Merja Halonen

Tiedonlouhinnan hyödyntäminen asiakkaan sitoutumisen tutkimisessa

Tietotekniikan pro gradu -tutkielma

4. kesäkuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Merja Halonen

Yhteystiedot: merja.k.a.halonen@student.jyu.fi

Ohjaaja: Tommi Kärkkäinen

Työn nimi: Tiedonlouhinnan hyödyntäminen asiakkaan sitoutumisen tutkimisessa

Title in English: Using Data Mining to Research Customer Engagement

Työ: Pro gradu -tutkielma

Suuntautumisvaihtoehto: Ohjelmistotekniikka

Sivumäärä: 130+5

Tiivistelmä: Pro gradu -tutkielma käsittelee Knowledge Discovery in Databases (KDD) -prosessin soveltamista asiakkaan sitoutumisen tutkimiseen asiakkuuden elinkaaren eri vaiheissa. Tavoitteena on selvittää, voidaanko suurista sivukyseludatoista ja sosiaalisen median datoista saada tiedonlouhinnalla hyödyllistä tietoa asiakkaan sitoutumisesta KDD-prosessia seuraten. Lisäksi tutkielmassa selvitetään, millaisia muita reaaliaikaisia datoja ja menetelmiä on käytetty sitoutumisen analysointiin. Empiiristen tulosten perusteella klusteroinnilla saadaan muodostettua asiakasryhmiä sitoutumisasteittain asiakkuuden elinkaaren eri vaiheissa tutkimalla reaaliaikaisten datojen erilaisia muunnoksia.

Avainsanat: Knowledge discovery, asiakkaan sitoutuminen, asiakkuuden elinkaari, klusterointi

Abstract: In this master's thesis the Knowledge Discovery in Databases (KDD) process and its usage with customer engagement in different stages of the customer life cycle are discussed. The aim is to find out, whether KDD process and data mining can help to discover useful information from customer engagement by using large clickstream and social media data. In addition, the thesis explains what kind of non-purchase data and methods are used for analyzing the engagement. Based on empirical results, customers can be grouped according to the state of engagement by different transformations of non-purchase data using clustering.

Keywords: Knowledge discovery, customer engagement, customer life cycle, clustering

Termiluettelo

AUC	Area Under The (ROC) Curve (pitoisuus-aikakäyrän pinta-ala) ROC-käyrän alle jäävä koko pinta-ala (“Classification: ROC Curve and AUC” 2019).
Big Data	Massadata Data (koneluettava tieto), joka on monimuotoinen, kooltaan suuri ja nopeasti kasvava (Finto 2019).
CEB	Customer Engagement Behavior (Asiakkaan sitoutumiskäyttäytyminen) Asiakkaan ja yrityksen välistä suhdetta tarkastellaan käyttäytymisen näkökulmasta (ei ostokäyttäytymisen). CEB on muuta kuin ostotapahtumia ja voidaan määritellä asiakkaan käyttäytymisen ilmentymänä (customer’s behavioral manifestation), jossa keskeisenä on yritys tai brändi ja tuloksena motivaatiolliset ajurit. (ks. Doorn ym. 2010, s. 254).
CLV	Customer Lifetime Value (Asiakkaan elinkaaren arvo) Kertoo, kuinka paljon asiakkuuden aikana asiakas tuottaa yritykselle liikevaihtoa tai liikevoittoa (ks. Pyyhtiä ym. 2013, s. 200).
CRM	Customer Relationship Management (Asiakkuudenhallinta) Viittaa käytäntöihin, strategioihin ja teknologioihin, joita yritykset käyttävät hallitsemaan ja analysoimaan asiakkaan vuorovaikutusta ja tietoja asiakkaan elinkaaren aikana, tavoitteen parantaa asiakassuhteita ja auttaa asiakkaiden säilyttämisessä sekä myynnin edistämisessä (“CRM (customer relationship management)” 2018).
DBSCAN	Density-Based Spatial Clustering of Applications with Noise Tiheyspohjaisen klusteroinnin menetelmä
eWOM	electronic Word-of-mouth (elektroninen suusanallinen viestintä)

	Suusanallinen viestintä, joka tapahtuu sähköisessä muodossa, yleensä Internetissä (“Termiharava” 2019).
KDD	Knowledge Discovery in Databases (Tietämyksen muodostus, tiedonlouhinta prosessi) Prosessi hyödyllisen tiedon löytämiseen datasta. Prosessin yhtenä vaiheena on tiedonlouhinta. Ks. tarkemmin luvusta 5
MDS	Multidimensional Scaling (Moniulotteinen skaalaus) Monidimensioisen datan dimensioiden vähentämistekniikka.
PCA	Principal Component Analysis (Pääkomponenttianalyysi) Monidimensioisen datan dimensioiden vähentämistekniikka.
ROC	Receiver Operating Characteristic Curve (ROC-käyrä, toimintaominaiskäyrä) Graafi, joka esittää luokitusmallin suorituskyvyn kaikilla luokituksen kynnysarvoilla (“Classification: ROC Curve and AUC” 2019).
SSE	Sum of Squared Errors (Virheiden neliösumma) Klusteroinnin algoritmien kriittisyyskriteeri.
TF-IDF	Term Frequency—Inverse Document Frequency (Tilastomenetelmä) Määrittää, kuinka monta kertaa termi esiintyy datassa eli eniten käytetyt sanat.
WOM	Word-of-mouth (Suusanallinen viestintä) Esimerkiksi suusanallinen markkinointi, hyvän asiakaskokemuksen kertominen toiselle.

Kuviot

Kuvio 1. Hankitun ja poistuneen asiakkaan vaikutus lähipiiriin (Kazienko, Brodka ja Ruta 2009) 19
Kuvio 2. Asiakkaan käytöstä kuvaavat sessioista muodostetut verkot (Baumann ym. 2017) 24
Kuvio 3. Asiakkaan session keskimääräinen aste sivuvierailuittain (Baumann ym. 2017) . 24
Kuvio 4. KDD-prosessin vaiheet (Fayyad, Piatetsky-Shapiro ja Smyth 1996a) 46
Kuvio 5. Klusterien havaitseminen (Xu ja Wunsch 2009) 54
Kuvio 6. Klusteroinnin vaiheet (Xu ja Wunsch 2009) 55
Kuvio 7. Asiakashankinnan asiakkaiden lukumäärän jakaumat session tiedoilla 69
Kuvio 8. Asiakashankinnan muunnoksen 1.1 kolmea klusteria tukevien validointi-indeksien kuvaajat 71
Kuvio 9. Asiakashankinnan muunnoksen 3.1 neljää klusteria tukevien validointi-indeksien kuvaajat 72
Kuvio 10. Asiakashankinnan muunnoksen 5.2 12 klusteria tukevien validointi-indeksien kuvaajat 72
Kuvio 11. Asiakashankinnan muunnoksen 6.2 viittä klusteria tukevien validointi-indeksien kuvaajat 73
Kuvio 12. Asiakashankinnan muunnoksen 1.1 muuttujien erottelevuus klustereittain 76
Kuvio 13. Asiakashankinnan muunnoksen 2.3 muuttujien erottelevuus klustereittain 78
Kuvio 14. Asiakashankinnan muunnoksen 3.1 muuttujien erottelevuus klustereittain 81
Kuvio 15. Asiakashankinnan muunnoksen 5.2 muuttujien erottelevuus klustereittain 83
Kuvio 16. Asiakashankinnan muunnoksen 6.2 muuttujien erottelevuus klustereittain 85
Kuvio 17. Asiakkuuden kehittämisen käyttäjien lukumäärien jakauma sitoutumistoiminnoittain 89
Kuvio 18. Asiakkuuden kehittämisen muunnoksen 1.1 muuttujien erottelevuus klustereittain 93
Kuvio 19. Asiakkuuden kehittämisen muunnoksen 1.1 julkaisuille tehtyjen toimintojen lukumäärä 95
Kuvio 20. Asiakkuuden kehittämisen muunnoksen 1.1 julkaisujen reaktiot klustereittain . 96
Kuvio 21. Asiakkuuden kehittämisen muunnoksen 1.4 muuttujien erottelevuus klustereittain 98
Kuvio 22. Asiakkuuden kehittämisen muunnoksen 1.4 julkaisujen reaktiot klustereittain . 99
Kuvio 23. Asiakkuuden kehittämisen muunnoksen 1.4 julkaisuille tehtyjen toimintojen lukumäärä 99
Kuvio 24. Asiakkuuden kehittämisen muunnoksen 2.1 muuttujien erottelevuus klustereittain 101
Kuvio 25. Asiakkuuden kehittämisen muunnoksen 2.1 julkaisujen reaktiot klustereittain . 102
Kuvio 26. Asiakashankinnan muunnoksen 1.1 hierarkkisen prototyyppipohjaisen klusteroinnin tulos 123
Kuvio 27. Asiakashankinnan muunnoksen 1.1 hierarkkisen klusteroinnin tulos 125

Taulukot

Taulukko 1. Luokittelukriteeri aikaisemmille tutkimusdatoille asiakkuuden elinkaaren vaiheittain 15
Taulukko 2. Yhteenveto asiakashankinnan tutkimuksista 27
Taulukko 3. Yhteenveto asiakkuuden kehittämisen tutkimuksista 38
Taulukko 4. Yhteenveto asiakkuuden säilyttämisen tutkimuksista 44
Taulukko 5. Sisäiset validointi-indeksit (Hämäläinen, Jauhiainen ja Kärkkäinen 2017) 61
Taulukko 6. Asiakashankintadatan kuvailu 66
Taulukko 7. Asiakashankinnan datamatriisi 66
Taulukko 8. Asiakashankinnan datan muuttujien kuvaus 67
Taulukko 9. Asiakashankinnan muunnoksen 1.1 klustereiden metadata 74
Taulukko 10. Muunnoksen 1.1 asiakkaiden sessioiden jakautuminen klustereittain 75
Taulukko 11. Asiakashankinnan muunnoksen 2.3 klustereiden metadata 77
Taulukko 12. Asiakashankinnan muunnoksen 3.1 klustereiden metadata 80
Taulukko 13. Asiakashankinnan muunnoksen 5.2 klustereiden metadata 82
Taulukko 14. Asiakashankinnan muunnoksen 6.2 klustereiden metadata 84
Taulukko 15. Asiakkuuden kehittämisen datamatriisi 87
Taulukko 16. Asiakkuuden kehittämisen muuttujat ja metadata 88
Taulukko 17. Asiakkuuden kehittämisen muunnoksen 1.1 klustereiden metadata 92
Taulukko 18. Asiakkuuden kehittämisen 1.1 muunnoksen klusterit sitoutumisjärjestyksessä 94
Taulukko 19. Asiakkuuden kehittämisen muunnoksen 1.4 klustereiden metadata 97
Taulukko 20. Asiakkuuden kehittämisen 1.4 muunnoksen klusterit sitoutumisjärjestyksessä 98
Taulukko 21. Asiakkuuden kehittämisen muunnoksen 2.1 klustereiden metadata 100
Taulukko 22. Asiakkuuden säilyttämisen heikosti sitoutuneiden osuudet sitoutuneimpien tykkäyksistä, muunnos 1.1 105
Taulukko 23. Asiakkuuden säilyttämisen heikosti sitoutuneiden osuudet sitoutuneimpien tykkäyksistä, muunnos 1.4 106
Taulukko 24. Asiakashankinnan muunnoksen 1.1 klustereiden metadata (hierarkkinen) 124
Taulukko 25. Asiakashankinnan muunnoksen 3.1 klustereiden metadata (hierarkkinen) 126
Taulukko 26. Asiakashankinnan muunnoksen 4.1 klustereiden metadata 127

Sisältö

1 JOHDANTO 1

2 ASIAKKAAN SITOUTUMINEN ASIAKKUUDEN ELINKAAREN VAIHEISSA 4
2.1 Asiakkaan sitoutuminen 4
2.1.1 Sitoutuminen ja sitouttaminen 5
2.1.2 Sitoutumistasot, -prosessit ja -mallit 7
2.2 Asiakkuuden elinkaari 10
2.2.1 Asiakashankinta 12
2.2.2 Asiakkuuden kehittäminen 12
2.2.3 Asiakkaan säilyttäminen 13

3 ASIAKKAAN SITOUTUMISEN TUTKIMUKSET ASIAKKUUDEN ELINKAAREN VAIHEESSA 15
3.1 Asiakashankinnan tutkimukset 15
3.1.1 Suusanallista viestintää hyödyntävät tutkimukset 16
3.1.2 Sivukyselydataa hyödyntävät tutkimukset 21
3.1.3 Yhteistyöhalukkuuteen liitettävät tutkimukset 26
3.2 Asiakkuuden kehittämisen tutkimukset 27
3.2.1 Suusanalliseen viestintään ja sosiaalisen verkoston kasvuun liittyvät tutkimukset 28
3.2.2 Brändiyhteisöön ja yhteistyön toimintaan liittyvät tutkimukset 31
3.2.3 Asiakasvalituksiin ja verkkosivuston dataan liittyvät tutkimukset 34
3.3 Asiakkuuden säilyttämisen tutkimukset 38
3.3.1 Sosiaaliseen verkostoon liittyvät tutkimukset 39
3.3.2 Kanta-asiakasohjelmaan liittyvät tutkimukset 41

4 KDD-PROSESSI 45
4.1 Prosessin vaiheet 45
4.2 Valinta 46
4.3 Esikäsittely 47
4.4 Muunnos 47
4.5 Tiedonlouhinta 48
4.5.1 Tehtävät ja mallit 49
4.5.2 Algoritmit 51
4.5.3 Tiedonlouhinta 51
4.6 Tulkinta / Arviointi 52

5 KLUSTEROINTI 53
5.1 Klusterit ja klusteroinnin vaiheet 53
5.2 Hierarkkinen klusterointi 55
5.3 Prototyyppipohjainen klusterointi 56
5.3.1 K:n keskiarvon klusterointimenetelmä (K-means) 57
5.4 Tiheyspohjainen klusterointi 58

5.5 Klusterin validointi 59

6 ASIAKKAAN SITOUTUMISEN LOUHINTA 64
6.1 Asiakashankinta ja sivukyselydata 64
6.1.1 Esikäsittely 65
6.1.2 Muunnos 68
6.1.3 Tiedonlouhinta 70
6.1.4 Tulokset: Muunnos 1.1 73
6.1.5 Tulokset: Muunnos 2.3 76
6.1.6 Tulokset: Muunnos 3.1 79
6.1.7 Tulokset: Muunnos 5.2 81
6.1.8 Tulokset: Muunnos 6.2 83
6.2 Asiakkuuden kehittäminen ja sosiaalisen median data 85
6.2.1 Esikäsittely 86
6.2.2 Muunnos 88
6.2.3 Tiedonlouhinta 90
6.2.4 Tulokset: Muunnos 1.1 91
6.2.5 Tulokset: Muunnos 1.4 96
6.2.6 Tulokset: Muunnos 2.1 100
6.3 Asiakkuuden säilyttäminen ja sosiaalinen verkosto 102
6.3.1 Tulokset 105

7 POHDINTA 107
7.1 Asiakashankinta 107
7.2 Asiakkuuden kehittäminen 109
7.3 Asiakkuuden säilyttäminen 110

8 YHTEENVETO 111

LÄHTEET 112

LIITTEET 123
A Asiakashankinnan muunnoksen 1.1 hierarkkinen prototyyppipohjainen klusterointi 123
B Asiakashankinnan muunnoksen 3.1 hierarkkinen prototyyppipohjainen klusterointi 125
C Asiakashankinnan muunnoksen 4.1 klusterit 127

1 Johdanto

Asiakkaiden sitouttaminen on syvällisemmän yhteyden luomista asiakkaisiin (Venkatesan 2017). Sitoutuminen on asiakkaan mielentila, joka voi olla kognitiivista, emotionaalista ja käyttäytymiseen liittyvää (Potdar ym. 2018). Käyttäytymisellä ei viitata asiakkaan ostokäyttäytymiseen, vaan se merkitsee asiakkaan osallistumista tai yhteyttä brändiin, yritykseen, yrityksen verkkosivuihin ja sosiaalisiin verkostoihin (Doorn ym. 2010). Asiakas voi sitoutua asiakkuutensa elinkaaren eri vaiheissa, jotka muodostuvat seuraavista vaiheista: asiakashankinta, asiakkuuden kehittäminen, asiakkuuden säilyttäminen (Bijmolt ym. 2010).

Asiakkaan sitouttaminen on tärkeää nykypäivän monipuolisessa palvelutarjonnassa ja informaatiomäärässä, sillä asiakkaan on mahdollista etsiä tietoa eri palveluista ja valita itselleen sopiva palvelu tai mahdollisesti vaihtaa palvelun tarjoajaa. Näin ollen pysyvämmän suhteen luominen asiakkaaseen voi olla haastavaa. Berryn, Linoffin ja Berryn (2004) mukaan pienen yrityksen on mahdollista rakentaa suhteensa asiakkaisiin ja parantaa palveluaan heitä kohtaan henkilökohtaisessa vuorovaikutuksessa huomaamalla asiakkaidensa tarpeet, muistamalla heidän mieltymyksensä ja oppimalla aikaisemmasta vuorovaikutuksesta. Suuremman yrityksen ei ole mahdollista luoda yhtä henkilökohtaista suhdetta asiakkaisiin, mutta asiakkaista kerätyn datan ja teknologian soveltamisella yrityksen on mahdollista päästä lähelle samanlaista tilannetta.

Yritysten keräämää dataa asiakkaista on paljon ja sen määrä kasvaa. Perinteisen historiallisen asiakasdatan, kuten ostotapahtumien lisäksi, ei perinteistä dataa pystytään keräämään asiakkaan klikkauksista yrityksen verkkosivuilla, erilaisista arvioista yrityksen sosiaalisessa mediassa sekä muusta suusanallisesta viestinnästä (Word-of-mouth, WOM). Kerätyllä datalla ei ole hyötyä yritykselle, jos sitä ei analysoida tai etsitä siitä uutta ja hyödyllistä tietoa. Varsinkin ei perinteisen datan analysoinnilla on hyötyä asiakkaan sitouttamisessa. Bijmoltin ym. (2010) mukaan suurimmaksi osaksi analyyttisiä malleja on muodostettu juuri datoista, jotka sisältävät asiakkaiden ostotapahtumia. Tällöin asiakkaan sitoutumisessa jää huomioimatta asiakkaan käyttäytyminen (engl. behavioral manifestations), millä on epäsuora vaikutus yrityksiin.

Knowledge Discovery in Databases (KDD) -prosessilla tunnistetaan uutta ja hyödyllistä tietoa suurista data-aineistoista (Fayyad, Piatetsky-Shapiro ja Smyth 1996a). KDD-prosessin erityinen vaihe on tiedonlouhinta, jolla analysoidaan data-aineistoja pyrkien löytämään odottamattomia suhteita ja tiivistämään dataa (Hand, Mannila ja Smyth 2001). Yksi tiedonlouhinnan menetelmistä on klusterointi, joka on datan ohjaamatonta luokittelua samanlaisiin ryhmiin (klustereihin) (Jain, Murty ja Flynn 1999).

Tässä tutkielmassa selvitetään, millaisia ei perinteisiä data-aineistoja ja metodeja on käytetty asiakkaan sitoutumisen analysoinnissa asiakkuuden elinkaaren eri vaiheissa. Lisäksi tutkielmassa sovelletaan KDD-prosessia asiakkuuden elinkaaren kolmeen eri vaiheeseen ja selvitetään, saadaanko kohdeaineistoista tiedonlouhinnan avulla hyödyllistä tietoa asiakkaan sitoutumisesta. Kohdeaineistoina käytetään avoimista datalähteistä saatuja sivukyselydataa ja sosiaalisen median dataa. Sivukyselydata (clickstream-data) on tallennettua online-toimintaa internet-käyttäjästä (Su ja Chen 2015; Bucklin ja Sismeiro 2009). Se tallentaa kaikki käyttäjän toiminnot, kuten selauspolut, ostetut tuotteet ja klikatut mainosbannerit (Su ja Chen 2015).

Tiedonlouhinnan menetelmänä käytetään klusterointia, jonka avulla pyritään löytämään eri tavoin sitoutuneita asiakas- tai käyttäjäryhmiä asiakkuuden elinkaaren eri vaiheissa. Kahdessa ensimmäisessä asiakkuuden vaiheessa käytetään eri data-aineistoja ja asiakkaan säilyttämisen vaiheessa analysoidaan asiakkuuden kehittämisessä esikäsiteltyä dataa tarkemmin. Tutkimuskysymys on seuraava:

- **Voidaanko tiedonlouhinnalla löytää hyödyllistä tietoa asiakkaan sitoutumisesta asiakkuuden elinkaaren eri vaiheissa?**

Tutkielma rakentuu siten, että toisessa luvussa kerrotaan asiakkaan sitoutumisesta asiakkuuden elinkaaren eri vaiheissa ja määritellään sitoutuminen tarkemmin. Kolmannessa luvussa käydään läpi aikaisempia tutkimuksia, jotka liittyvät asiakkaan sitoutumiseen ja ei perinteisiin data-aineistoihin, asiakkuuden elinkaaren eri vaiheittain. Neljännessä luvussa kuvataan tarkemmin KDD-prosessi vaiheineen. Klusteroinnin menetelmistä ja vaiheista kerrotaan tarkemmin luvussa viisi. Tutkielman empiirisen osan muodostaa luku kuusi, jossa data-aineistojen käsittely, tulokset ja klusterien arviointi käsitellään asiakkuuden elinkaaren eri

vaiheiden ja KDD-prosessin mukaisesti. Luvussa seitsemän pohditaan tarkemmin empiirisen osan tuloksia peilaten aikaisempiin tutkimuksiin. Kahdeksannes ja viimeinen luku sisältää yhteenvedon tutkimuksesta sisältäen arvioinnin tutkielmasta ja mahdollisia tulevaisuuden tutkimuskohteita.

2 Asiakkaan sitoutuminen asiakkuuden elinkaaren vaiheissa

Asiakkaan sitoutuminen voidaan määritellä eri tavoin ja ajatella sen muodostuvan prosessin tai tasojen kautta. Venkatesan (ks. 2017, s. 289) määrittelee asiakkaan sitouttamisen lyhyesti strategiaksi, jolla luodaan entistä syvempi yhteys asiakkaaseen. Perinteisemmät tavat, joilla sitoutumista on selvitetty, ovat liittyneet historialliseen dataan kuten asiakkaan ostotapahtumiin. Teknologian kehittyminen ja sosiaalisen median tuleminen ovat mahdollistaneet reaaliaikaisen datan käytön asiakkaan sitoutumisen tutkimisessa. Sitoutumista voidaan myös tarkastella asiakkuuden elinkaaren eri vaiheiden kautta. Eri vaiheissa voidaan käyttää perinteistä tai reaaliaikaista dataa asiakkaan sitoutumisesta.

2.1 Asiakkaan sitoutuminen

Asiakkaan sitoutumisen mittaaminen on lisääntynyt viimeisten kymmenen vuoden aikana ja sitouttamisesta on tullut päätavoite yrityksille (Ryan 2017). Okazakin ym. (2015) mukaan asiakkaan rooli on muuttunut 90-luvulta tiedon vastaanottajasta ja etsijästä sekä arvon luojasta asiakkaaksi, jolla on vaikutusvaltaa. Internetin ja sosiaalisen median kehittyessä on siirrytty asiakkaan uskollisuudesta suhdemarkkinoinnin kautta sosiaalisen median markkinointiin ja asiakkaan sitouttamiseen. Perinteisen ajatuksen mukaan asiakkaan sitoutumisella maksimoidaan yrityksen asiakaskunnan arvoa, jolloin asiakkaan sitoutuminen on sidoksissa asiakasarvon johtamiseen (Bijmolt ym. 2010; Verhoef, Doorn ja Dorotic 2007). Asiakasarvo on asiakkaan näkemys yrityksen luomasta arvosta, kuten hinta. Asiakasarvon johtamisessa asiakkaan arvoa katsotaan suoraan asiakkaan nykyisistä ja tulevaisuuden ostotapahtumista (Bijmolt ym. 2010).

Yrityksien on kuitenkin otettava huomioon asiakkaiden epäsuorat vaikutukset esimerkiksi brändiinsä ja tällöin asiakkaan sitoutumisessa on huomioitava asiakkaan käyttäytyminen (Bijmolt ym. 2010). Asiakkaan käyttäytymistä voi tutkia esimerkiksi suusanallisen viestinnän, asiakkaan ja yrityksen yhteistoiminnan, asiakkaan palvelun parantamisehdotuksien, asiakkaan mielipiteiden, brändiyhteisöihin osallistumisen tai yritykselle haitallisten toimin-

tojen kautta (Bijmolt ym. 2010). Brändi ei ole pelkästään logo tai nimi vaan summa siitä, mitä yritys on ja mitä se edustaa (Roberts ja Alpert 2010). Nykyisessä massadata maailmassa myös Kunz ym. (2017) ovat sitä mieltä, että asiakkaan sitoutuminen vaatii yrityksiä siirtymään ostotapahtumien johtamisesta asiakkaan ja heidän arvojen laajempaan ymmärtämiseen ja johtamiseen. Pansarin ja Kumarin (2017) mukaan yritykset keskittyvätkin heidän ja asiakkaan välisen suhteen laatuun sekä muihin yrityksille hyödyllisiin asioihin eikä pelkästään asiakkaan ostoihin.

2.1.1 Sitoutuminen ja sitouttaminen

Roberts ja Alpert (ks. 2010, s. 198) määrittelevät sitoutuneen asiakkaan seuraavasti: "Sitoutunut asiakas on lojaali brändille ja hän suosittelee aktiivisesti tuotteita ja palveluita toisille". Heidän mielestään sitoutunut asiakas ei ole pelkästään lojaali, vaan hän on myös uskottava, luotettava ja ennen kaikkea nykypäivän tehokas myynti- ja viestintäkanava. Keskeisiin asioihin, joilla asiakas sitoutuu, kuuluu Robertsin ja Alpertin (2010) mukaan brändi ja asiakkaan kokemus. Kokemus on heidän mukaan sitä, miten asiakas kokee vuorovaikutuksen fyysisesti ja emotionaalisesti. Näiden lisäksi keskeisiä asioita ovat asiakasarvon esitys (engl. customer value proposition) ja sisäiset kulttuurielementit. Asiakasarvo on syy, miksi asiakkaan pitäisi ostaa yritykseltä eikä kilpailijalta. Kulttuuri on tapa, joka muodostuu yrityksen tavoitteista, arvoista ja sisäisestä toiminnasta. Eniten asiakkaan suosittelua ja kannatusta seuraa loistavasta asiakaskokemuksesta ja vahvasta asiakasarvosta, jolla on toiminnallisia ja emotionaalisia ulottuvuuksia.

Asiakkaiden sitouttaminen on Sashin (2012) mukaan sitä, että keskitytään tyydyttämään asiakas tarjoamalla kilpailijoihin verrattuna ylivertaisempaa palvelua, jotta voidaan rakentaa luottamuksellinen suhde ja pitkäaikainen sitoutuminen. Sitoutuneesta asiakkaasta tulee yhteistyökumppani yritykselle, joka tekee yrityksen kanssa yhteistyötä, jotta yritys tyydyttäisi paremmin asiakkaiden tarpeet. Sosiaalisen median kautta käytävä vuorovaikutus helpottaa suuresti luotettavan ja sitoutuneen suhteen syntymisen. Asiakkaan sitouttamisen keskiössä on asiakas ja emotionaalisen siteen luominen asiakkaisiin.

Lisäksi Sashin (2012) mukaan sitoutuneisuutta on kahdenlaista: laskelmoivaa ja tunnepoh-

jaista. Laskelmoiva on rationaalista ja sitä esiintyy tarjonnan puuttuessa. Esimerkiksi kiinalaisesta ruuasta pitävät henkilöt joutuvat käymään kaupungin ainoassa kiinalaisessa ravintolassa, jolloin he luovat pysyvämmän suhteen ravintolaan ja tulevat lojaaleiksi asiakkaiksi. Tunnepohjaista sitoutumista muodostuu luottamuksesta ja vastavuoroisesta suhteesta. Esimerkiksi ravintolaan, missä asiakas käy mielellään paljon, voi muodostua läheinen suhde. Kumpaakin sitoutuneisuutta voi olla yhtä aikaa, jolloin asiakas on lojaali ja haltioitunut. Lojaalius ja haltioituneisuus ovat välttämätöntä asiakkaan sitoutumisessa. Haltioitunut asiakas voi tukea yritystä jakamalla sosiaalisessa verkostossaan mielihyvänsä yritystä kohtaan, jolloin asiakas myös sitoutuu. Tällöin sitoutuminen vaatii sekä laskelmoivaa että tunteellista sitoutuneisuutta tai luottamusta ja sitoutuneisuutta yrityksen ja asiakkaan välillä. Asiakas on sitoutunut, kun asiakkaalla on vahva emotionaalinen side yritykseen. Sitoutunut asiakas puolestaan muuttuu yrityksen ihailijaksi, joka pysyy suhteessa niin myötä kuin vastamäessä.

Doorn ym. (2010) huomioi käyttäytymisen asiakkaan sitoutumisessa, joka ei ole ostokäyttäytymistä, vaan sen keskiössä on brändi tai yritys. Asiakkaan sitoutumiskäyttäytymiseen (Customer Engagement Behavior, CEB) liittyy muun muassa suusanallinen viestintä, suosittelut ja arvostelujen kirjoittaminen. Doornin ym. (2010) mukaan asiakkaan sitoutumiskäyttäytyminen voi olla positiivista, kuten positiivisen brändiviestin julkaiseminen blogissa, tai negatiivista, kuten julkisen toiminnan organisointi yritystä vastaan. Lisäksi asiakkaan sitoutumiskäyttäytyminen ei välttämättä kohdistu pelkästään yritykseen tai brändiin, vaan se voi kohdistua myös laajempaan verkostoon, johon kuuluvat nykyiset ja potentiaaliset asiakkaat, toimittajat, yleisö, sääntelyviranomaiset sekä yrityksen työntekijät. Doornin ym. (2010) mielestä sitoutumiseen liittyvää käyttäytymistä on myös yhteistyö, joka sisältää esimerkiksi asiakkaiden tekemät ehdotukset kuluttajakokemuksien parantamiseksi, palveluntuottajien auttamisen ja ohjaamisen sekä toisten asiakkaiden auttamisen kuluttamaan paremmin.

Ryan (2017) pitää myös käyttäytymisen seuraamista tärkeänä ja hänen mielestään ensisijaisesti on seurattava sosiaalisen median mittareita, kuten sisällön jakamista, viipymisaikaa sivustolla, palautteita ja yhteisön toimintaa. Rayanin (ks. 2017, s. 399) mukaan sitoutumisen selvittämiseksi data voidaan pilkkoa osiin seuraavilla kysymyksillä: *Who* (kuka), *Where* (missä), *When* (milloin), *What* (mitä) ja *Why* (miksi). Esimerkiksi pelkästään sisällön lukemisen mittarilla ei voi sanoa, että pitääkö asiakas sisällöstä, vaan lisäksi tarvitaan muun

muussa tunteiden analysointia. Verkkoyhteisöjen lisääntyessä Verhoefin, Reinartzin ja Krafftin (2010) mielestä myös käyttäytymisen seuraaminen tulee tärkeämmäksi lähitulevaisuudessa, sillä sosiaalisen verkoston ja muiden uusien medioiden kautta asiakkaat voivat helposti olla vuorovaikutuksessa toisten asiakkaiden ja yrityksen kanssa.

Pansari ja Kumar (2017) huomioivat asiakkaan sitoutumisessa myös ostokäyttäytymisen. Heidän mukaan asiakkaan sitoutuminen on mekanismiksi, joka lisää asiakkaan arvoa yritykselle, joko suorien tai epäsuorien vaikutuksien kautta (ks. Pansari ja Kumar 2017, s. 295). Suoria ovat asiakkaan ostotapahtumat ja epäsuoria ovat asiakkaan palkitseminen, vaikutus ja tieto. Asiakas sitoutuu yritykseen, kun asiakkaan ja yrityksen välillä on emotionaalinen side ja suhde perustuu luottamukseen sekä sitoutuminen tyydyttää asiakasta (Pansari ja Kumar 2017).

Yhteenvetona voidaan todeta, että asiakkaan roolin muuttumisen myötä sitoutumisessa ei enää huomioida ostotapahtumia, vaan asiakkaan käyttäytyminen. Käyttäytymiseen liittyy brändi, yhteistyö, asiakkaan ja yrityksen välinen suhde, emotionaalinen side sekä asiakkaan kokemus, tyytyväisyys ja lojaalius. Varsinkin sosiaalinen media mahdollistaa asiakkaan sitoutumisen selvittämisen käyttäytymisen kautta, sillä sitoutunut asiakas tuottaa aktiivisesti sisältöä sosiaalisessa verkostossa. Sosiaalisen median myötä asiakkaan sitoutuminen määritelläänkin siten, että sitoutunut asiakas suosittelee yritystä muille ja tekee yrityksen kanssa yhteistyötä. Sitoutuneella asiakkaalla on syvällinen ja pitkäaikainen yhteys yritykseen. Sitouttaminen on puolestaan strategia, jolla kyseinen yhteys muodostetaan.

Kasvusuuntana asiakkaiden lähestymisessä on se, että oikea sisältö ilmestyy asiakkaalle oikeaan aikaan ja paikkaan. Tällöin sitoutuminen tullaan määrittelemään, siten kuinka paljon asiakas on todella sitoutunut annetulle sisällölle oikeaan aikaan (Ryan 2017). Tässä tutkielmassa asiakkaan sitoutuminen määritellään käyttäytymisen, kuten suosittelujen ja yhteistyön kautta.

2.1.2 Sitoutumistasot, -prosessit ja -mallit

Asiakkaan sitoutumista voidaan tarkastella sitoutumistasojen, -prosessien tai erilaisten mallien kautta. Doorn ym. (2010) on luonut havainnemallin ymmärtämään asiakkaan sitoutu-

miskäyttäytymistä. Havainnemallissa asiakaslähtöisyyden, yrityslähtöisyyden ja kontekstipohjaisuuden tekijät voivat vaikuttaa asiakkaan sitoutumiskäyttäytymiseen. Yhtenä asiakaslähtöisyyden tekijänä on asiakkaan tavoitteet, jotka voivat vaikuttaa brändiin. Tavoitteena asiakkaalla voi olla maksimoida kuluttamisen tai asiakkaan ja yrityksen välisen suhteen hyödyt. Doornin ym. (2010) mukaan sitoutumistasoon vaikuttaa myös asiakkaan luonteenpiirteet, taipumukset ja positiiviset tai negatiiviset kokemukset yrityksestä tai sen kilpailijoista. Lisäksi myös asiakkaan aika, vaiva ja raha vaikuttavat asiakkaan sitoutumiskäyttäytymiseen.

Tärkein yrityslähtöisyyden tekijä on brändi ja sen todelliset sekä asiakkaan itse kokemat ominaisuudet, sillä arvostetun brändin epäonnistuminen voi aiheuttaa suurta pettymystä asiakkaissa (Doorn ym. 2010; Roehm ja Brady 2007). Muita tekijöitä Doornin ym. (2010) mukaan ovat erilaiset alustat ja prosessit, joilla hoidetaan esimerkiksi valitukset ja ideat, erilaiset Online Chatit, kilpailut sekä erilaiset palkitsemisjärjestelmät. Kontekstipohjaisia tekijöitä ovat poliittiset, lailliset, ekonomiset, ympäristölliset, sosiaaliset ja teknologiset näkökulmat, kilpailijat toiminnoillaan ja markkinoiden kilpailukyky.

Lisäksi Doorn ym. (2010) esittää havainnemallissaan sitoutuneen asiakkaan vaikutukset ja seuraukset asiakkaisiin, yritykseen ja muihin. Yritykset voivat vaikuttaa asiakkaan sitoutumiskäyttäytymiseen esimerkiksi palkitsemisella. Käyttäytymisen näkökulmasta tällaiset onnistuneet toimet sitouttavat asiakkaita useammin ja syvällisemmin heidän sitoutumiskäyttäytymiseensä kuten suositteluun. Esimerkiksi sitoutunut asiakas tuottaa aktiivisesti sisältöä asiakasyhteisölle. Asiakkaan sitoutumiskäyttäytymiseen liittyvät toiminnot (mm. suosittelut, suusanallinen viestintä) vaikuttavat yrityksen talouteen ja maineeseen, sillä sitoutuminen vaikuttaa asiakkaan oman ostokäyttäytymisen lisäksi toisten asiakkaiden ostokäyttäytymiseen ja brändin tunnettavuuden lisääntymiseen. Lisäksi Doornin ym. (2010) mukaan sitoutunut asiakas voi auttaa yritystä muun muassa tuotteen ideoinnissa tai kehityksessä. Asiakkaiden sitouttamisella voi olla myös tärkeä rooli yrityksen toiminnan seuraamisessa sekä lainsäädännön ja säännellyn ympäristön parantamisessa.

Sashin (2012) mukaan sitoutuminen rakentuu prosessina asiakkaan sitoutumisympyrän kautta, jonka vaiheita ovat: yhdistäminen, vuorovaikutus, tyytyväisyys, säilyttäminen, sitoutuneisuus, kannattaminen ja sitoutuminen. Yhdistäminen voi tapahtua perinteisesti myyjän tai digitaalisesti sosiaalisen verkoston kautta. Asiakas voi tyydyttää tarpeensa olemassa olevien

yhteyksien tai uusien yhteyksien kautta. Lisäksi myyjä voi luoda yhteyksiä etsimällä uusia asiakkaita. Kun yhteys on luotu, asiakas voi olla vuorovaikutuksessa viestien, sähköpostin, blogien ja sosiaalisen verkoston kautta. Sitoutumisympyrässä edetään lähemmäksi sitoutumista, jos vuorovaikutus on tyydyttänyt kumpaakin. Sashi (2012) painottaakin, että tyytyväisyys on välttämätön edellytys asiakkaan sitoutumisessa, mutta se ei yksin riitä sitouttamaan asiakasta. Tyytyväisyydestä voi seurata asiakkaan säilyminen, mutta säilyminen voi olla seurausta myös muista erittäin positiivista tunteista. Uudelleen ostoa syntyy kuitenkin tyytyväisyyden kautta ja se luo pitkäaikaisen suhteen asiakkaan ja yrityksen välille, mutta siitä ei välttämättä synny emotionaalista sidettä asiakkaan ja yrityksen välille.

Potdar ym. (2018) ovat puolestaan muodostaneet asiakkaan sitoutumisprosessin sosiaalisessa mediassa, joka huomioi neljä näkökulmaa: psykologisen, kognitiivisen, käyttäytymisen ja sosiaalisen. Prosessi alkaa, kun asiakas huomaa yrityksen tarjonnan sosiaalisessa mediassa ja päättyy, kun asiakkaasta on tullut yrityksen kannattaja. Prosessissa on seitsemän vaihetta sisältäen kokemusvaiheen, jota ei ole muissa prosesseissa itsenäisenä vaiheena. "Kokemus eroaa vuorovaikutuksesta, koska se luo käsityksen asiakkaan mielessä, johon liittyy käyttäjän tunteita ja asenteita tietystä tuotteesta tai palvelusta."(ks. Potdar ym. 2018, s. 597). Prosessin vaiheita ovat kommunikointi, vuorovaikutus, kokemus, tyytyväisyys, jatkuva osallistuminen, läheisen suhteen luominen ja suositus.

Pääpiirteissään Potdarin ym. (2018) muodostama sitoutumisprosessi etenee niin, että yritys kommunikoi asiakkaalle sosiaalisen median kautta, josta seuraa vuorovaikutus asiakkaan ja yrityksen välillä kommenttien, tykkäyksien ja julkaisujen avulla. Jos vuorovaikutus on positiivista, siitä voi seurata positiivisia kokemuksia, jotka puolestaan voivat johtaa asiakkaan tyytyväisyyteen ja sitä kautta jatkuvaan osallistumiseen. Jatkuva osallistuminen voi puolestaan luoda emotionaalisen siteen, jolloin asiakas alkaa suositella yritystä toisille ja siten asiakkaasta tulee yrityksen kannattaja. Kommunikoinnin tarkoituksena on yhdistää sosiaalisen median käyttäjät brändiin luomalla brändin näkyvyyttä ja tietoisuutta.

