hieman eri tavalla kuin trendiviivat antaisivat olettaa – Instagram on ollut pienellä marginaalilla alusta, jossa on ollut eniten Joensuu-mainintoja itse tapahtuman aikana heinäkuussa jo vuonna 2015, mutta on jäänyt kuitenkin vielä muina kuukausina jälkeen Facebookin mainintamääristä.
- Twitterissä esiintyvät maininnat Joensuusta ovat pysyneet koko tarkastelujakson ajan huomattavan tasaisina.
- Foorumien ja blogien mainintamäärät Joensuun osalta ovat olleet koko tarkastelujakson ajan lievässä laskussa, mikä heijastelee luultavasti yleistä trendiä Internetin käyttötottumusten muutoksessa. News, eli uutismaininnat, näyttäisivät olevan trendiviivan mukaan nousussa, mutta varsinaisten mainintojen määrää kuvaavan viivan perusteella datasetti on ollut mahdollisesti tämän muuttujan osalta jossain määrin epäluotettava – viivan kulku ei noudattele muiden viivojen kulkemaa polkua lähes lainkaan, eivätkä heinäkuiset piikit ilmene siinä muiden viivojen tapaan.
- Pohjois-Karjalan osalta merkittävimmät sosiaalisen median kanavat olivat Twitter ja Facebook. Mainintojen kokonaismäärä oli tarkastelujaksolla suhteellisen vähäinen sekä Joensuuhun verrattuna.



Kuva 1. Joensuu-mainintoja kussakin sosiaalisessa mediassa kuukausittain logaritmisella asteikolla. Katkoviivat ovat mainintojen määrien trendiviivoja.



Kuva 2. Pohjois-Karjala-mainintoja sosiaalisissa medioissa kuukausittain lineaarisella asteikolla. Katkoviivat ovat mainintojen määrien trendiviivoja.

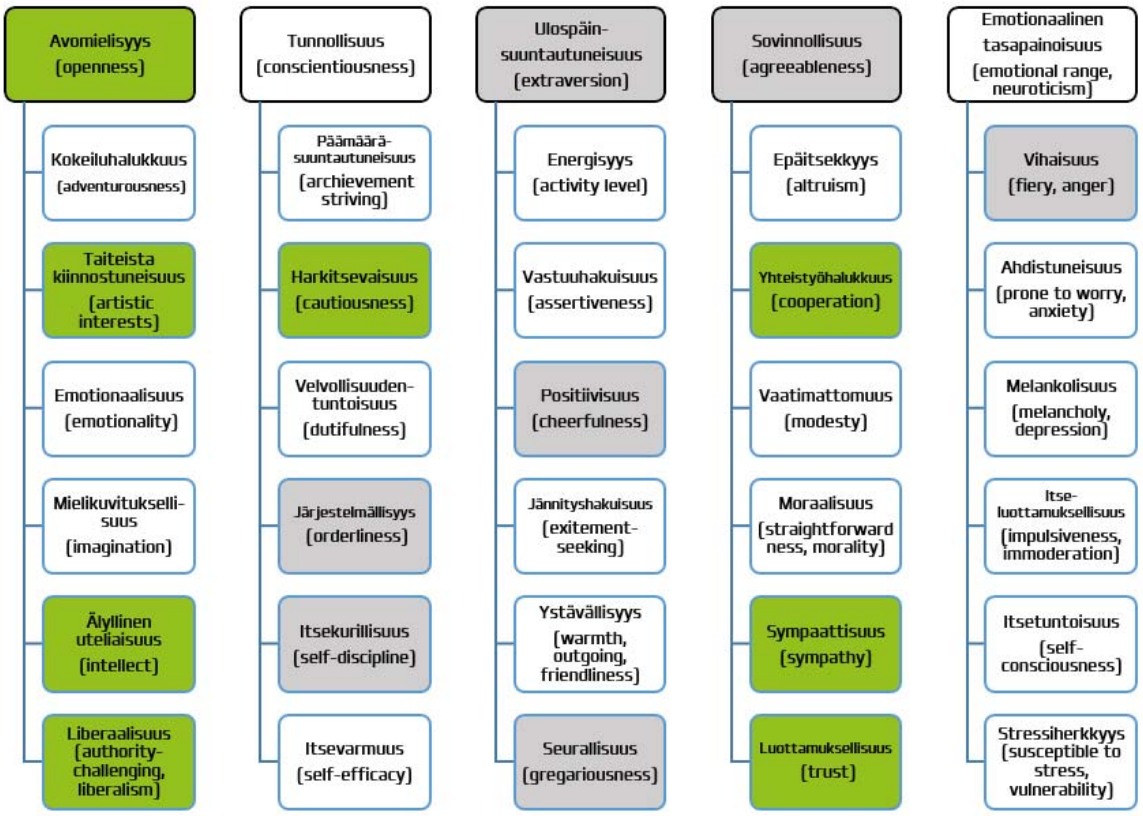
4 Psykografiset tiedot

Viestidatan psykografisten tietojen (Luku 4.1 ja Luku 4.2) analysointiin on käytetty *KoDa – Kokonaisvaltainen data hallinnointi ja hyödyntäminen* -projektin kesäkuussa 2018 julkaisemassa selvityksessä *Teksti- ja kuvanäkemykset - sentimentit, persoonallisuuspiirteet, kulutusmieltymykset ja tunteet* (Koponen ja Hotti, 2018) kuvattuja menetelmiä, pois lukien Facebookin Graph-APIn kautta saatujen kuvien analysointi sekä sentimenttianalyysi. Selvityksessä ja sen rinnalla laaditussa artikkelissa *Behavioral Interventions from Trait Insights* (Gain, Koponen, & Hotti, 2018) käytettiin Jukolan viestin -datasettiä, joka on rakenteeltaan samanlainen kuin Ilosaarirock-datasetti, jonka rakenne on kuvattu luvussa 3.