Tyytyväisyys on Potdarin ym. (2018) mukaan kokemuksen tulos eli se, miten asiakas kokee tarjonnan. Jos asiakas on tyytyväinen, hän sitoutuu yritykseen sosiaalisen median kautta. Tämä puolestaan voi johtaa emotionaalisiin siteisiin ja asiakkaan uskollisuuteen yritystä kohtaan. Jatkuva osallistuminen on prosessi, jossa pidetään asiakkaan kiinnostusta yllä, jotta

asiakas pysyisi sitoutuneena. Asiakkaan tyytyväisyys, luottamus ja sitoutuneisuus johtaa läheiseen suhteeseen asiakkaan ja brändin välillä. Myös Roberts ja Alpert (2010) painottavat asiakkaan tyytyväisyyttä.

Sitoutuminen voidaan nähdä muodostuvan myös tasojen kautta. Roberts ja Alpert (2010) erittelevät viisi eri sitoutumistasoa, joista käy myös selville suosittelu ja kannattaminen. Jokainen taso rakentuu edellisestä tasosta seuraavasti:

1. asiakas ostaa tuotteen tai palvelun
2. asiakas on uskollinen tuotteelle tai palvelulle
3. asiakas ostaa mielellään toisen tuotteen tai palvelun
4. asiakas suosittelee tilaisuuden tullen tuotetta tai palvelua toisille
5. asiakas on kannattaja ja promoaa tuotetta tai palvelua kaikissa mahdollisissa tilanteissa.

Robertsin ja Alpertin (2010) mukaan asiakas on sitoutunut jo silloin, kun hän ostaa mielellään toisen tuotteen. Asiakkaan ollessa tasoilla 3-5, hän on sitoutunut.

Asiakkaan sitoutumista voidaan näin ollen selvittää erilaisten mallien, tasojen tai prosessien kautta. Sitoutumisprosessi alkaa asiakkaan havaitsemisesta ja päättyy kokemuksen ja tyytyväisyyden kautta sitoutumiseen. Asiakkaan sitoutumistasoon vaikuttaa muun muassa asiakkaan omat tavoitteet ja yrityksen brändi. Tässä tutkielmassa asiakkaan sitoutumista tarkastellaan sitoutumisasteittain heikosti sitoutuneesta vahvasti sitoutuneeseen, joihin vaikuttavat asiakkaan käyttäytyminen.

2.2 Asiakkuuden elinkaari

Mäntyneva (2003, s. 126) määrittelee asiakkuuden elinkaaren (engl. customer lifecycle) seuraavasti: "Asiakkuuden elinkaari alkaa asiakkuuden alkaessa, ja sen pituus vaihtelee asiakkuuksittain. Teoreettisesti tarkasteltuna asiakkuuden elinkaari jakaantuu neljään vaiheeseen: asiakkuuden alkaminen, haltuunotto, kehittäminen ja päättyminen". Kamakuran, Melan, Ansarin ym. (2005, s. 281) mukaan: "Asiakkaan elinkaari merkitsee sitä, että jokaisella asiakkaalla on arvoa hänen toimikautensa aikana". Asiakkuuden elinkaari ja asiakkuudenhallinta

(Customer Relationship Management, CRM) liittyvät yhteen, sillä asiakkuuden elinkaarta voidaan käyttää kuvaamaan asiakkuudenhallinnan rakennetta (ks. Kamakura, Mela, Ansari ym. 2005, s.281).

Asiakas voi sitoutua asiakkuutensa eri vaiheissa: asiakashankinta, asiakkuuden kehittäminen ja asiakkaan säilyttäminen (Bijmolt ym. 2010). Venkatesanin (2017) mukaan asiakassuhteet kehittyvät karkeasti hankinnan, kasvun, säilyttämisen ja takaisin voittamisen kautta. Näissä jokaisessa vaiheessa asiakas sitoutuu toistuvien kokemuksien kautta yritykseen ja sosiaaliseen verkostoon. Kokemukset auttavat asiakasta etenemään vaiheista toiseen, mikä johtaa eri tasoiseen sitoutumiseen. Näin ollen asiakkaan sitoutumisstrategian kehys sisältää asiakkuuden vaiheiden kartoittamisen asiakkaan ostokaaren (ennen ostoa, osto ja oston jälkeen) ohella, sillä näiden yhdistelmällä yritys voi tunnistaa ihanteelliset asiakkaat ja heille tarjottavat kokemukset (Venkatesan 2017). Esimerkiksi jos yritykset tavoittelevat pitkäkestoisia asiakkaita, yrityksien pitäisi tarjota asiakkaille erilainen kokemus ennen ostoa kuin taas asiakkaille, joiden mahdollisesti odotetaan vaikuttavan muihin asiakkaisiin (ks. Venkatesan 2017, s.290).

Malthouse ym. (2013) puolestaan huomioivat sosiaalisen median ja asiakkuuden vaiheiden vaikutuksen sitoutumisessa. Heidän mukaan sosiaalinen media vaikuttaa asiakkaan sitoutumisasteeseen (heikko tai vahva), mutta selvää rajaa asteiden välillä ei ole. Sitoutumisaste vaikuttaa siihen, miten yritys lähestyy asiakasta asiakkuuden vaiheissa. Myös yrityksen lähestyminen vaikuttaa asiakkaan sitoutumisasteeseen.

Lehtinen (2004) kuvaa myös asiakkuuden vaiheet kolmen näkökohdan kautta: asiakkuuden syntyminen, jalostaminen ja päättyminen. Pyyhtiä ym. (2013) puolestaan jakaa asiakkuuden elinkaaren vaiheet asiakkaan roolien mukaisesti: potentiaalinen asiakas, ostava asiakas, palveltava asiakas ja poistunut asiakas. Potentiaalinen ja ostava asiakas voidaan luokitella asiakkuuden alkamiseen ja haltuunottoon tai asiakashankintavaiheeseen. Palveltava asiakas liittyy kehittämisvaiheeseen, koska käyttäessään yrityksen palveluita tai tuotteita asiakas tarvitsee tukea (Pyyhtiä ym. 2013).

2.2.1 Asiakashankinta

Perinteisesti asiakashankintavaiheessa valitaan hankintakampanjaa varten asiakkaat, jotka suurimmalla todennäköisyydellä sopivat kampanjaan tai tulevat ostamaan tuotteita. Asiakkuudenhallinnan periaatteiden mukaan asiakkaiksi voidaan valita ne, joille odotetaan suurta asiakkaan elinkaaren arvoa (Customer Lifetime Value, CLV) (Bijmolt ym. 2010). Mäntynevan (2003) mukaan hankintavaiheen tavoitteena on potentiaalisten asiakkuuksien hankinta ja haltuunottovaiheen tavoitteena on uusien asiakkuuksien kannattavuus muiden tuotteiden lisämyynnillä. Lisäksi hänen mukaansa lähestymisperusteena näissä vaiheissa on demografinen profiili ja aiemmat ostot. Myös Kamakuranin, Melanin, Ansarin ym. (2005) mukaan asiakashankintavaiheen strategiana on saavuttaa enemmän tuottavia asiakkaita.

Asiakkaan sitoutumisen näkökulmasta asiakkaiden valinnassa voidaan Bijmoltin ym. (2010) mukaan hyödyntää asiakkaan käyttäytymistä ja ennustaa esimerkiksi suusanallisen viestinnän määriä eikä asiakkaan ostoja. Asiakkaiden valinnan jälkeen on mietittävä, miten resurssit kohdennetaan markkinointiin näiden asiakkaiden hankkimiseksi. Vaihe on olennainen kasvun aikaansaamisen ja kaikkien asiakaskantojen kierron vuoksi (Lehtinen 2004).

Malthousen ym. (2013) mukaan sosiaalisessa mediassa asiakkaan “tykkääminen” sisällöstä tai sen jakaminen voi auttaa yrityksiä luomaan tietoisuus mahdollisesta asiakkaasta, mutta tämä on heikon tason sitoutumista, sillä ne eivät aktiivisesti sitouta asiakasta tukemaan yrityksen markkinointitoimia. Asiakkaan luodessa sisältöä brändille hän on vahvasti sitoutunut. Sosiaalisessa mediassa vahvasti sitoutuneita asiakkaita ei voi eritellä heikosti sitoutuneista asiakkaista, koska vahvasti sitoutunut asiakas saattaa jakaa tuotteen arvostelun tai paljastaa promootiotarjoukset toisille heikosti sitoutuneille asiakkaille, jotka eivät yrityksen mukaan kuulu samaan vahvasti sitoutuneiden kategoriaan (Malthouse ym. 2013).

2.2.2 Asiakkuuden kehittäminen

Asiakkuuden kehittämisen vaiheessa asiakkaan elinkaariarvoa stimuloidaan erilaisten markkinointitoimintojen kautta (Bijmolt ym. 2010). Mäntynevan (2003) mukaan suunnitelmat ja toimintamallit voidaan tehdä myös ryhmäkohtaisella tasolla, mutta niiden tarkoituksena on kuitenkin syventää asiakkuuksia ja pyrkiä sitä kautta lisäämään yrityksen osuutta asiakkaan

kokonaisostoksista. Lähestymistapana on potentiaalin realisointi todennäköisyyslaskennan näkökulmasta.

Kehittämisvaiheeseen kuuluu myös olemassa olevien asiakkaiden tuottojen kasvattaminen esimerkiksi ristiinmyynnillä, lisämyynnillä ja kanavanhallinnalla (Kamakura, Mela, Ansari ym. 2005). Asiakkaan sitoutumisen kannalta kehittämisvaiheessa puolestaan keskitytään siihen, miten suusanallinen viestintä, yhteistyötoiminta asiakkaan ja yrityksen välillä sekä valitukset vaikuttavat asiakkaan elinkaaren arvoon (Bijmolt ym. 2010). Lehtisen (2004) mukaan kehittämisvaiheessa keskeistä on ympäristön ja paradigman muutoksien analysointi makro- ja mikrotasolla.

2.2.3 Asiakkaan säilyttäminen

Asiakkaan säilyttämisessä pyritään siihen, ettei asiakas poistuisi yrityksestä tai asiakassuhde ei muuten vaurioituisi. Asiakkaan ja yrityksen välinen suhde voi perustua sopimukseen tai voi olla, että suhteesta ei ole tehty mitään sopimusta, jolloin asiakkaan mahdollinen poistuminen on vaikeampaa havaita (Bijmolt ym. 2010). Mäntynevan (2003) mukaan vaihe sisältää asiakkaan syvällisen ymmärtämisen, sillä asiakkaat ja heidän todelliset tarpeensa on tunnettava, jotta pystytään mallintamaan asiakkaat, jotka ovat poistumasta yrityksestä. Lähestymistapana on asiakkaan ostohistoria ja profiili, mutta myös asiakkaan tekemät valitukset (Mäntyneva 2003).

Kamakuranin, Melanin, Ansarin ym. (2005) mukaan asiakkaan säilyttämisen kannalta on tärkeää ymmärtää asiat, jotka vaikuttavat asiakkaan poistumiseen. On myös pystyttävä ennustamaan asiakkaat, joihin poistumistekijät vaikuttavat. Heidän mielestään säilyttämisvaihe vaikuttaa merkitsevästi yrityksen kannattavuuteen. Myös Bijmoltin ym. (2010) mukaan perinteinen tapa säilyttää asiakas on ollut ennustaa ketkä asiakkaat ovat todennäköisemmin poistumasta yrityksestä. Hänen mielestään tapa ei kuitenkaan huomioi asiakkaan sitoutumista, joka pitäisi ottaa huomioon asiakkaan säilyttämisessä. Malthousen ym. (2013) mukaan sosiaalisen median kautta yritys pystyy todentamaan asiakkaat, jotka kannattaa säilyttää, vaikka muuten he eivät olisi säilytettäviä asiakkaita. Näillä asiakkailla voi olla on sellaisia sosiaalisia kontakteja, joista yrityksen puolestaan kannattaa pitää kiinni.

Yhteenvetona voidaan todeta, että asiakas voi sitoutua asiakkuuden elinkaaren eri vaiheissa ja asiakkaan sitoutumisaste (heikko tai vahva) vaikuttaa siihen, miten yritys kannattaa lähestyä asiakasta eri vaiheissa. Asiakashankintavaiheessa mahdollinen asiakas voidaan havaita käyttäytymisen kautta, esimerkiksi yrityksen sosiaalisen median sisällön tykkäämisellä tai jakamisella. Asiakkuuden kehittämisen vaiheessa asiakkuuksia syvennetään suusanallisen viestinnän, yhteistyön tai valituksien avulla. Esimerkiksi suusanallisen viestinnän ja yhteistyön avulla passiivisia asiakkaita voidaan saada aktiivisemmaksi. Säilyttämisvaiheessa pyritään havaitsemaan poistuvat asiakkaat ja ehkäisemään niiden poistuminen esimerkiksi sosiaalisen verkoston avulla. Tässä tutkielmassa sitoutumista tutkitaan asiakkuuden eri vaiheittain ja sitoutumisasteen avulla määritellään asiakkaat, jotka ovat huomion arvoisia kyseisessä vaiheessa.

3 Asiakkaan sitoutumisen tutkimukset asiakkuuden elinkaaren vaiheessa

Aikaisemmasta luvusta käy ilmi, että asiakas voi sitoutua elinkaarensa eri vaiheissa: asiakashankinta, asiakkuuden kehittäminen ja asiakkuuden säilyttäminen. Suuremmaksi osaksi tutkimukset ovat keskittyneet analysoimaan asiakkaan sitoutumista perinteisen datan, kuten ostohistorian kautta. Tutkielmaa varten löydettiin tutkimuksia, joissa sitoutumista analysoidaan enemmän reaaliaikaisemman sosiaalisen median datan avulla. Aikaisemmat tutkimukset on luokiteltu asiakkuuden elinkaaren vaiheisiin: asiakashankinta, asiakkuuden kehittäminen ja asiakkuuden säilyttäminen. Jos artikkelista ei selvästi käynyt ilmi asiakkuuden elinkaaren vaihetta, artikkelit luokiteltiin siten, miten Bijmolt ym. (2010) olivat tutkimuksessaan luokitelleet artikkelinsa (Taulukko 1). Tällöin artikkeli luokiteltiin siinä käytetyn aineiston mukaan.

Taulukko 1: Luokittelukriteeri aikaisemmille tutkimusdatoille asiakkuuden elinkaaren vaiheittain (Bijmolt ym. 2010)

Asiakashankinta	Asiakkuuden kehittäminen	Asiakkuuden säilyttäminen
Suusanallinen viestintä	Brändiyhteisö ja verkkosivuston data	Kanta-asiakasohjelman data
Sivukyselydata	Suusanallinen viestintä	Sosiaalisen verkoston data
Yhteistyön halukkuus	Sosiaalisen verkoston kasvu	Aikasarjadata
	Yhteistyön toiminta	
	Valitukset	

3.1 Asiakashankinnan tutkimukset

Perinteisesti asiakashankintavaiheessa on käytetty ostohistoriadataa sekä perinteisiä malleja, joilla on ennustettu kontaktoivat henkilöt, jotka tulevat sopimaan kampanjaan tai vastaamaan tiettyä ostotasoa. Bijmoltin ym. (2010) mukaan asiakkaan sitoutumisen näkökulmasta ostohistoriadata voidaan korvata asiakkaan käyttäytymisdatalla, johon puolestaan käytetään

perinteisiä malleja. Jos oletetaan, että positiivinen suusanallinen viestintä vaikuttaa yrityksen tuloihin, voidaan suusanallisen viestinnän määrää ennustaa tulevaisuuden ostojen sijasta. Tällöin yritykset voivat kohdentaa toimiaan asiakkaisiin, joilla on taipumusta suusanalliseen viestintään. Asiakkaan suusanallista viestintää voisi tutkia myös pitkältä ajalta, koska viestinnän oletetaan muuttuvan asiakkuuden elinkaaren aikana.

Bijmoltin ym. (2010) mukaan asiakkaan halukkuutta ja mahdollisuutta yhteistyöhön voidaan analysoida esimerkiksi scoring-mallilla. Internetissä tapahtuvaan yhteistyöhalukkuuteen selvittämiseen voidaan käyttää perinteisiä malleja, joissa käytetään sivukyselydataa. Sivukyselydata on hyödyllisin lähde yksilöiden käyttäytymisen ymmärtämiseksi (Su ja Chen 2015). Myös yhteistyöhalukkuuden kohdalla yhteistyön kehittymistä ja sen vaikutusta asiakkaan elinkaaren arvoon voisi tutkia pitkällä aikavälillä (Bijmolt ym. 2010).

3.1.1 Suusanallista viestintää hyödyntävät tutkimukset

Asiakashankintavaiheessa suusanallisen viestinnän selvittämiseen voidaan hyödyntää erilaisia asiakaskyselyitä. Bowman ja Narayandas (2001) analysoivat yli 60 amerikkalaisen brändin asiakkaita, jotka olivat ottaneet yhteyttä asiakaspalveluun. Asiakkailta kysyttiin puhelinkyselyllä, olivatko he kertoneet joillekin heidän brändikokemuksestaan, ja arviota, kuinka monelle henkilölle he olivat kertoneet. Brändikokemuksen kertomiseen Bowman ja Narayandas (2001) käyttivät logistista regressiomallia ja negatiivista binomiaalista regressioanalyysia (engl. truncated-at-zero negative binomial regression, NBD) siihen, kuinka monelle asiakas oli kertonut brändikokemuksestaan.

Wangenheim ja Bayon (2007) tutkivat myös suusanallista viestintää selvittämällä asiakkaan tyytyväisyyden, suusanallisen viestinnän ja uuden asiakkaan hankinnan epälineaarista suhdetta. He tutkivat logistisella regressiomallilla, millainen suusanallinen viestintä johtaa uuden asiakassuhteen syntymiseen. Tyytyväisyyden ja suusanallisen viestinnän yhteyttä mallintamaan, he käyttivät zero-inflation Poison mallia (ZIP). Heidän datansa koostui puhelinkyselyistä saksalaisen energiayhtiön kahdelle asiakasryhmälle (vaihtajat ja pysyjät) sen jälkeen, kun energiamyynnin markkinat olivat vapautuneet. Vapautumisen myötä asiakkaiden oli mahdollista vaihtaa energiayhtiötä, jolloin vaihtajat olivat luoneet uuden asiakassuhteen

toisen energiayhtiön kanssa. Wangenheimin ja Bayonin (2007) tuloksien mukaan hiljattain hankittu asiakas todennäköisesti kertoo kokemuksistaan muille riippumatta tyytyväisyystasostaan, jolloin heidän suusanallisen viestintänsä lisäämiseen olisi panostettava.

Bijmoltin ym. (2010) mukaan yrityksien olisi myös ymmärrettävä, miten suusanallisen viestinnän kautta hankitulla asiakkaalla on vaikutusta elinkaaren arvoon. Villanueva, Yoo ja Hanssens (2008) vertasivat asiakashankinnassa yrityksen markkinoinnin vaikutusta suusanalliseen viestintään asiakaspääoman kasvussa. He määrittelevät asiakaspääoman kaikkien nykyisten ja tulevien asiakkaiden elinkaariarvojen summaksi. Asiakaspääoma kasvaa muun muassa asiakkaan negatiivisen tai positiivisen suusanallisen viestinnän kautta. He käyttivät mittarina asiakashankinnan tehokkuutta suhteessa asiakaspääomaan. Mittarin mallintamiseen he käyttivät vektoriautoregressiiviä (VAR). Data oli kerätty Internet-yritykseltä, joka tarjosi ilmaista Web hosting -palvelua 70 viikon ajaksi. Rekisteröinnin kohdalla asiakkailta oli kysytty demografisia tietoja sekä sitä, miten he olivat kuulleet yrityksestä.

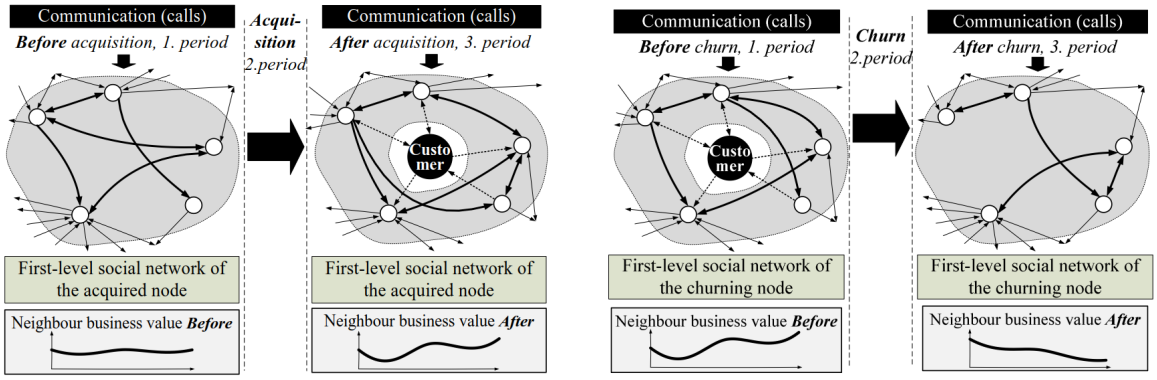
Asiakashankinnassa on myös mahdollista huomioida sosiaalinen verkosto, vaikka Bijmolt ym. (2010) luokittelivat sosiaalisen verkoston asiakkuuden kehittämiseen ja säilyttämiseen. Kazienkon, Brodkanin ja Rutanin (2009) mukaan asiakashankinnan näkökulmasta uudella asiakkaalla voi olla merkittävä vaikutus sen lähipiirin suhteisiin, joihin asiakas liittyy. Näin siksi, että uusi asiakas voi joko kasvattaa tai vähentää kommunikointia vanhojen asiakkaiden välillä. Asiakashankinnan vaikutuksen lisäksi Kazienko, Brodka ja Ruta (2009) tutkivat asiakkaan poistumisen (ks. luku 3.3.1) vaikutusta asiakkaiden lähipiiriin. He tutkivat, kuinka paljon asiakkaiden käytös vaikuttaa muihin asiakkaisiin, ja voidaanko vaikutusta arvioida datasta, joka kuvaa asiakkaiden keskinäisiä yhteyksiä.

Tutkimuksessa käytettiin kahden modernin televiestintäpalvelujen tarjoajan dataa, sillä heidän asiakkaat luovat epäsuorasti vuorovaikutteisen sosiaalisen verkoston. Lisäksi puhelut sisältävät paljon tietoa asiakkaan toiminnasta. Kazienko, Brodka ja Ruta (2009) muodostivat televiestinnän sosiaalisen verkoston niin, että asiakkaat ovat solmuja ja asiakkaiden välillä on yhdistävä suhde. Asiakas liittyy yhteen puhelinnumeroon, joka on suhteessa sosiaaliseen kokonaisuuteen eli ihmiseen, ihmisryhmään tai yritykseen. Asiakas ja sen lähipiiri voi kuulua eri sosiaaliseen kokonaisuuteen. Analysointimittareina tutkimuksessa käytettiin soitettujen puheluiden kokonaismäärää ja kestoa sekä vastaanotettujen ja soitettujen puheluiden

kokonaismäärää ja kestoa. Näiden lisäksi käytettiin sosiaalisen aseman mittaria, asiakkaan lähipiiriä ja lähipiirin sosiaalista arvoa (ks. Kazienko, Brodka ja Ruta 2009, s.492-493).

Kazienkon, Brodkan ja Rutan (2009) mukaan televiestintäpalveluyrityksissä kommunikoinnin muutokset näkyvät lähipiirin arvon yksilöllisissä muutoksissa ja sitä voidaan tarkastella yhteiskunnallisen arvon dynamiikan puitteissa. Asiakas muodostaa lähipiirin niiden kanssa, joille hän on soittanut, ja jotka ovat soittaneet hänelle ainakin kerran elämän aikana. Nämä yhteydet muodostavat dynaamisen arvon. Jokainen yhteys eli asiakkaan sosiaalinen arvo lisää verkoston arvoa.

Kazienkon, Brodkan ja Rutan (2009) sosiaalisen lähipiirin analysointiprosessi eteni vaiheittain. Ensimmäisessä vaiheessa tunnistettiin yhteydet asiakkaiden välillä. Tämän jälkeen etsittiin analysoitavat asiakkaat, jotka oli hankittu keskivaiheessa analyysia tai asiakkaat, jotka olivat jo poistuneet. Kun asiakkaiden väliset suhteet ja analysoitavat asiakkaat oli tunnistettu, muodostetiin lähipiiri poistumista ennen olevista asiakkaista ja hankkimisen jälkeisistä asiakkaista. Tämän jälkeen mitattiin kokonaisliiketoiminta-arvo lähipiiristä aikasarjana pisteestä ennen tapahtumaa (asiakashankinta/asiakkaan poistuminen) tapahtuman jälkeiseen pisteeseen. Tutkimusdata oli kuukauden ajalta ja he olivat jakaneet datan kolmeen 10 päivän osaan. Ensimmäistä dataosiota käytettiin lähipiirin tunnistamiseen sekä lähipiirin arvonlaskemiseen ennen asiakashankintaa tai asiakkaan poistumista. Datan toista osiota hyödynnettiin hankitun asiakkaan tai poistuneen asiakkaan tunnistamiseen. Viimeisestä osiosta laskettiin tapahtuman jälkeiset arvot ja niiden muutokset. Kuviossa 1 on esitetty hankitun sekä poistuneen asiakkaan vaikutukset lähipiiriinsä ja lasketun liiketoiminta-arvon muutos.



Kuvio 1: Vasemmalla on hankitun asiakkaan vaikutus lähipiiriin ja oikealla poistuneen asiakkaan vaikutus lähipiiriin (Kazienko, Brodka ja Ruta 2009)

Lähipiirin lisäksi asiakasvalituksia ja -palautteita voidaan hyödyntää asiakashankinnassa. Faed, Hussain ja Chang (2014) rakensivat käsitteellisen viitekehyksen vastaamaan asiakkuudenhallintaan liittyvän teknologian puutteisiin ja ratkaisemaan kysymykset, jotka liittyvät lojaliteetin parantamiseen ja asiakashankintaan. Heidän mielestään valituksilla ja niiden käsittelyllä on tärkeä rooli asiakassuhteissa, sillä jos valitukset käsitellään tehokkaasti, on todennäköistä, että asiakas on tyytyväinen ja pysyy lojaalina. Tutkimuksellaan Faed, Hussain ja Chang (2014) haluavat pienentää aikaa asiakkaan valituksen tekemisestä siihen, kun yritys saa valituksen.

Bijmolt ym. (2010) luokittelevat valitukset asiakkuuden kehittämisvaiheeseen, mutta tässä tutkielmassa kyseinen tutkimus luokitellaan asiakashankintaan, koska Faed, Hussain ja Chang (2014) osoittavat tutkimuksessaan asiakkaan tyytyväisyyden suhteen lojaalisuuteen ja asiakashankintaan. Heidän mukaansa asiakasvalituksia ei ole aikaisemmin tutkittu lojaalisuuden ja asiakashankinnan näkökulmasta. Heidän mielestään asiakashankinnalla on suuri vaikutus asiakkuudenhallintaan ja se luo tehokkaita asiakassuhteita. Lisäksi tutkimuksessa selvitetään vuorovaikutteisuuden ja havaitun arvon vaikutusta asiakkaan tyytyväisyyteen. Luvussa 2.1 kerrottiin, että osana asiakkaan sitoutumista ovat asiakkaan lojaalius ja tyytyväisyys.

Faed, Hussain ja Chang (2014) arvioivat asiakkaan tyytyväisyyttä ja lojaaliuutta Länsi-Australian sataman asiakkaille ja työntekijöille tehdyn verkkokyselydatan sekä haastatteluiden pe-

rusteella. Haastattelussa käytettiin tekstinlouhintaa, jolla tunnistettiin yhdeksän suurta ongelmaa: johtaminen, kirjanpidonjärjestelmä, terveys ja turvallisuus, kontin sisällön lajittelu, kulut, lainkäyttövalta, ahtaajien suorituskyky ja ajanhallinta. Lisäksi Faed, Hussain ja Chang (2014) käyttivät lineaarista optimointia yksinkertaistamaan päätöksentekomenettelyä. Päätöksentekomenettelyyn liittyy asiakkaiden useita valituksia, joille annettiin eri koodit tunnistamaan ja käsittelemään niitä edelleen. Valituksien koodaamiseen käytettiin tekstilouhintaa. Yksinkertaistamisen jälkeen Faed, Hussain ja Chang (2014) halusivat saada selville, onko yhdeksällä ongelmalla yhteyttä lojaaliuteen ja asiakashankintaan.

Tutkimuksessa pääkomponenttianalyysia, klusterointia ja tietojenkeruuanalyysiä (engl. data envelopment analysis, DEA) käytettiin tunnistamaan ja kategorioimaan asiakkaat tärkeisiin ja erittäin tärkeisiin asiakkaisiin. Tunnistamisen ja kategorioinnin jälkeen Faed, Hussain ja Chang (2014) analysoivat asiakkaiden suhteita lineaarisella mallilla. Suhteiden muodostamiseen he käyttivät sumeaa logiikkaa (engl. fuzzy logic). Suhteiden lineaarisessa analyysissä oli kaksi vaihetta: tekijöiden määritys ja lineaarisen mallin arviointi asiakkaan tyytyväisyyden, havaitun arvon ja vuorovaikutteisuuden välillä sekä lojaaliuuden, asiakashankinnan ja asiakkaan tyytyväisyyden välillä. Tekijöiden määrityksessä kullekin asiakkaalle laskettiin neljässä osassa (havaittu arvo, tyytyväisyys, asiakashankinta, lojaaliuus) pääkomponenttianalyysilla yleinen sopimusaste (engl. overall degree of agreement, ODA). Sumealla logiikalla Faed, Hussain ja Chang (2014) todistivat suhteet ja tuottivat samalla sääntöjä, joita yrityksen on otettava huomioon. Lineaarisen analyysin avulla asiakkaan tyytyväisyyttä yritykseen ja asiakashankintaan ei voitu taata. Toisin oli sumean logiikan kohdalla, sillä se voi antaa säännöt valitusten käsittelemiseksi ja kunkin sovelletun muuttujan tason nostamiseksi.

Asiakkaan tyytyväisyyttä voidaan selvittää myös sitoutumispisteytyksen kautta, johon hyödynnetään suusanallista viestintää. Vanderveldin ym. (2016) tutkimus käsitteli suurelle globaalille verkkokaupalle (Groupon) tehtyä asiakkaan elinkaaren arvon mallia, joka analysoi ei-ostavia asiakkaita sitouttamispisteytyksellä, koska ostotietoja ei ole mahdollista käyttää analysointiin. Sitoutumispisteytys ennustaa myös nopeammin muutoksia asiakkaan elinkaaren arvossa kuin ostotietoihin perustuvat mallit, kuten asiakkaan tyytymättömyyden. Tutkimuksessa sitouttamispisteytys muodostettiin sähköposteista ja Groupon Mobile App -mobiilisovelluksesta.

Vanderveldin ym. (2016) malli hyödyntää kuitenkin käyttäjän yleistä ostokäyttäytymistietoa, sillä se jakaa käyttäjät ostotietojen perusteella kuuteen ryhmään. Jokainen ryhmä mallinnetaan erikseen, jotta muuttujien painotukset voidaan määrittää. Mallia opetetaan jokaisen kuuden käyttäjäryhmän ja kolmen aikaikkunan (lyhyt, keskikokoinen ja pitkä; neljännesvuosittaisesta vuoteen) yhdistelmällä. Mallissa käytetään satunnaista metsää (engl. random forest) ja sen algortimia. Algoritmissa muuttujista muodostettaan satunnaisesti metsäehdokkaat. Tutkimuksen tarkoitusta vastasi parhaiten algoritmi, jossa muuttujia oli vähän ja luotuja puita oli paljon. Mallin ensimmäisessä vaiheessa ennustetaan käyttäjän ostoaikomusta aikaikkunassa kahden luokan luokittelulla (engl. binary classification) eli ennustetaan käyttäjän ostavan tai jättävän ostamatta. Toisessa vaiheessa käyttäjälle, jonka ennustetaan ostavan, ennustetaan asiakkaan arvo dollareina.

Yhteensä Vanderveldin ym. (2016) malli käyttää yli 40 muuttujaa, kuten käyttäjän demografisia tietoja ja yleistä suhdetta Grouponiin. Sitouttamispisteytys on käytöksellä painotettu summa jokaisesta yksittäisestä alustasta. Käyttäytymistä voidaan mitata esimerkiksi Grouponin lähettämien sähköpostiviestien avaamisten ja klikkausten määrillä sekä mobiilisovelluksen näyttökerroilla ja haulla. Käyttäjäkokemusta mitataan asiakaspalvelun avulla, esimerkiksi palautuksien ja asiakaspalvelun tikettien määrällä, asiakaspalvelun puhelimen ja sähköpostin odotusajalla sekä sillä antaako asiakas positiivista vai negatiivista palautetta asiakaspalvelusta jälkikäteen tehdyssä kyselyssä.

3.1.2 Sivukyselydataa hyödyntävät tutkimukset

Asiakashankintavaiheessa asiakkaan kiinnostusta voidaan selvittää sivukyselydatan avulla. Su ja Chen (2015) hyödynsivät sivukyselydataa ja muodostivat Kiinan suurimman verkkokaupan datasta toimintamalleja, jotka kuvaavat verkkokaupan käyttäjien kiinnostusta käyttäjien arvioita paremmin ja tarkemmin. Kehitetyn algoritmin avulla käyttäjät voidaan jakaa ryhmiin käytöksen mukaan ja siten määrittää kiinnotuksen toimintamalli. Su ja Chen (2015) huomioivat, että käyttäjillä on erilaisia kiinnostuksen kohteita ja he vierailevat verkkokaupan erilaisissa kategorioissa sekä useissa tuotteissa. Tällöin käyttäjien selailupolku, vierailujen määrä ja sivustolla käytetty aika vaihtelee. He myös ottivat huomioon ajan kynnysarvona (engl. threshold) sen, että käyttäjä voi avata sivun, mutta ei selaile sivua koko aikaa. Käyt-

täytymisen mallintaminen auttaa asiakkaan personoinnissa ja siten esimerkiksi yhteistyöhaluukkuuden tarkastelussa.

Sunin ja Chenin (2015) algoritmi pohjautuu klusterointiin ja koska yksi käyttäjä voi kuulua erilaisten kiinnostuksiensa vuoksi moneen klusteriin, he integroivat algoritmiin myös rough set teorian. Tutkimuksessa mitattiin kahden käyttäjän samanlaisuutta vierailujen määrän ja keston suhteen. Tämän jälkeen polkujen samankaltaisuus määriteltiin kahden käyttäjän sekvenssinä. Näistä kolmesta samankaltaisuuksista muodostettiin yhteinen käyttäjäparin samankaltaisuus. Yleinen kynnysarvo kontrolloi käyttäjien samanlaisuutta jokaisessa ryhmässä ja rough kynnysarvo mahdollistaa käyttäjän olemisen monessa klusterissa. Lisäksi tutkimuksessa verrattiin kehitettyä algoritmia (engl. rough leader clustering algorithm), perinteistä klusteriointialgoritmia (engl. traditional leader clustering algorithm) ja k:n medoidin klusterointimenetelmää. Sisäisten klustereiden validointi-indeksien mukaan kehitetty algoritmi on parempi verrattuna perinteiseen klusterointialgoritmiin ja k:n medoidin klusterointimenetelmään. Lisäksi se on paljon tehokkaampi kuin k:n medoidin klusterointimenetelmä. Tutkimusessa datajoukko oli suuri (198 325 tietuetta) ja moniulotteinen.

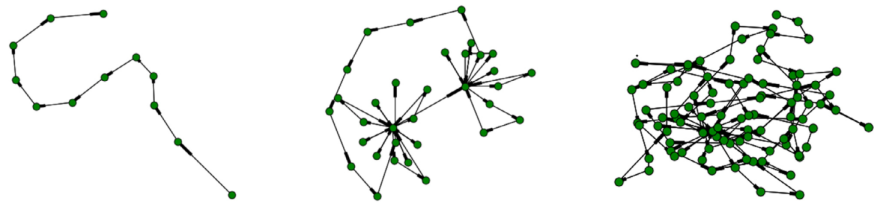
Asiakkaan verkkokäyttäytymistä tutkivat myös Raphaeli, Goldstein ja Fink (2017) Israelilaisen verkkokaupan datan avulla. Tutkimuksessa vertailtiin, miten asiakkaiden käytös eroaa PC- ja mobiililaitteiden välillä. Tutkimuksen kiinnostuksen kohteena olivat ostomotivaatio ja tavoitteet, jotka Raphaelin, Goldsteinin ja Finkin (2017) mielestä perustuvat selailukäyttäytymiseen. Käytöstä he luonnehtivat sitoutumismittauksilla ja navigointimallien (engl. navigation patterns) sarjana käyttäen lähestymistapaa, mikä yhdistää jalanjäljen visualisoinnin sekä peräkkäisten assosiaatiosääntöjen louhinnan (engl. sequential association rule mining). Jalanjäljen kuvaaja (engl. footstep graph) on yleinen visualisointi tekniikka kuvaamaan käyttäjän session aikaista selailupolkua, missä vaaka-akselilla on viipymisaika sivulla ja pystyakselilla on vierailtu sivu (Raphaeli, Goldstein ja Fink 2017). Tutkimuksessa sessioiden navigointimallit muutettiin sarjaksi, joita visualisoitiin jalanjäljen kuvaajina.

Tutkimus oli kaksivaiheinen sisältäen käytöslouhinnan (engl. usage mining) ja mallin analysoinnin (engl. pattern analysis). Ensimmäinen vaihe keskittyi mobiili- ja PC-laitteiden selailukäyttäytymisen eroavaisuuksien ilmentämiseen. Sitoutumismittareiden eroavaisuuksiin käytettiin T-testiä. Sitoutumismittareita olivat session kesto, keskimääräinen sivun kesto ja

vierailtujen sivujen määrää. Session korkeampi kesto ja alhaisempi vierailtujen sivujen määrä viittasi vahvempaan sitoutumiseen. Peräkkäisten assosiaatiosääntöjen louhinnalla Raphaeli, Goldstein ja Fink (2017) tutkivat navigointimallin sarjan toistuvuutta ja sitä kautta eroja. Lift-arvolla he ennustivat ostamista. Mallin analysoinnissa he jakoivat datan neljään alaotokseen: mobiililaitteella ostaminen, PC-laitteella ostaminen, ei ostoa mobiililaitteilla ja ei ostoa PC-laitteella. Jokaiselle alaotokselle suoritettiin erikseen peräkkäisten assosiaatiosääntöjen louhinta, jotta pystyttiin tunnistamaan esimerkiksi, miten navigointisäännöt eroavat mobiililaitteiden ja PC-laitteiden sessioiden välillä.

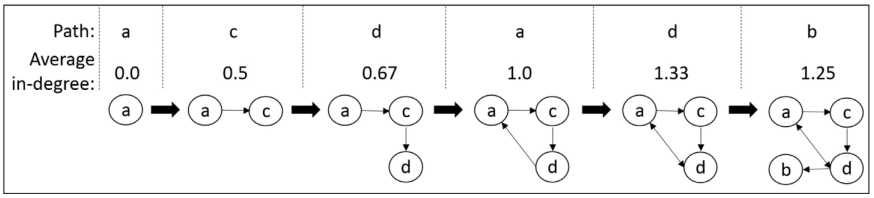
Baumann ym. (2017) puolestaan tutkivat verkkokäyttäytymistä sivukyselydatan avulla vaihtoehtoisesta näkökulmasta ja ei-perinteisten mittareiden avulla. Tutkimuksessa selvitettiin, pystyykö verkkokäyttäytymistä ennustamaan hyödyntämällä käyttäjäverkkoja (engl. user graph). Verkot ovat metodologinen lähestymistapa, joka perustuu verkkoteoriaan ja ne rakentuvat solmuista ja kaarista (ks. Baumann ym. 2017, s.137). Verkon jokainen solmu on verkkosivusto, jossa käyttäjä on vieraillut session aikana. Verkon kaari muodostuu verkkosivustojen välille, jos käyttäjät ovat vierailleet niissä peräkkäin. Kaaret muodostuvat siinä järjestyksessä, missä käyttäjä vierailee sivustoilla. Verkko muodostuu inkrementaalisesti solmujen eli verkkosivujen vierailujen lisääntyessä. Baumann ym. (2017) muodostivat verkon jokaisesta käyttäjän sessiosta, jotta pystyivät johtamaan kovarianssit ennustamaan ostoja.

Baumannin ym. (ks. 2017, s.140) mukaan verkon rakenne kuvaa käyttäjän käyttäytymistä ja sitä kautta verkon mittareilla voidaan ennustaa käyttäytymisen muutokset (ks. kuvio 2). Sivukyselydata kuvaa käyttäjän selailupolut, joka on suoraan verrattavissa käyttäjän vierailun tarkoitukseen. Muutokset kuvaavat puolestaan käyttäjän aikomuksia. Esimerkiksi suora polku solmusta toiseen kuvaa vahvaa tavoitteellista käytöstä, koska polulla ei ole palattu takaisin (kuvion vasen verkko). Täten suora polku voidaan yhdistää ostoon liittyväksi. Verkosta tulee polkua monimutkaisempi, jos käyttäjä etsii useaa tuotetta eri kategorioista eli asiakkaan verkkokäyttäytyminen on vertailevaa (kuvion keskimmäinen verkko). Selailukäytöstä puolestaan kuvaa verkko, jossa on monta keskisolmua, esimerkiksi hakutulosten yleiskatsaussivut, joihin käyttäjä palaa tarkasteltuaan tuotteita (kuvion oikeanpuoleinen verkko).



Kuvio 2: Asiakkaan käytöstä kuvaavat sessioista muodostetut verkot (Baumann ym. 2017)

Tutkimuksessa muodostettiin verkosta rakenteellisia, keskeisiä ja välimatkallisia mittareita, joita käytettiin ennustemalleissa (ks. Baumann ym. 2017, s.140). Jokaiselle sivuvierailulle muodostettiin uusi kuvaaja, joista mittarit laskettiin inkrementaalisesti keskimääräisen asteen (average in-degree) avulla. Keskimääräinen aste on solmuihin yhteydessä olevien kaarien määrä jaettuna solmujen määrällä (ks. kuvio 3). Rakenteellisilla mittareilla kuvataan verkkoa, kuten kaikkien solmujen ja sivujen lukumäärää. Verkon kehät ja silmukat kuvaavat siirtymiskäyttäytymistä nykyiseltä verkkosivulta edelliselle sivulle tai siirtymistä samalla verkkosivulla. Transitiivisuus ilmaisee kolmiomuodostelman lukumäärää ja verkkokäyttäytymisen näkökulmasta mittari ilmaisee sitä, että käyttäjä miettii ja vaihtelee kahden tuotteen välillä. Käyttäjän tarkoituksen selvittämiseen voi käyttää myös tiheysmittaria. Pienet tiheysarvot viittaavat käyttäjän tavoitteelliseen toimintaan, sillä tällöin käyttäjä liikkuu sivuilta sivuille ilman pyörimistä tai takaisin menemistä. Solmujen liitettävyyden arvo on keskimääräinen lukumäärä jokaisesta erillisestä solmuparista. Sen korkea arvo merkitsee sitä, että huomattava määrä solmuja on yhteen punoutunut. Tämä puolestaan kuvastaa sitä, että käyttäjän verkkokäyttäytyminen ei ole järjestelmällistä.



Kuvio 3: Asiakkaan session keskimääräinen aste sivuvierailuittain (Baumann ym. 2017)

Baumann ym. (ks. 2017, s. 140) kuvaavat välimatkamittareilla verkon laajuutta. Lyhyin polkumittari on verkon kaikkien erillisten solmuparien lyhimmän reitin pituuden keskiarvo. Se

liitettynä verkon epäkeskisyyteen (muuttuja eccentricity), koko verkon epäkeskisyyden minimiin ja maksimiin (muuttujat diameter ja radius), solmujen lukumäärään, joiden epäkeskisyyden arvo on sama kuin epäkeskisyyden minimi ja maksimi (muuttujat center ja periphery) kertovat, miten monipuolisesti käyttäjä selailee verkkosivuja. Esimerkiksi korkea arvo kertoo, että käyttäjä katselee peräkkäin verkkosivuja ilman taaksepäin selailua.

Baumann ym. (ks. 2017, s. 140) kuvaavat solmujen tärkeyttä keskeisyysmittareilla, sillä tietyn solmun keskeisyyden aste voi kuvata sitä, miten monesti käyttäjä on palanut takaisin tietylle sivulle muilta sivuilta. Läheisyysmittari kuvaa läheisyyden keskiarvon eli verkon kaikkien solmujen keskiosan. Vireysmittari kuvaa kaikkien solmujen läheisyyden muutoksen, jos yksi solmu on poistettu verkosta. Näiden kahden mittarien arvo on korkea, jos jokin solmu on keskellä verkkoa. Verkkokäyttäytymisen näkökulmasta korkea arvo kuvastaa sitä, että käyttäjä on vierailut lukuisia kertoja tietyllä verkkosivustolla istunnon aikana.

Tutkimuksessaan Baumann ym. (2017) käyttivät kahden suuren verkkokaupan vaatteiden ja kenkien myyntiä kuvaavaa dataa kahdelta kuukaudelta ja jakoivat datat kuukauden mukaisesti opetusjoukkoon ja testausjoukkoon. Tutkimuksessa käytettiin kolmea luokittelualgoritmia ennustamaan johtaako käyttäjän sessio ostotapahtumaan: yleistettyä lineaarista logistista regressiomallia (GLM), satunnaismetsää (RF) ja gradient boosting -menetelmää (GB). Muuttujina malleissa käytettiin verkosta muodostettuja mittareita. Ennustemallin arvioimiseen käytettiin area-under-the-precision-recall-curves:ia (AUC-PR) ja lift-indeksiä.

Datan esikäsittelyssä Baumann ym. (2017) poistivat käyttäjän sessiot, joissa oli alle neljä klikkausta sivuilla, koska vähintään neljä tarvitaan ostojen tekemiseen. Tämän jälkeen verkon mittareista muodostettiin korrelaatiomatriisi ymmärtämään, mitkä muuttujat sisältävät samanlaista tietoa käyttäjän selailun rakenteesta verkkosivustolla. Korrelaatiomatriisin tuloksena oli 13 verkon mittaria, joiden ennustettavuutta vertailtiin luokittelualgoritmilla muissa tutkimuksissa käytettyihin perinteisiin mittareihin (mm. vierailtujen sivujen määrä, session kokonaisaika jne.) Ostojen ennustamiseen tärkeimpiä muuttujia ovat vireys, silmukoiden lukumäärä, kehien lukumäärä ja koko verkon epäkeskisyyden minimi. Malleista GLM ja RF keskeisillä ja välimatkamittareilla ovat tehokkaita ennustamaan ostokäyttäytymistä.

3.1.3 Yhteistyöhalukkuuteen liitettävät tutkimukset

Sosiaalisen median kautta on mahdollista seurata tuotteista käyttävää keskustelua ja siten havaita uusia asiakkaita. Du ym. (2013) ovat kehitelleet työkalun, jolla voidaan hyödyntää twiittien sisältöä ja sosiaalisia suhteita, löytämään potentiaalisia asiakkaita erilaisille tuotteille. Työkalu eroaa muista siinä, että siinä otetaan huomioon käyttäjän läheiset offline-ystävät, joiden kautta on mahdolista huomata esimerkiksi henkilön aikomus ostaa lahja läheiselle ystävälleen. Tällöin yritys voi reagoida lahjanostoaikeiseen kohdistetulla markkinoinnilla. Työkalu muodostaa merkityksellisyyspisteen jokaiselle käyttäjälle sekä sosiaalisen verkoston tekstitietojen että ihmissuhteiden avulla. Merkityksellisyyspiste kertoo käyttäjän mahdollisuudesta ostaa tuote, sillä se kuvaa käyttäjän keskustelun volyymiä tuotteesta. Lisäksi pisteen avulla saa selville lojaalit asiakkaat, koska se perustuu käyttäjän kaikkiin twiitteihin.

Työkalun data koostuu Twitterin aikajanasta, käyttäjän profiilin tiedoista ja verkostosta. Merkityksellisyyspisteen algoritmissa Du ym. (2013) käyttävät hyödyksi tiedonhakupalvelun perusideaa eli löytyykö termi dokumentista vai ei löydy. Jos termi löytyy useammin, niin sitä korkeampi merkityksellisyyspiste. Offline ystäväverkoston algoritmi tuottaa lähimmäisyyspisteen käyttäjien välillä hyödyntäen strukturaalista verkostoa ja soveltaen iteratiivisesti satunnaista kävelyä. Visuaaliseen analyysiin työkalu käyttää esimerkiksi tutkakaaviota, joka kertoo käyttäjän merkityksellisyyspisteestä eri ulottuvuuksissa.

Taulukoon 2 on koottu yhteenveto asiakashankintaan liittyvistä tutkimuksista. Taulukosta näkee tutkimuksissa hyödynnetyt menetelmät ja datat. Tutkimukset hyödynsivät suusanallisen viestinnän dataa, sivukyselydataa ja Twitter-dataa, joka liittyi yhteistyöhalukkuuden selvittämiseen. Suusanallisen viestinnän datat muodostuivat asiakkaille tehdyistä erilaisista kyselyistä, joista tutkittiin brändikokemusta, asiakkaiden keskinäisiä yhteyksiä, asiakasvalituksien ja asiakassuhteiden välistä yhteyttä ja sitoutumispisteytyksellä asiakkaan tyytymättömyyttä. Sivukyselydatojen kohdalla tutkittiin asiakkaiden kiinnostusta vierailujen määrällä ja kestolla sekä asiakkaiden käytöstä esimerkiksi käyttäjäverkkojen avulla. Verkkokäyttäytymisestä huomattiin, että session korkeampi kesto ja alhaisempi vierailtujen sivujen määrä kuvastaan asiakkaan vahvempaa sitoutumista. Tässä tutkielmassa asiakkaan sitoutumista tarkastellaan sivukyselydatasta session kestolla, sessiossa katsottujen eri tuotteiden määrällä sekä ostoskoriin siirrettyjen eri tuotteiden ja ostojen määrällä.

Taulukko 2: Yhteenveto asiakashankinnan tutkimuksista

Tutkimus	Luokittelukriteeri	Metodi	Data
Bowman ja Narayandas (2001)	Suusanallinen viestintä	Katkaistu NBD-malli	Puhelinkysely
Wangenheim ja Bayon (2007)	Suusanallinen viestintä	ZIP-malli	Puhelinkysely
Villanueva, Yoo ja Hanssens (2008)	Suusanallinen viestintä	VAR-malli	Internet-kysely
Kazienko, Brodka ja Ruta (2009)	Suusanallinen viestintä	Sosiaalisen verkoston analyysi	Televiestintäpalveluyrityksen data
Faed, Hussain ja Chang (2014)	Suusanallinen viestintä	Tekstinlouhinta, PCA, DEA, sumea logiikka	Verkkokysely ja haastattelu
Vanderveld ym. (2016)	Suusanallinen viestintä	Random Forest ja kahden luokan luokittelu	Eri sovellusalustojen data ja asiakaspalveludata
Su ja Chen (2015)	Sivukyselydata	Rough pohjainen klusterointi	Verkkokaupan data
Raphaeli, Goldstein ja Fink (2017)	Sivukyselydata	Käytösloulhinta ja peräkkäisten assosaatiosääntöjen louhinta	Verkkokaupan data
Baumann ym. (2017)	Sivukyselydata	Käyttäjäverkot, yleistetty lineaarinen logistinen regressiomalli, satunnainen metsä ja XGBoost	Verkkokaupan data
Du ym. (2013)	Yhteistyö halukkuus	Hakualgoritmi, struktuaalinen verkko ja satunnainen kävely	Twitter-data

3.2 Asiakkuuden kehittämisen tutkimukset

Asiakkuuden kehittämisen mallit ovat perinteisesti keskittyneet asiakkaan elinkaaren arvon sekä kuluttajien marginaalin ennustamiseen ja ristiinmyyntiin. Marginaalin ennustamiseen on Bijmoltin ym. (2010) mukaan käytetty ekonometrisia menetelmiä ja ristiinmyyntiin logit-mallia, erotteluanalyysiä (engl. discriminant analysis) ja neuroverkkoja. Datana malleissa on käytetty suurimmaksi osaksi asiakkaan ostotietoja. Sitoutumisen näkökulmasta asiakkuuden kehittäminen vaatii asiakkaan käytöksen ymmärtämistä ja analysointia, esimerkiksi suusanallisen viestinnän ja yhteistyön vaikutusta asiakkaan elinkaaren arvoon. Lisäksi Bijmoltin ym. (2010) mukaan yrityksille yhteistyön hallinnointi ja stimulointi ovat tärkeitä tehtäviä, sillä brändiyhteisöt voivat luoda arvoa yritysmaailman verkottuneiden toimijoiden keskuudessa.

Asiakkaan sitoutumisen ilmentymä, joka vaikuttaa asiakkaan elinkaaren arvoon, ovat asiakasvalitukset. Bijmoltin ym. (2010) mukaan yritykset tiedostavat, että valituksien kautta voidaan korjata tuotteisiin sekä palveluihin liittyviä ongelmia ja vaikuttaa siten positiivisesti asiakkaiden käyttäytymiseen. Monet yritykset eivät kuitenkaan tiedosta, että valituksien tekeminen ja varsinkin valitusten hallinta voi vaikuttaa asiakkaan elinkaaren arvoon. Bijmoltin ym. (2010) mielestä valituksien mallintaminen ja yrityksen valituksien käsittely on hyödyllinen tulevaisuuden tutkimusalue asiakkaan sitoutumisessa.

3.2.1 Suusanalliseen viestintään ja sosiaalisen verkoston kasvuun liittyvät tutkimukset

Suusanallisen viestinnän vaikutusta yrityksen voittoon voidaan tutkia sosiaalisen verkoston avulla. Goldenberg ym. (2007) tutkivat agenttipohjaisella mallilla (engl. agent-based model), joka on the Small World mallin laajennus, miten yksilön ja verkoston negatiivinen suusanallinen viestintä vaikuttaa yrityksen voittoon. The Small World -mallin solmut kuvaavat potentiaalisia asiakkaita neljässä vaiheessa: hyväksynyt tuotteen, ei hyväksynyt tuotetta, tyytyväinen tuotteeseen tai tyytymätön tuotteeseen. Siteillä kuvataan asiakkaan yhteyttä toiseen. Side on vahva, jos toinen asiakas on liitettynä asiakkaan verkostoon ja heikko, kun yhteyksiä on verkon ulkopuolisiin. Malli muodostaa renkaan solmuista ja suorista siteistä (vahvat ja heikot). Lisäksi mallissa on oikoteitä solmujen välillä renkaan lävitse, joita Goldenberg ym. (2007) muuttivat siten, että ne vaihtelivat ajan ja sosiaalisen siteen vahvuuden mukaan. Tutkimuksessa analysointiin sosiaalisen järjestelmän tietorakenteen muutoksia, voimakkaiden ja heikkojen siteiden intensiteettiä, markkinoiden vaikutusta, negatiivisen suusanallisen viestinnän vaikutusta verrattuna positiiviseen suusanalliseen viestintään ja pettyneiden asiakkaiden lukumäärän vaikutusta yrityksen myyntiin ja kassavirran nettonykyarvoon.

Mallissa käytetty data oli MBA-opiskelijoille järjestetystä kyselystä, jota kautta selvitettiin sosiaalisen verkoston yhteyksiä ja keskinäistä viestintää. Verkoston vahvaa sidosta Goldenberg ym. (2007) etsivät tiedustelemalla muun muassa henkilöiden osoitekirjassa olevien ihmisten määrää. Heikon sidoksen tutkimiseksi opiskelijoita pyydettiin arvioimaan, kuinka moneen he olivat viimeisen viikon aikana olleet satunnaisesti yhteydessä. Tutkimuksessa saatiin selville, että negatiivinen suusanallinen viestintä vaikuttaa yrityksen nettonykyarvoon merkittävästi, vaikka tyytymättömien asiakkaiden määrä olisikin pieni (Goldenberg ym. 2007; Bijmolt ym. 2010).

Trusov, Bucklin ja Pauwels (2009) tutkivat suusanallisen viestinnän vaikutusta sosiaalisen median verkoston kasvuun. He myös vertasivat sen vaikutusta perinteiseen markkinointiin. Tutkimuksessa kehitettiin ja arvioitiin malli, joka ottaa huomioon dynaamiset suhteet asiakashankinnan, suusanallisen viestinnän ja perinteisten markkinointitoimien välillä. Asiakashankintaa mitattiin uusilla kirjautumisilla sivustolle. Käytetty data oli yksi sosiaalisen verkoston tärkein sivusto, jonka nimeä tutkimuksessa ei tuotu julki.

Ensimmäisenä Trusov, Bucklin ja Pauwels (2009) testasivat endogeenisuuden olemassa oloa uusilla kirjautumisilla, tapahtumamarkkinoinnilla (maksettu markkinointi sosiaalisessa mediassa), median esiintymisillä (suhdetoiminnan aiheuttamia) ja suusanallisella viestinnällä. Tämän jälkeen he hyödynsivät vektoriautoregressiivistä (VAR) mallia, joka kuvaa endogeenisuutta ja dynaamista vastausta sekä vuorovaikutusta markkinointimuuttujien ja tuloksien välillä. Lisäksi tutkimuksessa verrattiin VAR-mallia vaihtoehtoisiin malleihin. Viimeiseksi tutkimuksessa arvioitiin kirjautumisia lyhyeltä sekä pitkältä ajalta suusanalliseen viestintään ja perinteiseen markkinointiin sekä laskettiin vastaava jousto η_{arc} (ks. kaava 3.1). Tutkimus osoittaa, että suusanallisen viestinnän jousto on noin 20 kertaa korkeampi kuin tapahtumamarkkinoinnin ja 10 kertaa korkeampi kuin median esiintymisen jousto (Bijmolt ym. 2010; Trusov, Bucklin ja Pauwels 2009).

$$\eta_{\text{arc}} = \frac{\Delta Y}{\sigma_X} \times \frac{\bar{X}}{\bar{Y}} \tag{3.1}$$

missä X on suusanallinen viestintä ja Y on kirjautumiset.

Sosiaalisen verkoston kasvuun vaikuttaa myös tiedonleviäminen. Zadeh ja Sharda (2014) tutkivat moniuloitteisen pisteprosessin (engl. multidimensional point process) avulla tiedon leviämistä Twitterissä ja sitä kautta joukon sitoutumisaktiivisuutta sekä yhteyksiä. Tutkimuksen päätarkoituksena oli analysoida yksittäisten brändijulkaisujen suosiota ja sitä, miten käyttäjien vuorovaikutukset verkkoystävien kanssa edistää brändijulkaisujen suosiota. Julkaisujen suosion tutkimiseen käyttivät ETAS-mallia (Epidemic Type Aftershock Sequences) ja Hawkes-pisteprosessia sekä ajan ja seuraajien määrän yhdistettyä todennäköisyys -funktiota. Tuloksien mukaan seuraajien määrä on yksi parhaimmista mittareista, jolla pystytään demonstroimaan vaikuttavien käyttäjien roolia sosiaalisissa verkostoissa. Seuraajien lisäksi on huomioitava tapahtumien esiintymisaika sosiaalisen dynamiikan mallintamisessa.

Okazaki ym. (2015) käsittelivät verkoston kasvua elektronisen suusanallisen viestinnän (electronic Word-of-mouth, eWOM) kannalta. He tutkivat asiakkaan sitoutumista brändin sosiaaliseen verkostoon. Tutkimuksessa analysoitiin viestejä tiedonlouhinnan avulla, jotka oli julkaistu Ikean Twitteriin tai twiitteihin. Tiedonlouhinnan avulla tunnistettiin elektronisen suusanallisen viestinnän käyttäytymismallit, joilla pystyi selventämään, minkä tyyppistä elekt-

roniseen suusanalliseen viestintään liittyvää käytöstä ilmenee, kun asiakas on sitoutunut brändiin sosiaalisessa verkostossa. Okazakin ym. (2015) mielestä sosiaalisessa verkostossa on kolme elektronisen suusanallisen viestinnän muotoa: objektiivisia ja subjektiivisia lausuntoja sekä tiedon jakamista. Objektiivinen lausunto perustuu faktoihin, kuten tietoihin ja kysymyksiin. Subjektiiviset lausunnot ovat luonteeltaan tuomitsevia, kuten kritiikit tai ylistykset. Tiedon jakamisessa viesti odotetaan jaettavan ja kierrätettävän julkisesti.

Lisäksi tutkimuksessa selvitettiin, miten tiedon jakaminen edistää kuluttajaverkostojen luomista Internetissä ja miten asiakkaat osallistuvat tuottajakuluttajuuteen (engl. prosumption) (Okazaki ym. 2015). Näin siksi, että viestien uudelleen twiittaaminen voidaan Okazakin ym. (2015) mukaan liittää suoraan yhteisön muodostamiseen. Tuottajakuluttajuus on palvelukeskeisen arvonluonnin logiikan (engl. service-dominant logic) käyttäytymisen näkökulma, minkä Okazaki ym. (2015, s. 417) määrittelevät "kuluttajan arvon luomistoiminnoiksi, jotka johtavat tuotteiden tuottamiseen ja lopulta tuotteiden kuluttamiseen sekä kulutuskokemuksen muodostumiseen". Sosiaalisessa mediassa tuottajakuluttajilla (engl. prosumers) on taipumus muodostaa monimutkaisempi sosiaalinen verkosto, mikä puolestaan heijastuu korkeampana sitoutumisena yritykseen. Monimutkaisessa verkostossa he eivät välttämättä jaa tietoa, mutta myös kritisoivat ja kyselevät toisiltaan, mistä muut ovat kiinnostuneita tai mistä yrityksen pitäisi olla huolissaan.

Tutkimuksessa käytettiin datana Ikean 300 twiittiä, jotka esikäsiteltiin ja normalisoitiin. Lisäksi data luokiteltiin kahteen luokkaan, jotka perustuivat asiakkaiden emotionaalisiin tiloihin (tyytyväisyys, tyytymättömyys, neutraali ja poissuljetut) ja vuoropuhelutoimintoihin (jakaminen, tieto, mielipide, kysymys, vastaus ja poissuljetut). Poissuljettuja ovat esimerkiksi yrityksen tai kolmannen osapuolen turhat vitsit, huono viestintä tai kaupalliset viestit. Okazaki ym. (2015) loivat mallin datasta löytääkseen hahmoja koneoppimisen menetelmillä: naiivi Bayes (NB), k:n lähimmän naapurin menetelmä (KNN), päätöspuu C4.5, tukivektorikone (SVM), neuroverkko, The uRules luokittelija (R) ja The Forest luokittelu (F). Luokittelun arviointiin tutkimuksessa käytettiin täsmällisyysparametria (precision), takaisinkutsua (recall) ja F-mittaa. Parhaan luokittelun antoi NB-luokittelija.

NB-luokittelija oli myös paras luokittelija 4 000 twiittiin, joita Okazaki ym. (2015) käyttivät tutkimuksen lopussa muodostamaan sosiaalisen verkostoanalyysin vuoropuheluluokan

perusteella. Sosiaalinen verkosto tuo esille käyttäjien elektronisen suusanallisen viestinnän mallit. Analyysi keskittyi tunnistamaan vaikutusvaltaisia toimijoita, joita ovat käyttäjät, joiden mielipide leviää helposti verkoston kautta. Verkoston solmuja ovat käyttäjät ja suhteita ovat uudelleen twiittaukset ja maininnat. Solmun väreillä kuvattiin asiakkaan emotionaalista tilaa ja kaarien kirjaimilla elektronisen suusanallisen viestinnän muotoa. Nuolella kuvattu reuna osoittaa ne käyttäjät, jotka twiittasivat uudelleen tai mainitsivat toisen käyttäjän twiitin. Verkoston analyysissä käytettiin PageRank-algoritmia, mitä Google Web -hakukone käyttää.

Tutkimuksessa sosiaalista verkostoa Okazaki ym. (2015) analysoivat siten, että esimerkiksi verkoston keskiössä olevan tuottajakuluttajan, joka jakoi aktiivisesti tunteitaan tai ajatuksiaan (paljon jakamista vuoropuhelutoiminnan luokassa), he katsoivat tyytyväiseksi asiakkaaksi. Tuottajakuluttajan katsottiin tällä tavoin osallistuvan myös aktiivisesti arvoyhteistyöhön Ikean kanssa. Tällöin vuoropuhelun jakamistoiminta vastaa elektronisessa suusanallisessa viestinnässä tiedon jakamista, joka voi johtaa tuottajakuluttajuuteen.

3.2.2 Brändiyhteisöön ja yhteistyön toimintaan liittyvät tutkimukset

Cvijikj ja Michahelles (2013) tutkivat asiakkaan sitoutumisen havaitsemista yrityksen Facebook-sivuilta. He kehittivät mallin, joka selittää yrityksen Facebook-sivuilleen laittaman sisällön mediatyypin, sisällön tyypin, julkaisun ajankohdan ja asiakkaan sitoutumisen tason suhdetta. Sitoutumista mitattiin tykkäyksien, kommenttien ja jakojen määrällä sekä vuorovaikutuksen kestolla. Cvijikjin ja Michahellesin (2013) mielestä brändiyhteisöjen sitoutumiseen vaikuttavien tekijöiden ymmärtäminen on varteenotettavaa, koska se voi lisätä suusanallista viestintää ja parantaa asenteita brändiä kohtaan ja sitä kautta mahdollisesti lisätä yrityksen tuloja. Asiakkaat voivat sitoutua yrityksen Facebook-sivuille julkaisemalla sisältöä seinälle, kommentoimalla, tykkäämällä tai jakamalla omalla seinällään yrityksen laittamaa julkaisua. Näistä muodostuu suusanallinen viestintä, sillä jokainen toiminta näkyy asiakkaan kavereille.

Tutkimuksen data koostui Top 100 -listalla olevien ruoka- ja juomayrityksien Facebook-sivujen toiminnoista. Data oli kerätty muokatulla scriptillä Facebook Graph API:n avulla. Sitoutumisen laskemiseen Cvijikj ja Michahelles (2013) muokkasivat Facebookin virallista

asiakkaan sitoutumisen mittaria eli palautteen astetta. Riippuvia muuttujia olivat tykkäyksien, kommenttien ja jakojen määrät sekä vuorovaikutuksen kesto. Datan kuvaamisessa käytetty jakauma oli negatiivinen binomijakauma.

Cvijikjin ja Michahellesin (2013) tuloksien mukaan asiakkaat sitoutuvat useammin sisällön tykkäyksellä, kuin kommentoimalla ja jakamalla. Lisäksi viihteellinen tai informatiivinen sisältö lisää sitoutumisen tasoa. Asiakkaat reagoivat positiivisesti palkkiota sisältävään sisältöön, mutta vain kommentoimalla. Eloisuus lisää sitoutumista ja vuorovaikutteisuus heikentää sitoutumista. Eloisuudella ja vuorovaikutteisuudella tutkimuksessa viitattiin mediatyypin asteisiin. Eloisuuden aste voi vaihdella tekstin ja videon välillä. Statuksella tai kuvalla ei ole vuorovaikutteisuutta, koska niitä voidaan vain lukea tai katsoa. Linkeillä ja videoilla on puolestaan korkea vuorovaikutteisuus, koska niitä on klikattava, jotta näkee koko sisällön. Arkipäivinä tehdyt julkaisut lisäävät kommenttien määrää ja puolestaan käyttäjien aktiivisimpana aikana tehdyt julkaisut vähentävät sitoutumista.

Schultz (2017) tutki myös kuluttajien sitoutumista brändijulkaisuissa ottamalla huomioon julkaisun ominaisuuksia: elävyys (tiedon eri aistien aste), vuorovaikutteisuus (aste, missä määrin julkaisu innostaa käyttäjää reagoimaan), sisällön tyyppi ja ajoitus (aika, jonka julkaisu on ylimpänä), pituus, seuraajien määrä ja toimialaan osallistumisen aste. Elävyys ilmenee eri mediatyypeistä, kuten teksti, kuva ja videot. Kuva on asteeltaan alhainen ja video korkea. Vuorovaikutteisuutta mitattiin brändijulkaisun tykkäyksien, kommenttien ja jakojen määrällä. Tutkimuksen tarkoituksena oli ymmärtää, miten brändijulkaisun piirteet lisäävät sitoutumista sosiaalisiin verkostoihin. Toisin sanoen, mitkä ominaisuudet ja sisältö saa asiakkaan tykkäämään, kommentoimaan ja jakamaan julkaisun.

Tutkimuksen data oli koottu kahdelta vaatteita ja elintarvikkeita myyvän brändin Facebookin julkaisuista. Tutkimuksessa reaktion astetta kuvaavan mallin parametrit estimoitiin pienimmän neliösumman menetelmän (PNS) avulla. Reaktion aste määrittää, kuinka eri julkaisun ominaisuudet vaikuttavat julkaisujen reaktioihin. Riippuvia muuttujia olivat vuorovaikutuksen ominaisuudet: tykkäyksien, kommenttien ja jakojen määrät. Schultzin (2017) tuloksien mukaan vuorovaikutteisuuden ominaisuuksilla on suurimmaksi osaksi positiivisia vaikutuksia sosiaaliseen vuorovaikutukseen. Elävyyden ominaisuuksilla on ristiriitaisia vaikutuksia, kuten kuvat ja videot vaikuttavat positiivisesti tykkäyksiin ja jakoihin, kun taas kommenttei-

hin ne vaikuttavat negatiivisesti.

He, Zha ja Li (2013) tutkivat asiakkaiden sitoutumista Facebook- ja Twitter-sivustoilla tykkäyksien, kommenttien ja jakojen määrillä. Tutkimuskohteena oli kolmen USA:n suurimman pitsaketjun Facebook- ja Twitter-sivustot. Tutkimusta tehtiin kilpailija-analyysin muodossa. Lisäksi he hyödynsivät tekstinlouhintaa analysoidakseen sivustojen sisältöä. Tutkimuksessa huomattiin, että asiakkaiden sitoutuminen on vahvempaa Facebookissa.

Mostafa (2013) tutki tekstilouhintaa käyttäen asiakkaiden mielipiteitä ja tunteita viiden brändin (Nokia, T-Mobile, IBM, KLM ja DHL) twiiteistä. Tutkimus hyödynsi sentimenttianalyysia, jossa laskettiin myös sentimentaalipisteet. Sanojen kategorisoimiseksi sentimenttianalyysia varten tutkimuksessa käytettiin Lexiconia, jonka sanaston avulla twiitit luokiteltiin positiiviseksi tai negatiiviseksi. Lisäksi sanat koodattiin kvantitatiiviseksi, jolloin saatiin laskettua eri sanojen lukumäärät ja analysoitua asiakkaiden tunteita brändiä kohtaan. Sentimentaalipisteiden laskennassa huomioitiin vain positiiviset (+1) ja negatiiviset (-1) sanat. Twiittien trendit visualisointiin tarkastelujakson aikana kaikille brändeille. Mostafan (2013) mukaan sentimenttianalyysejä voidaan käyttää esimerkiksi asiakkaiden palautteiden käsittelyyn.

Moro ym. (2016) keskittyivät mallintamaan kosmetiikkayrityksen suorituskykymittareita Facebooksivujen julkisista julkaisuista tiedonlouhinnan avulla. Tutkimuksen tavoitteena oli esimerkiksi ennustaa julkaisujen vaikutusta niiden ominaisuuksien kautta, saada selville, mitkä syötteet (engl. input) vaikuttavat mittareihin ja miten nämä syötteet vaikuttavat jokaiseen julkaisuun. Tutkimuksessa datana käytettiin maailmanlaajuisesti tunnetun kosmetiikkabrändin Facebookin julkisia julkaisuja, joita oli 790.

Tutkimuksen data oli jaoteltu neljään muuttujaryhmään tietotyypin perusteella: tunnistaminen, sisältö, kategoria ja toiminta (ks. Moro ym. 2016, s.3342-3344), joita voi käyttää mittaamaan julkaisun suorituskykyä. Näistä muuttujaryhmistä seitsemän (mm. aikaan liittyvät) toimivat syötteinä ja loput 12 toimintamuuttujaa olivat tulosteita (engl. output), jotka mallinnettiin. Toimintamuuttujat sisälsivät mittarit, joilla mitattiin julkaisun vaikutusta. Esimerkiksi *kaikki näyttökerrat* -muuttuja kertoi, kuinka monta kertaa julkaisu oli ladattu käyttäjän selaimeen. *Sitoutuneet käyttäjät* -muuttuja huomioi kaikkien julkaisujen klikkauksien tyypit

ja alkuperän. Muuttuja kertoi myös käyttäjien määrän, jotka klikkasivat julkaisua (yksittäiset käyttäjät). Lisäksi oli muuttuja, joka huomioi henkilöt, jotka tykkäsivät sivustosta sekä klikkasivat julkaisua. Tutkimuksen käsitekartan mukaan sitoutumisia voidaan myös mitata muilla vuorovaikutuksilla, kuten kommenttien tykkäyksillä ja kommentoijan nimen klikkauksella (ks. konseptuaalinen kartta Moro ym. 2016, s.3343).

Tutkimuksessa Moro ym. (2016) käyttivät tukivektorikonetta. Normaalijakautumisen arvioimiseksi he käyttivät Shapiro-Wilk-testiä. Jokaiselle 12 toimintamuuttujalle tuotettiin malli ja mallin ennustamaa arvoa verrattiin kyseisen mittarin todelliseen arvoon. Malliltaan parhaimman suorituskyvyn antoi muuttuja, joka kuvasi henkilöt, jotka tykkäsivät sivustosta sekä klikkasivat jossakin kohtaa julkaisua. Parhaimman suorituskyvyn antoi myös *julkaisua klikkanneet käyttäjät* -muuttuja (engl. lifetime post consumer), joka kuvasi käyttäjien määrää, jotka klikkasivat minne tahansa julkaisua. Lisäksi tutkimuksessa käytettiin datapohjaista herkkyysanalyysia saamaan tietoa julkaisuista klikanneista käyttäjistä ja miten muuttujaryhmän seitsemän syötettä vaikuttaa niihin. Julkaisun sisällön tyyppi oli mallin kannalta kaikkein tärkein syötemuuttuja.

3.2.3 Asiakasvalituksiin ja verkkosivuston dataan liittyvät tutkimukset

Bhatia ym. (2014) kuvailivat kehittämäänsä järjestelmää, jolla voi monitoroida ja analysoida asiakkaiden sosiaalisen median palautetta. Järjestelmän avulla asiakas voidaan sitouttaa ja ratkaista asiakkaiden kokemia ongelmia. He käyttivät crawler hakurobottia keräämään sosiaalisesta mediasta julkaisut, jotka ovat relevantteja ja sen jälkeen käyttivät koneoppimiseen pohjautuvaa sentimentaalista luokittelijaa jakamaan julkaisut negatiivisiin, positiivisiin ja neutraaleihin. Lisäksi he olivat kehitelleet reaaliaikaisen tai tietyn hetken *event*-havaitsijan tilastollisten (mm. termien frekvenssejä historiallisesta datasta) ja poikkeavien tapahtumien havaitsemisen metodeja hyödyntäen. Poikkeavia tapahtumia ovat tiettynä aikana äkillisesti alkaneet tapahtumat, joita voi olla havaittu aikaisemminkin.

Tutkimuksessa hyödynnettiin Term Frequency—Inverse Document Frequency:ä (TF-IDF) tärkeiden sanojen määrittämiseen, josta Bhatia ym. (2014) muokkasivat ajan normalisoivan version eli viimeaikaisia havaintoja painottavan version. Lisäksi järjestelmää testattiin

kolmeen eri kaupalliseen brändiin. Järjestelmän avulla jokaisesta brändistä saatiin kerättyä poikkeavia ja uusia tapahtumia, jotka kuvasivat brändin Twitterissä olevia trendejä. Lukemalla tarkemmin näitä tapahtumia ja niihin liittyviä tweettejä brändit voivat nopeasti saada selville, mistä valitukset tulevat.

Ordenes ym. (2014) analysoivat asiakkaan kokemuksia asiakaspalautteista lingvistispohjaisella tekstinlouhinnalla. Heidän mielestään asiakaskokemuksen monimutkaisen luonteen vuoksi asiakkaan kokeman palvelukokonaisuuden mittaaminen on haastavaa. Lisäksi haasteena on se, että nykyään teknologia mahdollistaa erilaisten asiakaspalautteiden keräämisen, mutta data on yleensä strukturoimatonta tekstidataa (mm. puhelinkeskustelut, sähköpostit, sosiaalisen median sisällöt), jota yrityksien on vaikeaa analysoida. Tutkimuksessaan Ordenes ym. (2014) loivat tähän parannetun viitekehyksen, joka sisältää tärkeitä elementtejä asiakaskokemuksesta, palvelumenetelmistä ja teorioista, kuten yhteistyöprosessista, vuorovaikutuksesta ja kontekstista. Näillä elementeillä asiakaspalautteista saa muodostettua syvällisemmän analyysin, joka kattaa kolme arvonmuodostuselementtiä: toimintaa kuvaavat verbit, tekijät tai objektit (sekä asiakkaiden että yrityksien) ja kontekstit (henkilökohtaisen ja tilanteellisen).

Tutkimuksen datana käytettiin Iso-Britanian lentokentällä parkkipaikkoja ja kuljetuspalveluja tarjoavan yrityksen asiakaspalautteita. Yritys lähettää online-kyselyn asiakkaille kahden päivän jälkeen palveluiden käytöstä. Kysely sisältää sekä strukturoitua että strukturoimatonta dataa. Strukturoimaton data koostuu avoimesta kysymyksestä: “Mikä on tärkein asia, jossa voisimme parantaa parkki- ja kuljetuspalvelujamme.” (Ordenes ym. 2014).

Tutkimuksen tekstinlouhinnan prosessin vaiheita olivat:

1. liiketoiminnan ja datan ymmärtäminen
2. mallin opettaminen tai kehittäminen
3. korpuksen (yritykseltä kerätty joukko asiakaspalautteita sisältäviä asiakirjoja) tuominen. Sisältää myös uusien käsitteiden, hahmojen ja tekstilouhinnan mallien määrittämisen ja arvioinnin.
4. testaus tai mallin arviointi.

Tekstilouhinta sisälsi kaksi iteraatiota: tekstinlouhintamallin viitekehyksen toteuttaminen ja

mallin parantaminen. Ensimmäisessä iteraatiossa Ordenes ym. (2014) suorittivat prosessin kaikki vaiheet satunnaisesti valitulle osadatalle, jonka jälkeen he esittivät ala- ja pääkategoriat asiakaspalvelujohtajalle, joka arvioi mallin pätevyyttä. Osadata oli luokiteltu manuaalisesti etukäteen. Iteraation testausvaiheessa käytettiin koko dataa tuottamaan ennusteita, johon sisältyi asiakkaiden kommenttien automaattista luokittelua kohteliaisuuksiin ja valituksiin sekä palautteessa olevien tekijöiden, objektien ja toimintaa kuvaavien verbien tunnistaminen. Ordenes ym. (2014) arvioivat mallia tiedonkeräyksellä (engl. data capture) ja luokitteluntarkkuudella. Tiedonkeräys kuvaa niiden asiakaspalautteiden lukumäärää, joista malli löytää yhden tai useamman hahmon. Luokittelutarkkuus kuvaa kerätyistä hahmoista mallin mukaisten oikeiden ennusteiden lukumäärää.

Toisessa iteraatiossa Ordenes ym. (2014) käyttivät deduktiivista lähestymistapaa yrityksen aikaisempaan manuaaliseen luokitteluun ja kehittelivät lingvistiset mallit käytettäväksi tekstinlouhintamallin avulla. Iteraation päätteeksi he kehittelivät lopullisen joukkon luokkia uusien lingvistisien mallien luomiseksi, joilla voidaan analysoida asiakaspalautteet. Toisen iteraation jälkeen käsitteet luokiteltiin ja kartoitettiin ehdotetun kehyksen pääluokkiin, mukaan lukien synonyymit. Tutkimuksessa kehitetyn mallin tarkkuus oli korkeatasoista asiakkaiden valituksien ja ehdotuksien sekä resurssien ja toimintojen ennustettavuudessa. Mallia voidaan myös muokata eri liiketoimintoihin ja siihen voidaan lisätä uutta asiakassanastoa, kun palvelun resurssit (tekijät ja objektit), toiminnot ja asiakkaan konteksti muuttuvat.