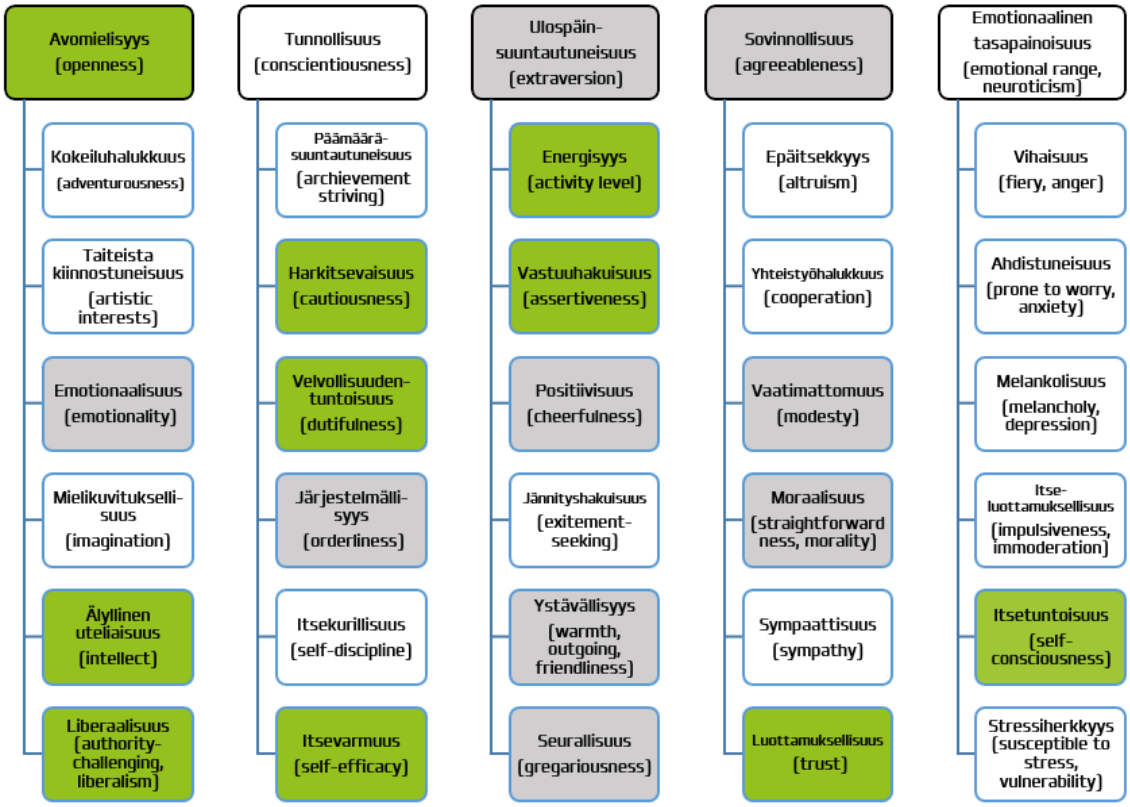
Facebook muutti ensimmäisen analyysin valmistumisen jälkeen käyttöehtojaan ja rajapintojaan siten, että kuvien lataaminen viestien `linkki alkuperäiseen` -muuttujan sisältämien URL-osoitteiden avulla ei ollut enää mahdollista Facebookin Graph API:n (Facebook, 2018) kautta. Koska muunlainen kuvien lataaminen esimerkiksi suoraan Facebookin Internet-selaimilla käytettäväksi tarkoitetun rajapinnan (web-sivusto) kautta olisi hyvin todennäköisesti rikkonut Facebookin käyttöehtoja, jouduttiin kuvien lataaminen sekä kaikki kuvien analysointiin liittyvä toiminnallisuus jättämään pois Ilosaarirock-datasetin osalta.

4.1 Persoonallisuustyypit ja niiden fasetit

Kuva 3 esittää Ilosaarirock-datasetin ja Kuva 4 Jukolan viesti -datasetin kirjoittajista toteutetut keskimääräiset persoonallisuustyypit ja niiden fasetit. Kukin persoonallisuustyyppi ja fasetti on väritetty siten, että mikäli kirjottajien keskimääräinen taipumus kyseiseen ominaisuuteen on analyysin mukaan korkea – 0,66 tai yli – on kyseinen ominaisuus väritetty vihreällä. Vastaavasti mikäli kirjottajien keskimääräinen taipumus kyseessä olevaan persoonallisuuspiirteeseen tai fasettiin on poikkeuksellisen matala, eli 0,33 tai vähemmän, on kyseinen ominaisuus värjätty harmaalla. Mikäli kirjoittajien keskimääräinen taipumus ominaisuuteen ei ole ollut suuresti keskimääräisestä poikkeava, on ominaisuus jätetty valkoiseksi.



Kuva 3. Ilosaarirockin persoonallisuustyypit ja niiden fasetit.



Kuva 4. Jukolan viestin persoonallisuustyypit ja niiden fasetit.

Persoonallisuustyyppien ja fasettien väliset erot pohjautuvat sellaisiin piirteisiin (ml. persoonallisuustyypit ja niiden fasetit), joiden erotuksen itseisarvo datasettien välillä on suurempi kuin 0,15 (Taulukko 3). Koska laskenta on toteutettu siten, että Jukolan viestin datasetin arvoista on vähennetty Ilosaarirockin vastaava arvo kutakin muuttujaa kohti, tarkoittaa negatiivisella etumerkillä varustettu erotus sitä, että kyseinen ominaisuus on ollut voimakkaampi Ilosaarirockin kirjoittajissa. Vastaavasti positiiviset luvut tarkoittavat sitä, että kyseinen ominaisuus on esiintynyt voimakkaampana Jukolan viestistä kirjoittaneilla.

Eroja Jukolan viestin ja Ilosaarirockin kirjoittajien välillä muodostuu analyysin perusteella Jukolan viestistä kirjoittaneiden itseluottamus ja vastuuhakuisuus sekä persoonallisuustyyppitasolla emotionaalinen tasapainoisuus. Vastaavasti Ilosaarirockista kirjoittaneet ovat esimerkiksi kiinnostuneimpia taiteista, seurallisempia, sympaattisempia ja kokeiluhaluisempia kuin Jukolan viestistä kirjoittaneet.

Taulukko 3. Piirteiden eroavaisuudet dataseteissä.

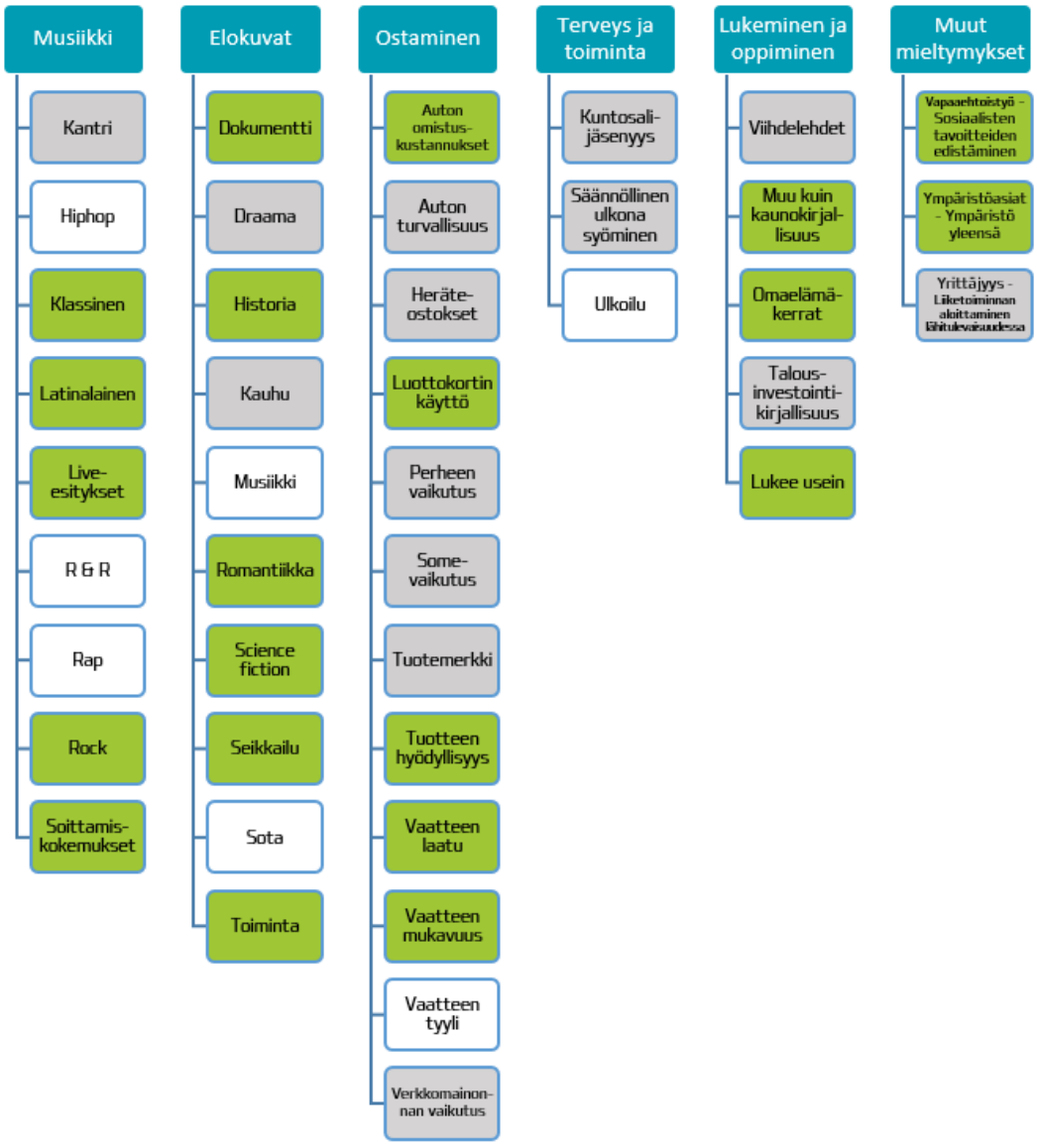
Piirteet	Erotuksen itseisarvo	Erotus
Itsevarmuus ('facet_self_efficacy'_'percentile')	0,2043	0,2043
Vastuuhakuisuus ('facet_assertiveness'_'percentile')	0,1712	0,1712
Emotionaalinen tasapainoisuus ('big5_neuroticism'_'percentile')	0,1543	0,1543
Kokeiluhalukkuus ('facet_adventurousness'_'percentile')	0,1568	-0,1568
Vaatimattomuus ('facet_modesty'_'percentile')	0,1591	-0,1591
Seurallisuus ('facet_gregariousness'_'percentile')	0,1653	-0,1653
Moraalisuus ('facet_morality'_'percentile')	0,1723	-0,1723
Yhteistyöhalukkuus ('facet_cooperation'_'percentile')	0,1829	-0,1829
Emotionaalisuus ('facet_emotionality'_'percentile')	0,2063	-0,2063
Sympaattisuus ('facet_sympathy'_'percentile')	0,2143	-0,2143
Itseluottamuksellisuus ('facet_immoderation'_'percentile')	0,2569	-0,2569
Taiteista kiinnostuneisuus ('facet_artistic_interests'_'percentile')	0,3434	-0,3434

4.2 Kulutusmieltymykset

Taulukko 4 on kuvattuna datasetteihin kuuluvien kirjoittajien keskimääräiset analyysin tuloksena saadut tarkat kulutusmieltymykset. Kulutusmieltymykset on jaettu kuuteen ryhmään (musiikki, elokuvat, ostaminen, terveys ja toiminta, lukeminen ja oppiminen, muut mieltymykset). Kuva 5 ja Kuva 6 on kuvattuna kirjottajien keskimääräiset arvioidut kulutusmieltymykset, jotka on väritetty samalla periaatteella kuin persoonallisuustyypit ja niiden fasetit.

Taulukko 4. Kulutusmieltymykset (consumption preferences)