Dasgupta, Dey ja Verma (2016) analysoivat ja luokittelivat intialaisen televiestintäyrityksen asiakkaiden tekstimuotoisia valituksia liiketoiminta-alueisiin. Yhden asiakkaan valitus voi kuulua moneen liiketoiminta-alueeseen, joten tutkimuksessa sovellettiin sumeaa monitunnuksista luokittelua (engl. fuzzy multi-label classification). Liiketoiminta-alueen osaaja oli luokitellut valitukset manuaalisesti, joka osoitti valituksen joko kuuluvan kyseiseen liiketoiminta-alueeseen tai ei kuuluvan.

Tutkimuksessa käytettiin sumeaa k:n lähimmän naapurin (engl. Fuzzy K Nearest Neighbor, FKNN) algoritmia, joka on perinteisen k:n lähimmän naapurin algoritmin sumea versio. FKNN ei suoraan luokittele testidataa tiettyihin luokkiin, vaan määrittää luokkaan kuuluvuuden arvot. Valituksen kuuluvuus tiettyyn liiketoiminta-alueeseen annettiin vektorina, jossa jokainen liiketoiminta-alue sai arvon 0.1:stä 0.9:sään sen mukaan, että valituksella ei

ollut yhteyttä tai oli korkea yhteys liiketoiminta-alueeseen. Tutkimuksessa datan monitunnuksen luonne määriteltiin kahdella mittarilla: tunnuksen tiheydellä (engl. label density) ja tunnuksen mahtavuudella (engl. label cardinality) (ks. Dasgupta, Dey ja Verma 2016, s.380).

Datan suuren kohinaisuuden vuoksi Dasgupta, Dey ja Verma (2016) eivät pystyneet käyttämään tavanomaisia NLP-työkaluja (natural language prosessing) valituksien merkityksellisien ominaisuuksien tunnistamiseen ja poistamiseen, kuten luonnollisen kielen jäsennintä. Tutkimuksessa puolestaan käytettiin seuraavanlaisia tilastollisia ominaisuuksia ja tekniikoita: unigram, bigram, TF-IDF, latent sematic analysis (LSA) ja point-wise mutual information (PMI) suorittamaan monitunnusluokittelua (ks. Dasgupta, Dey ja Verma 2016, s.381). Liiketoiminta-alueeseen kuuluvien ennustettujen arvojen arviointiin käytettiin muun muassa ennustetun ja todellisen arvon välimatkan mittaamista. PMI ja unigram muuttujien yhdistelmä palautti parhaan mahdollisen tuloksen.

Riaz ym. (2017) tutkivat sentimenttianalyysillä tuotteiden asiakasarvioista, mitkä ovat positiivisia, negatiivisia ja neutraaleja arvioita. Tutkimuksen menetelmää voidaan käyttää analysoimaan tuotteista asiakkaiden mieltymyksiä, tarpeita ja käyttäytymisiä. Tutkimus eteni niin, että ensin kerättiin tuotteiden arvosteluja eri verkkokaupan sivustoilta (mm. Amazon, eBay, Alibaba). Arvosteluille tehtiin muunnos, jotta ne saatiin samaan muotoon. Tämän jälkeen tutkimuksessa hyödynnettiin sentimenttianalyysia sanakirjapohjaisella (Lexicon) lähestymistavalla. Sanakirjapohjaisella lähestymistavalla määritetään asiakirjan sisältö käyttämällä asiakirjan sanojen ja lauseiden semanttista suuntaa. Ensimmäisenä analyysissa sanat merkittiin positiiviseksi ja negatiiviseksi hyödyntämällä MPQA-korpusta ja sanat myös pakattiin. Tämän jälkeen vähennettiin datan kohinaa useilla esikäsittelytekniikoilla. Sanat, joita ei merkattu negatiiviseksi tai positiiviseksi, tulkittiin neutraaleiksi.

Vaiheessa, jossa poimittiin avainsanoja datasta, Riaz ym. (2017) käyttivät avainsanojen poimimisen tekniikkaa. Avainsanojen avulla laskettiin absoluuttinen TF, jossa hyödynnettiin TF-IDF:ää. Viimeisessä vaiheessa laskettiin sentimentin polaarisuuden voimakkuus mittaamalla sentimentaalinen vahvuus hyödyntäen TF:ää. Lisäksi sovellettiin k:n keskipisteen klusterointimenetelmän algoritmia klusteroimaan data sentimentaalisen vahvuuden perusteella. Näin tutkimuksella saatiin datasta yhteenveto ja parempi analyysi.

Tutkielmassa käsitellyt asiakkuuden kehittämisen tutkimukset on koottu taulukkoon 3, josta näkee tutkimuksissa hyödynnetyt menetelmät ja datat. Tutkimukset keskittyivät tutkimaan suusanallista viestintää, sosiaalisen verkoston kasvua, brändiyhteisöjä, yhteistyötoimintaa ja asiakasvalituksia. Tutkittiin esimerkiksi negatiivisen suusanallisen viestinnän vaikutusta yrityksen voittoon ja suusanallisen viestinnän vaikutusta sosiaalisen median verkostoon sekä asiakkaan sitoutumista verkostoon. Brändiyhteisöjen tutkimuksissa asiakkaan sitoutumista tutkittiin julkaisujen tykkäyksien, kommenttien ja jakojen määrällä sekä jossain määrin vuorovaikutuksen kestolla. Suurin osa tutkimuksista keskittyivät tutkimaan sitoutumista yksittäisten julkaisujen näkökulmasta eikä yksittäisten asiakkaiden näkökulmasta. Lisäksi asiakasvalituksiin liittyvissä tutkimuksissa analysoitiin ja luokiteltiin asiakkaiden palautteita sekä valituksia. Tässä tutkielmassa asiakkuuden kehittämisen näkökulmana on brändiyhteisön yksittäiset asiakkaat. Heidän sitoutumista selvitetään julkaisujen tykkäyksien ja kommenttien määrällä sekä vuorovaikutuksen kestolla.

Taulukko 3: Yhteenveto asiakkuuden kehittämisen tutkimuksista

Tutkimus	Luokittelukriteeri	Metodi	Data
Goldenberg ym. (2007)	Suusanallinen viestintä	Agent-pohjainen malli	Kyselylomake
Trusov, Bucklin ja Pauwels (2009)	Suusanallinen viestintä ja sosiaalisen verkoston kasvu	VAR-malli	Sosiaalinen verkosto
Zadeh ja Sharda (2014)	Sosiaalisen verkoston kasvu	Multiuloitteinen pisteprosessi	Twitter-data
Okazaki ym. (2015)	Suusanallinen viestintä ja sosiaalisen verkoston kasvu	Eri koneoppimisen menetelmiä, naiivi Bayes	Twitter-data
Cvijikj ja Michahelles (2013)	Brändiyhteisö	Negatiivinen binomiaalinen arviometodi	Facebook-data
He, Zha ja Li (2013)	Brändiyhteisö	Tekstinlouhinta	Facebook- ja Twitter-data
Mostafa (2013)	Brändiyhteisö	Tekstinlouhinta ja sentimentaalianalyysi	Twitter-data
Moro ym. (2016)	Brändiyhteisö	Tukivektorikone	Facebook-data
Schultz (2017)	Brändiyhteisö	PNS-menetelmä	Facebook-data
Bhatia ym. (2014)	Asiakasvalitukset	Sentimentaalianalyysi	Sosiaalisen median asiakaspalautteet
Ordenes ym. (2014)	Asiakasvalitukset	Lingvistispohjainen tekstinlouhinta	Kyselylomake
Dasgupta, Dey ja Verma (2016)	Asiakasvalitukset	Suomea k:n lähin naapuri	Tekstimuotoiset asiakasvalitukset
Riaz ym. (2017)	Verkkosivuston data	Sentimentaalianalyysi ja klusterointi	Verkkokaupan tuotearvostelut

3.3 Asiakkuuden säilyttämisen tutkimukset

Perinteisesti asiakkuuden säilyttämisessä on haluttu mallintaa todennäköisyyttä sille, että asiakas pysyy asiakkaana. Bijmoltin ym. (2010) mukaan malleina on käytetty todennäköisyysmalleja, kuten negatiivista binomijakaumaa (NBD). Asiakkaan poistumisen näkökulmasta on puolestaan haluttu ennustaa, onko asiakas todennäköisesti poistumassa tai ei ole poistumassa määrätyllä ajanjaksolla. Tätä on pidetty perinteisesti binäärisenä ongelmana ja

hyödynnetty binäärisiä malleja, kuten logistista regressioanalyysia, päätöspuita ja erotteluanalyysia (engl. discriminant analysis, DA). Koneoppimisen menetelmiä, kuten neuroverkkoja ja tukivektorikonetta, on myös käytetty parantamaan ennustamista perinteisten mallien sijasta. Asiakassuhteen kestoa on myös mallinnettu todennäköisyyden näkökulmasta, jolloin on arvioitu asiakkaiden siirtymien todennäköisyyttä, esimerkiksi Markov-mallilla. Perinteiset lähestymistavat asiakkaiden säilyttämisessä ovat Bijmoltin ym. (2010) mukaan keskittyneet ennustamaan ketkä asiakkaat ovat todennäköisesti poistumassa ja sen jälkeen on yritetty saada heidät jäämään. Lähestymistavoissa ei ole kuitenkaan huomioitu asiakkaan sitoutumista yritykseen.

Bijmoltin ym. (2010) mukaan sitoutumisen näkökulma tulee huomioitua, jos asiakkaan poistumisen ennustemalleissa otetaan huomioon asiakkaiden vuorovaikutukset toisiinsa, sillä toisen asiakkaan poistuminen voi johtaa myös toisen asiakkaan poistumiseen. Sosiaalisen verkoston tiedot voitaisiin huomioida mallinnettaessa asiakkaiden välisiä vaikutuksia, esimerkiksi televiestintäyrityksissä asiakkaiden väliset soitot. Lisäksi asiakkaat todennäköisesti sitoutuvat eri tasolla yritykseen, jolloin olisi hyödyllistä ottaa huomioon asiakaskunnan heterogeenisuus asiakkaiden sitoutumisessa. Tulevaisuudessa asiakkaan poistumisen huomioimiseen olisi myös mahdollista liittää asiakkaan sitoutuminen ja asiakkaan arvo yritykselle.

Lisäksi Bijmoltin ym. (2010) mukaan olisi huomioitava se, että poistumisen aiheuttavat tekijät voivat muuttua dynaamisesti asiakkaan elinkaaren aikana. Tämä voitaisiin huomioida esimerkiksi käyttämällä aikamuuttujakerrointa tai dynaamista lineaarista mallia. Käytettäessä Markovin mallia, todennäköisyys asiakkaan poistumiseen voi riippua suhteesta tai sitoutumisen asteesta, jonka asiakas kokee tietyssä tilanteessa. Siirtyminen tilojen välillä voi riippua ajan mukaan vaihtelevilla kovariansseilla (Bijmolt ym. 2010; Netzer, Lattin ja Srinivasan 2008). Tutkimusdatana Netzer, Lattin ja Srinivasan (2008) käyttivät tyypillisiä ostotietoja.

3.3.1 Sosiaaliseen verkostoon liittyvät tutkimukset

Backiel, Baesens ja Claeskens (2014) huomioivat perinteisissä asiakkaan säilyttämisen ennustemalleissa sosiaalisen verkoston, joilla he ennustivat asiakkaan poistumista sosiaalisen verkoston kautta. Tutkimuksen mielenkiinto ei ollut mallissa vaan muuttujissa, jotka Backie-

lin, Baesensin ja Claeskensin (2014) mukaan on otettava huomioon mallin sijasta. Esimerkiksi sosiaalisen verkoston puhelutiedoista voidaan erottaa verkoston muuttujia, joita voidaan puolestaan käyttää useissa perinteisissä ennustemalleissa. Tutkimuksessa erotettiin ja hyödynnettiin kahta poistumisen ennustamismallia: perinteinen malli ja verkkomalli. Perinteisissä malleissa, joissa käytetään yksittäisen asiakkaan tietoja (sopimustyyppiä, maksuja jne.) tai käytöstä, ennustavat tietyn asiakkaan lähtemisaikomusta. Verkostomallit ennustaessaan poistuvia asiakkaita ottavat puolestaan huomioon verkoston asiakkaat sekä niiden suhteet ja odottavat suhteessa olevien asiakkaiden käyttäytyvän samalla tavalla. Tutkimuksessa käytettiin Belgialaisen Telco puhelinoperaattorin dataa. Asiakkaiden väliset soitot mallinnettiin sosiaalisena verkostona, jossa jokainen solmu edusti asiakasta ja yksi ylimääräinen solmu edusti niitä, jotka eivät olleet asiakkaita. Puheluiden kokonaiskestolla painotettu reuna (suunnaton viiva) kuvasi asiakkaiden välisiä soittoja. Verkosta poimittiin kuusi mittarilinkkiä muuttujiksi, kuten naapureiden määrä ja puhelujen kesto naapureiden kanssa. Asiakas todettiin poistuneen, jos hän ei soittanut tai vastaanottanut puheluja 30 päivän aikana (Backiel, Baesens ja Claeskens 2014).

Sosiaalisesta verkostosta poimittuja muuttujia, asiakkaan omista tiedoista saatuja muuttujia ja näiden yhdistelmiä Backiel, Baesens ja Claeskens (2014) testasivat logistisella regressiolla ja Cox-regressiolla (Cox PH). Malleja he opettivat 70 %:lla datasta ja loppuosaa käyttivät arviointiin. Arviointia he tekivät muun muassa toimintaominaisuuskäyrällä (Receiver Operating Characteristic Curve, ROC) ja pitoisuus-aikakäyrän pinta-alan (Area Under The (ROC) Curve, AUC) arvoilla. Verkostosta poimitut muuttujat parantavat ennustamista verrattuna yksilön muuttujiin, jolloin verkoston tiedot voivat olla dynaamisempia ja heijastavan enemmän nykyhetken muutoksia asiakkaan käytöksessä. Lisäksi tutkimuksen käytännönläheisempi tulos oli Java:lla muodostettu toteutus, joka prosessoi yrityksen yhden kuukauden datan ja tekee ennustukset seuraaville viikoille tai kuukausille. Tällä toteutuksella yritykset voivat identifioida asiakkaat, jotka ovat vaarassa poistua.

Guelman, Guillen ja Perez-Marin (2012) eivät keskittyneet tutkimuksessaan ennustamaan, kuka asiakas on poistumassa, vaan tutkivat satunnaisen metsän ja uplift-mallin avulla, ketkä vakuutusyhtiön asiakkaista reagoivat positiivisesti yrityksen asiakkuuden säilyttämisen strategiaan. Asiakkaan konkreettisen menetyksen lisäksi asiakkaan poistuma johtaa asiakkaan

tuottaman voiton häviämiseen. Vakuutusyhtiöt voivat ennustaa asiakkaan elinkaaren arvon vakuutustuotteille ja asiakkaan sitoutumista voidaan puolestaan kasvattaa myymällä enemmän lisävakuutuksia.

Uplift-mallia Guelman, Guillen ja Perez-Marin (2012) käyttivät tutkimuksessa arvioimaan jokaisen asiakkaan poistumisen pienentymistä, kun asiakkuuden säilyttämisen strategiaa sovelletaan asiakkaaseen. Näin säilyttämisen strategiaa voidaan soveltaa niihin asiakkaisiin, joihin strategia vaikuttaa positiivisesti. Uplift-mallin algoritmi pohjautui satunnaiseen metsään (ks. Guelman, Guillen ja Perez-Marin 2012, s.127-128). Tutkimuksessa algoritmia testattiin kanadalaisen vakuutusyhtiön asiakkaisiin.

Kazienko, Brodka ja Ruta (2009) tutkivat lähipiiri vaikutusta asiakkaan säilyttämiseen asiakashankinnan (ks. luku 3.1.1) lisäksi. Heidän mukaan lähtevällä asiakkaalla on vaikutusta ainakin lähipiiriinsä, joihin hän on suoraan yhteydessä. Vaikutus voi vaihdella eri ääripäiden välillä häipymisestä, uudelleen suunnatusta tai elvytetystä toiminnasta poistumisen seurantaan. Tutkimuksen mukaan asiakkaat, jotka ovat poistumassa, vähentävät toimintaansa lähipiirissä ennen lähtöään. Toiminnan vähentämisen kautta voidaan ennustaa lähtevät asiakkaat, sillä tutkimuksessa poistuvan asiakkaan lähipiirissä havaittiin sosiaalisen aseman mittarin merkittävää pienentymistä. Sosiaalisen aseman mittari voidaan arvioida puheluiden keston ja määrien osalta iteratiivisesti (ks. Kazienko, Brodka ja Ruta 2009, s.492). Sosiaalisen aseman pienemmät arvot osoittavat poistuvat asiakkaat ennen kuin he poistuvat.

3.3.2 Kanta-asiakasohjelmaan liittyvät tutkimukset

Bahari ja Elayidom (2015) kehittivät tehokkaan tiedonlouhintaviitekehyksen asiakkuudenhallintaan ja tutkivat kahdella luokittelumallilla asiakkaan käytöksen ennustettavuutta asiakkuudenhallintasovelluksessa parantaakseen päätöksentekoa arvokkaiden asiakkaiden säilyttämiseksi. Luokittelumalleina käytettiin monikerroksista perseptronineuroverkkoa (engl. Multilayer Perception Neural Network, MLPNN) ja naiivi Bayes:ia (NB). Lisäksi työkaluna oli Weka (The Waikato Environment for Knowledge Analysis). Tutkimuksen datana oli UC Irvine Machine Learning Repository -sivuston hyvin tunnetun portugalilaisen pankin markkinointidata. Data sisälsi 16 muuttujaa, joista kahdeksan liittyivät asiakkaaseen, neljä markki-

nointikampanjan viimeiseen yhteydenottoon ja loput neljä kampanjaan. Tulosmuuttuja vastaa kampanjan tulosta, joka on binäärimuotoinen ja kuvaa, onko asiakas sitoutunut vai sitoutumaton talletusjärjestelmään.

Tutkimuksessa Bahari ja Elayidom (2015) jakoivat alkuperäisen datajoukon luomalla satunnaisesti kymmenen erillistä alijoukkoa. Opetusjoukko oli näiden yhdeksän alijoukon yhdistelmä. Jäljelle jääneestä alijoukosta muodostettiin testausjoukko luokittelumallien suorituskerroille. Tuloksena saatiin kaksi luokkaa: asiakkaiden positiivinen vastaus ja negatiivinen vastaus. Mallien suorituksen arviointiin käytettiin luokituksen tarkkuutta, herkkyyttä ja spesifisyyttä. Tarkkuus kuvaa, millä prosenttimäärällä tapahtumat on luokiteltu oikein. Herkkyys ilmaisee, millä prosenttimäärällä tapaukset on luokiteltu oikein positiiviseksi vastauksiksi ja spesifisyys puolestaan oikein negatiiviseksi vastauksiksi luokiteltujen prosenttimäärää. (ks. Bahari ja Elayidom 2015, s.730). Korkean ennustettavuuden suorituksen antoi MLPNN-malli 88,63% tarkkuudella.

Parantaakseen olemassa olevan asiakkaan poistumismallia Vo ym. (2018) käyttivät koneoppimista analysoidakseen asiakkaan luonteenpiirteitä ja käytöstä strukturoimattomasta datasta (äänipuheluista). Heidän mukaan koneoppimista on käytetty aikaisemminkin ennustamisessa, mutta datana on ollut vain strukturaalinen data, kuten demograafiset tiedot ja ostohistoria. Lisäksi heidän mielestään tekstimuotoisessa datassa on tunnepisteitä (positiivisen ja negatiivisen sanan ilmaiseva numero) enemmän tietoa, jolloin mielipiteenlouhinta ja sentimentaalianalyysi eivät pelkästään riitä. Vonin ym. (2018) malli hyödynsi asiakkaiden puhelulokien käännösskriptien tekstitietoja ja strukturaalista dataa (mm. asiakkaiden demografisia tietoja) ennustaakseen, onko asiakas poistumassa vai ei ole poistumassa.

Tutkimuksessa Vo ym. (2018) sovelsivat eläkerahastoyrityksen dataan kolmea tekstinlouhintamenetelmää, joiden avulla saatiin kolme erilaista tekstin muuttujien sarjaa: semaattiset tiedot (engl. sematic information), sanan merkitys (engl. word importance) ja sanan upotus (engl. word emdedding). Semaattisten tietojen poimintaan käytettiin sentimenttianalyysia. Sanan merkityksen selvittämiseen hyödynnettiin TF-IDF:ää. TF:ää käytettiin mittaamaan puheluiden termien esiintymiskertoja ja IDF:ää selvittämään, onko termi yleinen vai harvoin esiintyvä. TF-IDF pienentää myös muuttujien dimensiota poistamalla yleisiä ja ei tärkeitä sanoja. Samanlaisten termien erilaisilla yhdistelmillä on erilaisia merkityksiä, joten tutki-

muksessa käytettiin WORD2Vec-mallia sanan upotuksessa sekä selvittämään termien asemat, suhteet ja paikat lauseissa. Lisäksi tutkimuksessa käytettiin koneoppimisen algoritmina Python paketin Extreme Gradient Boost (XGBoost) -luokittelijaa. Mallin arvioimiseen käytettiin AUC-arvoja ja ROC-käyrää. Näiden lisäksi mallin ennustamia arvoja asiakkaan poistumisesta arvioitiin yrityksen antamiin todellisiin arvoihin. Tuloksien mukaan strukturoimattoman ja strukturoidun datan yhdistäminen parantaa poistumisen ennustamisen tarkkuutta.

Eläkerahaston dataa hyödynsivät myös Culbert ym. (2018). He ehdottavat tutkimuksessaan uutta lähetystapaa asiakkaan poistumisen ennustamiseen, varsinkin näkymättömien asiakaspoistumisien takia, mitä ilmenee Australian eläkerahastossa. Lisäksi he ehdottivat ratkaisua vähäiseen asiakkaan sitoutumisdataan. Asiakkaan näkymätöntä poistumista ilmenee, kun asiakastilistä tulee uinuva pakollisten työnantajamaksujen lopettamisen jälkeen. Eläkerahaston sitoutumisdatan vähäisyyteen on vaikuttanut se, että lainsäädäntö on mahdollistanut, että piittaamattomat asiakkaat voivat lykätä päätösprosessia eläkerahaston suhteen eläkeikäänsä asti.

Tutkimuksessa Culbert ym. (2018) vertasivat koneoppimisen päätöspuun, satunnaisen metsän ja extreme gradiant boostingin (XGBoost) malleja ja logistista regressiota. Jokainen koneoppimisen malli jakoi puupohjaisen rakenteensa, mutta jokaisella oli omanlainen hyvyysfunktio. Data oli jaettu kolmeen osaan tuoteryhmien mukaisesti. Opetusjoukko muodostui 80%:sta jokaisesta datan osasta. Opetusjoukon muuttujina tutkimuksessa käytettiin asiakkaan tietoja (demograafiset tiedot), välittäjätietoja (asiakas- ja tilitietojen vuorovaikutustiedot), tilitietoja (mm. vakuutukset, lainat), taloustietoja (mm. maksut, omaisuuden muutokset). Sitoutumismuuttuja, jota myös käytettiin, koostui asiakaskeskuksen kyselyistä ja verkkopankkitoiminnoista. Lisäksi muuttujana oli kontrastisen peräkkäismallin louhinnan merkkijonot (engl. contrast sequential pattern mining), jotka yhdistivät taloudelliset ja sitoutumistiedot. Mallien suorituskyvyn mittaamiseen tutkimuksessa käytettiin muun muassa pitoisuus-aikakäyrän pinta-alaa. Tutkimuksen mukaan XGBoost oli paras malli jokaisessa tuoteryhmässä. Lisäksi tulokset osoittivat ennustemallin suorituskyvyn parantumista, kun peräkkäismallin muuttujia yhdistetään demograafisiin ja tilitietojen muuttujiin.

Asiakkaan poistumisesta löytyy hyvin tutkimuksia, mutta suurimmassa osassa niissä tutkitaan asiakkaan säilyttämistä perinteisestä näkökulmasta. Tässä tutkielmassa käsitellyt tut-

kimukset huomioivat asiakkaan sitoutumisen. Tutkimukset on koottu taulukkoon 4. Sosiaaliseen verkostoon liittyvissä tutkimuksissa asiakkaan poistumista ennustettiin sosiaalisen verkoston kautta. Lähtevällä asiakkaalla on yleensä vaikutusta lähipiiriinsä, joten tätä kautta pystytään ennustamaa lähtevät asiakkaat. Poistuvan asiakkaan ennustamisen lisäksi tutkimuksissa ennustettiin, ketkä asiakkaista reagoivat postiivisesti yrityksen asiakkuuden säilyttämisen strategiaan. Kanta-asiakasohjelmaan liittyvissä tutkimuksissa tutkittiin luokittelumallilla asiakkaan käytöksen ennustettavuutta asiakkuudenhallintasovelluksessa. Lisäksi strukturoimattomasta datasta (äänipuheluista) analysoitiin asiakkaan luonteenpiirteitä ja käytöstä poistumismalleja varten. Tässä tutkielmassa ei ole mahdollista soveltaa verkkoteoriaa klusteroinnin lisäksi, joten klusteroitua tietoa analysoidaan asiakkaan säilyttämisen näkökulmasta.

Taulukko 4: Yhteenveto asiakkuuden säilyttämisen tutkimuksista

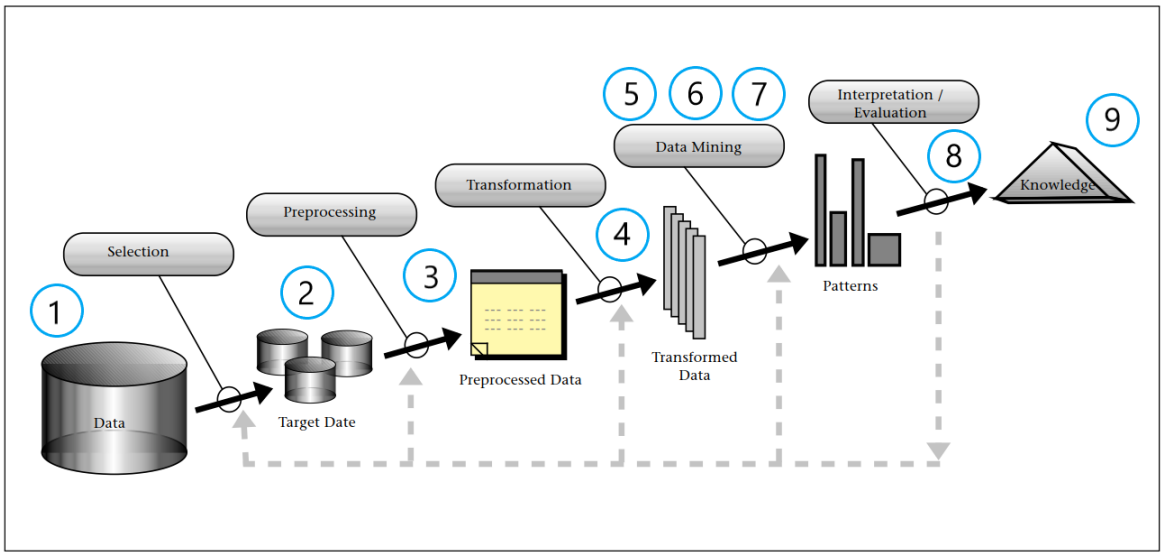
Tutkimus	Luokittelukriteeri	Metodi	Data
Backiel, Baesens ja Claeskens (2014)	Sosiaalinen verkosto	Logistinen regressio ja Cox PH	Puhelinoperaattorin data
Guelman, Guillen ja Perez-Marin (2012)	Sosiaalinen verkosto	Uplift-malli ja satunnainen metsä	Vakuutusyhtiön asiakkaat
Bahari ja Elayidom (2015)	Kanta-asiakasohjelman data	Monikerroksisen havainnoinnin neuroverkko ja Naiivi Bayes	Pankin markkinointidata (UCI)
Vo ym. (2018)	Kanta-asiakasohjelman data	Koneoppiminen; sentimentaalianalyysi ja XGBoost	Eläkerahaston asiakkaiden äänipuhelut
Culbert ym. (2018)	Kanta-asiakasohjelman data	Koneoppiminen; XGBoost	Eläkerahaston asiakasdata

4 KDD-prosessi

Vuosikymmeniä sitten digitaalisen datan nopea kasvaminen vaikeutti ja hidasti datan manuaalista käsittelyä, joten tarvittiin laskennallisia teorioita ja työkaluja. Vuonna 1989 pidettyssä KDD-työpajassa keksittiin ilmaus Knowledge Discovery in Databases (KDD) (Piatetsky-Shapiro 1990). Ilmaus korosti sitä, että tieto on datapohjaisen tutkinnan lopputuote (Fayyad, Piatetsky-Shapiro ja Smyth 1996a). Fayyad, Piatetsky-Shapiro ja Smyth (1996b, 1996a, 1996c) esittelivät KDD:n kokonaisvaltaisena prosessina, jolla etsitään hyödyllistä tietoa datasta. He määrittelivät KDD-prosessin seuraavasti: “Monimutkainen prosessi tunnistamaan valideja, uusia, potentiaalisia, hyödyllisiä ja ennen kaikkea ymmärrettäviä hahmoja (engl. pattern) datasta” (ks. Fayyad, Piatetsky-Shapiro ja Smyth 1996a, s.40-41).

4.1 Prosessin vaiheet

KDD-prosessi on interaktiivinen ja iteratiivinen koostuen erilaisista vaiheista (ks. kuvio 4). Prosessin vaiheita ovat datan valinta, datan esikäsittely, datan muunnokset, tiedonlouhinta ja tulosten tulkinta sekä arviointi. Näiden vaiheiden kautta saavutetaan tieto datasta. Vaiheet sisältävät lisäksi yhdeksän askelta etenemällä sovellusalueen tuntemuksesta (askel 1) saavutetun tiedon käyttöön (askel 9). Fayyadin, Piatetsky-Shapiron ja Smythin (1996a) mukaan prosessin erityinen vaihe on tiedonlouhinta, jonka menetelmillä etsitään hahmoja (tai malleja) datasta. Hahmo on ilmaisu, joka kuvaa datan osajoukkoa tai osajoukkoon sovellettavaa mallia. Handin, Mannilan ja Smythin (2001) mukaan hahmo on lokaali eli koskee vain tiettyjä datapisteitä tai muuttujia. Malli (engl. model) on puolestaan kattava, abstrakti esitys datasta ja globaali eli koskee kaikkia datapisteitä.



Kuvio 4: KDD-prosessin vaiheet (Fayyad, Piatetsky-Shapiro ja Smyth 1996a)

Tutkielmassa sovelletaan KDD-prosessia, koska halutaan saada selville uutta ja hyödyllistä tietoa asiakkaan sitoutumisesta. Tilastollista tutkimusta ei voida soveltaa, koska siinä on oltava etukäteen tiedossa, mitä hypoteesien avulla etsitään, ja se myös keskittyy tilastolliseen merkitsevyyteen. Lisäksi tutkielmassa halutaan ymmärtää paremmin, millaiset asiakkaat sitoutuvat yritykseen ja miten he sitoutuvat. Tätä ymmärtämistä lisää juuri KDD-prosessilla saatavat hahmot. KDD-prosessi soveltuu myös parhaiten tutkielmassa käytettyihin datoihin, koska ne ovat reaaliaikaisia ja suuria, jolloin manuaalinen käsittely tai tilastollinen käsitelly ei ole tehokasta, vaan tarvitaan algoritmeja käsittelemään dataa.

4.2 Valinta

Ennen KDD-prosessin *valinta*-vaihetta ensimmäisellä askeleella luodaan ymmärrys sovellusalueesta ja asiaan liittyvästä aiemmasta tiedosta. Lisäksi tunnistetaan ja määritellään KDD-prosessin tavoite. Toisella askeleella edetään *valinta*-vaiheeseen. Vaiheessa luodaan datajoukko, josta tieto lopulta muodostetaan, joko valikoimalla datajoukot, tai keskittymällä muuttujien osajoukkoon tai otokseen datasta (Fayyad, Piatetsky-Shapiro ja Smyth 1996a). Datajoukko esitetään usein $n \times d$ matriisina (n rivejä ja d sarakkeita) (Zaki ja Meira Jr. 2014). Rivit kuvaavat datan havaintoja ja sarakkeet datan muuttujia. Muuttujat voidaan luokitella

luokittelu- ja järjestysasteikollisiin sekä välimatka- ja suhdeasteikollisiin. Matriisin jokainen rivi sisältää havaintojen muuttujien arvot. Datan ei välttämättä tarvitse olla matriisimuodossa vaan se voi olla monimutkaisempi, kuten lukujono, teksti, aikasarja, kuva, ääntä tai video (Zaki ja Meira Jr. 2014).

4.3 Esikäsittely

Datajoukon muodostamisen jälkeen seuraavassa vaiheessa data esikäsitellään ja puhdistetaan (askel 3 kuviossa 4). Yleensä datasta poistetaan mahdollinen kohina. Päätetään, mitä tehdään puuttuville arvoille, ja käsitellään datatyypit (Fayyad, Piatetsky-Shapiro ja Smyth 1996b, 1996a, 1996c). Kantardzicin (2011) mukaan puuttuvat havainnot voidaan esimerkiksi poistaa kokonaan, jos yksi arvo puuttuu tai arvo voidaan korvata esimerkiksi asiantuntijan arviolla, vakioarvolla tai muuttujan keskiarvolla. Datatyyppien käsittelyssä kategoriset muuttujat voidaan esimerkiksi muuttaa binäärisiksi numeromuuttujiksi.

4.4 Muunnos

Datan *muunnos* -vaihe sisältää neljännen askeleen eli datan tiivistämisen ja projektion. Fayyadin, Piatetsky-Shapiron ja Smythin (1996b, 1996a, 1996c) mukaan tarkoituksena on etsiä muuttujat, jotka kuvaavat dataa tavoitteiden mukaisesti, joko poistamalla muuttujia tai valitsemalla niitä. Tällöin datalle voidaan tehdä ulottuvuuksien (muuttujien) vähentäminen tai käyttää muunnosmenetelmiä. Monen tiedonlouhinnan asiantuntijan mukaan muunnosvaihe on prosessin kriittisin vaihe (Kantardzic 2011). Esimerkiksi, jos datan on oikealle vino, voidaan Handin, Mannilan ja Smythin (2001) mukaan ottaa logaritminen muunnos, jotta jakaumasta tulee symmetrisempi.

Kantardzicin (2011) mukaan datan ulottuvuuksien vähentäminen parantaa laskennallista tehokkuutta ja datan analysoinnin tarkkuutta. Muuttujien valinta vähentämiskeinona perustuu esimerkiksi ennakkotietoon ja siihen, mitä halutaan tiedonlouhinnalta saavuttaa. Summamuuttujia on myös mahdollista luoda. Lisäksi alkuperäinen muuttujien joukko voidaan muuntaa uudeksi pienemmäksi muuttujajoukoksi lineaarisilla ja epälineaarisilla menetelmillä. Tavallisimpia lähetystapoja ovat pääkomponenttianalyysi (Principal Component Analy-

sis, PCA) ja moniuloitteinen skaalaus (Multidimensional Scaling, MDS). Pääkomponenttianalyysissä muuttujat yhdistetään ja muutetaan uudeksi muuttujajoukoksi toivoen, että ne säilyttävät alkuperäisen tietosisältönsä pienennetyssä muodossa. Zakin ja Meira Jr:in (ks. 2014, s.187-191) mukaan pääkomponenttianalyysi etsii kannan pienemmässä ulottuvuudessa, jossa datalla on suurin varianssi. Suurimman projisoidun varianssin omaava suunta on ensimmäisessä pääkomponentissa, ja niin edelleen. Moniuloitteinen skaalaus esittää muuttujat pienemmässä ulottuvuudessa niin, että havaintopisteiden etäisyys uudessa ulottuvuudessa on sama kuin alkuperäisessä ulottuvuudessa (Kantardzic 2011).

Muunnosmenetelmänä voidaan käyttää muun muassa normalisointia, joka on tärkeä varsinkin, kun lasketaan havaintopisteiden etäisyyksiä, jottei muuttujat, joiden luvut ovat suuria, painoita etäisyyden mittausta liikaa (Kantardzic 2011). Normalisointina voidaan käyttää min-max-skaalausta, jossa muuttujan X arvot ($x_i = \{(\mathbf{x}_i, j = 1, ..., n\}$) skaalataan esimerkiksi välille $[0, 1]$ seuraavasti (Zaki ja Meira Jr. 2014):

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}} \tag{4.1}$$

Muuttujan X arvot voidaan normalisoida myös keskiarvon ja keskihajonnan mukaan seuraavasti (Zaki ja Meira Jr. 2014):

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}} \tag{4.2}$$

missä $\hat{\mu}$ on otoskeskiarvo ja $\hat{\sigma}$ on keskihajonta muuttujan X arvoista. Tällöin normalisoitujen muuttujien keskiarvo on nolla ja keskihajonta on yksi.

4.5 Tiedonlouhinta

Tiedonlouhinta on KDD-prosessin vaihe, jossa käytetään datan analysointi- ja etsintäalgoritmeja, jotka hyväksyttävien laskennallisten tehokkuusrajoitusten puitteissa tuottavat tietyn hahmon (tai malleja) datasta (Fayyad, Piatetsky-Shapiro ja Smyth 1996b, 1996a, 1996c). Handin, Mannilan ja Smythin (2001) mukaan tiedonlouhinta on “havaintoaineistojen (usein

suurien) analysointia löytääkseen odottamattomia suhteita, sekä datan tiivistämistä uusilla tavoilla, jotka ovat ymmärrettäviä ja hyödyllisiä datan omistajalle.”

4.5.1 Tehtävät ja mallit

Tiedonlouhinnan-vaiheessa KDD-prosessin viidentenä askeleena (ks. kuvio 4) on yhdistää prosessin tavoitteet, jotka ensimmäisellä askeleella määriteltiin, tiettyyn tiedonlouhinnan tehtävään (Fayyad, Piatetsky-Shapiro ja Smyth 1996b, 1996a, 1996c). Fayyadin, Piatetsky-Shapiron ja Smythin (2001) ja Äyrämön (2006) mukaan tiedonlouhinta voidaan täten luokitella tehtävittäin siten, että ne vastaavat analysoijan tavoitteita. Tehtävät voidaan jakaa kuvailevaan ja ennustavaan mallintamiseen. Kuvailevan mallin tavoitteena on kuvata korkeaulotteinen data kokonaisuudessaan ilman vahvoja oletuksia taustalla olevista luokista tai rakenteista muun muassa tiheysarviolla (engl. density estimation), klusterianalyysillä tai segmentoinnilla. Fayyadin, Piatetsky-Shapiron ja Smythin (1996c) mukaan kuvailevaa mallia tulkitaan todellisuuden heijastukseksi. Tässä tutkielmasssa hyödynnetään klusterointia, koska tarkoituksena on kuvata data kokonaisuudessaan ilman, että keskitytään mihinkään tiettyy muuttujaan. Lisäksi taustalla olevista luokista tai rakenteista ei ole tietoa. Klusterointi on esitelty tarkemmin luvussa 5.

Handin, Mannilan ja Smythin (2001) mukaan kuvailevassa mallissa ei ole mitään muuttujaa keskiössä, kun puolestaan ennustavassa on joku muuttuja, esimerkiksi markkina-arvo tai sairauden luokka. Ennustavassa mallintamisessa tavoitteena on rakentaa malli, jolla voi ennustaa tietyn muuttujan arvon muiden tunnettujen muuttujien arvoista. Ennustava malli keskittyy ennustettavuuden tarkkuuteen ja tehokkuuteen, jolloin mallin käyttäjä ei välitä siitä kuvastaako malli todellisuutta (Fayyad, Piatetsky-Shapiro ja Smyth 1996c). Fayyadin, Piatetsky-Shapiron ja Smythin (1996c) mukaan todellisuudessa KDD-sovellukset vaativat sekä kuvailevaa että ennustavaa mallintamista. Ennustettavia malleja ovat luokittelu (engl. classification) kategorisille muuttujille ja regressio kvantitaviisille muuttujille (Hand, Mannila ja Smyth 2001). Seuraavassa nämä on esitelty lyhyesti:

- **Luokittelu:** Bramerin (2013) mukaan luokittelun tehtävänä on ennustaa tunnus tai luokka annetulle luokittelemattomalle havaintopisteelle. Luokittelija on malli tai funk-

tio M, joka ennustaa luokan tunnuksen $\hat{y}$ annetusta syötteestä x, joka on $\hat{y} = M(x)$ ja jossa $\hat{y} \in \{c_1, c_2, ..., c_k\}$. Luokittelumallin rakentamiseen vaaditaan opetusjoukko eli havaintopisteet, joilla on oikeat luokkatunnukset. Mallin M jälkeen voidaan automaattisesti ennustaa luokka uudelle havaintopisteelle. Luokittelu malleja on erilaisia, kuten päätöspuut, naiivi Bayes ja tukivektorikone.