Kulutusmieltymys (alkuperäinen englanninkielinen nimi IBM:n Personality Insights -rajapinnasta)	Jukolan viesti	Ilosaarirock
Ostaminen - Auton omistuskustannukset (Likely to be sensitive to ownership cost when buying automobiles)	0,914396887	0,732142857
Ostaminen - Auton turvallisuus (Likely to prefer safety when buying automobiles)	0,079766537	0,228316327
Ostaminen - Vaatteen laatu (Likely to prefer quality when buying clothes)	0,972762646	0,841836735
Ostaminen - Vaatteen tyyli (Likely to prefer style when buying clothes)	0,338521401	0,484693878
Ostaminen - Vaatteen mukavuus (Likely to prefer comfort when buying clothes)	0,846303502	0,81122449
Ostaminen – Tuotemerkki (Likely to be influenced by brand name when making product purchases)	0,365758755	0,239795918
Ostaminen - Tuotteen hyödyllisyys (Likely to be influenced by product utility when making product purchases)	0,8307393	0,756377551
Ostaminen - Verkkomainonnan vaikutus (Likely to be influenced by online ads when making product purchases)	0,091439689	0,229591837
Ostaminen - Some-vaikutus (Likely to be influenced by social media when making product purchases)	0,015564202	0,117346939
Ostaminen - Perheen vaikutus (Likely to be influenced by family when making product purchases)	0,044747082	0,153061224
Ostaminen – Heräteostokset (Likely to indulge in spur of the moment purchases)	0,114785992	0,174744898
Ostaminen - Luottokortin käyttö (Likely to prefer using credit cards for shopping)	0,988326848	0,857142857
Terveys ja toiminta - Säännöllinen ulkona syöminen (Likely to eat out frequently)	0,166342412	0,161989796
Terveys ja toiminta – Kuntosalijäsenyys (Likely to have a gym membership)	0,377431907	0,214285714
Terveys ja toiminta – Ulkoilu (Likely to like outdoor activities)	0,732490272	0,631377551
Muut mieltymykset - Ympäristöasiat – ympäristö yleensä (Likely to be concerned about the environment)	0,844357977	0,832908163
Muut mieltymykset - Yrittäjyys – liiketoiminnan aloittaminen lähitulevaisuudessa (Likely to consider starting a business in next few years)	0,305447471	0,198979592
Muut mieltymykset - Vapaaehtoistyö – sosiaalisten tavoitteiden edistäminen (Likely to volunteer for social causes)	0,821011673	0,770408163
Elokuvat – Romanttiikka (Likely to like romance movies)	0,007782101	0,073979592
Elokuvat – Seikkailu (Likely to like adventure movies)	0,990272374	0,857142857
Elokuvat – Kauhu (Likely to like horror movies)	0,007782101	0,079081633
Elokuvat – Musiikki (Likely to like musical movies)	0,418287938	0,642857143
Elokuvat – Historia (Likely to like historical movies)	0,935797665	0,887755102
Elokuvat – Science fiction (Likely to like science-fiction movies)	0,935797665	0,887755102
Elokuvat – Sota (Likely to like war movies)	0,871595331	0,489795918
Elokuvat – Draama (Likely to like drama movies)	0,085603113	0,285714286
Elokuvat – Toiminta (Likely to like action movies)	0,994163424	0,908163265
Elokuvat – Dokumentti (Likely to like documentary movies)	0,972762646	0,918367347
Musiikki – Rap (Likely to like rap music)	0,500972763	0,432397959
Musiikki – Kantri (Likely to like country music)	0,281128405	0,192602041
Musiikki – R & B (Likely to like R&B music)	0,456225681	0,409438776
Musiikki – Hiphop (Likely to like hip hop music)	0,428988327	0,479591837
Musiikki – Live-esitykset (Likely to attend live musical events)	0,941634241	0,931122449
Musiikki – Soittamiskokemukset (Likely to have experience playing music)	0,791828794	0,849489796
Musiikki – Latinalainen (Likely to like Latin music)	0,76848249	0,762755102
Musiikki – Rock (Likely to like rock music)	0,95233463	0,896683673
Musiikki – Klassinen (Likely to like classical music)	0,820038911	0,846938776
Lukeminen ja oppiminen - Lukee usein (Likely to read often)	0,898832685	0,817602041
Lukeminen ja oppiminen – Viihdelehdet (Likely to read entertainment magazines)	0,009727626	0,12755102
Lukeminen ja oppiminen - Muu kuin kaunokirjallisuus (Likely to read non-fiction books)	0,922178988	0,869897959
Lukeminen ja oppiminen -Talousinvestointikirjallisuus(Likely to read financial investment books)	0,591439689	0,272959184
Lukeminen ja oppiminen - Omaelämänkerrat (Likely to read autobiographical books)	0,910505837	0,803571429



Kuva 5. Ilosaarirockin kulutusmieltymykset.

Kuva 6. Jukolan viestin kulutusmieltymykset.

Kulutusmieltymysten osalta suurimmat erot (Taulukko 5) näyttävät ilmenevän seuraavissa asioissa:

- Jukolan viestistä kirjoittaneet ovat todennäköisemmin kiinnostuneita sotaelokuvista verrattuna kuin Ilosaarirockista kirjoittaneisiin.
- Jukolan viestistä kirjoittaneet olivat myös todennäköisemmin kiinnostuneita talousinvestointikirjallisuudesta.
- Vaikka sekä Jukolan viestistä että Ilosaarirockista kirjoittaneilla oli molemmilla voimakas taipumus ottaa omistuskustannukset huomioon auto-ostoksia

tehdessään, ilmeni kyseinen ominaisuus voimakkaammin Jukolan viestistä kirjoittaneilla.

– Ilosaarirockista kirjoittaneet pitävät kohtalaisessa määrin enemmän musiikki-kategorian elokuvista (Taulukko 4) verrattuna Jukolan viestistä kirjoittaneisiin – tämä on uskottava tulos, sillä he ovat kirjoittaneet suhteellisen paljon, yli 600 sanaa musiikkitapahtumasta, mikä antaa syytä olettaa, että he saattaisivat olla yleisestikin keskimääräistä enemmän kiinnostuneita musiikista ja sen eri ilmenemismuodoista.

Taulukkoon on otettu mukaan vain merkitsevimmät erot (Taulukko 5), eli itseisarvoltaan 0,15 tai suuremmat. Myös arvojen etumerkit ovat merkitykseltään samat, eli positiiviset tarkoittavat ominaisuuden olevan voimakkaampi Jukolan viestistä kirjoittaneilla, ja vastaavasti negatiiviset arvot kertovat ominaisuuden ilmenevän voimakkaammin Ilosaarirockista kirjoittaneilla.

Taulukko 5. Kulutusmieltymyseroavaisuudet dataseteissä.

Kulutusmieltymys	Erotuksen itseisarvo	Erotus
Pitää todennäköisesti sotaelokuvista ('Likely to like war movies')	0,3818	0,3818
Lukee todennäköisesti sijoitusoppaita (' Likely to read financial investment books')	0,3185	0,3185
On todennäköisesti herkkä auton omistuskustannuksille ostaessaan autoa ('Likely to be sensitive to ownership cost when buying automobiles')	0,1823	0,1823
On todennäköisesti kuntosalin jäsen ('Likely to have a gym membership')	0,1631	0,1631
Pitää todennäköisesti draamaelokuvista ('Likely to like drama movies')	0,2001	-0,2001
Pitää todennäköisesti musikaaleista ('Likely to like musical movies')	0,2246	-0,2246

5 Sanavektorien hyödyntäminen kirjoittajien klusteroinnissa

Psykografisten tietojen lisäksi selvityksen yhteydessä hyödynnettiin word embedding-teknologioita eri datasetteihin jakautuneiden käyttäjien erottamiseksi toisistaan. Koska datasetit ovat muodostuneet siten, että kuhunkin datasettiin on sisällytetty jostakin tietystä yksittäisestä aihepiiristä sosiaalisessa mediassa kirjoittaneita käyttäjiä, on oletettavaa, että eri datasetteihin kuuluvien käyttäjien ominaisuudet eroavat keskimäärin jollakin tavalla toisistaan. Tätä olettamusta erosta eri datasetteihin kuuluvien käyttäjien välillä tukee luvussa 4 toteutettu analyysi, jossa datasetteihin sisältyvien käyttäjien välillä oli havaittavissa psykografisia eroja. Tässä luvussa on pyritty selvittämään voisiko eri datasetteihin kuuluvien käyttäjien välillä havaita eroavaisuuksia käyttäen ns. word embedding -menetelmiä (sanaupotusmenetelmiä).