- **Regressio:** D. C. Montgomeryn ym. (2015) mukaan regressiota voidaan käyttää ennustamaan muun muassa yrityksen seuraavan kuukauden myyntiä. Tilastollisesta näkökulmasta mallinnetaan ennustettavan muuttujan ehdollista jakaumaa. Tutkitaan ja mallinnetaan muuttujien välisiä suhteita. Lineaarisessa regressiossa muuttujien välistä lineaarista riippuvuutta kuvataan suoralla viivalla, $y = \beta_0 + \beta_1 x$, jossa x on selittävä muuttuja ja y on selitettävä muuttuja.

Ennustavaa mallia kutsutaan myös ohjatuksi oppimiseksi, koska siinä hyödynnetään datan luokittelua (engl. labelled) (Bramer 2013). Ohjatussa oppimisessa on mallin haluttua ulostuloa kuvaavat muuttujat, kuten luokkien tunnukset. Bramerin (2013) mukaan kuvailevassa mallissa käytettävä data ei ole luokiteltu (engl. unlabelled) eli siinä ei ole mitään tiettyä muuttujaa, jota yritetään uudelle havainnolle ennustaa, vaan yritetään saada kaikki mahdollinen tieto, mitä datasta on saatavilla. Kuvailevaa mallia käyttävää tiedonlouhintaa kutsutaan myös ohjaamattomaksi oppimiseksi.

Muita tiedonlouhinnan tehtäviä ovat (Hand, Mannila ja Smyth 2001; Äyrämö 2006):

- **Eksploratiivinen data-analyysi** (engl. Exploratory Data Analysis, EDA): Tavoitteena on tutkia dataa ilman mitään selvää käsitystä siitä, mitä ollaan etsimässä. EDA:n tyypillisiä tekniikoita ovat interaktiivinen ja visuaalinen. Lisäksi on olemassa monia tehokkaita graafisia menetelmiä suhteellisen pienimuotoisille datajoukoille.
- **Mallien ja sääntöjen löytäminen** (engl. Discovery of Patterns and Rules): Tavoitteena on löytää suuresta datajoukosta kiinnostavia sääntöjä ja suhteita. Tähän tekniikoita ovat assosiaatiosääntö ja peräkkäishahmoja (engl. sequential patterns).
- **Hakeminen sisällön mukaan** (engl. Retrieval by Content): Tehtävän käyttäjää kiinnostaa tietynlaiset hahmot ja tavoitteena on löytää samanlaisia hahmoja datajoukosta. Käytetään yleisesti teksti- ja kuvadatoihin. Hahmo voi olla avainsanojen joukko ja halutaan löytää relevantti dokumentti monien dokumenttien joukosta (esim. Web-sivut).

4.5.2 Algoritmit

Tiedonlouhinta-vaihe sisältää myös tiedonlouhinnan algoritmin valinnan (6. askel kuviossa 4). Fayyadin, Piatetsky-Shapiron ja Smythin (1996b, 1996a, 1996c) mukaan askel sisältää metodin tai metodien valinnan, jolla etsitään datan hahmoja. Siinä on päätettävä, mitkä mallit ja parametrit voisivat olla tarkoituksenmukaisia, esimerkiksi mallit kategoriselle datalle ovat erilaisia kuin etäisyyspohjaiset mallit. Lisäksi on täsmättävä valittu metodi KDD-prosessin yleisten kriteerien kanssa, esimerkiksi käyttäjän kiinnostus malliin eikä sen ennustusmahdollisuuksiin.

Fayyad, Piatetsky-Shapiro ja Smyth (1996b, 1996a, 1996c) esittelevät algoritmien koostuvan pääosin jonkin tietyn kolmen komponentin yhdistelmästä:

- **Malli:** Sisältää kaksi oleellista tekijää: mallin tehtävän (esim. luokittelu ja klusterointi) ja mallin esitysmuodon (esim. useiden muuttujien lineaarinen funktio ja Gaussin todennäköisyystiheysfunktio). Lisäksi malli sisältää parametrejä, jotka määritetään datasta. Analysoijan on täysin ymmärrettävä esitettävät oletukset, jotka voivat olla luontaisia tietyssä menetelmässä.
- **Prefenssikriteeri:** Perusta yhden mallin tai parametrisarjan preferenssille toisesta mallista tai parametrisarjasta. Kriteeri sille, kuinka hyvin tietty malli tai parametri täyttää KDD:n tavoitteet. Tavallisesti kriteeri on jonkinlainen mallin yhteensopivuuden aste (engl. goodness-of-fit) dataa varten, jolla esimerkiksi lievitetään ylisovittumista. Esimerkiksi ennustettava malli arvioidaan sen ennustustarkkuudella ja kuvailevaa mallia voidaan arvioida uudenlaisuuden, hyödyllisyyden ja ymmärrettävyyden mukaan.
- **Etsintäalgoritmi:** Algoritmien määrittely tiettyjen mallien ja parametrien, annetun datan, mallin ja preferenssikriteerin löytämiseksi. Kun aikaisemmat komponentit on sovitettu, niin etsintäalgoritmin tehtävänä on puhtaasti optimointi.

4.5.3 Tiedonlouhinta

Tiedonlouhinnan tehtävien ja algoritmien määrittämisen jälkeen seitsemäntenä askeleena on suorittaa tiedonlouhinta valitulla algoritmilla mallin saavuttamiseksi. Tiedonlouhinnalla etsitään mielenkiintoisia hahmoja tietystä esitysmuodosta tai sen joukosta, sisältäen luokitte-

lusäännöt tai -puut, regression ja klusteroinnin (Fayyad, Piatetsky-Shapiro ja Smyth 1996a). Fayyadin, Piatetsky-Shapiron ja Smythin (1996a) mukaan tiedonlouhintamenetelmän toimintaa voi edistää merkittävästi suorittamalla edelliset vaiheet oikein. Se, kuinka hyvän tuloksen prosessi antaa, riippuu kuitenkin kaikissa vaiheissa tehdyistä valinnoista.

4.6 Tulkinta / Arviointi

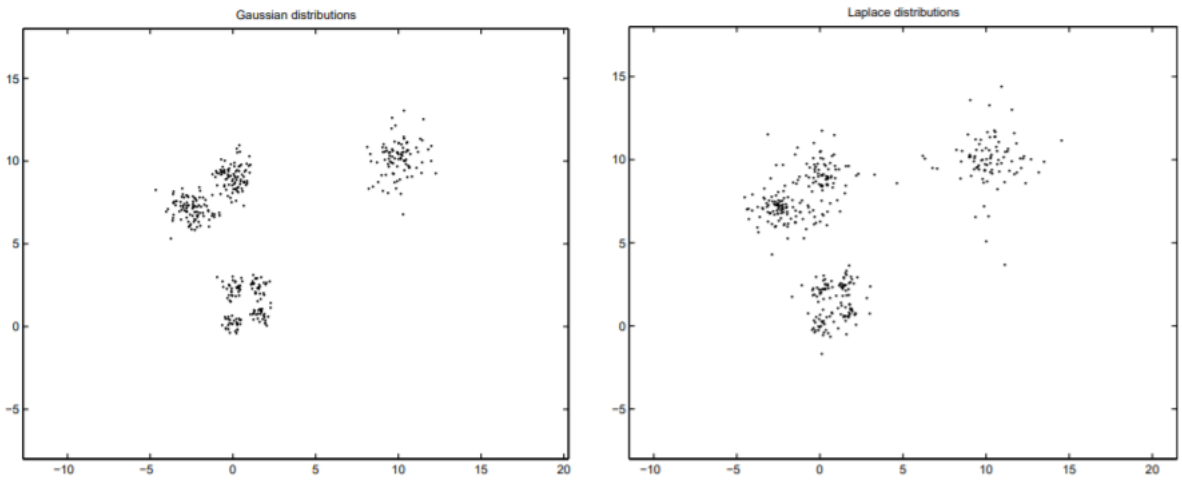
KDD-prosessin viimeinen vaihe sisältää tulkinnan (8. askel kuviossa 4) löytyneistä hahmoista ja mahdollistaa paluun aikaisempiin askeliin. Lisäksi askel sisältää datan visualisoinnin saaduilla hahmoilla, tarpeettomien sekä merkityksettömien hahmojen poistamisen ja hyödyllisien hahmojen muokkaamisen ymmärrettävään muotoon (Fayyad, Piatetsky-Shapiro ja Smyth 1996b, 1996a, 1996c). Tulkinnan jälkeen on tärkeää hyödyntää löydettyä tietoa (9. askel kuviossa 4). Fayyadin, Piatetsky-Shapiron ja Smythin (1996b, 1996a, 1996c) mukaan tietoa hyödynnetään joko suoraan, sisällyttämällä tietämys toiseen järjestelmään, jossa sitä voidaan hyödyntää, tai yksinkertaisesti raportoimalla löydöistä kiinnostuneita osapuolia. Lisäksi tässä vaiheessa tarkistetaan tulokset ja ratkaistaan mahdolliset ristiriidat aikaisempien tietojen perusteella.

5 Klusterointi

Klusterointi on havaintojen, datan tietoalkioiden tai muuttujavektoreiden eli hahmojen ohjaamatonta luokittelua ryhmiin eli klustereihin (Jain, Murty ja Flynn 1999). Zakin ja Meira Jr:in (2014) mukaan klusterointi jakaa havaintopisteet luonnollisiksi klustereiksi siten, että havaintopisteet ovat samanlaiset klusterien sisällä ja erilaiset suhteessa toisten klustereiden havaintopisteisiin. Klusteroinnista on erilaisia mallinnusparadigmoja, kuten jakava eli prototyyppipohjainen, hierarkkinen, tiheyspohjainen, graafinen ja spektriin liittyvä klusterointi. Mallien valinta riippuu käytettävästä datasta ja halutun klusterin ominaisuuksista. Täten erilainen klusterikriteeri, algoritmi tai jopa sama algoritmi, mutta erilaiset parametrit, voivat johtaa täysin erilaisiin klusterointituloksiin (Xu ja Wunsch 2009).

5.1 Klusterit ja klusteroinnin vaiheet

Xun ja Wunschin (2009) mukaan klusteri on havaintopisteiden yhdistelmä testiavaruudessa siten, että kahden pisteen etäisyys klusterissa on vähemmän kuin etäisyys minkä tahansa pisteen välillä klusterissa tai pisteen välillä, mikä ei ole samassa klusterissa. Estivill-Castron (2002) mukaan klusterialgoritmeja on monia, koska klustereita ei voi tarkasti määritellä vaan klusterointi on katsojan silmässä. Kuviossa 5 on esitetty, miten klustereiden määrä ja muoto ei aina ole selvä, sillä katsojasta riippuen klustereita voi olla kolme tai seitsemän.



Kuvio 5: Vasemmalla klusterit on muodostettu datasta normaalijakauman sekoituksena ja oikealla Laplace-jakauman sekoituksena. (Äyrämö 2006)

Klusterointi ei ole vain pelkästään tietyn tekniikan käyttämistä datan tutkimiseen, vaan se vaatii useita vaiheita, jotka voivat riippua toisistaan. Xun ja Wunschin (2009) mukaan klusterointi sisältää seuraavat neljä vaihetta (ks. kuvio 6):

1. **Muuttujien valinta ja poistaminen:** Muuttujien valinta ja poistaminen ovat hyvin tärkeitä klusteroinnin tehokkuuden kannalta. Valinnassa valitaan tunnusomaisia muuttujia ehdokasjoukosta tai poistossa muutetaan alkuperäisiä muuttujia hyödyllisiksi ja uusiksi muuttujiksi.
2. **Klusteroinnin algoritmin suunnittelu tai valinta:** Yleensä päätetään sopiva läheisyysmittari ja rakennetaan kriteerifunktio. Melkein jokainen klusterialgoritmi on nimenomaisesti tai epäsuorasti yhteydessä tiettyyn läheisyysmittauksen määrittelyyn. Läheisyysmittarin päätöksen jälkeen klusterointi voidaan ajatella optimointiongelmaksi tietyn kriteerifunktion avulla.

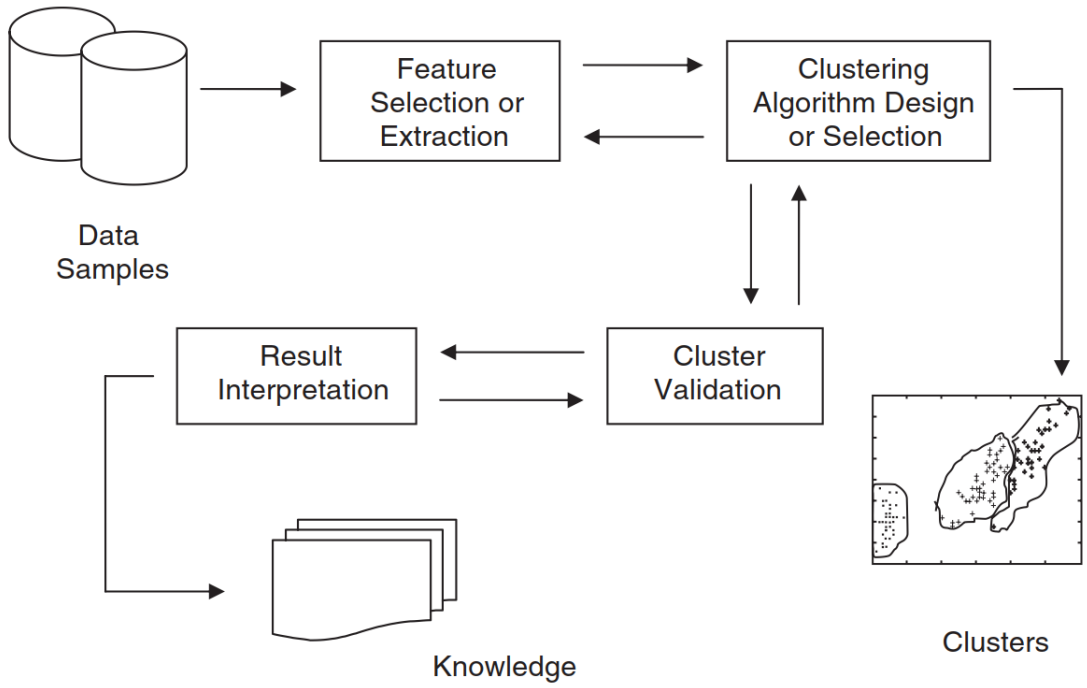
   Yleisin läheisyysmittari jatkuville muuttujille on euklidinen etäisyys, joka voidaan yleistää erityistilanteeksi Minkowskin etäisyydestä, joka tunnetaan l_p normina,

$$D(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{d} |x_{il} - x_{jl}|^p \right)^{1/p}. \tag{5.1}$$

   Kun $p = 2$, niin etäisyys on euklidinen etäisyys ja $p = 1$, niin on city-block eli Man-

hattan etäisyys eli L_1 normi. Xu ja Wunsch (ks. 2009, luku 2) käyvät läpi lisää läheisyysmittareita eri muuttujille.

3. **Klusterin validointi:** Kuten jo aikaisemmin todettiin, niin erilaiset klusteroinnin lähestymistavat johtavat erilaisiin klustereihin. Lisäksi klusterointialgoritmi voi aina tuottaa jokinlaisen jaon, vaikka todellisuudessa jaottelua ei olisikaan. Täten klusterien tehokkaat arviointistandardit ja -kriteerit ovat erittäin tärkeitä, jotta klusterointitulokset olisivat luotettavia. Luvussa 5.5 on tarkemmin klusterien validoinnista.
4. **Tulosten tulkinta:** Klusteroinnin tavoitteena on tuottaa alkuperäisestä datasta merkityksellisiä näkemyksiä, jotta voidaan ymmärtää dataa ja ratkaista ongelmia. Klusterit itsessään eivät ole lopputulos, vaan hahmotelma, joita pitää vielä tulkita.



Kuvio 6: Klusteroinnin vaiheet (Xu ja Wunsch 2009)

5.2 Hierarkkinen klusterointi

Hierarkkinen klusterointi luo sisäkkäisten osioiden sarjan n pisteistä d-ulotteisessa avaruudessa, mitä voidaan visualisoida hierarkiana klustereista (dendrogrammina) (Zaki ja Meira Jr. 2014). Hierarkian alhaisin taso muodostuu jokaisesta pisteestä, jotka ovat omia klusterei-

ta. Ylimmällä tasolla kaikki pisteet muodostavat yhden klusterin. Rakenteen horisontaalisilta tasoilta näkee klustereiden määrän (Zaki ja Meira Jr. 2014).

Xun ja Wunschin (2009) mukaan hierarkkinen klusterointi voidaan tehdä, joko kokoavan hierarkkisen klusteroinnoin (engl. agglomerative hierarchical clustering) tai hajottavan hierarkkisen klusteroinnin (engl. divisive hierarchical clustering) mukaan. Agglomeratiivinen klusterointi alkaa N klustereista, jotka sisältävät yhden havaintopisteen, jotka yhdistetään toisiinsa siten, että lopussa on yksi klusteri, joka sisältää kaikki havaintopisteet. Hajottava klusterointi lähtee jakamaan koko datan (kaikki havaintopisteet) sisältävää klusteria päätyen siihen, että jokaiset havaintopisteet ovat omia klustereita.

5.3 Prototyyppipohjainen klusterointi

Prototyyppipohjaisessa klusteroinnissa tarkoituksena on jakaa n havaintopisteinen data d-ulotteisessa avaruudessa haluttuihin K klustereihin eli suorittaa *klusterointi*, $\mathscr{C} = \{C_1, C_2, ..., C_K\}$. Jokaiselle C_i on olemassa edustava piste (prototyyppi), joka summaa klusterin. Yleisin valinta prototyypiksi on kaikkien pisteiden keskiarvo $\boldsymbol{\mu}_i$,

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{x_j \in C_i} \mathbf{x}_j, \tag{5.2}$$

missä $n_i = |C_i|$ on klusterin C_i pisteiden lukumäärä (Zaki ja Meira Jr. 2014).

Koska klusteroinnin tarkoituksena on jakaa dataa siten, että samassa klusterissa olevat havaintopisteet ovat homogeeniset ja klusterit keskenään ovat erilaiset, niin tällöin jakoa on arvioitava jälkikäteen kriteerifunktiolla (Xu ja Wunsch 2009). Tavalisimmin käytetty kriteerifunktio on virheiden neliösumma (Sum of Squared Errors, SSE) (Zaki ja Meira Jr. 2014),

$$SSE(\mathscr{C}) = \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2. \tag{5.3}$$

Tavoitteena on löytää klusterointi, joka minimoi SSE:n arvon (Zaki ja Meira Jr. 2014),

$$\mathscr{C}^* = \arg\min_{\mathscr{C}} \{SSE(\mathscr{C})\}. \tag{5.4}$$

Funktio $\arg\min()$ palauttaa klusteroinnin $\mathscr{C}$, jolla pienin SSE:n arvo.

5.3.1 K:n keskiarvon klusterointimenetelmä (K-means)

K:n keskiarvon klusterointimenetelmä (K-means) etsii optimaalista datan jaottelua minimoimalla SSE-kriteerin. Tämä menetelmä on tunnetuin ja käytetyin klusterioinnin algoritmi. Sen perusmenettely on seuraavanlainen (Zaki ja Meira Jr. 2014; Xu ja Wunsch 2009):

1. Alustetaan K klusterin prototyypit satunnaisesti tai perustuen johonkin etukäteistietoon.
2. Määritetään datan jokainen havaintopiste lähimpään prototyyppiin. Tällöin saadaan jokainen havaintopiste $\mathbf{x}_j$ osoitettua klusteriin C_{j*}, jossa

$$j^* = \arg\min_{i=1}^{k} \left\{ ||\mathbf{x}_j - \boldsymbol{\mu}_i||^2 \right\} \tag{5.5}$$

3. Lasketaan uusiksi prototyypit.
4. Toistetaan askeleet 2 ja 3 kunnes prototyypit ja klusterit eivät enää muutu.

K:n keskiarvon klusterointimenetelmä on Xun ja Wunschin (2009) mukaan helppo toteuttaa ja se toimii hyvin käytännön ongelmissa, varsinkin jos klusterit ovat kompakteja ja pallomaisia (engl. hyperspherical). Aikavaativuus on lähes lineaarinen, jolloin sillä voi klusteroida isoja datajoukkoja. K:n keskiarvon klusterointimenetelmän ongelmana voi joskus olla esimerkiksi suppeneminen ja alkujaottelu. Iteratiivisesti tehty optimaalinen menettely ei voi taata suppenemista globaaliin optimiin. Se kuitenkin suppenee paikalliseen optimiin, jolloin erilaiset alkupisteet (prototyypit) johtavat erilaisiin suppemisprototyyppeihin. Tällöin on tärkeää aloittaa sopivalla alkujaottelulla, jotta saavuttaa laadukkaan klusterointituloksen. Teoriassa ei ole kuitenkaan olemassa mitään tehokasta ja yleistä menetelmää aloitusjaottelun päättämiseen. Stokastisilla optimaalisilla hakutekniikoilla on mahdollista ratkaista monimutkaisia ongelmia tehokkaasti ja löytämään globaalin tai lähes globaalioptimin. Yleisin strategia ongelman suhteen on kuitenkin se, että algoritmia ajetaan useita kertoja satunnaisella

alkujaottelulla.

Xun ja Wunschin (2009) mukaan ongelmana on myös klustereiden lukumäärä eli K. K:n keskiarvon klusterointimenetelmä olettaa, että klustereiden määrä on tiedossa, vaikka yleensä näin ei ole. Lukumäärän valintaan ei ole myöskään mitään tehokasta tai yleistä menetelmää. Lisäksi K:n keskiarvon klusterointimenetelmä on herkkä poikkeaville havainnoille ja kohinalle, koska keskiarvon laskemisessa otetaan huomioon kaikki havaintopisteet myös poikkeavat havainnot. Tällöin klusterista kaukana oleva havaintopiste on prototyypin laskennassa mukana, jolloin se vääristää klusterin muotoa.

Xun ja Wunschin (2009) mukaan keskiarvon laskennan takia K:n keskiarvon klusterointimenetelmän käyttäminen rajoittuu numeerisiin muuttujiin jättäen kategoriset muuttujat pois. Toisaalta myös numeerisilla muuttujilla keskiarvo voi olla vaikeaa tulkita. Tällöin K:n medoidin klusterointimenetelmä on luonnollinen valinta, kun laskennallista keskiarvoa ei voi käyttää, koska medoidi on aina olemassa. Medoid on piste, jolla on pienin keskietäisyys saman klusterin sisällä oleviin kaikkiin muihin havaintopisteisiin. Estivill-Castro ja Yang (2000) ovat esitelleet K:n medoidin klusterointimenetelmään pohjautuvan algoritmin, jossa käytetään diskreettiä mediaania klusterin prototyyppinä.

Tutkielmassa hyödynnetään K:n keskiarvon klusterointimenetelmää, koska varsinkin asiakashankintavaiheen datajoukko on suuri, joten kyseinen menetelmä on tehokkain ja helpoin toteuttaa. Lisäksi datojen muuttujat ovat numeerisia, jolloin K:n keskiarvon klusterointimenetelmää voidaan suoraan hyödyntää.

5.4 Tiheyspohjainen klusterointi

Monet yleisimmin käytetyt klusterointialgoritmit tuottavat Aggarwalin ja Reddyn (2014) mukaan pyöreän muotoisia klustereita eivätkä pysty käsittelemään datajoukkoja, jotka todellisuudessa eivät ole pyöreän muotoisia. Lisäksi esimerkiksi K:n keskiarvon klusterointimenetelmässä on oltava jonkinlainen käsitys klustereiden määrästä. Tiheyspohjainen klusterointi voi puolestaan muodostaa muodoltaan mielivaltaisia klustereita ja lisäksi se ei tee oletuksia klustereiden määrästä. Siten se ei ole herkkä poikkeaville havainnoille tai kohinalle (Subramania ym. 2011). Tiheyspohjainen klusterointi käyttää etäisyyden sijasta pistei-

den paikallista tiheyttä päättääkseen klusterit (Zaki ja Meira Jr. 2014). Klusterit ovat alueita, joissa on korkea tiheys, joita ympäröi alhaisen tiheyden alueet eli kohina (Subramania ym. 2011).

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) on eniten käytetty tiheyspohjaisen klusteroinnin menetelmä (Ester ym. 1996). Subramanian ym. (2011) mukaan siinä tiheys määritellään paikallisesti jokaisen pisteen naapurustossa eli lähimpien pisteiden joukossa. Se vaatii kaksi parametriä: pisteiden naapuruston määrittelyyn etäisyyden ε (Eps) ja minimi määrän pisteitä naapurustossa μ (MinPts). ε voidaan määritelmä seuraavasti: $N_\varepsilon(\mathbf{x}) = \{\mathbf{y}|\delta(\mathbf{x},\mathbf{y} \leq \varepsilon)\}$ (Zaki ja Meira Jr. 2014).

Tiheyspohjainen klusterointi erottelee kolmen tyyppistä pistettä (Zaki ja Meira Jr. 2014; Aggarwal ja Reddy 2014):

- **Ydin piste** (engl. core point): Piste, jos sillä on vähintään MinPts:n verran pisteitä naapurustossaan (tiheä naapurusto), $|N_\varepsilon(\mathbf{x})| \geq \mu$.
- **Reunapiste** (engl. border point): Piste, joka kuuluu naapurustoon tai klusteriin jonkun ydin pisteen mukaan, mutta sillä itsellä ei ole MinPts:n verran pisteitä naapurustossa (naapurusto ei ole tiheä), $|N_\varepsilon(\mathbf{x})| < \mu$.
- **Kohinapiste** (engl. noise point): Piste, joka ei ole ydin piste tai rajapiste eli ei kuulu naapurustoon tai klusteriin.

5.5 Klusterin validointi

Prototyyppipohjainen klusterointi, kuten K:n keskiarvon klusterointimenetelmä olettaa, että klustereiden määrä on tiedossa, koska se jakaa datan etukäteen määriteltyihin lukumääriin. Täten klusteroinnin tulosten kannalta on tärkeä tietää, mikä on klustereiden optimaalinen määrä. Rendonin ym. (ks. 2011, 27) mukaan klusterin validointi on tekniikka, jolla löydetään klustereiden lukumäärä eli joukko selviä klustereita, ilman tietoa luokittelusta. Klusterin validointi-indeksit voivat perustua rakenteellisesti ulkoiseen (engl. external), sisäiseen (engl. internal) tai suhteelliseen (engl. relative) tietoon. Ulkoinen validointi-indeksi perustuu aikaisempaan tietoon datasta ja sisäinen validointi perustuu pelkästään tietoihin datasta. Suhteellisessa kriteerissä klusteroinnin tulosta vertaillaan muihin klusterointituloksiin. Täs-

sä tutkielmassa käytetään sisäisiä indeksejä, koska tarkempaa aikaisempaa tietoa datasta ei ole ja valinta halutaan tehdä datan pohjalta.

Rendonin ym. (2011) mukaan klusterin validointi-indeksit ovat yleensä määritelty yhdistämällä kompaktius ja erotettavuus. Kompaktius mittaa klusterin sisällä olevien elementtien läheisyyttä ja sen yleisin mittari on varianssi. Eroavaisuus mittaa kahden klusterin erillisyyttä eli kahden klusterin välimatkaa. Hämäläisen, Jauhiaisen ja Kärkkäisen (2017) mukaan klusteroinnin yleinen tavoite on siis, miten hyvin klusterin samanlaisuus eli pisteiden etäisyys klusterin prototyypistä (Intra) sekä eri klustereiden välinen eroavaisuus eli esimerkiksi klustereiden keskinäinen etäisyys (Inter) toteutuvat. Riippuen mittarista Intran ja Interin alhaisemmat tai korkeammat arvot ovat hyviä. Yleensä validoinnin mittarin arvo muodostuu Intran ja Interin osamääränä ja optimaalinen arvo on joko pienin tai suurin arvo riippuen jakolaskun muodosta.

Hämäläisen, Jauhiaisen ja Kärkkäisen (2017) ja Everittin ym. (2010) mukaan on hyvä valita yhden sijasta useita validointi-indeksejä, joilla arvioida klustereiden määrää, jotta saa luotettavimman klusterin validoinnin. Indeksien valintaan vaikuttaa data, mutta valinnassa on myös hyvä huomioida indeksien algoritmien kertaluokka. Sisäisiä validointi-indeksejä on esitetty taulukossa 5, joka on muodostettu Hämäläisen, Jauhiaisen ja Kärkkäisen (ks. 2017, s.5) tutkimuksen taulukon mukaisesti. Tutkimusta varten he muokkasivat indeksien laskentaa siten, että klusterien optimaalisen lukumäärän osoittaa indeksin pienin arvo. Ainoastaan Wemmert-Ganqarski -indeksi on jätetty alkuperäiseen muotoon, jolloin sen suurin arvo osoittaa klusterien optimaalisen lukumäärän. Hämäläinen, Jauhiainen ja Kärkkäinen (ks. 2017, s.5) valitsivat kyseiset indeksit, koska Jauhiaisen ja Kärkkäisen (2017) tutkimuksessa kyseiset indeksit suoriutuivat parhaiten. Indeksit on esitetty taulukossa kertaluokkien mukaisessa järjestyksessä. Lisäksi taulukossa on viittaukset alkuperäisiin artikkeleihin, joista löytyvät indeksirakenteiden perustelut ja tarkemmat kuvaukset.

Taulukko 5: Sisäiset validointi-indeksit (Hämäläinen, Jauhiainen ja Kärkkäinen 2017)

Nimi	Intra	Inter	Kaava
KCE	$K \times J_K$		Intra
(Jauhiainen ja Kärkkäinen 2017)			
WB-indeksi, WB	$K \times J_K$	$\sum_{k=1}^{K} n_k \|\|\mathbf{c}_k - m\|\|_p^q$	$\frac{\text{Intra}}{\text{Inter}}$
(Zhao ja Fränti 2014)			
Calinski–Harabasz, CH	$(K-1) \times J_K$	$(N-K) \times \sum_{k=1}^{K} n_k \|\|\mathbf{c}_k - m\|\|_p^q$	$\frac{\text{Intra}}{\text{Inter}}$
(Calinski ja Harabasz 1974)			
Davies–Bouldin, DB	$\frac{1}{n_k} J_K^k + \frac{1}{n_{k'}} J_K^{k'}$	$\|\|\mathbf{c}_k - \mathbf{c}_{k'}\|\|_p^q$	$\frac{1}{K} \sum_{k=1}^{K} \max_{k \neq k'} \frac{\text{Intra}(k,k')}{\text{Inter}(k,k')}$
(Davies ja Bouldin 1979)			
Pakhira, Bandyopadhyay, Maulik, PBM	$K \times J_K$	$\max_{k \neq k'} (\|\|\mathbf{c}_k - \mathbf{c}_{k'}\|\|_p^q) \times J_1$	$(\frac{\text{Intra}}{\text{Inter}})^2$
(Pakhira, Bandyopadhyay ja Maulik 2004)			
Ray-Turi, RT	$\frac{1}{N} \times J_K$	$\min_{k \neq k'} \|\|\mathbf{c}_k - \mathbf{c}_{k'}\|\|_p^q$	$\frac{\text{Intra}}{\text{Inter}}$
(Ray ja Turi 2000)			
Wemmert-Gançarski, WG	$\|\|\mathbf{x}_i - \mathbf{c}_k\|\|_p^q$	$\min_{k \neq k'} \|\|\mathbf{x}_i - \mathbf{c}_{k'}\|\|_p^q$	$\frac{1}{N} \sum_{k=1}^{K} \max(0, n_k - \sum_{i \in I_k} \frac{\text{Intra}(i)}{\text{Inter}(i)})$
(Desgraupes 2013)			

Taulukossa 5 K on klustereiden määrä ja $\{\mathbf{c}_k\}$ on prototyyppipohjaisella klusteroinnilla saavutetut parhaat prototyypit sekä $\mathbf{C}_k$ prototyyppien läheisyydessä olevien pisteiden joukot eli klusterit, $k = 1, ..., K$ (ks. Hämäläinen, Jauhiainen ja Kärkkäinen 2017, s.5). Kun $K = 1$, niin prototyyppi on koko datan keskiarvo, mediaani tai spatiaalinen mediaani, m, riippuen käytetystä etäisyysmitasta eli l_p-normista ja potenssista q. Esimerkiksi keskiarvossa $p = q = 2$. J_K on koko datan klusterointivirhe (engl. clustering error), $J_K^k = \sum_{\mathbf{x}_i \in \mathbf{C}_k} ||\mathbf{x}_i - \mathbf{c}_k||_p^q$ (ks. Hämäläinen, Jauhiainen ja Kärkkäinen 2017, s.3).

Taulukossa 5 puuttuu yleinen ja hyvin toimiva Silhouette-indeksi. Hämäläinen, Jauhiainen ja Kärkkäinen (2017) olivat jättäneet testeistä kyseisen indeksin pois, koska sen kertaluokka on $\mathcal{O}(N^2 n)$ ja siten isolle datalle suoritettuna se voi olla liian raskas. Rousseeuwin (1986) mukaan indeksin graafisessa muodossa jokaista klusteria edustaa niin sanotut *silhouetit*, jotka perustuvat klusterin tiiviyteen ja eroavaisuuteen. Lisäksi graafisesta muodosta näkee hyvin, mitkä havaintopisteet on jaettu oikeisiin klustereihin ja mitkä saattaisivat kuulua toiseen klusteriin. Keskimääräinen silhouette-leveys auttaa arvioimaan klusterin validiutta ja siten sitä voidaan käyttää valitsemaan sopiva määrä klustereita.

Silhouette-arvo (ks. Rousseeuw 1986, s.55-57) muodostetaan siten, että $a(\mathbf{x}_i)$ on pisteen $\mathbf{x}_i$ keskimääräinen eroavaisuus (yleensä euklidinen etäisyys) saman klusterin muihin pisteisiin. C on toinen lähin klusteri pisteelle $\mathbf{x}_i$ ja $b(\mathbf{x}_i)$ on pisteen $\mathbf{x}_i$ pienin (minimi) keskimääräinen eroavaisuus muihin pisteisiin C:ssä. Tämän jälkeen jokaiselle pisteelle $\mathbf{x}_i$ muodostetaan arvo $s(\mathbf{x}_i)$, joka kuvaa pisteen silhouette-leveyttä:

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}} \tag{5.6}$$

Jokaiselle pisteelle $\mathbf{x}_i$ silhouette leveys $s(\mathbf{x}_i)$ on välillä $[-1, 1]$. Kun arvo on lähellä 1, niin piste $\mathbf{x}_i$ on sille kuuluvassa klusterissa. Arvon ollessa lähellä 0, niin ei ole selvää, onko piste $\mathbf{x}_i$ oikeassa klusterissa vai kuuluisiko se toiseen lähimpänä olevaan klusteriin. Arvon ollessa lähellä -1, piste $\mathbf{x}_i$ on väärässä klusterissa. Täten pisteiden arvojen ollessa positiivisia, klusterointi on sopiva. Lopullinen indeksi (silhouette-leveyden kokonaiskeskiarvo) kuvaamaan validiutta muodostetaan $s(\mathbf{x}_i)$ arvojen keskiarvosta koko datan kaikille pisteille $\mathbf{x}_i$:

$$\bar{s}(k) = \frac{1}{N}\sum_{i=1}^{N} s(\mathbf{x}_i). \tag{5.7}$$

Klustereiden määräksi voidaan valita se k, jonka $\bar{s}(k)$ (Silhouette-indeksi) on suurin (Rousseeuw 1986).

Jauhiaisen ja Kärkkäisen (2017) tutkimuksen mukaan Silhouette-indeksi oli paras yleinen indeksi. MATLAB:in evalclusters-funktion indeksit (DB, CH, Silhouette) olivat selkeästi parhaita, sillä 12 datasta oikean määrä klustereita ne antoivat 9 datan kohdalla. Parhaimpia indeksejä tutkimuksen mukaan olivat CH, Silhouette, PBM ja WG, sillä ne antoivat myös 9 datan kohdalla oikean määrän klustereita. Lisäksi tutkimuksessa ehdotettu KCE indeksi toimi kaikille datoille, jopa haastaville (Sim52, Sim510).

Hämäläisen, Jauhiaisen ja Kärkkäisen (2017) tutkimus jatkoi edellä olevaa tutkimusta. Tutkimuksen mukaan K:n keskiarvon klusterointimenetelmässä neliöllisen euklidisen etäisyyden kanssa parhaimmat indeksit ovat: WG, KCE ja CH. KCE:n lisäksi WB löysi oikean määrän klustereita synteettiselle datalle (Sim5). Tutkimuksen testit eivät suosittele DB ja RT indeksejä. Yleisesti DB, RT ja lisäksi CH epäonnistuvat useammin, kun klustereiden määrä kasvaa.

6 Asiakkaan sitoutumisen louhinta

Tutkielman tutkimusosio on jaettu kolmeen osaan asiakkuuden elinkaaren vaiheiden mukaisesti. Jokaisessa osassa hyödynnettiin KDD-prosessia ja sen tiedonlouhinnan menetelmänä klusterointia, minkä tavoitteena oli selvittää asiakkaiden eri asteista sitoutumista. Asiakashankinnan ja sivukyselydatan kohdalla aikaisempien tutkimuksien pohjalta tiedetään, että sitoutuneimpia asiakkaita ovat ne, joiden verkkokaupan vierailtujen sivujen määrä on vähäinen, mutta sivustolla vietetty aika on kestoltaan pitempi (Raphaeli, Goldstein ja Fink 2017). Asiakkuuden kehittämisen ja sosiaalisen media kohdalla Malthousen ym. (2013) mukaan asiakas on puolestaan sitoutuneempi luodessaan sisältöä brändille kuin tykkäämällä tai jakamalla brändin sisällön julkaisua. Asiakkuuden säilyttämisessä asiakkaiden sosiaaliset kontaktit ovat keskeisiä, sillä asiakkaan poistumisella voi olla vaikutusta sosiaaliseen verkostoonsa (Malthouse ym. 2013; Bijmolt ym. 2010). Kahteen ensimmäiseen elinkaaren vaiheeseen etsittiin erilaisista avoimista tietovarastoista (engl. data repository) samankaltaista dataa, mitä aikaisemmissa tutkimuksissa oli käytetty. Asiakkuuden säilyttämisessä puolestaan analysoitiin asiakkuuden kehittämisen dataa.

6.1 Asiakashankinta ja sivukyselydata

Asiakkaan sitoutumisen tutkimiseen asiakashankintavaiheessa käytettiin sivukyselydataa, joka oli peräisin Kaggle-sivustolta (Retailrocket 2018). Datajoukko koostui kolmesta tiedostosta, joista tässä tutkimuksessa käytettiin vain käytösdataa sisältävää tiedostoa (events.csv). Data oli kerätty olemassa olevalta verkkokaupan sivustolta neljän ja puolen kuukauden ajalta vuodelta 2015. Events.cvs sisälsi asiakkaan tekemiä tapahtumia: view, addtocart, transaction. Lisäksi se sisälsi myös ajankohta, milloin asiakas oli klikannut jotain tapahtumaa sivustolla. Aika oli muunnettu UNIX-ajaksi. Asiakkaan, tuotteen ja tapahtuman yksilöintiin oli asiakkaan id, tuotteen id ja tapahtuman id.

6.1.1 Esikäsittely

Esikäsittelyvaiheessa käytettiin MATLAB:n (R2018a) table-taulukkoa, joka myöhemmin muutettiin matriisiksi. UNIX-aika muutettiin gregoriaanisen kalenterin ajaksi, joka sisälsi päivämäärän sekä kelloajan. Datasta poistettiin tapahtuman id, koska se yksilöi numerolla vain oston ja sitä ei sitoutumisen näkökulmasta tarvittu.

Datasta poistettiin sivustolla alle neljä klikkausta tehneet asiakkaat, koska Baumannin ym. (2017) mukaan vähintään neljä klikkausta tarvitaan ostojen tekemiseen. Sitoutumisen näkökulmasta Robertsin ja Alpertin (2010) mukaan asiakas on sitoutunut, kun hän ostaa mielellään toisen tuotteen, jolloin alle neljä klikkausta tehneet asiakkaat voitiin poistaa. Tätä soveltaen esikäsittelyssä poistettiin myös asiakkaat, jotka olivat ostaneet vähintään kaksi eri tuotetta, sillä he voidaan katsoa jo sitoutuneeksi, jolloin heitä ei tarvitse huomioida asiakashankintavaiheessa. Taulukossa 6 on kuvaukset alkuperäisestä datasta ja datoista, joista on poistettu vähäiset klikkaukset sekä sitoutuneet.

Klikkausajoista laskettiin jokaiselle päivälle asiakkaan vierailun kesto sekunneissa. Joidenkin asiakkaiden kohdalla päivittäinen kesto saattoi olla hyvinkin suuri, sillä asiakkaissa oli sellaisia, jotka olivat klikkailleet sivustoa pitkin päivää. Pitkien kestojen kohdalla voitiin olettaa, etteivät he olleet vierailleet sivustolla yhtä jaksoisesti koko aikaa. Tästä syystä päivittäisistä kestoista muodostettiin sessiot sessioiden heuristiikan mukaisesti, tarkemmin aikasuuntautuneen heuristiikan sessionkestopohjaisella metodilla (Raphaeli, Goldstein ja Fink 2017; ks. Berendt ym. 2001, s.730; Liu ja Keselj 2006). Session ensimmäisen klikkauksen ja seuraavan klikkauksen keston ylittäessä 1800 sekuntia eli 30 minuuttia klikkauksen katsottiin aloittavan uuden session. Lisäksi session viimeisen tapahtuman klikkaukseen lisättiin Liun ja Keseljin (ks. 2006, s. 310) mukaisesti kyseisen session keskimääräinen kesto, koska viimeisen tapahtuman klikkauksen kesto ei ollut tiedossa. Viimeisen klikkauksen kesto otettiin huomioon 30 minuutin raja-arvossa laskettaessa session kestoa. Näin ollen kaikki sessiot ovat kestoltaan enintään 30 minuuttia.

Raphaelin, Goldsteinin ja Finkin (2017), A. L. Montgomeryn ym. (2004) ja Liun ja Keseljin (2006) mukaisesti tutkielmassa huomioitiin sessiot, joissa oli vähintään kaksi tapahtumaklikkausta. Jos käyttäjällä oli vain yksi tapahtumaklikkaus sessiossa, se poistettiin käyttäjältä,

koska kestoa ei voitu laskea eikä haluttu käyttää käyttäjän tai kaikkien käyttäjien keskimääräistä kestoa. Lisäksi asiakkailta poistettiin sessiot, joiden kesto oli 10 sekuntia tai alle, koska siinä ajassa asiakas ei ehdi sitoutumismielessä tutkia tuotetta. Taulukon 6 viimeinen sarake kuvaa lopullisen datan rivien ja asiakkaiden määriä.

Taulukko 6: Asiakashankintadatan kuvailu

	Alkuperäinen data	Klikkauksia ≥ 4	Ostot< 2	Klikkaukset ≥ 2, kestot ≥ 10 s
Rivejä	2 756 101	1 103 721	947 767	171 789
Asiakkaita	1 407 580	120 416	117 864	109 888

Kestojen lisäksi jokaiselle käyttäjälle laskettiin eri tuotteiden lukumäärä sessioittain. Datasta laskettiin myös session aikana ostoskoriin siirrettyjen ja ostettujen eri tuotteiden lukumäärät. Koska datasta poistettiin asiakkaat, jotka olivat ostaneet vähintään kaksi tuotetta, ostojen lukumäärä oli joko 0 tai 1 eli ostoista muodostui dikotominen muuttuja. Näistä muuttujista ja asiakkaan yksilöivästä id:stä muodostettiin datamatriisi taulukon 7 mukaisesti, jossa kaksi viimeistä saraketta sisältää metadataa: ostettu ja asiakas. Datamatriisin riveillä on asiakkaiden jokainen sessio. Matriisin muuttujat on kuvattu tarkemmin taulukkoon 8.

Taulukko 7: Asiakashankinnan datamatriisi

Kesto	Tuotteet	Ostoskori	Ostettu	Asiakas
1780	4	0	0	2
1270	2	0	0	6
274	1	0	0	37
285	2	0	0	37
⋮	⋮	⋮	⋮	⋮
43	2	0	0	155
474	6	6	0	155
600	1	1	1	186

Taulukko 8: Asiakashankinnan datan muuttujien kuvaus

Muuttuja	Kuvaus
Kesto	Asiakkaan session kokonaiskesto [s]
Tuotteet	Eri tuotteiden yhteismäärä, joita asiakas on klikannut sessiossa
Ostoskori	Eri tuotteiden yhteismäärä, joita asiakas on lisännyt ostoskoriinsa sessiossa
Ostettu	*Metadata*: Ostettujen eri tuotteiden yhteismäärä (vain 0 tai 1 kpl)
Asiakas	*Metadata*: id asiakkaan yksilöintiin

Datan havainnoista 87 %:lla oli *ostoskori*-muuttujan arvo nolla ja suurin arvo oli 55, jolloin min-max-skaalaus välille $[0,1]$ jättää lähes kaikkien muuttujien arvot edelleen nollaksi. Datan seitsemässä muunnoksessa *ostoskori*-muuttujaa esikäsiteltiin tekstinlouhinnassa dokumentin sanojen tärkeyden selvittämiseen käytettävällä TF-IDF-menetelmällä. Muuttujaa esikäsiteltiin erikseen TF:llä ja IDF:llä sekä näiden yhdistelmällä eli TF-IDF:llä (Manning, Raghavan ja Schutze 2009; Luhn 1957; Jones 1972).

TF määriteltiin seuraavasti:

$$1 + \log(f_{t,d}) \tag{6.1}$$

missä $f_{t,d}$ on ostoskoriin siirrettyjen tuotteiden lukumäärien frekvenssi koko datan osalta.

IDF:ssä asiakasta ajateltiin dokumenttina ja se määriteltiin seuraavasti:

$$\log \frac{N}{n_t} \tag{6.2}$$

missä N on koko datan asiakkaiden lukumäärä yhteensä ja n_t kuvaa asiakkaiden lukumäärän, joilla on kyseistä ostoskoriin siirrettyjen tuotteiden lukumäärää.

TF-IDF:ssä yhdistettiin TF hieman muutettuna ja IDF seuraavasti:

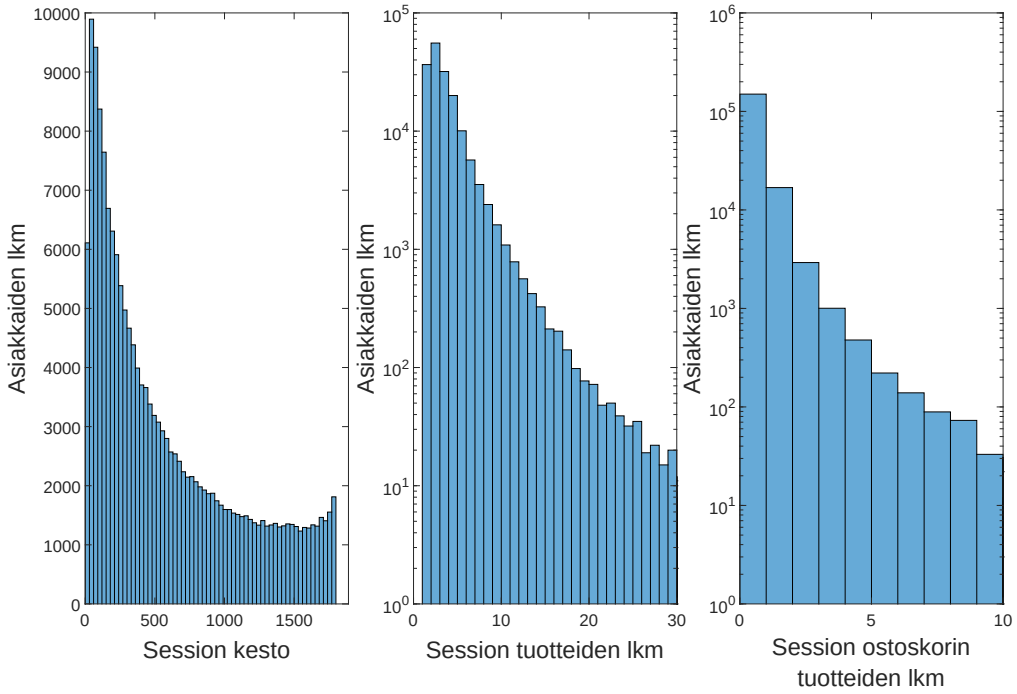
$$f_{t,d} \times \log \frac{N}{n_t} \tag{6.3}$$

missä $f_{t,d}$ on yksittäisen asiakkaan ostoskoriin siirretyn tuotteen lukumäärän frekvenssi. $\log \frac{N}{n_t}$ on sama kuin ylhäällä. Yksittäisen lukumäärän frekvenssi ($f_{t,d}$) kerrotaan vastaavalla lukumäärän IDF:llä ($\log \frac{N}{n_t}$).

6.1.2 Muunnos

Sitoutumista tutkittiin ensin sessioittain (muunnos 1) datamatriisin (kts. taulukon 7) mukaisesti. Muunnoksen data sisälsi 171 789 havaintoja ja kolme muuttujaa. Muunnokselle tehtiin myös min-max-skaalaus välille $[0, 1]$.

Datan jokainen muuttuja oli oikealle vinoutunut (kts. kuviosta 7), joten muunnoksen 1 *kesto* ja *tuotteet* -muuttujille tehtiin logaritminen muunnos, jotta datasta tulisi enemmän symmetrinen muodostaen muunnoksen 1.1. Kahden viimeisen kuvaajan häntä ulottuu pitkälle oikealle, koska muuttujat sisältävät yksittäisiä käyttäjiä, joilla tuotteiden määrät ovat suuria. Visualisointia varten kuvion 7 kuvaajien häntiä on katkaistu pienemmäksi ja pystyakselille on tehty logaritmiskaalaus.



Kuvio 7: Asiakkaiden lukumäärien jakauma session tiedoilla

Suurin osa datamatriisin *ostoskori*-muuttujan arvoista oli nolla, joten muunnoksessa 2.1 muuttujaan testattiin luvussa 6.1.1 määriteltyä TF:ää ja muunnoksessa 2.2 IDF:ää. TF-IDF-yhdistelmää testattiin *ostoskori*-muuttujaan muunnoksessa 2.3. Lisäksi näissä muunnoksissa *kesto* ja *tuotteet* -muuttujille tehtiin logaritminen muunnos ja muunnoksien kaikki muuttujat (ml. *ostoskori*-muuttuja) skaalattiin välille $[0, 1]$.

Lisäksi selvitettiin, miten sitoutuminen muuttuu, kun sessioiden sijasta tutkitaan asiakkaita. Näin ollen jokaisen asiakkaan sessioiden kestot, eri tuotteiden ja ostoskoriin siirrettyjen eri tuotteiden lukumäärät laskettiin yhteen. Tämän jälkeen data sisälsi 109 888 havaintoa ja kolme muuttujaa. Ostojen lukumäärät laskettiin myös yhteen, mutta lukumäärät pysyivät edelleen dikotomisena. Lisäksi datalle tehtiin min-max-skaalaus välille $[0, 1]$ (muunnos 3). Sessioiden yhteenlaskeminen ei muuttanut muuttujien vinoutumista, joten muunnoksessa tehtiin logaritmimuunnos *kesto* ja *tuotteet* -muuttujille (muunnos 3.1).

Asiakkaiden sessioiden tietojen yhteenlaskemisen jälkeen selvitettiin sessioiden lukumäärän vaikutusta klustereiden muodostumiseen. Täten yhdeksi muuttujaksi muodostettiin asiak-

kaan sessioiden lukumäärä, jolloin data sisälsi 109 888 havaintoja ja neljä muuttujaa. Muuttujat skaalattiin vielä välille $[0,1]$ (muunnos 4). Lisäksi muunnoksen *kesto* ja *tuotteet* -muuttujille tehtiin jälleen logaritminen muunnos (muunnos 4.1).

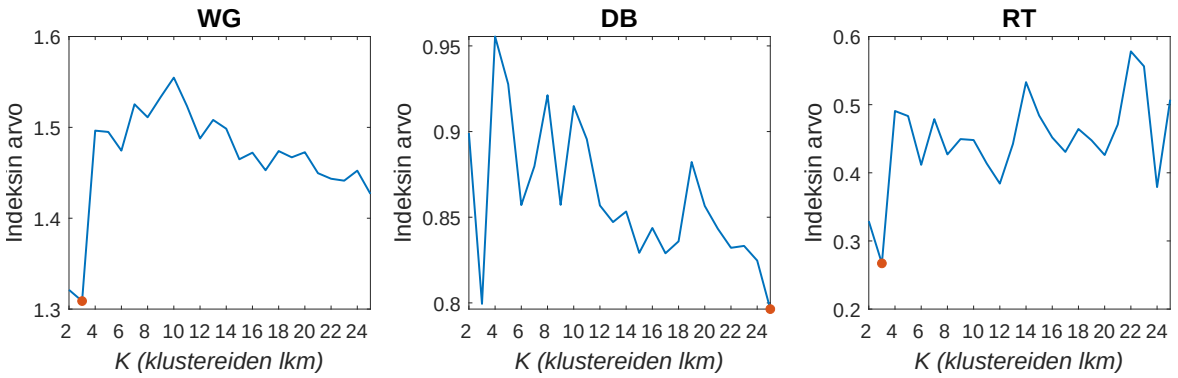
Ostoskori-muuttujan skaalaamista TF:llä (muunnos 5.1) ja TF-IDF:llä (muunnos 5.2) testattiin myös dataan, jossa sessiot olivat yhteenlaskettu. IDF:ää ei testattu, koska yhteenlaskettujen sessioiden kohdalla se oli sama kuin TF-IDF. Näiden kahden muunnoksen lisäksi huomioitiin sessioiden *lukumäärä* -muuttuja, joka skaalattiin TF:llä (6.1) ja TF-IDF:llä (muunnos 6.2) *ostoskori*-muuttujan lisäksi. Muille muuttujille tehtiin logaritminen muunnos ja kaikki muuttujat skaalattiin välille $[0,1]$. Näin ollen muodostettiin yhteensä 13 muunnosta.

6.1.3 Tiedonlouhinta

Asiakashankintavaiheen sitoutumisen tiedonlouhintaan käytettiin MATLAB:in k-means++ -algoritmia ja sen oletuksena käyttämää neliöllistä euklidista etäisyyttä. Klusterointi ja validointi-indeksien testaaminen suoritettiin klusterien määrällä $K = 2-25$. Klustereiden optimaalisen lukumäärän valinnassa hyödynnettiin seitsemää klusterien validointi-indeksiä: WB-indeksi, Calinski-Harabasz, Wemmert-Gançarski, KCE, PBM, Davies-Bouldin ja Ray-Turi. Indeksit valittiin Hämäläisen, Jauhiaisen ja Kärkkäisen (2017) ja Jauhiaisen ja Kärkkäisen (2017) tutkimuksien perusteella. Kaikki muut paitsi kolme validointi-indeksiä toteutettiin MATLAB:iin taulukon 5 kaavojen mukaisesti. Calinski-Harabasz ja Davies-Bouldin indeksien laskennassa hyödynnettiin MATLAB:in evalclusters-funktiota, jossa parametrina annettiin näiden indeksien nimet. Jos indeksin laskenta oli tehty niin, että klustereiden optimaalinen lukumäärä oli indeksin maksimiarvo, arvoista otettiin käänteisarvot, jolloin ne olivat vertailukelpoisia muihin indekseihin.

Muunnoksen 1 ja 2 kohdalla validointi-indeksit ehdottivat optimaalisiksi klustereiksi $K = 2$ tai $K = 25$, joten klusterien määrää ei voitu todeta. Klusterin määrää ei voitu myöskään selvästi havaita indeksien arvoista muodostettujen käyrien kuvaajista. Muunnokselle 1 tehdyn logaritmisen muunnoksen jälkeen (muunnos 1.1) WG, DB ja Ray-Turi indeksien arvojen perusteella klustereiden määräksi valittiin $K = 3$. Näiden indeksien arvot eri klustereiden määrien kohdalla on esitetty kuvion 8 kuvaajissa. Klustereiden optimaalinen määrä on kohdassa,

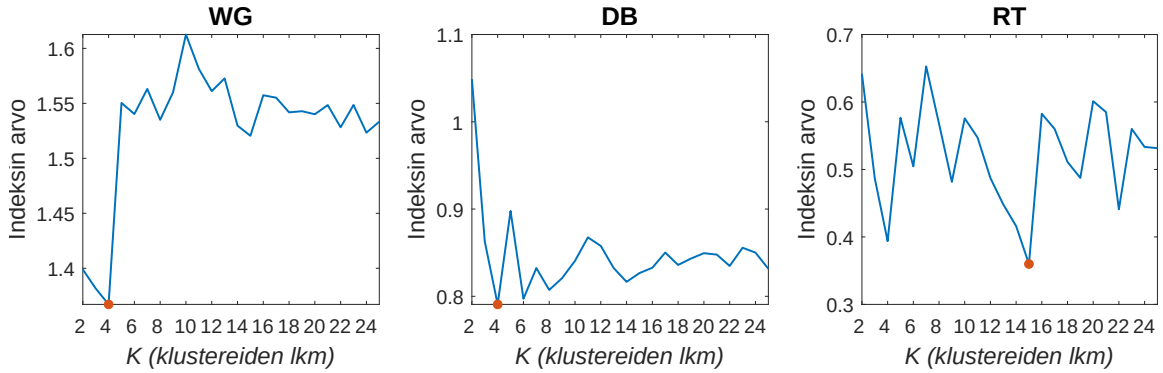
jossa on validointi-indeksin minimiarvo eli käyrän lokaali minimi (merkitty kuvaajaan punaisella pisteellä). DB-indeksi ehdotti määräksi $K = 25$, mutta käyrän kuvaavasta voi havaita, että selkeä "knee-point"-kohta on $K = 3$ kohdalla, joka on todella lähellä minimiä.



Kuvio 8: Muunnoksen 1.1 kolmea klusteria ($K = 3$) tukevien validointi-indeksien kuvaajat

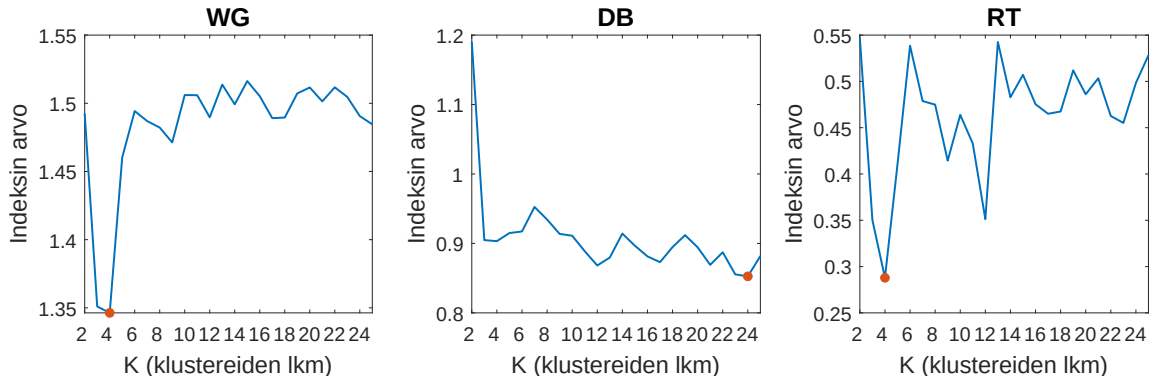
Sessioittain olevan datan *ostoskori*-muuttujan skaalaaminen TF:llä (muunnos 2.1) validointi-indeksit CH, WG, KCE, DB ja Ray-Turi ehdottivat klustereiden määräksi $K = 3$. Myös IDF:llä skaalaamisessa (muunnos 2.2) WG, CH, KCE ja Ray-Turi ehdottivat klusterin määräksi $K = 3$. Muunnoksien klustereiden määrät ovat samat muunnoksen 1.1 kanssa, joten niitä ei tutkittu tarkemmin. TF-IDF skaalauksen jälkeen (muunnos 2.3) CH-, KCE- ja PBM-indeksien ehdotuksen perusteella klusterien määräksi valittiin $K = 12$.

Muunnos 3.1, jossa asiakkaiden sessiot oli laskettu yhteen ja tehty logaritmimuunnos, WG-, DB- ja RT-indeksien arvojen perusteella klustereiden määräksi valittiin $K = 4$. Muunnoksen kohdalla RT-indeksi ehdotti klusterin määräksi $K = 15$, mutta indeksin kuvaajasta näkee selkeän "knee-point"-kohdan $K = 4$ kohdalla (kts. kuvio 9).



Kuvio 9: Muunnoksen 3.1 neljää klusteria ($K = 4$) tukevien validointi-indeksien kuvaajat

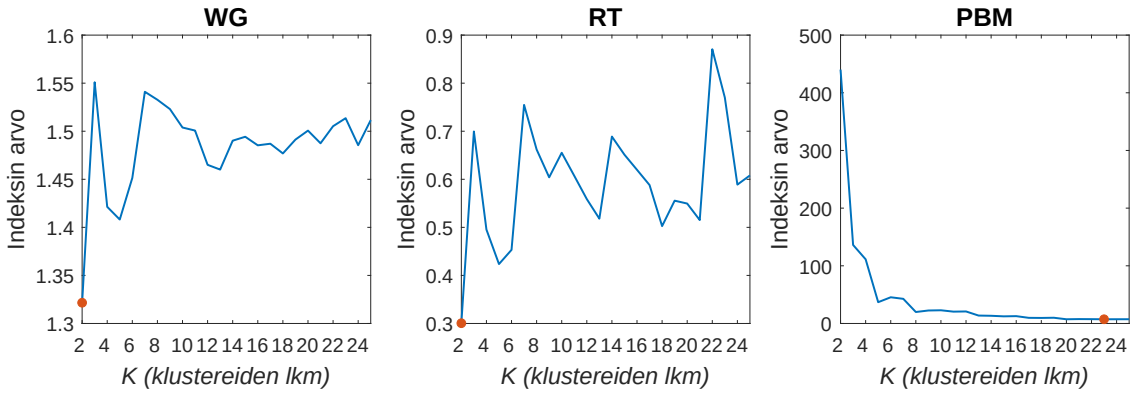
Sessioiden yhteenlasketun datan *ostoskori*-muuttujan TF-skaalauksen (muunnos 5.1) jälkeen CH ja KCE ehdottivat klusterien määräksi $K = 3$ sekä WG ja RT $K = 4$. Klusterien määrät olivat näin ollen lähellä muunnoksen 3.1 määriä, joten muunnoksen tuloksia ei tutkittu. *Ostoskori*-muuttujan TF-IDF-skaalauksen (muunnos 5.2) jälkeen validointi-indeksien kuvaajista pystyi todentamaan klusterien määräksi $K = 12$. Kuviosta 10 RT-indeksissä näkee selkeän "knee-point"-kohdan $K = 12$ kohdalla.



Kuvio 10: Muunnoksen 5.2 12 klusteria ($K = 12$) tukevien validointi-indeksien kuvaajat

Klusterien määrää ei voitu todeta, kun otettiin huomioon sessioiden lukumäärät muunnoksessa 4. Kun *kesto* ja *tuotteet* -muuttujille tehtiin logaritminen muunnos, niin validointi-indeksit CH, PBM ja KCE kuvaajan perusteella antoi klusterien määräksi $K = 16$ (muunnos 4.1). Näistä klustereista seitsemän oli kooltaan niin pieniä, että ne voi luokitella poikkeaviksi klustereiksi ja jättää huomioimatta. Näillä klustereilla kestot, tuotteiden ja sessioiden määrät olivat suuria (kts. liite C). Näin ollen tuloksia ei tutkittu tarkemmin. WG, RT ja

PBM validointi-indeksien kuvaajista pystyttiin puolestaan havaitsemaan klusterien määräksi $K = 5$, kun sessioiden lukumäärä skaalattiin TF-IDF:llä (muunnos 6.2). Kuviosta 11 näkee WG- ja RT-indeksien kohdalla selkeät “knee-point”-kohdat.



Kuvio 11: Muunnoksen 6.2 viittä klusteria ($K = 5$) tukevien validointi-indeksien kuvaajat

6.1.4 Tulokset: Muunnos 1.1

Asiakkaita sessioittain käsittelevän muunnoksen kohdalla löydettiin kolme klusteria. Klustereista kaksi olivat kooltaan huomattavasti suurempia, joten nämä klusteroitiin uudelleen hierarkkisesti (kts. (Wartiainen ja Kärkkäinen 2015)). Ensimmäisen kooltaan suuremman klusterin klusteroinnissa indeksit Ray-Turi ja WG antoivat klustereiden määräksi $K = 20$ ja kolmannen kohdalla Ray-Turi antoi klustereiden määräksi $K = 21$. Näin ollen hierarkkinen prototyyppipohjainen klusterointi antoi yhteensä 42 klusteria. Klustereiden tietoja nopeasti silmäillen huomattiin, että aliklustereiden tiedoissa ei ole merkittäviä eroja, jolloin uudelleenklusterointi ei tuo asiakkaiden sitoutumisesta lisää tietoa, eikä tulosta tulkittu enempää. Taulukko ja kuvio aliklustereista on esitetty liitteessä A.

Taulukossa 9 on metadata kolmesta klustereista ja koko datasta. Klusterit on merkitty C-kirjaimella ja Asiakkaita-rivi kuvaa klusterin yksittäisten asiakkaiden määrää. Ostoskori M_{d} -rivillä on ostoskoriin siirrettyjen eri tuotteiden lukumäärän mediaani. Tuoteita ostoskorissa 1 kpl ja Ostoja 1 kpl -rivien prosenttiosuudet kuvaavat havaintojen osuutta niistä, jotka ovat siirtäneet yhden eri tuotteen ostoskoriin tai ostaneet yhden eri tuotteen.

Taulukko 9: Muunnoksen 1.1 klustereiden metadata

	Koko data	C1	C2	C3
Koko	171 789	104 556	2 098	65 135
Asiakkaita	109 888	75 219	1 839	51 071
Kesto M_{d} [mm:ss]	6:46	12:28	14:17	1:55
M_{o}	0:18	5:13	29:37	0:18
Max	30:00	30:00	30:00	5:09
Tuote M_{d}	2	3	5	2
M_{o}	2	2	4	2
Max	186	60	186	28
Ostoskori M_{d}	0	0	4	0
M_{o}	0	0	3	0
Max	55	2	55	3
Tuotteita ostoskorissa 1 kpl	10%	11%	0%	7%
Tuotteita ostoskorissa > 2 kpl	3%	2%	100%	1%
Ostoja 1 kpl	3%	5%	3%	1%
Tuotteita ostoskorissa 2 kpl & Ostoja 1 kpl	0.22%	0.27%	3.15%	0.06%

Muunnoksen sessioiden kestojen ja eri tuotteiden lukumäärän mediaanit ja moodit (M_{o}) vaihtelevat selvästi keskenään. Session kestoltaan klusteri C2 näyttäisi olevan sitoutuneempi. Toisaalta huomioitaessa tuotteiden määrät, jotka kyseisellä klusterilla ovat suuret, Raphaelin, Goldsteinin ja Finkin (2017) mukaan klusterin asiakkaat eivät ole vahvasti sitoutuneita, koska vierailtujen sivujen määrän (tässä tutkielmassa tuotteiden määrän) pitäisi olla vähäinen. Klusterilla C2 ostoskoriin siirrettyjen eri tuotteiden määrät ovat myös suurempia, jotka vaikuttavat sitoutumisen asteeseen. Toisaalta ostojen osuus tällä klusterilla on toiseksi suurin. Tarkastellessa osuutta, jossa ostoskoriin on siirretty kaksi eri tuotetta ja ostettu yksi tuote, niin klusteri on prosenttiosuudeltaan suurin.

Huomioitaessa ostokoriin siirrettyjen tuotteiden lukumäärät ja ostot asiakkaiden voidaan ajatella jakautuneen klustereihin kolmen sitoutumisasteen mukaan: heikosti (C3), keskivertoi-

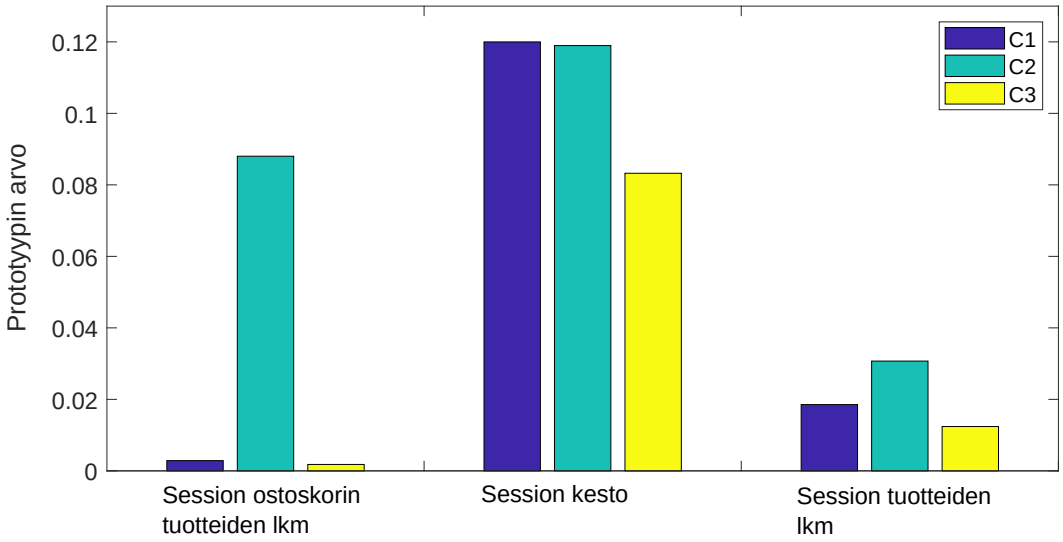
sesti (C1) ja vahvasti (C2) sitoutuneet. Vaihtoehtoisesti klusterin C1 asiakkaat voidaan ajatella vahvasti sitoutuneiksi ja klusterin C2 asiakkaat keskivertoisesti sitoutuneiksi. Tarkasteltaessa klustereiden kokoja ja niiden asiakkaiden määriä huomataan, että yksittäisen asiakkaan sessio voi olla monessa klusterissa. Tutkimalla tätä tarkemmin huomataan, että suurin osa (83,7 %) asiakkaiden sessioista on vain yhdessä klusterissa (kts. taulukko 10).

Taulukko 10: Muunnoksen 1.1 asiakkaiden sessioiden jakautuminen klustereittain

	Yhdessä klusterissa	Kahdessa klusterissa	Kolmessa klusterissa	Yhteensä
Sessioita	91 941	17 653	294	109 888
%-osuus	83,7 %	16,3 %	0,3 %	100 %

Tulkinta- ja arviointivaiheessa selvitettiin muuttujien järjestys klustereita eniten erottelevasta muuttujasta vähiten erottelevaan muuttujaan. Järjestys määriteltiin jokaisen muuttujan kohdalla yhteenlaskemalla klustereiden prototyyppien keskinäiset erotukset, kuten Jauhiainen (2017) oli yleistänyt pro gradussaan Saarelan ja Kärkkäisen (2015) ehdottaman lähestymistavan. Ensimmäiseksi klustereiden prototyypit järjestettiin eniten sitoutuneimmasta vähiten sitoutuneeseen siten, että jokaisen muuttujan arvot lajiteltiin nousevaan järjestykseen. Jokaisen muuttujan kohdalla laskettiin lajiteltujen prototyyppien väliset itseisarvolliset erotukset, jotka summattiin muodostaen muuttujan kokonaiserotuksen. Muuttujien kokonaiserotukset lajiteltiin laskevaan järjestykseen, jolloin ensimmäisenä on muuttuja, joka erotteli eniten klusterit toisistaan.

Muunnoksen 1.1 muuttujien järjestys klustereita eniten erottelevasta muuttujasta vähiten erottelevaan muuttujaan on kuviossa 12. Sessiossa ostoskoriin siirrettyjen eri tuotteiden määrät erottelevat eniten klusterit toisistaan ja session aikana katsottujen eri tuotteiden määrät erottelevat vähiten. Eniten erotteleva muuttuja erottelee huomattavasti juuri klusterin C2 muista klustereista.



Kuvio 12: Muunnoksen 1.1 muuttujien erottelevuus klustereittain

6.1.5 Tulokset: Muunnos 2.3

Muunnoksen 2.3, jossa sessioittain olevan datan *ostokori*-muuttuja skaalattiin TF-IDF menetelmällä, kohdalla löydettiin $K = 12$ klusteria. Taulukossa 11 on nähtävissä näiden klustereiden metadata. Klustereilla C4, C10, C11 ja C12 on huomattavasti pienemmät sessioiden kestot kuin muilla klustereilla. Session katseltujen eri tuotteiden lukumääriä verrattaessa näillä klustereilla on myös tuotteita vähemmän, lukuun ottamatta klusteria C10. Keston mukaan näitä klustereita voitaisiin pitää heikosti sitoutuneina, mutta tuotteiden määrät eivät tue sitä (kts. (Raphaeli, Goldstein ja Fink 2017)). Lisäksi klusterilla C12 on ostoskoriin siirrettyjä tuotteita ja myös ostoja, kun muilla saman ryhmän klustereilla ei ole. Tämä klusteriryhmä on kooltaan ja asiakasmäärältään isoin.

Huomattavaa on, että klusteri C2 koostuu kokonaan yhdestä asiakkaasta, kuten taulukosta 11 voi havaita. Tällä klusterilla keston mediaani on korkea, mutta moodi ei ole suuri. Näiden perusteella klusterin C2 kanssa samaan ryhmään voidaan jaotella klusterit C3, C5 ja C8. Tämä klusteriryhmä on kooltaan ja asiakas määrältään pienin. Ryhmää voisi juuri session keston ja tuotteen perusteella pitää keskivertoisesti sitoutuneina tai ne voi luokitella poikkeaviksi

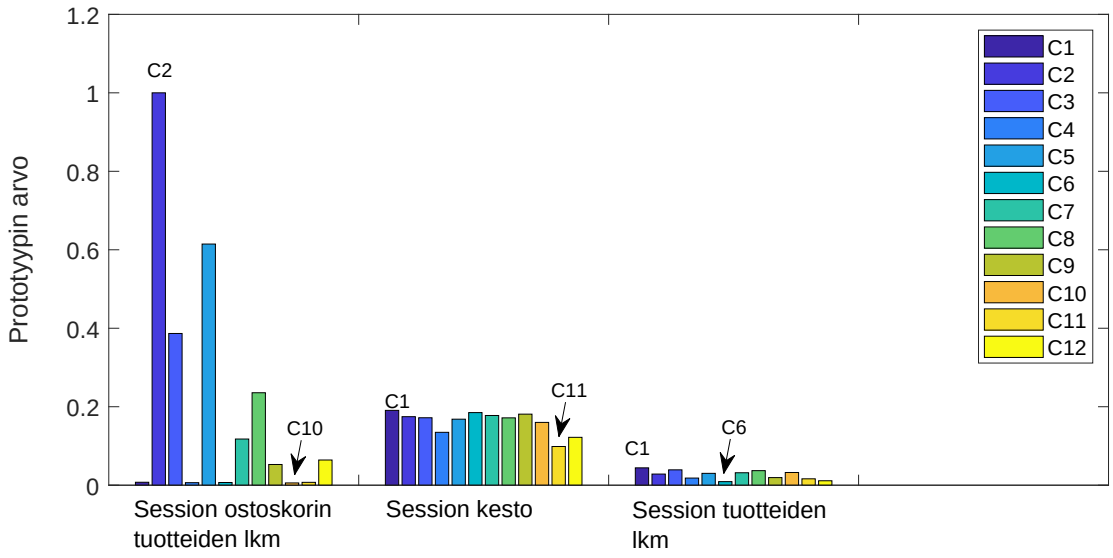
Taulukko 11: Muunnoksen 2.3 klustereiden metadata

	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Koko	171 789	29 067	317	1 003	33 418	397	27 353	6 754	2 370	13 630	34 258	17 453	5 769
Asiakkaita	109 888	24 497	1	42	29 758	5	23 259	4 436	595	11 164	31 692	15 419	4 706
Kesto M_{d} [mm:ss]	6:46	18:10	16:35	12:23	2:18	12:24	14:54	12:48	13:37	13:30	5:53	00:42	1:45
M_{o}	0:18	12:30	0:15	0:18	1:18	00:33	8:57	4:13	0:24	5:10	4:21	0:18	0:12
Max	30:00	30:00	29:59	30:00	5:17	29:56	30:00	30:00	30:00	30:00	11:32	1:14	5:10
Tuote M_{d}	2	5	2	4	2	3	1	3	3	2	3	2	1
M_{o}	2	3	2	2	2	2	1	2	2	1	3	2	1
Max	186	90	43	186	12	26	2	98	60	32	72	13	16
Korit	0	0	0	0	0	0	0	2	0	1	0	0	1
M_{o}	0	0	0	0	0	0	0	2	0	1	0	0	1
Max	55	1	0	31	0	14	0	5	55	1	0	1	4
Koria 1	10%	0%	0%	7%	0%	0%	0%	24%	7%	79%	0%	0%	71%
Koria > 2	3%	0%	0%	9%	0%	3%	0%	56%	33%	0%	0%	0%	8%
Ostoja 1	3%	0%	0%	0%	1%	0%	2%	9%	1%	28%	0%	0%	9%
Koreja 2 & Ostoja 1	0.22%	0%	0%	0.10%	0%	0%	0%	5.29%	0.34%	0%	0%	0%	0.35%

klustereiksi, koska asiakkaiden määrät ovat pieniä. Tällä ryhmällä on kuitenkin eniten klustereita, joilla on tuotteita siirretty ostoskoriin. Varsinkin klusterin C8 ostoskoriin siirrettyjen tuotteiden määrän osuus on toiseksi suurin.

Klusterit, joilla on korkea session keston mediaani ja myös moodi on suurempi, voidaan ajatella vahvasti sitoutuneiksi. Näitä ovat klusterit C1, C6, C7 ja C9. Klusteri C6 voidaan ajatella olevan vahvimmiten sitoutunut, koska sillä on korkea session kesto ja tuotteiden määrä on pieni. Klusterin asiakkaat eivät ole kuitenkaan siirtäneet tuotteita ostoskoriin ja ostoja on vähän. Jos vahvasti sitoutumisessa otetaan huomioon myös ostoskoriin siirretyt ja ostettu tuotteet, niin silloin klusteri C9 on vahvimmin sitoutunut. Lisäksi kyseisellä klusterilla on eniten ostoja. Klusterilla C7 on puolestaan eniten asiakkaita, jotka ovat siirtäneet tuotteita ostoskoriin. Klusterin C1 voidaan ajatella sisältävän asiakkaita, jotka eivät osaa päättää, mitä ostaisivat, koska session keston ja katsottujen tuotteiden määrän mukaan he selailevat verkkokaupan sivustoa.

Erottelevat muuttujien järjestys on sama kuin muunnoksen 1.1 kohdalla (kts. kuvio 13). Eniten erotteleva muuttuja on sessiossa ostoskoriin siirrettyjen eri tuotteiden lukumäärä. Muuttujan kohdalla klusterien välillä on selkeää eroa. Suurimmat prototyypin arvot eniten erottelevan muuttujan kohdalla saavat taulukon 11 perusteella ryhmitellyt keskivertoisesti sitoutuneet klusterit.