Word embedding -menetelmien avulla on mahdollista muuntaa tekstissä esiintyviä sanoja numeeriseen muotoon. Yleinen esitystapa numeromuotoon muunnetuille sanoille on satoja reaalilukuja sisältävä vektori, joka kuvaa sanan sijaintia vektoriavaruudessa. Eräitä algoritmeja sanavektoreiden muodostamiseksi ovat esimerkiksi Facebookin kehittämä FastText, sekä Googlen kehittämä Word2Vec. (Bornstein, 2019)

Teknologian tarkoituksena olisi saada kuvattua sanan semanttista merkitystä mahdollisimman tarkasti numeerisessa muodossa siten, että sanoilla voi esimerkiksi suorittaa erilaisia laskutoimituksia. Nykyiset word embedding teknologiat onnistuvat tässä tavoitteessa jo kohtalaisesti, kuten voidaan päätellä siitä, että esimerkiksi sanoilla king (kuningas), man (mies), woman (nainen) ja queen (kuningatar) voidaan tehdä semanttisia laskutoimituksia, kunhan ne muutetaan sanavektorimuotoon. Kun kyseisten sanojen sanavektoreilla suoritetaan laskutoimitus $king - man + woman$, saadaan lopputuloksena sanavektori, joka kuvaa sanaa *queen*. Tässä laskutoimituksessa siis poistetaan ensin kuninkaan ($hallitsija + mies$) käsitteestä *mies*, ja tilalle lisätään *nainen*, jolloin lopputuloksena on $hallitsija + nainen$, eli *queen*. Sanojen muuntaminen numeeriseen muotoon lisää myös huomattavasti niiden käyttökelpoisuutta erilaisia koneoppimisalgoritmeja silmällä pitäen. (Emerging Technology from the arXiv, 2019)

Luvussa 5.1 on esitelty analyysissä käytetyt datasetit. Luvussa 5.2 on esitelty analyysissä käytetyt menetelmät. Luvussa 5.3 on esitelty tulokset datasetin sanojen muuntamisesta sanavektorimuotoon sekä kahden eri sanavektorialgoritmin kattavuus datasetissä esiintyvien sanojen suhteen. Luvussa 5.4 on esitelty keskiarvoistettujen sanavektoreiden klusteroinnin tulokset.

5.1 Datasetit

Luvussa 4 esitettyjen datasettien lisäksi vertailuun otettiin mukaan kolmas datasetti, jonka avainsanana oli ”tekoäly”. Datasetti on noudettu Futusomen palvelusta luvussa 4 esitetyllä tavalla, ja on rakenteeltaan samanlainen. Erottavana tekijänä datasettien muotoilun välillä on se, että datasettien noutohetkien välillä vaikuttaisi viestilähteiden erottelu tarkentuneen Google+ -palvelun osalta siten, että Google+ -julkaisut ja niihin liittyvät kommentit erotellaan omiksi kategorioikseen. Tämä ei vaikuta suoritetun analyysin tuloksiin, sillä kaikkia viestilähteitä käsitellään analyysissä tasavertaisina. Datasetin sisältämien viestien lähteiden jakauma on kuvattuna Taulukko 6. Taulukosta ilmenee myös viestien kokonaismäärä.

Taulukko 6. Tekoäly-datasetin viestien alkuperä.

Tyyppi	Tyypin selite	Lukumäärä
blog_answer	Vastaus blogiviestiin	
blog_comment	Kommentti blogiviestiin	528
blog_post	Blogiviesti	802
facebook_comment	Facebook-kommentti	156
facebook_event	Facebook-tapahtuma	41
facebook_link	Facebook-linkki	11
facebook_photo	Facebook-valokuva	299
facebook_post	Facebook-viesti	1094
facebook_status	Facebook-tila	
facebook_video	Facebook-video	73
forum_post	Keskustelupalstan viesti	2844
googleplus_post	Google+ -viesti	17
googleplus_post_comment	Google+ -kommentti	1
instagram_image	Instagram-kuva	178
instagram_image_comment	Kommentti Instagram-kuvaan	11
news_comment	Kommentti uutisessa	1030
pinterest_pin	Pinterestin pin, eli mediaviesti	
twitter_retweet	Twitter-viestin uudelleenlähetys	12210
twitter_tweet	Twitterin viesti	10791
youtube_video	Youtube-video	64
youtube_video_comment	Kommentti Youtube-videoon	17
Yhteensä		**30167**

5.2 Käytetyt menetelmät

Luvussa 5.2.1 käydään läpi viestien muuntaminen sanavektoreiksi ja sanavektorien yhdistäminen käyttäjäkohtaiseksi keskiarvoksi. Luvussa 5.2.2 esitellään, kuinka käyttäjäkohtaiset keskiarvot on klusteroitu.

5.2.1 Viestien muuntaminen sanavektoreiksi ja vektoreiden yhdistäminen

Datan käsittelemiseksi word embedding -menetelmiä hyväksikäyttäen, datasetissä esiintyvät sanat muutettiin sanavektorimuotoon. Sanavektoriksi muuntamisessa käytettiin kahta erilaista algoritmia, jotta niiden kattavuutta tutkimusaineiston sisältämien sanojen suhteen voitiin vertailla – joskin käytetyllä sanakirjalla on myös vaikutus lopputulokseen. Käytetyt algoritmit olivat Word2Vec ja FastText. Sanavektoreiden tuottaminen tutkimusaineistosta toteutettiin alla kuvatulla tavalla:

1. Datasetistä kerättiin kaikki erilliset kirjoittajien nimimerkit, ja erillisille nimimerkeille kuuluvat viestit kohdistettiin ja yhdistettiin kunkin kirjoittajan nimikkeen alle.
2. Viestit tokenisoitiin erillisiksi sanoiksi käyttäen Pythonin NTLK-kirjaston *nltk.tokenize.casual.TweetTokenizer -luokkaa.*
3. Kunkin kirjoittajan yhteyteen yhdistetyistä tokenisoiduista sanoista kerättiin kaikki uniikit sanat, sekä kunkin uniikin sanan esiintymien määrä. Kunkin kirjoittajan viesteissä esiintyvistä sanoista muodostettiin siis ns. sanapussit (bags-of-words)
4. Kuhunkin käyttäjään yhdistetystä sanapussista löytyville sanoille suoritettiin muunnos vektorimuotoon. Muunnos suoritettiin kahdella eri työkalulla: Word2Vec ja FastText. Word2Vecin tapauksessa muunnokseen käytettiin Turku BioNLP Groupin ”Finnish Internet Parsebank” -sivustolta löytyvää valmista sanavektoritiedostoa, joka sisältää sanavektorimuunnokset suurelle osalle suomenkielisistä sanoista. FastText-muunnoksessa käytettiin puolestaan Facebookin tutkimusryhmän Github-sivulta löytyvää sanavektoritiedostoa.
5. Kullekin käyttäjälle laskettiin käyttäjän viesteistä luotujen sanavektoreiden keskiarvo. Keskiarvot laskettiin siten, että kaikki käyttäjän sanapussista löytyvien vektoreiden arvot laskettiin yhteen dimensiokohtaisesti ottaen

huomioon sanojen esiintymismäärä, jonka jälkeen saatu tulos jaettiin käyttäjän sanavektoripussin sisältämien sanojen määrällä. Koska sanavektorit olivat käytetyissä sanavektoritiedostoissa 300-ulotteisia, oli lopputuloksena kullekin käyttäjälle 300 reaalilukua sisältävä vektori, joka kuvaa käyttäjän kirjoittamia viestejä kokonaisuudessaan.

6. Lopuksi käyttäjää kuvaavat keskiarvoistetut vektorit sijoitettiin kaksiulotteiseen matriisiin siten, että yksi rivi vastaa yhtä käyttäjää. Matriisin sarakkeet sisältävät kunkin käyttäjän viestien sanoista muodostettujen sanavektorien dimensioiden keskiarvot.

5.2.2 Käyttäjien klusterointi

Käyttäjien klusteroinnilla pyrittiin selvittämään, olisiko eri dataseteissä sijaitsevien kirjoittajien keskiarvoistettujen sanavektoreiden välillä havaittavissa klustereita, sekä olisiko tiettyyn datasettiin kuuluvien kirjoittajien keskiarvoistetuilla sanavektoreilla taipumus sijaita sanavektoriavaruudessa lähempänä toisiaan kuin muissa dataseteissä sijaitsevia kirjoittajia. Samaan datasettiin kuuluvien kirjoittajien keskiarvoistettujen sanavektoreiden sijaitseminen lähempänä toisiaan verrattuna muiden datasettien kirjoittajiin vaikuttaisi tässä tapauksessa intuitiivisesti todennäköiseltä, mikäli menetelmä, jolla kirjoittajat on tiivistetty sanavektoreiksi, kuvaa jollakin käytännön arvoa tarjoavalla tavalla kirjoittajia.

Käyttäjien klusterointiin käytettiin K-Means- sekä Gaussian Mixture -klusterointialgoritmeja. Klusteroinnit suoritettiin sekä FastTextillä, että Word2Vecillä tuotetuille keskiarvoistetuille käyttäjien sanavektoreille, jotta word embedding -algoritmien tuottamien tulosten klusteroitavuutta voitiin vertailla. Klusterointi tapahtui K-means-klusterointialgoritmin tapauksessa Pythonin Scikit-Learn-kirjaston *sklearn.cluster.KMeans*-luokalla. Vastaavasti Gaussian Mixture -klusterointialgoritmin tapauksessa käytettiin käytännön toteutuksessa Scikit-Learn-kirjaston *sklearn.mixture.GaussianMixture*-luokkaa. Molempia klusterointitoteutuksia käytettiin analyysissä vakioparametreilla, sillä parametrien muuttamisella ei havaittu olevan klusteroinnin laadulle merkittävää positiivista vaikutusta.

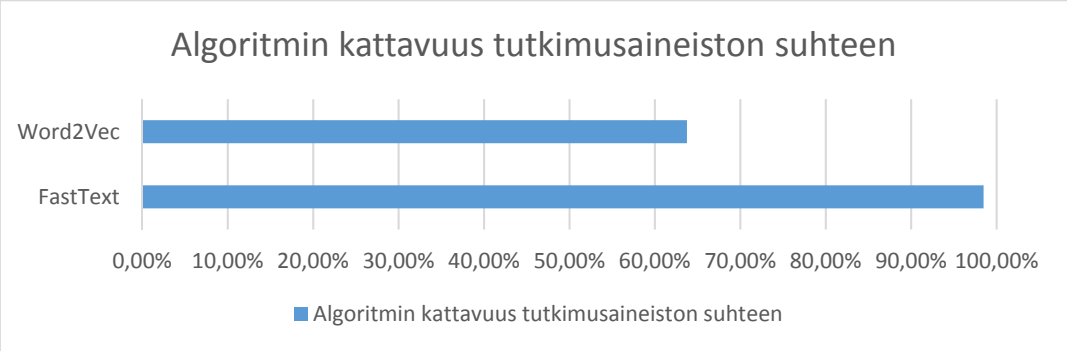
Kirjoittajista muodostettujen klustereiden datasettikohtaista jakaumaa – eli kuinka suuri osuus klusterin sisältämistä kirjoittajista on peräisin kustakin datasetistä – verrattiin sellaiseen tilanteeseen, että kirjoittajat olisivat jakautuneet klustereihin

tasaisesti edustamansa datasetin sisältämän kirjoittajamäärän mukaisesti. Tällaiset klusterit olisivat merkki siitä, että eri dataseteissä esiintyvien kirjoittajien välillä ei ole merkittäviä eroja – tai ainakaan sellaisia, jotka ilmenisivät klusterointialgoritmeja hyödyntämällä.

Datasettikohtainen odotettu tasainen jakauma laskettiin vertailemalla datasettien suhteellisia kirjoittajamääriä klusterin kokoon. Esimerkiksi datasettien kirjoittajamäärien ollessa 750 ja 250, eli suhdeluku 3:1, ja klusterin sisältäessä 200 kirjoittajaa, tulisi klusterissa olla 150 kirjoittajaa ensimmäisestä datasetistä ja 50 toisesta datasetistä, mikäli klusterissa esiintyvät kirjoittajat ovat jakautuneet tasaisesti. Muodostuneiden klustereiden datasettikohtaista jakaumaa verrattiin odotettuun tasaiseen jakaumaan jakamalla klusterissa esiintyvien kirjoittajien todellinen määrä odotetulla tasaisen jakauman määrällä.

5.3 Sanavektoreiden tuottaminen

Luvussa 5.2.1 kuvatulla tavalla kaikista tutkimusaineistona toimineista kolmesta datasetistä eristettiin yhteensä 462 619 erillistä sanaa, jotka kuvaavat tutkimusaineiston sanastoa käytetyn menetelmän rajoissa mahdollisimman kattavasti. FastText Facebook Researchin Github -sanakirjalla suoriutui sanavektoreiden tuottamisessa merkittävästi paremmin kuin Word2Vec. FastText kykeni tuottamaan 461 858 sanavektoria tutkimusaineistosta eristetyistä sanoista, siinä missä Word2Vec kykeni tuottamaan 294 940 sanavektoria. Prosentuaaliseksi sanaston kattavuudeksi muunnettuna FastText kykeni tuottamaan sanavektorin 99,84 % aineistossa esiintyvistä sanoista, siinä missä Word2Vecin kattavuus oli 63,75 % (Kuva 7).