Kuvio 13: Muunnoksen 2.3 muuttujien erottelevuus klustereittain

6.1.6 Tulokset: Muunnos 3.1

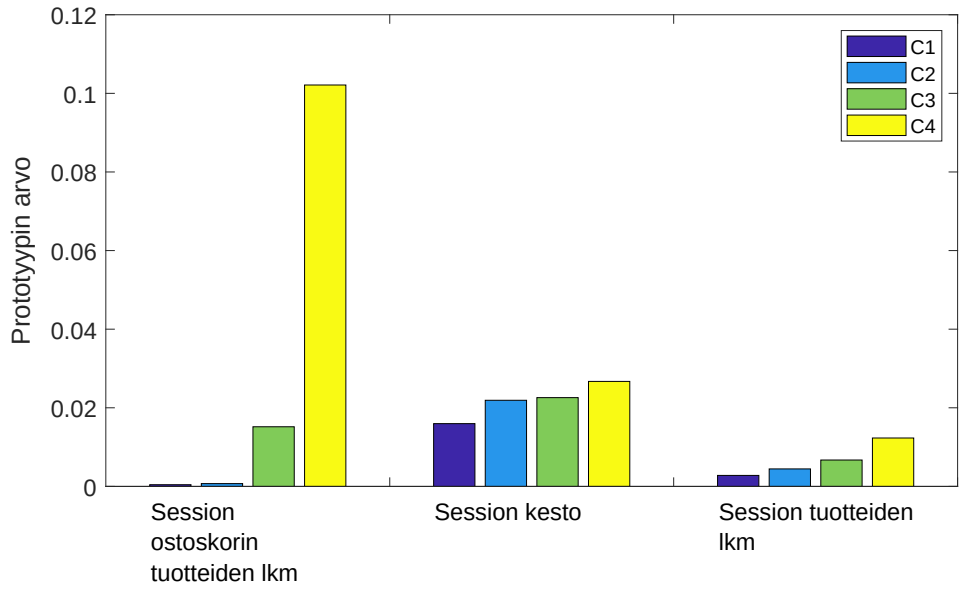
Muunnoksen 3.1 kohdalla, jossa jokaisen asiakkaan session tiedot oli laskettu yhteen, löydettiin neljä klusteria. Näistä kaksi klusteria olivat myös kooltaan huomattavasti isompia kuin kaksi muuta, joten klusterit klusteroitiin uudelleen hierarkkisesti. Ensimmäisen klusterin klusteroinnissa indeksit Ray-Turi ja DB ehdottivat klustereiden määräksi $K = 23$ ja toisen isoimman klusterin kohdalla Ray-Turi ehdotti $K = 16$ klusteria. Hierarkkisella prototyyppipohjaisella klusteroinnilla saavutettiin yhteensä 41 klusteria, joiden tulkinta ei ole mielekästä. Aliklustereiden tiedot eivät myöskään eronneet merkittävästi toisistaan. Liitteessä B on esitetty kuvio hierarkkisesta klusteroinnista ja taulukko klustereiden metadatasta.

Taulukosta 12 näkee, että muunnoksessa 3.1 on myös niin sanottu päättämättömien klusteri C4. Tällä klusterilla on korkein session kesto ja tuotteita on katsottu eniten. Toisaalta klusterin asiakkaat ovat siirtäneet enemmän kuin kaksi tuotetta ostoskoriin ja klusterin C3 kanssa sillä on eniten ostoja. Klusterilla C2 on korkein session keston moodi ja sen asiakkaat ovat katsoneet tuotteita vähemmän kuin klusterit, joilla session kesto on korkeampi. Näiden perusteella klusteri C2 voidaan pitää vahvasti sitoutuneena ja se on kooltaan isoin klusteri. Sillä on myös toiseksi eniten ostoja, joka tukee vahvasti sitoutumista. Toiseksi eniten asiakkaita on klusterilla C1, jolla session kesto on todella pieni ja ostokoriin tehtyjä siirtoja ja ostoja on vain hieman.

Taulukko 12: Muunnoksen 3.1 klustereiden metadata

	Data	C1	C2	C3	C4
Koko	109 888	45 720	61 841	2 264	63
Asiakkaita	109 888	45 720	61 841	2 264	63
Kesto M_{d} [[t]:mm:ss]	9:34	3:09	17:02	23:43	1:30:16
M_{o}	1:13	0:58	9:48	5:43	2:41
Max	83:04:13	9:33	83:04:13	46:10:13	12:00:14
Tuote M_{d}	3	3	4	8	42
M_{o}	2	2	4	4	20
Max	1 864	26	1 864	1 350	292
Ostoskori M_{d}	0	0	0	4	25
M_{o}	0	0	0	3	21
Max	118	4	2	18	118
Tuotteita ostoskorissa 1 kpl	13%	10%	16%	0%	0%
Tuotteita ostoskorissa > 2 kpl	5%	2%	3%	100%	100%
Ostoja 1 kpl	5%	3%	7%	9%	5%
Tuotteita ostoskorissa 2 kpl, Ostoja 1 kpl	0.74%	0.17%	0.84%	9.36%	4.76%

Sessioiden tietojen yhteenlaskiminen ei vaikuttanut erottelevien muuttujien järjestykseen. Kuvio 14 mukailee taulukon 12 tietoja. Kuviosta huomaa, miten klusterin C4 sessiossa ostoskoriin siirrettyjen tuotteiden lukumäärät eroavat selvästi muista.



Kuvio 14: Muunnoksen 3.1 muuttujien erottelevuus klustereittain

6.1.7 Tulokset: Muunnos 5.2

Sessioiden tietojen yhteenlaskeminen ja *ostoskori*-muuttujan skaalaaminen TF-IDF:llä antoi indeksien kuvaajien perusteella klustereiden määräksi $K = 12$. Vaikka osalla klustereissa on korkeampi kesto kuin muilla, niin näistä klustereista ei voi sanoa, että niiden asiakkaat ovat vahvasti sitoutuneita, koska näiden klustereiden eri tuotteiden katsomisien määrät ovat myös suuria ja ostokoriin siirrettyjen tuotteiden määrät vaihtelevat. Osalla näistä klustereista ostojen prosenttiosuus on myös suuri.

Taulukosta 13 näkee, että klustereiden C3, C4, C5 ja C9 sessioiden kestot vaihtelevat, jolloin kestojen ja katseltujen tuotteiden määrän puolesta ne on mahdollista luokitella sitoutumisasteen mukaisesti. Sen sijaan huomioitaessa ostoskoriin siirrettyjen tuotteiden ja ostojen prosenttiosuudet, näitä klustereita voitaisiin pitää saman asteisesti sitoutuneina, vaikka sessioiden kestot ovat erilaiset.

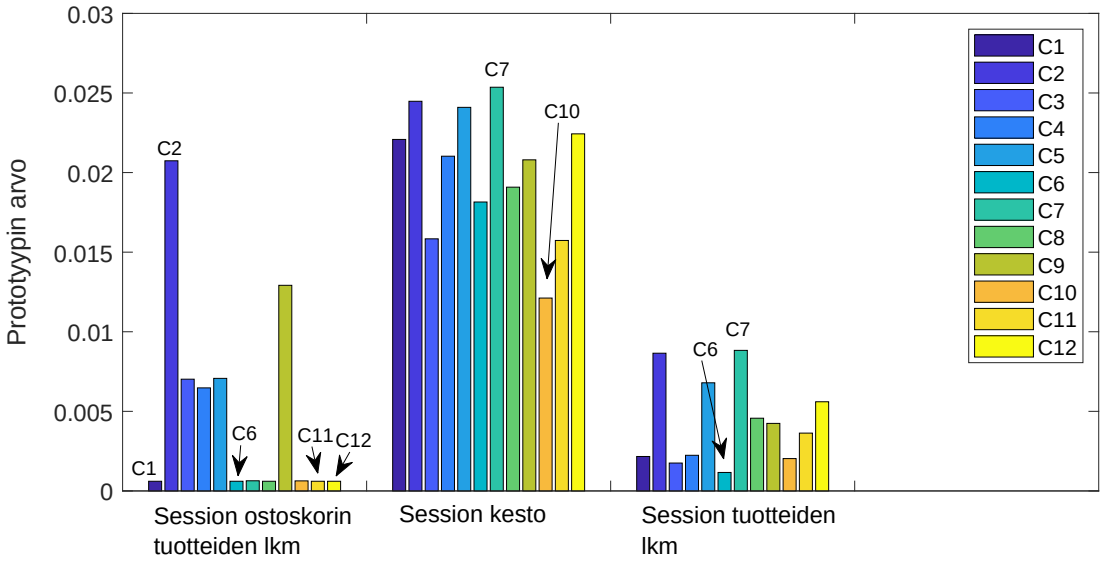
Klusterilla C5 on kolmanneksi korkein kesto, mutta pienempi tuotteiden mediaani kuin muilla kestoiltaan korkeilla klustereilla (C2, C7). Klusterilla C5 on myös ostoja ja ostoskoriin siirrettyjä tuotteita, jolloin se voidaan nähdä vahvemmin sitoutuneeksi. Toisaalta huomioitaessa

Taulukko 13: Muunnoksen 5.2 klustereiden metadata

	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Koko	109 888	14 426	1 197	3 730	7 394	3 679	9 948	4 554	20 680	3 276	8 268	15 917	16 819
Asiakkaita	109 888	14 426	1 197	3 730	7 394	3 679	9 948	4 554	20 680	3 276	8 268	15 917	16 819
Kesto M_d [[t]:mm:ss]	9:34	18:37	42:28	3:01	13:10	34:09	5:30	48:47	7:16	13:40	0:53	2:33	20:32
M_o	1:13	12:00	22:06	2:52	7:36	25:20	3:30	41:37	4:36	5:37	0:58	1:52	14:33
Max	83:04:13	1:12:38	46:10:13	6:08	56:47	25:57:42	11:26	83:04:13	13:10	1:31:40	1:54	4:28	1:23:42
Tuote M_d	3	2	15	2	2	8	2	15	4	4	2	3	6
M_o	2	2	8	1	1	5	2	12	4	3	2	3	4
Max	1 864	3	785	16	10	1 350	2	1 864	32	26	12	26	28
Korit M_d	0	0	6	1	1	1	0	0	0	2	0	0	0
M_o	0	0	5	1	1	1	0	0	0	2	0	0	0
Max	118	0	118	4	1	2	0	1	0	11	1	0	0
Koria 1	13%	0%	0%	89%	100%	88%	0%	1%	0%	0%	0%	0%	0%
Koria > 2	5%	0%	100%	11%	0%	12%	0%	0%	0%	100%	0%	0%	0%
Ostoja 1	5%	1%	9%	16%	40%	33%	1%	1%	0%	17%	0%	0%	1%
Koreja 2 & Ostoja 1	0.74%	0%	9.44%	0.97%	0%	3.23%	0%	0%	0%	16.58%	0%	0%	0%

Raphaelin, Goldsteinin ja Finkin (2017) tutkimus, klusterin C1 asiakkaiden sitoutuminen on vahvinta, koska sillä on vähemmän katseltuja tuotteita, vaikka sen sessioiden kesto on pienempi. Klusterilla ei kuitenkaan ole ostoja tai ostoskoriin siirrettyjä tuotteita. Klusteri C7 sisältää tietojensa perusteella sellaisia asiakkaita, jotka eivät osaa tehdä ostopäätöstä.

Klustereita erottelevien muuttujien järjestys ei muutu, kun *ostoskori*-muuttuja skaalataan TF-IDF:llä (kts. kuvio 15). Tässäkin session ostokorin tuotteiden lukumäärän kohdalla kaksi klusteria poikkeavat huomattavasti toisista. Huomioitavaa on se, että ostokorin tuotteiden lukumäärän kohdalla on neljä klusteria, joilla prototyypin arvot ovat samat (ei siirtoja ostoskoriin). Tällöin muut kuin session kestot ja katseltujen tuotteiden määrät erottelevat klusterit toisistaan. Jolloin näiden klustereiden sitoutumisasteen voidaan ajatella määräävän session kesto ja session tuotteiden lukumäärä.



Kuvio 15: Muunnoksen 5.2 muuttujien erottelevuus klustereittain

6.1.8 Tulokset: Muunnos 6.2

Huomioitaessa tiedonlouhinnassa sessioiden lukumäärämuuttuja, validointi-indeksien kuvaajien perusteella saatiin muodostettua viisi klusteria, joiden metadata on taulukossa 14. Taulukosta huomataan, että näiden klustereiden kohdalla toistuu selkeämmin se, että keston ollessa korkea myös tuotteiden määrät ovat korkeat. Lisäksi sessioiden lukumäärä on suuri

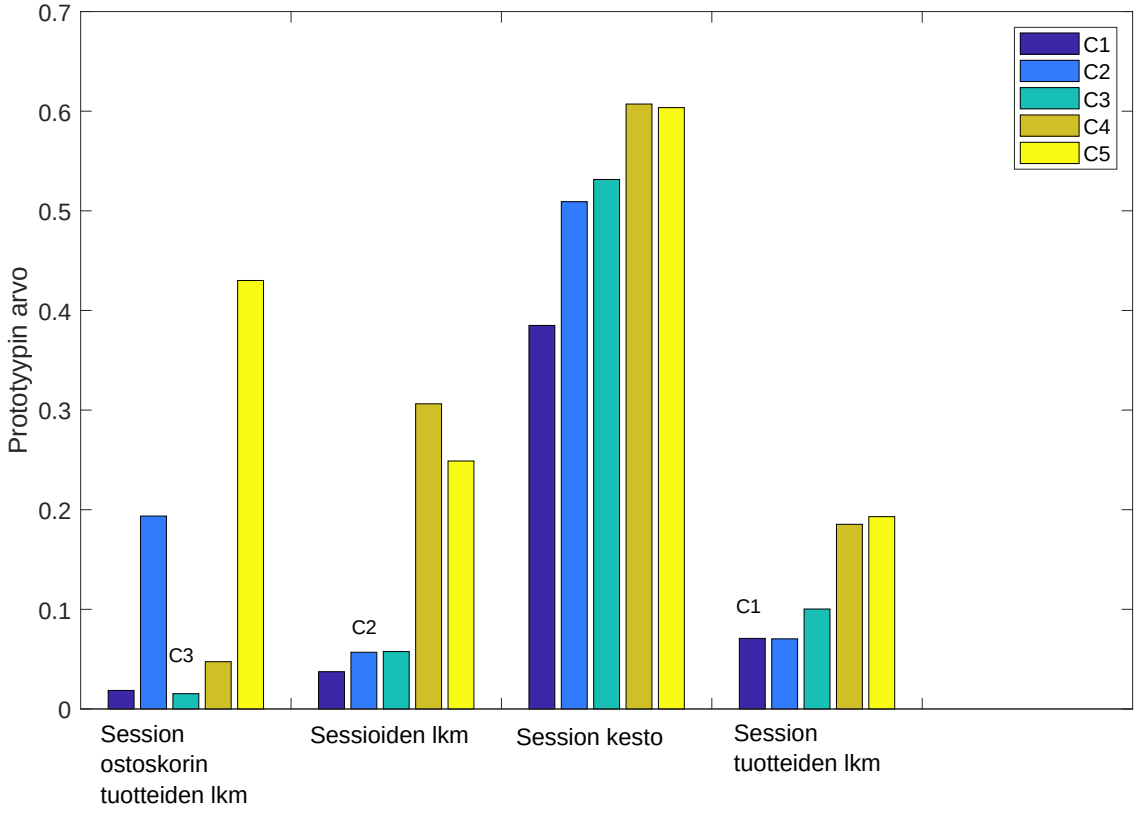
niillä klustereilla, joilla kestot ja tuotteet ovat suuria. Ostoskorin kohdalla ei ole havaittavissa näin selkeää jakaumaa. Raphaelin, Goldsteinin ja Finkin (2017) mukaisesti keston ja tuotteiden katselujen perusteella klustereita ei voi muodostaa sitoutumisasteen mukaisesti.

Taulukko 14: Muunnoksen 6.2 klustereiden metadata

	Data	C1	C2	C3	C4	C5
Koko	109 888	34 163	14 302	50 216	8 832	2 375
Asiakkaita	109 888	34 163	14 302	50 216	8 832	2 375
Kesto M_d [[t]:mm:ss]	9:07	2:30	11:20	13:27	37:12	35:34
M_o	1:13	0:58	2:52	6:21	14:44	16:43
Max	83:04:13	6:54	58:22	59:33	83:04:13	46:10:13
Tuote M_d	3	3	2	4	9	10
M_o	2	2	2	4	6	6
Max	1 864	26	49	61	1 864	785
Korit M_d	0	0	1	0	0	4
M_o	0	0	1	0	0	2
Max	118	1	7	0	2	118
Sessioiden lkm M_d	1	1	1	1	3	3
Max	318	3	3	3	318	217
Korit 1	13%	2%	80%	0%	21%	0%
Korit > 2	5%	0%	20%	0%	0%	100%
Ostoja 1	5%	0%	30%	1%	9%	16%
Korit 2 & Ostoja 1	0.74%	0%	2.93%	0%	0.06%	16.29%

Klusteria C2 voitaisiin pitää vahvasti sitoutuneena, koska sillä ei ole pienin kesto ja tuotteiden ja sessioiden määrä on myös pieni. Lisäksi sillä on paljon ostoskoriin siirrettyjä tuotteita ja eniten ostoja. Klusterin C4 voitaisiin ajatella sisältävän asiakkaita, jotka eivät osaa tehdä päätöksiä korkean keston ja tuotteiden sekä sessioiden suurien lukumäärien perusteella. Toisaalta osa näistä asiakkaista on ostanut tuotteita, mutta osuus on pienempi kuin ostoskoriin siirrettyjen osuus.

Erottelevia muuttujia tarkastellessa sessioiden lukumäärä on toiseksi eniten erotteleva muuttuja (kts. kuvio 16). Tällöin asiakkaiden sessioiden määrissä on suuria eroja, mutta kuitenkin ostoskoriin siirrettyjen tuotteiden lukumäärä erottelee eniten asiakkaat toisistaan. *Kesto* ja *tuotteet* -muuttujien järjestys on sama kuin aikaisempien muutoksien kohdalla.



Kuvio 16: Muunnoksen 6.2 muuttujien erottelevuus klustereittain

6.2 Asiakkuuden kehittäminen ja sosiaalisen median data

Asiakkuuden kehittämisen vaiheessa asiakkaan sitoutumisen tutkimiseen käytettiin Kagglesivustolta saatua Cheltenhamin julkisen yhteisön Facebookin julkaisuja vuoden 2008 heinäkuun ja vuoden 2016 kesäkuun väliseltä ajalta (Retailrocket 2019). Datajoukko koostui neljästä tiedostosta, joita jokaista hyödynnettiin tutkielmassa. Jokainen tiedosto sisälsi käyttäjien id:t ja tiedostot oli yhdistetty toisiinsa julkaisujen id:itten ja kommenttien id:itten avulla. Datassa oli myös julkaisujen ja kommenttien sisällöt sekä henkilöiden nimet. Käytetyn da-

tan julkaisut käyttäjittäin mahdollisti sitoutumisen tutkimisen henkilöittäin. Muiden avointen tietovarastojen sosiaalisen median datat olivat julkaisuittain, mikä olisi mahdollistanut tutkimaan vain, millaiset julkaisut sitouttavat käyttäjiä.

6.2.1 Esikäsittely

Esikäsittelyvaiheessa jokaisesta datajoukon tiedostoista muodostettiin MATLAB:n (R2018a) table-taulukot, koska datajoukot sisälsivät sekä numeerista että tekstidataa. Table-taulukoista tarvittavat tiedot laskettiin matriisiin, jolla suoritettiin klusterointi. Post-taulukon Tykkäykset-sarakkeessa oli tykkäyksien lukumäärät tekstimuotoisena, joten sarake muutettiin reaaliluvuksi. Sarakkeen tyhjät rivit, joissa käyttäjällä ei ollut tykkäyksiä, korvattiin nollalla. Samanlainen muutos tehtiin Like-taulukon kommenttien id:ille. Lisäksi Comment-taulukon käyttäjien, jotka olivat kommentoineet ensimmäisenä julkaisua, id:itten NaN-rivit muutettiin nollaksi.

Esikäsittelyn aikana huomattiin, että Post-, Comment- ja Like-taulukoissa on sellaisia id:eitä, joita ei ole Member-taulukossa. Tällöin lopulliseen matriisiin käyttäjien id:eitä ei voitu ottaa suoraan Member-taulukosta, vaan id:t muodostettiin Post-, Comment- ja Like -taulukoiden yksilöllisistä id:eistä. Lisäksi näistä kolmesta taulukosta laskettiin jokaiselle käyttäjälle julkaisujen, kommenttien ja tykkäyksien lukumäärät. Ryanin (2017) mukaan sosiaalisen median seurannassa on oleellista myös käyttäjän viipymisaika sivustolla. Täten käyttäjittäin laskettiin, kuinka monena eri päivänä käyttäjä on luonut julkaisuja tai kommentoinut.

Alkuperäisen datan kuvauksen mukaan Comment-tiedoston rid-sarake kertoo käyttäjän, joka on kommentoinut kommenttia. Id:t eivät kuitenkaan vastanneet toisten tiedostojen (Post-, Comment- ja Member) id:eitä, joten datan perusteella ei ollut mahdollista muodostaa lukumäärää siitä, kuinka moneen kommenttiin käyttäjä on kommentoinut. Datan perusteella pääteltiin, että rid-sarakkeen nolla (alkuperäisessä tiedostossa NaN) merkitsee ensimmäistä kommenttia julkaisuun. Tästä laskettiin julkaisuihin tehtyjen kommenttien lukumäärää käyttäjittäin.

Doornin ym. (2010) mukaan asiakkaan sitoutumisasteeseen vaikuttaa asiakkaan kokemukset brändistä. Like-taulukosta laskettiin erilliseen Like-tietorakenteeseen (engl. struct) tulkintaa

varten julkaisuihin ja kommentteihin tehtyjen eri reaktioiden (vihainen, haha, tykkää, tykkää (kommenttien), ihastu, surullinen ja vau) lukumäärät. Koska yhteisön toimintaan olivat osallistuneet muutkin käyttäjät kuin jäsenet (Member-tiedostossa olevat), niin esikäsittelyssä muodostettiin dikotominen muuttuja kuvaamaan jäsenyyttä.

Alkuperäisen datan kuvauksesta ilmeni, että käyttäjän nimi on käyttäjän pysyvä yksilöijä eikä käyttäjän id. Facebook vaihtaa käyttäjän id:een, jos käyttäjä vaihtaa profiilikuvansa, mutta käyttäjän nimi säilyy koko ajan samana. Lopullisessa matriisissa muutamien käyttäjien tietoja oli useammalla rivillä. Näiden käyttäjien tiedot laskettiin yhteen ja id:eksi laitettiin ainoastaan käyttäjän yksi id. Yhteenlaskun seurauksena käyttäjien id:eitten määrä pieneni yhdeksällä.

Taulukoista laskettiin myös, montako kommenttia käyttäjän kommentti on saanut. Lisäksi laskettiin käyttäjän julkaisujen tykkäyksien, jakojen, kommenttien ja näiden kommenttien määrät. Lasketuista muuttujista ja metadatoista sekä käyttäjän yksilöivästä id:stä muodostettiin datamatriisi taulukon 15 mukaisesti. Lisäksi matriisista tarkastettiin sisältääkö se kokonaan nollarivejä (ei metadatan kohdalla) eli käyttäjiä, joilla ei olisi mitään sitoutumistoimintaa (julkaisuja, kommentteja ja tykkäyksiä). Tällaisia ei ollut, koska id:t muodostettiin Post-, Comment-, Like -taulukoista. Matriisin muuttujat on kuvattu tarkemmin taulukkoon 16. Matriisin viimeiset sarakkeet sisältävät metadataa. Metadataa oli myös erillisenä Like-tietorakenteena olevat reaktioiden määrät.

Taulukko 15: Asiakkuuden kehittämisen datamatriisi

Julkaisut	Kommentit	Julkaisujen tykkäykset	…	Jäsen	ID
0	0	1	…	0	1012
1	4	6	…	0	1021
⋮	⋮	⋮	…	⋮	⋮
0	0	1	…	0	1939
0	2	8	…	0	5033

Taulukko 16: Asiakkuuden kehittämisen datan muuttujat ja metadata

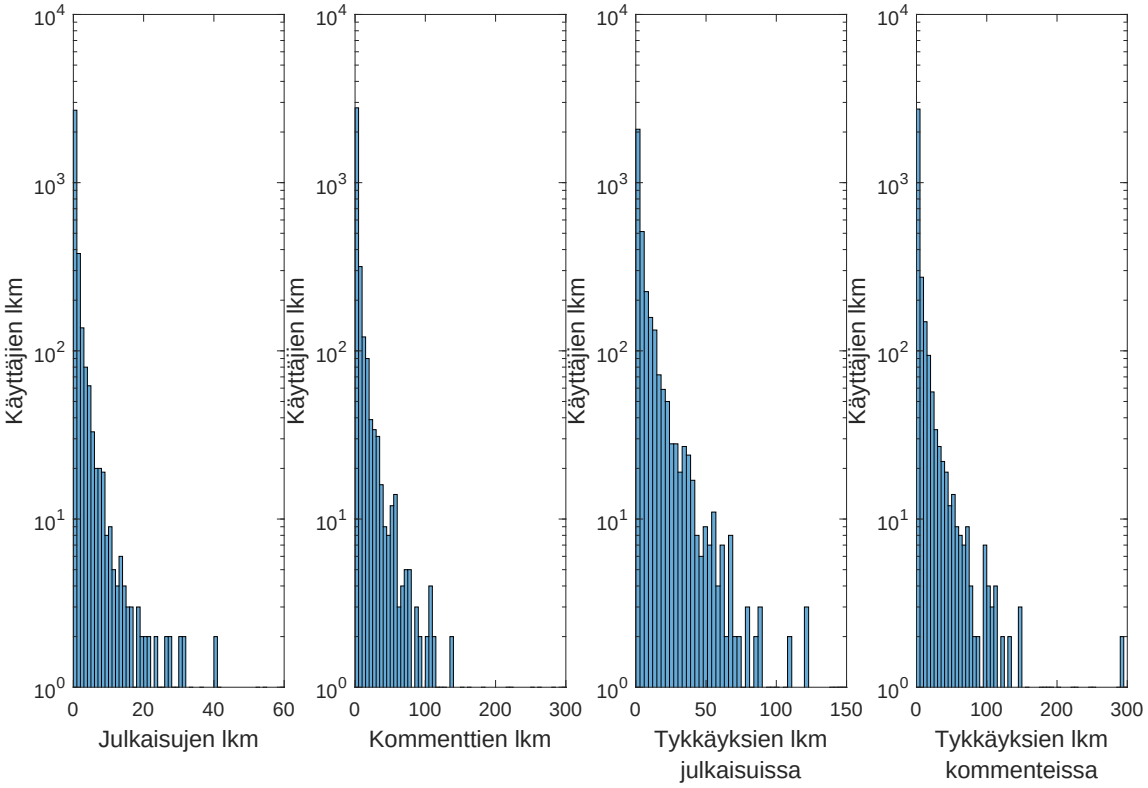
Muuttuja	Kuvaus
Julkaisut	Käyttäjän julkaisujen lukumäärä
Kommentit	Käyttäjän kommenttien lukumäärä
Julkaisujen tykkäykset	Käyttäjän tykkäämien julkaisujen lukumäärä
Kommenttien tykkäykset	Käyttäjän tykkäämien kommenttien lukumäärä
Julkaisujen pvm lkm	Käyttäjän tehtyjen julkaisujen päivien lukumäärä
Kommenttien pvm lkm	Käyttäjän tehtyjen kommenttien päivien lukumäärä
Kommenttien kommentit	*Metadata*: Käyttäjän kommentteihin tehdyt kommentit
Julkaisujen tykkäykset	*Metadata*: Käyttäjän julkaisujen tykkäyksien lukumäärä
Julkaisujen jaot	*Metadata*: Käyttäjän julkaisujen jakojen lukumäärä
Julkaisujen kommentit	*Metadata*: Käyttäjän julkaisujen kommenttien määrä
Kommenttien kommentit	*Metadata*: Kommenttiin tehtyjen kommenttien määrä
Jäsen	*Metadata*: Kuvaa käyttäjän id:een löytymistä Member-taulukosta (1 jäsen, 0 ei jäsen)
ID	*Metadata*: id käyttäjän yksilöintiin

6.2.2 Muunnos

Sitoutumista tutkittiin käyttäjittäin hyödyntämällä esikäsittelyn aikana muodostettua matriisia, jonka rivit muodostuivat jokaisesta yksittäisestä käyttäjästä. Muuttujina käytettiin matriisin kuutta muuttujaa (kts. taulukon 16 ensimmäiset muuttujat). Muuttujien valinta tehtiin aikaisempien tutkimuksien pohjalta, joissa sitoutumista oli tutkittu brändin julkaisujen tykkäyksien, kommenttien ja jakojen määrällä sekä jossain määrin vuorovaikutuksen kestolla (Cvijikj ja Michahelles 2013; He, Zha ja Li 2013; Schultz 2017; Moro ym. 2016). Muunnoksen data sisälsi 3 523 havaintoa ja 6 muuttujaa. Lisäksi datalle tehtiin min-max-skaalaus välille $[0, 1]$ (muunnos 1).

Datan jokainen muuttuja oli oikealle vinoutunut, joten $[0, 1]$ välille skaalatulle datalle tehtiin logaritminen muunnos, jotta datasta tulisi enemmän symmetrinen (muunnos 1.1). Kuviosta 17 huomaa, miten vinoutuneita neljä ensimmäistä muuttujaa ovat. Kuvaajien häntä

ulottuu pitkälle oikealle, koska jokainen muuttuja sisältää yksittäisiä käyttäjiä, jotka ovat tehneet paljon sitoutumistoimintaa. Visualisointia varten kuvion 17 kuvaajien häntiä on katkaistu pienemmäksi ja pystyakselille on tehty logaritmiskaalaus. Päivien lukumäärää kuvaavat muuttujat olivat yhtä vinoutuneita.



Kuvio 17: Käyttäjien lukumäärien jakauma sitoutumistoiminnoittain

Kuten kuviosta 17 voi havaita, niin data sisältää pienen osan sellaisia käyttäjiä, jotka ovat tehneet paljon sitoutumistoimintaa. Nämä käyttäjät voidaan nähdä joko voimakkaasti sitoutuneina käyttäjinä tai markkinointimielessä toimivina. Tutkimalla tarkemmin eniten julkaisuja tehneen käyttäjän julkaisujen sisältöä, huomattiin hänen julkaisunsa koskevan erilaisten tapahtumien ilmoituksia. Käyttäjä voi näin ollen toimia Facebook-sivustolla jonkun julkisen tai yksityisen yhteisön edustajana.

Paljon sitoutumistoimintaa tehneet käyttäjät jätettiin huomioimatta, joten datasta poistettiin ensin yli 15 julkaisua tehneet käyttäjät. Poiston jälkeen datassa oli vielä joitain käyttäjiä, joiden kommenttien määrät olivat suuria. Toisessa poistossa datasta otettiin yli 60 kommenttia

tehneet käyttäjät. Datassa oli vielä pieniä määriä yksittäisiä käyttäjiä, jotka olivat tykänneet julkaisuista tai kommenteista paljon. Täten poistettiin vielä käyttäjät, joilla oli yli 70 julkaisujen tykkäystä ja yli 90 kommenttien tykkäystä. Paljon sitoutumistoimintaa tehneiden käyttäjien poiston jälkeen data sisälsi 3 426 havaintoja ja 6 muuttujaa. Lisäksi datan arvot skaalattiin välille $[0,1]$ (muunnos 2). Koska data oli edelleen vinoutunut, niin datalle tehtiin logaritminen muunnos (muunnos 2.1).

Kaikissa aikaisemmissa tutkimuksissa sitoutumisessa ei oltu huomioitu vuorovaikutuksen kestoa, joten selvitettiin, millainen vaikutus julkaisujen ja kommenttien päivien lukumäärä muuttujien (muunnokset 1 ja 2) poistolla on. Muuttujien poistamisen jälkeen muunnoksen 1.3 data sisälsi 3 523 havaintoja ja 4 muuttujaa, ja muunnoksen 2.3 jälkeen 3 426 havaintoja ja 4 muuttujaa. Näille muunnoksille tehtiin lisäksi logaritmiset muunnokset (muunnokset 1.4 ja 2.4). Kaiken kaikkiaan asiakkuuden kehittämisen vaiheen datalle tehtiin kahdeksan muunnosta.

6.2.3 Tiedonlouhinta

Asiakkuuden kehittämisen vaiheen sitoutumisen tiedonlouhinnassa käytettiin myös MATLAB:in k-means++ algoritmia ja sen oletuksena käyttämää neliöllistä euklidista etäisyyttä. Klusterointi ja validointi-indeksien testaaminen puolestaan suoritettiin klusterien määrällä $K = 2-20$. Klustereiden optimaalisen lukumäärän valinnassa hyödynnettiin kahdeksaa klusterien validointi-indeksejä, joista seitsemän oli samoja indeksejä kuin asiakashankinnan vaiheessa (kts. 6.1.3). Koska datan havaintojen määrä oli huomattavasti pienempi kuin asiakashankinnan vaiheessa, pystyttiin hyödyntämään myös Silhouette-indeksiä. Silhouette-indeksin laskennassa hyödynnettiin MATLAB:in evalclusters-funktiota, jossa parametrina annettiin indeksin nimi.

Muunnoksen 1 kohdalla validointi-indeksit ehdottivat optimaalisiksi klustereiksi $K = 2$ tai $K = 20$ ja PBM ehdotti klusteriksi $K = 4$, joten klusterien määrää ei voinut todeta. Klusterin määrää ei voinut myöskään selvästi havaita indeksien arvoista muodostettujen käyrien kuvaajista. Logaritmimuunnoksen jälkeen (muunnos 1.1) WG-, Silhouette- ja Ray-Turi-indeksien arvojen perusteella klustereiden määräksi valittiin $K = 9$.

Muunnoksen 2 kohdalla validointi-indeksien arvot antoivat klustereiksi $K = 2$ lukuun ottamatta WB- ja PBM-indeksejä ($K = 19$ ja $K = 6$) eikä käyrien kuvaajista voinut selvästi havaita klusterien määrää. Logaritminen muunnos (muunnos 2.1) vaikutti validointi-indeksien tulokseen siten, että myös WG-, Silhouette- ja Ray-Turi-indeksien arvojen optimaalinen klusterin määrä oli $K = 10$ eli yksi klusteri enemmän kuin datasta, josta ei oltu poistettu yhtään käyttäjää.

Julkaisujen ja kommenttien päivien lukumäärän jättäminen pois klusteroinnista ei vaikuttanut siihen, että validointi-indeksien perusteella voisi päätellä klusterien määrän datasta, joka oli vain skaalattu (muunnokset 1.3 ja 2.3). Näiden kahden muuttujan jättäminen ja logaritminen muunnos vaikuttivat koko datan (muunnos 1.4) kohdalla siihen, että WG-, Silhouette- ja Ray-Turi-indeksit ehdottivat klustereiden määräksi yhden enemmän kuin muunnoksessa 1.1 eli $K = 10$.

Muunnoksen 2.4 kohdalla, josta oli poistettu paljon sitoutumistoimintoja tekevät käyttäjät ja vuorovaikutuksen keston muuttujat, samat validointi-indeksit antoivat klustereiden määräksi saman kuin muunnoksen 2.1 kohdalla ($K = 10$), vaikka päivien lukumäärä muuttujien hännät lyhenivät samoin kuin *julkaisu*- ja *kommentti* -muuttujien kohdalla. 2.4 muunnosta ei nähty tarpeelliseksi käsitellä, koska 1.4 käsiteltiin ja 2.1 muunnos ei poikennut paljoa 1.1 muunnoksesta. Näin ollen vain kolme muunnosta käsiteltiin tuloksissa: muunnos 1.1, muunnos 1.4 ja muunnos 2.1.

6.2.4 Tulokset: Muunnos 1.1

Muunnoksen 1.1, joka käsitteli kaikkia käyttäjiä ja muuttujia, kohdalla löydettiin yhdeksän klusteria. Näiden yhdeksän klusterin metadata on nähtävissä taulukosta 17. Taulukon sarakkeet koostuvat koko datasta ja yhdeksästä klusterista (merkitty C-kirjaimella ja numerolla). Taulukon riveillä on koko datan ja klustereiden koot, jokaisen muuttujan sitoutumistoimintojen lukumäärän mediaani (M_{d}) ja muuttujan maksimilukumäärä (Max). Lisäksi taulukossa on Jäseniä-rivi, joka kuvaa kuinka monta prosenttia käyttäjistä on jäseniä (Member-tiedostossa olevat) datassa ja klustereissa.

Taulukko 17: Muunnoksen 1.1 klustereiden metadata

	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9
Koko	3523	124	1173	223	560	536	263	167	287	190
Jäseniä	71%	90%	43%	96%	96%	96%	28%	91%	80%	92%
Julkaisut M_{d}	0	1	0	0	0	2	0	1	0	0
Max	134	20	0	0	0	134	0	30	0	0
Kommentit M_{d}	0	0	0	1	3	11	0	3	0	1
Max	393	0	0	18	137	393	0	53	0	25
Tykkäykset julkaisut M_{d}	2	1	1	3	7	12	0	1	2	0
Max	258	48	40	67	120	258	0	27	66	0
Tykkäykset kommentit M_{d}	1	0	0	0	5	13	1	0	2	0
Max	737	19	0	0	109	737	8	15	37	16
Julkaisut pvm M_{d}	0	1	0	0	0	2	0	1	0	0
Max	129	20	0	0	0	129	0	30	0	0
Kommentit pvm M_{d}	0	0	0	1	3	8	0	2	0	1
Max	210	0	0	14	90	210	0	42	0	20

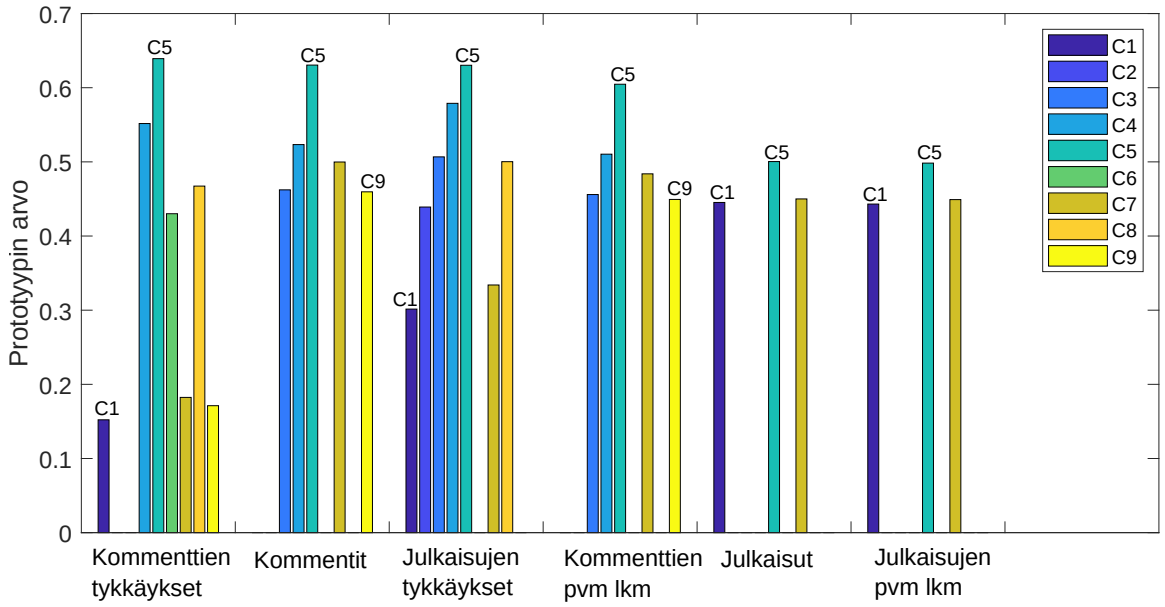
Klusteri C2 on kooltaan melkein kolmasosa koko datasta. Tätä klusteria ei kannattanut klusteroida lisää hierarkkisesti Wartiaisen ja Kärkkäisen (2015) tutkimuksen mukaisesti, sillä klusterin käyttäjien sitoutumistoimintaa oli ainoastaan julkaisujen tykkäykset. Näin ollen klusterin C2 käyttäjät ovat heikoiten sitoutuneita, koska sisällöstä tykkääminen on Malthousen ym. (2013) mukaan alhaisen asteen sitoutumista. Julkaisun tykkäämistä voidaan pitää vielä vähempiarvoisena tykkäämisenä kuin kommenttien, koska kommenttien tykkäämisessä vuorovaikutuksen kesto on pitempi.