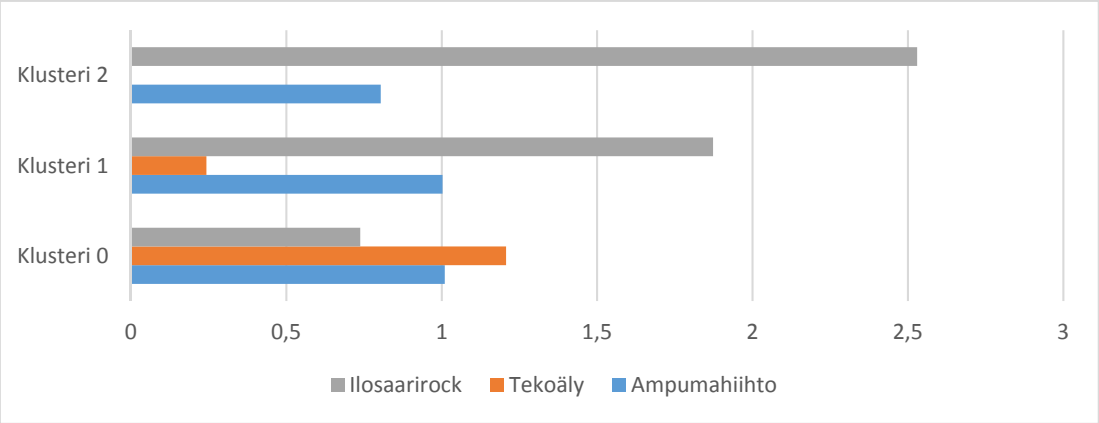
Kuva 7. Algoritmin kattavuus tutkimusaineiston suhteen.

Todennäköinen syy tälle erolle on, että Word2Vec käsittelee kokonaisia sanoja; mikäli haettavaa sanaa ei ole esiintynyt sanakirjatiedoston luomiseen käytetyssä aineistossa, ei sanalle ole mahdollista tuottaa sanakirjatiedoston avulla sanavektoria (Bornstein, 2019). FastText -algoritmi perustuu puolestaan n-grammeihin. Sanoja ei FastTextin tapauksessa tallenneta sanakirjaan kokonaisena, vaan sanat pilkotaan halutun pituisiksi siivuiksi, joista jokaisella on oma vektori. Lopullinen sanavektori sanalle muodostetaan näiden n-grammien yhdistelmästä. FastTextin tapauksessa muunnettavan sanan ei siis ole tarvinnut välttämättä esiintyä sanakirjatiedoston luomiseen käytetyssä aineistossa sellaisenaan, jotta sen voisi muuntaa sanakirjatiedoston avulla sanavektoriksi – riittää, että sanakirjatiedostosta löytyy vektorit kaikille n-grammeille, jotka muunnettava sana sisältää. (Gomez, Gibert, Gomez, & Karatzas, 2019)

5.4 Klusterointi

Tasaisimmat kokoiset klusterit tuottaneesta algoritmiyhdistelmästä (FastText, Gaussian Mix) tuotettiin uusi klusterointi sellaisista kirjoittajista, jotka olivat kirjoittaneet yli 600 sanaa. Tämän klusteroinnin tuloksista laskettiin klustereiden datasettikohtaiset edustukset, sekä erotus tasaiseen datasettikohtaiseen jakaumaan luvussa 5.2.2 kuvatulla tavalla.

Kuva 8 voimme nähdä, että datasettikohtaiset jakaumat eivät ole klustereissa tasaisia suhteessa datasettien suhteellisiin kokoihin. Näin ollen se, mihin datasettiin kirjoittaja on alun perin kuulunut, vaikuttaa siihen, mihin klusteriin kirjoittaja tulee klusteroinnin seurauksena päätymään. Tämä mahdollistaa sen, että koulutettua klusteroijaa hyödyntäen voitaisiin ennustaa jollakin todennäköisyydellä kirjoittajan kuuluminen tiettyyn datasettiin – eli siis välillisesti kirjoittajan mahdollinen kiinnostus klusteroijan kouluttamiseen käytettyjen datasettien aihepiirien välillä. Kuva 8 odotusarvo kullekin datasetille on 1 – palkin erotuksen itseisarvo 1:stä kertoo, kuinka suuri datasetin poikkeama on tasaiseen jakaumaan nähden. Näin ollen yli 1:n oleva arvo kertoo siitä, että datasetin suhteellinen edustus klusterissa on tasaista jakaumaa korkeampi. Vastaavasti alle 1:n oleva arvo kertoo siitä, että suhteellinen edustus klusterissa on tasaista jakaumaa matalampi.



Kuva 8. Datasettikohtainen jakauma verrattuna tasaiseen jakaumaan.

6 Johtopäätökset

Jatkuvasti käyttökustannuksiltaan edullistuva laskentateho sekä saatavilla olevan datan määrän, hienojakoisuuden ja reaaliaikaisuuden lisääntyminen ovat johtaneet siihen, että datafikaatio on nykypäivänä merkittävä ilmiö. Datan hienojakoinen tarkastelu otospohjaisen tarkastelun sijasta on mahdollistanut – ja edellyttänyt – sekä data-analyysin työkalujen, toimintatapojen, että data-analyysin menetelmin mahdollistuvien lopputulosten laadullista muutosta. Laskentatehon ja datan määrälliselle kasvulle ei ole toistaiseksi näkyvissä loppua, ja nykymuotoisen datafikaation aiheuttamien laadullisten muutoksen lisäksi on uskottavaa, että nykyisen trendin mukainen tulevaisuudessa tapahtuva datan sekä laskentatehon määrällisten ominaisuuksien muutos tulee aiheuttamaan laadullisia muutoksia niin data-analytiikan työkalujen, kuin myös sen mahdollistamien lopputulosten osalta myös tulevaisuudessa.

Data-analytiikan työkalut tulevat luultavasti tulevaisuudessa hyötymään samanlaisista muutoksista, joita datafikaatio on nykyisellään mahdollistanut esimerkiksi asiakaskokemukseen ja tuotteiden räätälöintiin, kuten Taulukko 1 (sivu 8) ilmenee. Tämä osaltaan johtaa abstraktiotason nousuun data-analytiikan työkaluissa – mikä tosin on yleinen trendi tietojenkäsittelyn saralla muutenkin.

Koneellisella persoonallisuusanalyysillä sekä sanavektoreihin pohjautuvilla word embedding -teknologioilla voidaan saavuttaa jo tälläkin hetkellä käyttökelpoisia lopputuloksia sosiaalisen median viestimassojen analysoinnissa. Persoonallisuuspiirteet ja etenkin niistä johdettavat kulutustottumukset ovat erinomainen työkalu markkinoinnin kohdentamiseen sekä tuotekehittelyyn.

Luvussa 5 esitetty menetelmä kirjottajien tuottaman tekstimassan tiivistämiseksi sanavektoriesitysmuotoon kaipaa jatkokehittelyä, mutta tässä selvityksessä on osoitettu menetelmän tuottavan tuloksia käytetyllä suhteellisen naiivilla toteutuksellakin. Kirjoittajia oli esitetyllä menetelmällä mahdollista erotella jossain määrin toisistaan sen perusteella, mistä he ovat kirjoittaneet.

Mikäli käytössä olisi dataa kirjottajien tuottaman tekstin lisäksi myös esimerkiksi kirjottajien kulutustottumuksista, voitaisiin menetelmän puitteissa tuottaa ennusteita

sellaisista kirjoittajakohtaisista ominaisuuksista, jotka eivät ilmene suoraan tekstistä. Tietyillä käytösmalleilla – esimerkiksi kulutustottumukset – olisi mahdollisesti korrelaatioita klusteriin kuulumisen kanssa. Tällöin voitaisiin arvioida kirjoittajan käyttäytymistä pelkästä tekstistä klustereihin kuulumisen perusteella. Menetelmä ei ole välttämättä nykyisellä naiivilla toteutuksella tarpeeksi tarkka, että sitä kannattaisi käyttää tähän tarkoitukseen esimerkiksi koneellisen persoonallisuusanalyysin sijasta – mutta mahdollisesti jonkinlaista käyttöarvoa voitaisiin saavuttaa persoonallisuusanalyysin tukena.

Menetelmän jatkokehityksen kannalta olisi aiheellista tutkia erilaisia algoritmeja kirjoittajia kuvaavan numeerisen esityksen muodostamiseksi, sekä kehittää kirjoittajien klusterointialgoritmia. Vaihtoehtoisesti voitaisiin pyrkiä löytämään klusterointialgoritmien lisäksi muita tapoja eriyttää käyttäjät toisistaan. Menetelmän hyödyllisyyttä silmällä pitäen käytetyn tavan tulisi olla sellainen, joka mahdollistaa klusteroinnin tavoin myös sellaisten kirjoittajien luokittelun, jotka eivät ole esiintyneet koulutusdatassa.

7 Lähteet

Bornstein, A. (26. 5 2019). *Beyond Word Embeddings Part 2*. Noudettu osoitteesta Towards Data Science: https://towardsdatascience.com/beyond-word-embeddings-part-2-word-vectors-nlp-modeling-from-bow-to-bert-4ebd4711d0ec

Cisco. (26. 11 2018). *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*. Haettu 25. 12 2018 osoitteesta https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html

ComputerWorld. (31. 12 2018). *Forgotten PC history: The true origins of the personal computer*. Noudettu osoitteesta ComputerWorld: https://www.computerworld.com/article/2532590/computer-hardware/forgotten-pc-history--the-true-origins-of-the-personal-computer.html?page=3

CPU-World. (31. 12 2018). *Intel 8008 (i8008) microprocessor family*. Noudettu osoitteesta CPU-World: http://www.cpu-world.com/CPUs/8008/

Cukier, K.;& Mayer-Schoenberger, V. (6 2013). The rise of big data. *Foreign Affairs*, ss. 28-40.

Emerging Technology from the arXiv. (05. 06 2019). *King - Man + Woman = Queen: The Marvelous Mathematics of Computational Linguistics*. Noudettu osoitteesta MIT Technology Review: https://www.technologyreview.com/s/541356/king-man-woman-queen-the-marvelous-mathematics-of-computational-linguistics/

Ericsson; Imperial College London; Sustainable Society Network. (2014). THE IMPACT OF DATAFICATION ON STRATEGIC LANDSCAPES.

Facebook. (29. 112 2018). *Graph ApPI*. Noudettu osoitteesta facebook for developers: https://developers.facebook.com/docs/graph-api/

Gain, U.;Koponen, M.;& Hotti, V. (2018). Behavioral Interventions from Trait Insights. *WELL-BEING IN THE INFORMATION SOCIETY – FIGHTING INEQUALITIES* (ss. 14-27). Turku: Springer Nature Switzerland AG, 2018.

Gomez, R.;Gibert J.;Gomez, L.;& Karatzas, D. (26. 5 2019). *Self-Supervised Learning from Web Data for Multimodal Retrieval e-prints arXiv: 1901.02004*. Noudettu osoitteesta https://gombru.github.io/2018/08/01/learning_from_web_data/

Koponen, M.;& Hotti, V. (2018). Teksti- ja kuvanäkemykset - sentimentit,. *KoDa – Kokonaisvaltainen data hallinnointi ja hyödyntäminen*.

Kurzweil, R. (31. 12 2018). *The Law of Accelerating Returns*. Noudettu osoitteesta Kurzweil Essays: http://www.kurzweilai.net/the-law-of-accelerating-returns

Microsoft. (29. 12 2018). *Correlation plot*. Noudettu osoitteesta Microsoft Power BI: https://community.powerbi.com/t5/R-Script-Showcase/Correlation-Plot/td-p/58462

Microsoft. (29. 12 2018). *DAX-perusteet Power BI Desktopissa*. Noudettu osoitteesta Microsoft Power BI: https://docs.microsoft.com/fi-fi/power-bi/desktop-quickstart-learn-dax-basics

Microsoft. (29. 12 2018). *Power BI*. Noudettu osoitteesta Microsoft: https://powerbi.microsoft.com/en-us/

Moore, G. (19. 4 1965). Cramming more components onto integrated circuits. *Electronics, Volume 38, Number 8*.

Newegg. (31. 12 2018). *Intel Core i7-8700K Coffee Lake 6-Core 3.7 GHz (4.7 GHz Turbo) LGA 1151 (300 Series) 95W BX80684I78700K Desktop Processor Intel UHD Graphics 630*. Noudettu osoitteesta Newegg: https://www.newegg.com/Product/Product.aspx?Item=N82E16819117827

Proclockers. (31. 12 2018). *Intel Core i7-8700K CPU Review: Page 4 of 6*. Noudettu osoitteesta Proclockers: https://proclockers.com/reviews/cpus/intel-core-i7-8700k-cpu-review/page/0/3

The Economist. (6. 5 2017). *The world's most valuable resource is no longer oil, but data*. Haettu 25. 12 2018 osoitteesta https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society, Volume 12, Issue 2*, 197-208.

Viitanen, J.;Paajanen, R.;Loikkanen, V.;& Koivistoinen, A. (2017). *Digitaalisen alustatalouden tiekartasto*.