Lisäksi klusterissa C2 on toiseksi vähiten käyttäjiä, jotka ovat jäseniä. Kun klusteria katsoi tarkemmin, niin sen jokainen käyttäjä on tykännyt ainakin mediaanin verran eli yhden kerran. Toisaalta klusterissa on ainakin yksi sellainen käyttäjä, joka on tykännyt 40 kertaa, joka on vain toiseksi pienin maksimiarvo klustereiden julkaisujen tykkäyksissä. Voidaan kuiten-

kin olettaa, että vähäisesti sitoutuneet eivät seuraa aktiivisesti sivustoa ja eivätkä ole jäseniä, koska klusterin C2 kohdalla jäsenien määrä oli pieni. Tätä tukee myös se, että klusteri C6 on jäsenmäärältä pienin ja sen käyttäjien sitoutumistoimintona on vain kommenttien tykkääminen.

Muunnoksen 1.1. kohdalla muuttujien järjestys klustereita eniten erottelevasta muuttujasta vähiten erottelevaan muuttujaan on nähtävissä kuviosta 18. *Kommenttien tykkäykset* -muuttuja erottelee eniten klusterit toisistaan ja puolestaan *julkaisujen pvm lkm* -muuttuja vähiten. Jokaista muuttujaa esiintyy enemmän klusterissa C5, jolloin klusterissa C5 olevat henkilöt ovat eniten sitoutuneita. Saman voi havaita taulukosta 17, koska jokaisen muuttujan mediaanit ovat suurempia klusterilla C5. Lisäksi maksimiarvot ovat samat koko datan kanssa.

Klusteri C5 ei kuitenkaan eroa huomattavasti muista klustereista kuvion 18 mukaan, kun katsotaan niitä klustereita, joilla esiintyy tarkasteltavia sitoutumistoimintoja. Ainoastaan tykkäyksien kohdalla ero on selkeämpi C1, C7 ja C9 klustereihin, jolloin kommenttien ja julkaisujen tykkäykset erottelevat klusterin C5 näistä klustereista. *Kommentit*-muuttujan kohdalla klusteri C5 ei eroa huomattavasti klusterista C9. Klusterin C9 käyttäjiä voidaankin pitää keskivertoisesti sitoutuneina.



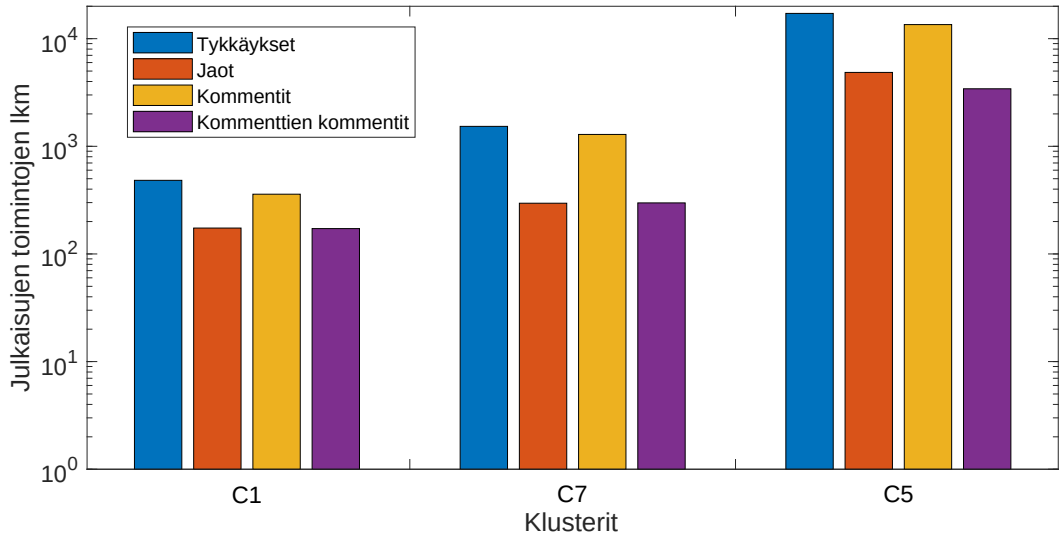
Kuvio 18: Muunnoksen 1.1 muuttujien erottelevuus klustereittain

Taulukon 17 ja kuvion 18 perusteella klusterit voidaan järjestää heikosti sitoutuneista vahvasti sitoutuneisiin julkaisujen, kommenttien ja tykkäyksien perusteella. Yhdeksästä klusterista voidaan järjestyksen perusteella muodostaa kolme sitoutumisasteen ryhmää taulukon 18 mukaisesti. Taulukossa M_{d}-sarake kuvaa sitoutumistoiminta-sarakkeen eli eri muuttujien mediaania. Näiden ryhmien mukaan suurin osa, melkein puolet yhteisön käyttäjistä on vähäisesti sitoutuneita. Käyttäjistä vain pieni määrä on voimakkaasti sitoutuneita. Toisaalta klusteri C4 voitaisiin luokitella vahvasti sitoutuneisiin sen perusteella, että siitä puuttuu vain yhtä muuttujaa. Tällöinkin vahvasti sitoutuneita olisi vähemmän kuin heikosti sitoutuneita.

Taulukko 18: Muunnoksen 1.1 klusterit sitoutumisasteen mukaisessa järjestyksessä

	Klusteri	Koko	Sitoutumistoiminta	M_{d}	Jäseniä
Heikko	C2	1173	Julkaisujen tykkäykset	1	90%
	C6	263	Kommenttien tykkäykset	1	28%
	C8	287	Julkaisujen ja kommenttien tykkäyksiä	2, 2	80%
Keskiverto	C3	223	Julkaisujen tykkäykset, kommentit	3, 1	96%
	C9	190	Kommenttien tykkäykset, kommentit	0, 1	92%
	C4	560	Ei julkaisuja + pvm lkm (muuten kaikki)	3-7	96%
Vahva	C1	124	Ei kommentteja + pvm lkm (muuten kaikki)	0-1	90%
	C7	167	Kaikki sitoutumistoiminnat	1-3	91%
	C5	536	Kaikki sitoutumistoiminnat	2-13	96%

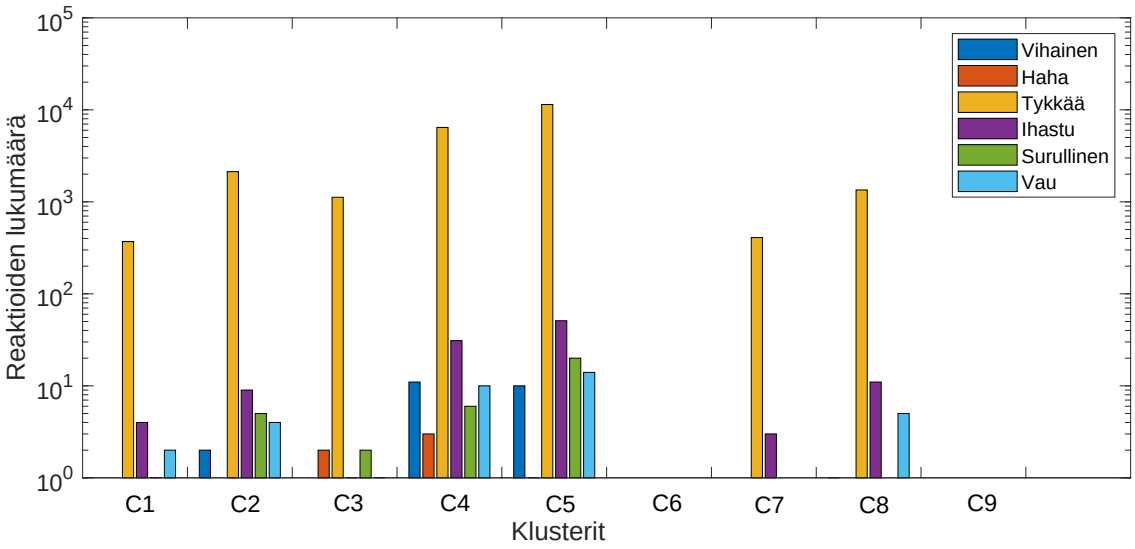
Vahvasti sitoutuneilla käyttäjillä on ainoastaan sisällön julkaisuja. Kuviossa 19 on jakaumat näiden klustereiden julkaisuille tehdyistä toiminnoista. Tarkastelemalla julkaisujen tykkäyksiä, jakoja, kommentteja sekä kommentteja näihin kommentteihin huomaa, että klusterin C5 käyttäjien julkaisuille tehtyjä toimintoja on selvästi enemmän kuin muilla. Tällöin klusterin C5 käyttäjiä saatetaan seurata enemmän.



Kuvio 19: Muunnoksen 1.1 julkaisuille tehtyjen toimintojen lukumäärä

Ryanin (2017) mukaan pelkkä tieto siitä, että käyttäjä on lukenut sisällön ei riitä, vaan on analysoitava tunteita. Myös Doornin ym. (2010) mukaan sitoutumistasoon vaikuttaa kokemukset brändistä. Näin ollen muunnoksen 1.1 julkaisujen reaktioiden määrät (Tykkää, Vihainen, Haha jne.) klustereittain on koottu kuvioon 20. Kuviossa ei ole klustereita C6 ja C9, koska niiden käyttäjät ovat reagoineet vain kommentteihin. Kommenttien kohdalla oli vain Tykkää-reaktioita, joten niitä ei ole tarve tutkia tarkemmin.

Klusterien käyttäjien reaktioista ei kuitenkaan pysty yksiselitteisesti tulkitsemaan heidän kokemuksiaan, koska julkaisu voi olla sisällöltään surullinen ja silloin käyttäjätkin vastaavat siihen surullisesti. Reaktioiden tarkemmassa tulkinnassa pitäisi tulkita myös julkaisujen sisällön tekstiä. Jos klusteri C4 luetaan mukaan vahvasti sitoutuneisiin, niin tällöin kuviosta 20 huomaa, että vahvasti sitoutuneilla on erilaisia reaktioita. Tarkasteltaessa taulukon 18 mukaista sitoutumisastejakoa reaktioissa huomaataan, että jokaisessa asteessa on Vihainen-reaktiota sisältävä klusteri (C2, C4, C5). Näillä on myös eniten erilaisia reaktioita. Sitoutumisen näkökulmasta näiden klustereiden käyttäjät ovat kiinnostavia, koska he ilmaisevat tunteitaan ja kokemuksiaan monipuolisesti.



Kuvio 20: Muunnoksen 1.1 julkaisujen reaktioiden lukumäärät klustereittain

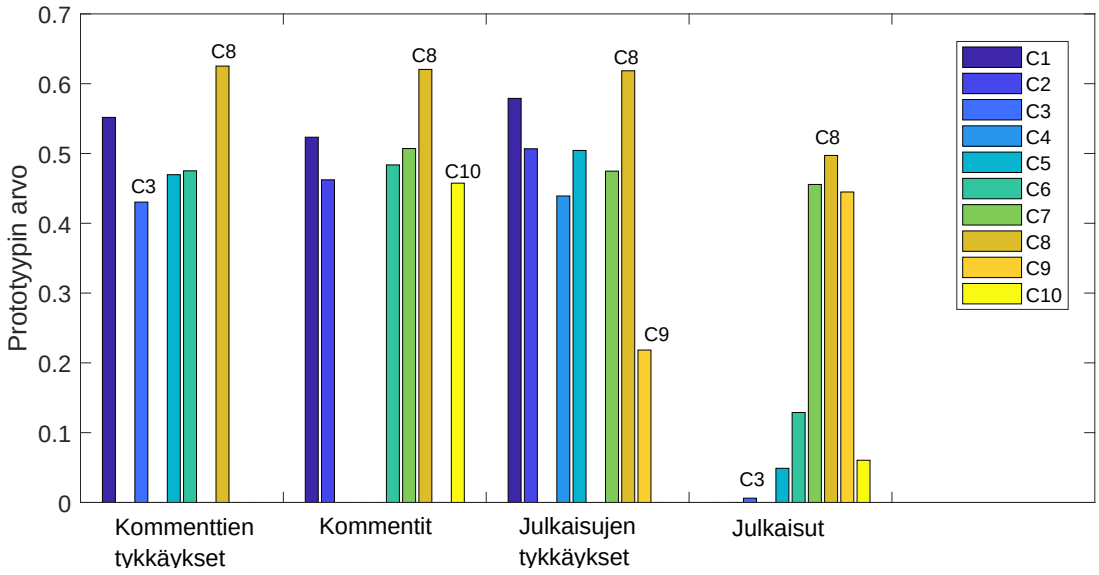
6.2.5 Tulokset: Muunnos 1.4

Muunnos 1.4 poikkesi muunnoksesta 1.2 siten, että julkaisujen ja kommenttien päivien lukumääriä ei otettu huomioon klusteroinnissa, jolloin löydettiin kymmenen klusteria. Taulukosta 19 näkee, että muunnos ei vaikuta yhteen klusteriin (C4), jonka tiedot säilyvät samanlaisena. Myös kaksi muuta klusteria (C2, C1) ovat muunnoksen 1.1 kanssa samanlaiset, toisessa ei vain ole kommenttien tykkäämistä. Lisäksi klustereissa on edelleen yksi vahvimmin sitoutunut klusteri (C8). Muuten sitoutumistoiminnot jakautuvat klustereiden kesken eri tavalla kuin muunnoksessa 1.1.

Taulukko 19: Muunnoksen 1.4 klustereiden metadata

	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Koko	3523	560	223	267	1173	322	97	80	575	85	141
Jäseniä	71%	96%	96%	29%	43%	81%	91%	91%	96%	91%	90%
Julkaisut M_{d}	0	0	0	0	0	0	0	1	2	1	0
Max	134	0	0	1	0	7	19	30	134	20	2
Kommentit M_{d}	0	3	1	0	0	0	2	3	10	0	1
Max	393	137	18	0	0	0	25	53	393	0	19
Tykkäykset julkaisut M_{d}	2	7	3	0	1	3	0	2	11	0	0
Max	258	120	67	0	40	66	0	27	258	17	0
Tykkäykset kommentit M_{d}	1	5	0	1	0	2	2	0	11	0	0
Max	737	109	0	8	0	37	16	0	737	0	0

Järjestys erottelevien muuttujien kohdalla ei muuttunut muunnoksesta 1.1. Kuviosta 21 näkee, että muunnos jakoi julkaisujen perusteella käyttäjiä seitsemään klusteriin kolmen sijasta. Klustereiden joukossa on klusteri C3, jolla on pieni määrä julkaisuja. Tutkimalla klusteria tarkemmin huomataan, että sillä on neljä käyttäjää, jotka ovat tehneet yhden julkaisun. Täten nämä käyttäjät voidaan pitää klusteroituna väärään klusteriin.



Kuvio 21: Muunnoksen 1.4 muuttujien erottelevuus klustereittain

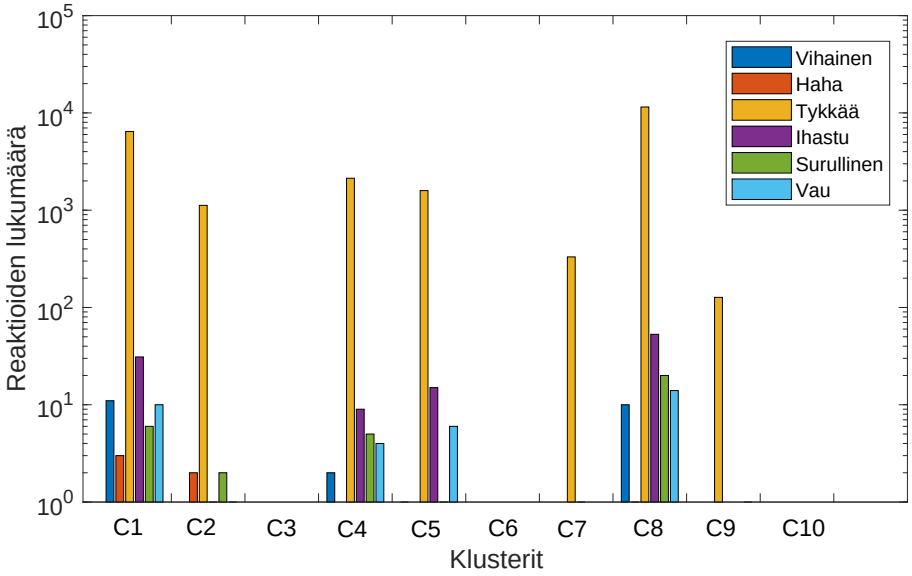
Järjestämällä klusterit sitoutumisasteen mukaan taulukkoon 20 huomataan, että klusterit eivät jakaannu yhtä tasaisesti asteisiin kuin muunnoksessa 1.1. Isoin klusteri on edelleen heikosti sitoutunut ja klusteri C3, joka sisälsi neljä väärin klusteroitua käyttäjää. Keskivertoisesti sitoutuneissa on säilynyt kaksi samanlaista klusteria. Vahvasti sitoutuneissa on kuusi klusteria, jossa klustereiden C5 ja C9 käyttäjät eivät ole kommentoineet ollenkaan. Lisäksi klusterit C10, C6 ja C7 eroavat muunnokseen 1.1 siinä, että niillä ei ole tykkäyksiä. Heikosti sitoutuneet ovat edelleen kooltaan suurin ryhmä, mutta vahvasti sitoutuneet ovat toiseksi suurin.

Taulukko 20: Muunnoksen 1.4 klusterit sitoutumisasteen mukaisessa järjestyksessä

	Heikko		Keskiverto		Vahva					
Klusteri	C4	C3	C2	C1	C5	C9	C10	C6	C7	C8
Koko	1173	267	223	560	322	85	141	97	80	575

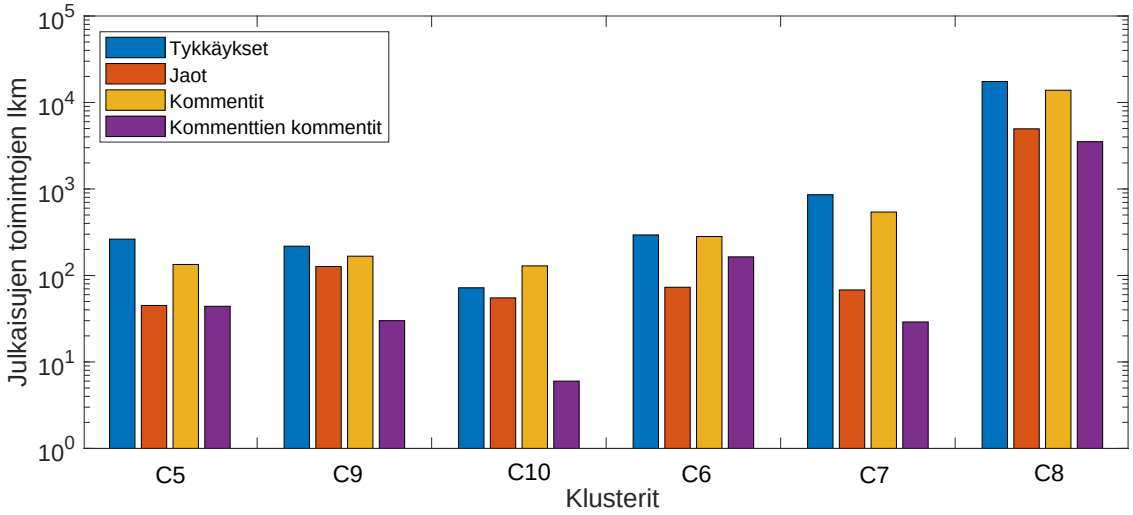
Julkaisujen reaktioiden jakaumat eivät poikkea paljoa muunnoksen 1.1 jakaumista (kuvio 22). Vahvasti sitoutuneissa klustereissa C7 ja C9 on pelkästään tykkäyksiä. Tällaisia klustereita ei ole muunnoksessa 1.1. Edelleen jokaisella sitoutumisasteella on klusterit, jotka reagoivat

julkaisuihin monipuolisesti.



Kuvio 22: Muunnoksen 1.4 julkaisujen reaktioiden lukumäärät klustereittain

Tarkastellessa jakaumia julkaisujen tykkäyksistä, jaoista ja kommenteista sekä kommenttien kommenteista vahvasti sitoutuneissa klustereissa huomataan, että julkaisujen päivien lukumäärän jättäminen pois ei vaikuta jakaumiin (kuvio 23). Edelleen on yksi klusteri, jonka käyttäjien julkaisuilla on paljon toimintoja. Vahvasti sitoutuneiden jakautuminen kuuteen klusteriin kolmen sijasta ei vaikuttanut tähän.



Kuvio 23: Muunnoksen 1.4 julkaisuille tehtyjen toimintojen lukumäärä

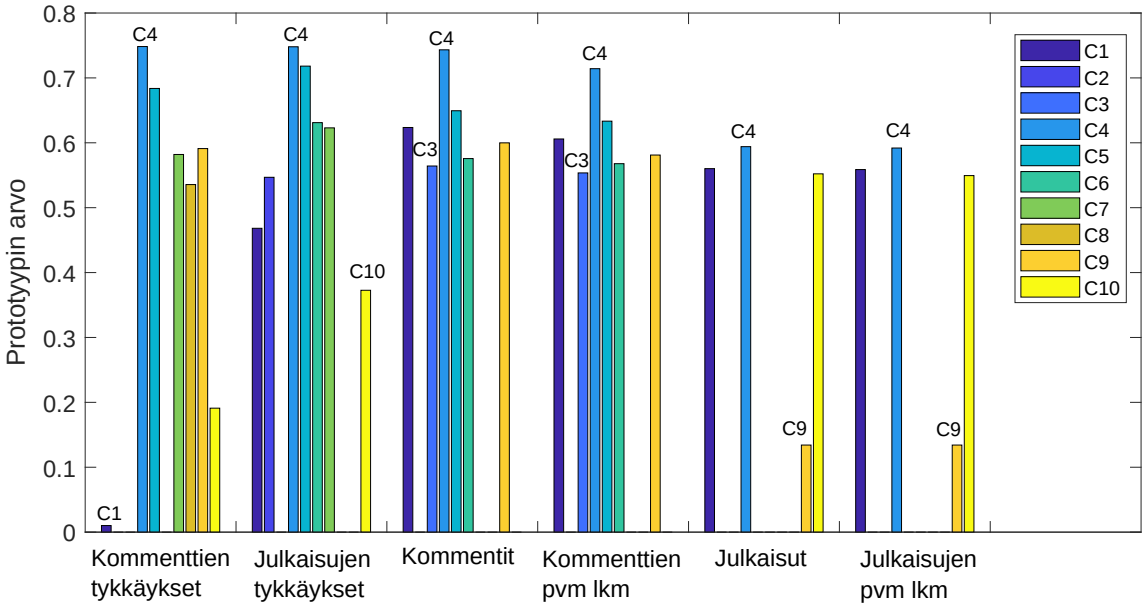
6.2.6 Tulokset: Muunnos 2.1

Muunnoksen 2.1, josta oli poistettu paljon sitoutumistoimintoa tehneet, kohdalla löydettiin kymmenen klusteria. Näiden kymmenen klusterin metadata on nähtävissä taulukosta 21. Muunnoksella 2.1 on kolme sitoutumistoiminnoiltaan ja kooltaan samanlaista klusteria (C2, C8, C7) muunnoksen 1.1 klustereiden (C2, C6, C8) kanssa. Odotetusti nämä klusterit ovat heikosti sitoutuneita. Taulukon 21 perusteella muunnos erotteli vahvasti sitoutuneita enemmän tuoden sinne yhden klusterin enemmän, vaikka paljon sitoutumistoimintaa tehneet poistettiin. Kuitenkin vahvasti sitoutuneet ovat edelleen pienin ryhmä. Jäsenien osuus koko datasta pysyi samana ja myös klustereiden välillä prosenttiosuudet pysyivät lähes samana.

Taulukko 21: Muunnoksen 2.1 klustereiden metadata

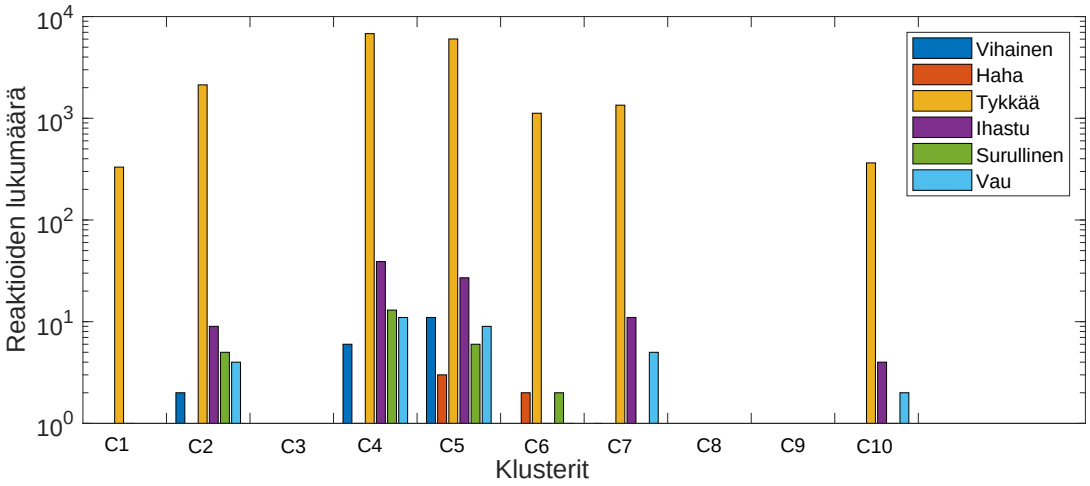
	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Koko	3426	101	1173	121	489	553	223	287	263	93	123
Jäseniä	71%	88%	43%	91%	96%	96%	96%	80%	28%	94%	89%
Julkaisut M_{d}	0	1	0	0	2	0	0	0	0	0	1
Max	15	12	0	0	15	0	0	0	0	3	8
Kommentit M_{d}	0	2	0	1	8	3	1	0	0	2	0
Max	59	53	0	19	59	58	18	0	0	25	0
Tykkäykset julkaisut M_{d}	1	1	1	0	10	7	3	2	0	0	1
Max	69	27	40	0	69	68	67	66	0	0	48
Tykkäykset kommentit M_{d}	1	0	0	0	9	5	0	2	1	2	0
Max	81	1	0	0	81	78	0	37	8	16	19
Julkaisut pvm M_{d}	0	1	0	0	2	0	0	0	0	0	1
Max	15	12	0	0	15	0	0	0	0	3	7
Kommentit pvm M_{d}	0	2	0	1	6	3	1	0	0	2	0
Max	46	42	0	12	46	45	14	0	0	20	0

Muunnoksen 2.1 kohdalla eniten ja vähiten erottelevat muuttujat eivät muuttuneet (kts. kuvio 24). Ainoastaan toiseksi ja kolmanneksi erottelevien muuttujien järjestys muuttui, jolloin eniten erottelevina muuttujina ovat julkaisujen ja kommenttien tykkäykset. Kuviosta huomaa myös sen, että vahvasti sitoutuneisiin on tullut yksi klusteri (C9) lisää, koska *julkaisut* ja *julkaisujen pvm lkm* -muuttujien kohdalla on C4 klusteria. Lisäksi nämä muuttujat erottelevat klusteri C9 selvästi vahvasti sitoutuneista C1, C4, ja C10 klustereista.



Kuvio 24: Muunnoksen 2.1 muuttujien erottelevuus klustereittain

Reaktioiden määrien jakaumat klustereittain pysyivät lähes samana muunnoksen 1.1 reaktioiden kanssa (kuviosta 25). Vahvasti sitoutuneita klustereita (C9, C10, C1, C4) tarkastellessa huomaa, että klusterissa C1 ei ole Ihastu-reaktiota, jota oli muunnoksen 1.1 vastaavassa klusterissa C7. Uudella klusterilla C9 ei ole reaktioita, koska sillä ei ole tykkäyksiä julkaisuista.



Kuvio 25: Muunnoksen 2.1 julkaisujen reaktioiden lukumäärät klustereittain

6.3 Asiakkuuden säilyttäminen ja sosiaalinen verkosto

Asiakkuuden säilyttämisen vaiheessa asiakkaan sitoutumista pystytään seuraamaan sosiaalisen verkoston avulla hyödyntämällä verkkoteoriaa. Tämän tutkielman rajoissa ei ollut mahdollista perehtyä tarkemmin verkkoteorioihin, joten tutkielmassa hyödynnettiin asiakkuuden kehittämisen vaiheessa käytettyä dataa Cheltenhamin julkisen yhteisön Facebookin julkaisuista. Dataa ei klusteroitu uudelleen, vaan MATLAB:in avulla luotiin algoritmit, joilla asiakkuuden kehittämisvaiheessa muunnoksen 1.1 ja 1.4 klustereita tutkittiin lisää. Klusterit mahdollistivat Bijmoltin ym. (2010) mainitseman asiakaskunnan heterogeenisuuden huomioimisen asiakkuuden säilyttämisessä.

Säilyttämisen kannalta kiinnostavia ovat heikoiten sitoutuneet klusterit, koska näiden käyttäjien toiminta on vähäistä, jolloin he voivat olla poistuvia käyttäjiä. Heikoiten sitoutuneilla käyttäjillä on vain tykkäyksiä, joten heitä seuraavia käyttäjiä ei ollut mahdollista identifioida. Näin ollen muodostettiin algoritmit (kts. algoritmit 1 ja 2), joilla selvitettiin vahvasti ja keskivertoisesti sitoutuneet käyttäjät, joiden julkaisuista tai kommenteista heikosti sitoutuneet käyttäjät ovat tykänneet. Zadehin ja Shardan (2014) mukaan seuraajien määrä on yksi parhaimmista mittareista, jolla pystytään huomaamaan vaikuttavien käyttäjien roolia sosiaa-

lisissa verkostoissa.

Algoritmi 1: Etsi käyttäjät, joita heikosti sitoutuneet seuraavat

Data: klusterit-tietorakenne, heikosti sitoutuneiden klusterien numerot

Tulos: seuratutKayttajat-tietorakenne klustereittain

for *j* = klusterienNrot

```
tiedot = klusterit(j).tiedot;     // käyttäjät riveittäin
nro = nro + 1;                                // alustettu 0
```

for *k* = 1: length(tiedot)

```
kayttajanID = tiedot(k, end);
seuratutKayttajat(nro).klusteri{k, 1} = kayttajanID;
```

1. Otetaan Like-taulukosta julkaisujen rivit, joista käyttäjä tykännyt ehdolla:

```
tykatytJulkaisut = Like.id == kayttajanID &
 Like.cid == 0;       // 0 == julkaisuista tykänneet
```

2. Kohdasta 1 muodostetusta käyttäjän tykatytJulkaisut-taulukosta otetaan julkaisujen yksittäiset id:t:

```
tykatytJulkaisut = unique(tykatytJulkaisut.pid);
```

3. Etsitään Post-taulukosta rivit, joissa kohdan 2 id:t esiintyvät:

```
rivitJulkaisut = ismember(Post.pid,
 tykatytJulkaisut);
```

4. Asetetaan seuratutKayttajat-tietorakenteeseen klusterin *nro* riville *k* sarakkeen 2 kohdalle Post-taulukon kohdan 3 rivien mukaiset yksittäiset id:t:

```
seuratutKayttajat(nro).klusteri{k, 2} =
 unique(Post.id(rivitJulkaisut));
```

5. Toistetaan askeleet 1-3 kommenttien tykkäykselle ehdolla:

```
tykatytJulkaisut2 = Like.id == kayttajanID &
 Like.cid == 0;      // >0 == kommenteista tykänneet
```

6. Asetetaan seuratutKayttajat-tietorakenteeseen klusterin *nro* riville *k* sarakkeen 3 kohdalle Post-taulukon kohdan 3 rivien mukaiset yksittäiset id:t:

```
seuratutKayttajat(nro).klusteri{k, 3} =
 unique(Post.id(rivitJulkaisut2));
```

end

end

Algoritmi 2: Laske seurattujen käyttäjien prosenttiosuudet

Data: seuratutKayttajat-tietorakenne klustereittain, klusterit-tietorakenne, seurattujen klustereiden numerot

Tulos: osuudetSeuratut-matriisi

Alustetaan osuudetSeuratut-matriisi

for j = 1: length(seuratutKayttajat)

Alustetaan julkaisuID- ja kommentitID-matriisit seuratutKayttajat-tietorakenteen j:nnen klusterin yksittäisistä id:eistä:

```
julkaisuID =
 unique(cell2mat(seuratutKayttajat(j).klusteri(:,
 2)));
kommentitID =
 unique(cell2mat(seuratutKayttajat(j).klusteri(:,
 3)));
```

for k = 1: length(klusterienNrot)

```
nro = klusterit(k);
```

1. Etsitään klusterin *nro* tiedoista rivit, joissa alustettujen matriisien id:t esiintyvät:

```
rivitJulkaisut = ismember(klusterit(nro).tiedot(:,
 end), julkaisutID);
rivitKommentit = ismember(klusterit(nro).tiedot(:,
 end), kommentitID);
```

2. Lasketaan rivitJulkaisut ja rivitKomentit matriisien Tosi-arvojen esiintymien lukumäärät
3. Lasketaan kohdan 2 julkaisujen lukumäärien osuus klusterin koosta ja asetetaan se `osuudetSeuratut`-matriisiin kohtaan (j, k)
4. Lasketaan kohdan 2 kommenttien lukumäärien osuus klusterin koosta ja asetetaan se `osuudetSeuratut`-matriisiin kohtaan (j, k + length(klusterienNro))

end

end

6.3.1 Tulokset

Asiakkuuden kehittämisen muunnoksesta 1.1, jossa oli kaikki käyttäjät sekä vuorovaikutusmuuttujat, muodostettujen heikosti sitoutuneiden klustereiden tykkäyksien osuudet jokaisesta vahvasti ja keskivertoisesti sitoutuneiden klustereiden kaikista käyttäjistä on kuvattu taulukkoon 22. Taulukosta näkee, että heikoiten sitoutuneen klusterin C2 käyttäjät ovat tykänneet klusterin C5 52 %:sta käyttäjien julkaisuista.

Klusterin C8 käyttäjät ovat tykänneet myös enemmän klusterin C5 julkaisuista sekä kommenteista kuin muista vahvasti sitoutuneiden klustereiden käyttäjien julkaisuista ja kommenteista. Näiden perusteella klusterissa C5 näyttäisi olevan seuratuimmat käyttäjät ja niiden sosiaalinen verkosto on laajempi. Klusterilla C5 oli enemmän julkaisuja verrattuna muihin klustereihin ja se on kooltaan isompi. Heikosti sitoutuneiden klustereiden tykkäyksien osuudet keskivertoisesti sitoutuneiden käyttäjien kommenteista jakaantuvat tasaisemmin. Klusterin C6 jakaumat klustereiden C3, C9 ja C4 välillä vaihtelevat 4 %:sta 9 %:tiin.

Taulukko 22: Muunnoksen 1.1 heikosti sitoutuneiden tykkäyksien osuudet vahvasti ja keskivertoisesti sitoutuneista

	Vahvasti sitoutuneet						Keskivertoisesti sitoutuneet		
	Julkaisut			Kommentit			Kommentit		
Klusteri	C1	C7	C5	C1	C7	C5	C3	C9	C4
C2	20%	28%	51%	0%	0%	0%	0%	0%	0%
C6	0%	0%	0%	0%	10%	20%	4%	6%	9%
C8	9%	22%	45%	0%	13%	39%	11%	9%	16%

Asiakkuuden kehittämisen muunnoksesta 1.4, josta oli jätetty vuorovaikutusmuuttujat pois, muodostettujen heikosti sitoutuneiden klustereiden tykkäyksien osuudet jokaisesta vahvasti ja keskivertoisesti sitoutuneiden klustereiden kaikista käyttäjistä on kuvattu taulukkoon 23. Heikoiten sitoutuneen klusterin C4 käyttäjät ovat tykänneet vahvimmin sitoutuneen klusterin käyttäjien julkaisuista. Toisaalta kolmanneksi suurin osuus on heikoimmin sitoutuneella klusterilla C9. Tämä klusteri on vahvasti sitoutuneista toiseksi pienin kooltaan. Kommenttien

tykkäyksissä myös vahvasti sitoutuneilla klusteilla on suurimmat osuudet.

Taulukko 23: Muunnoksen 1.4 heikosti sitoutuneiden tykkäyksien osuudet vahvasti ja keskivertoisesti sitoutuneista

	Vahvasti sitoutuneet												Keskivertoiset	
	Julkaisut						Kommentit						Kommentit	
Klusteri	C5	C9	C10	C6	C7	C8	C5	C9	C10	C6	C7	C8	C2	C1
C4	3%	18%	1%	10%	21%	51%	0%	0%	0%	0%	0%	0%	0%	0%
C3	0%	0%	0%	0%	0%	0%	0%	0%	6%	8%	10%	20%	4%	9%

7 Pohdinta

KDD-prosessilla ja klusteroinnilla saatiin muodostettua asiakkuuden kehittämisen vaiheessa käyttäjäryhmät (klusterit) sitoutumisasteittain taulukkoon 18, jotka auttavat muodostamaan käsityksen asiakkaiden sitoutumisen asteista ja niiden välisistä eroista. Lisäksi käyttäjäryhmiin tarkemmin perehtyminen mahdollistaa yksittäisen henkilön sitoutumisen havaitsemisen ja mahdollisesti sen kehittämisen. Samanlaisia selkeitä käyttäjäryhmiä sitoutumisasteittain ei saatu muodostettua asiakashankinnan vaiheessa, vaan klustereiden sitoutuminen vaihteli eri tulkintojen mukaisesti. Toisin kuin aikaisempien tutkimuksien yksittäisten julkaisujen tutkiminen asiakkuuden kehittämisen vaiheessa, tutkielmassa käytetyn yhteisön sosiaalisen median data mahdollisti juuri yksittäisen käyttäjän sitoutumisen selvittämisen. Kehittämisen vaiheen käyttäjäryhmät mahdollistivat myös asiakaskunnan heterogeenisuuden huomioimisen asiakkuuden säilyttämisen vaiheessa Bijmoltin ym. (2010) mukaisesti.

7.1 Asiakashankinta

Raphaelin, Goldsteinin ja Finkin (2017) tutkimuksessa asiakkaan session korkea kesto ja alhainen vierailtujen sivujen määrä viittasi vahvasti sitoutuneeseen asiakkaaseen. Tämän tutkielman tuloksista ei voitu havaita samanlaista selkeää käytöstä varsinkaan muunnoksissa, joissa *ostoskori*-muuttujaa ei ollut skaalattu TF-IDF:llä. Kyseiset muunnokset tuottivat vain muutamia klustereita, joista sessioltaan pitkäkestoisilla klustereilla oli myös eniten eri tuotteiden katsomisia.

Pitkäkestoisia ja eniten tuotteita katselleita asiakkaita voidaan pitää vahvasti sitoutuneina, kun huomioidaan näiden asiakkaiden ostoskoriin siirrettyjen eri tuotteiden määrät ja ostot, sillä Robertsin ja Alpertin (2010) mukaan asiakas on sitoutunut, kun hän ostaa mielellään toisen tuotteen. Asiakkaat, jotka ovat ostaneet yhden tuotteen ja tekevät paljon siirtoja ostoskoriin, voidaan ajatella ostavan uudelleen. Toisaalta klustereista löydettiin sellaisia, joilla oli alhainen sessioiden kesto ja tuotteita oli katsottu vähäisesti, mutta ostoja ja ostoskoreja oli suhteellisen paljon. Täten näitä voidaan pitää myös vahvasti sitoutuneena, koska he tekevät lyhyessä ajassa ostopäätöksiä.

Ostoskori-muuttujan skaalaaminen TF-IDF:llä lisäsi klusterien määrää ja toi esille klustereita, joilla session kesto oli korkea ja katsottujen tuotteiden määrät olivat pieniä. Näin oli varsinkin muunnoksessa, jossa havaintorivit muodostuivat asiakkaiden sessioista. Täten näistä klustereista pystyi selvemmin havaitsemaan kaksi vahvimmin sitoutunutta klusteria varsinkin, kun huomioon otettiin ostoskoriin siirrettyjen eri tuotteiden määrät ja ostot. TF-IDF-skaalauksen myötä klustereista paljastui myös klusteri, jonka asiakkaat eivät osaa tehdä ostopäätöstä. Tämän klusterin session kestot olivat korkeita, katseltujen tuotteiden määrät olivat suuria, niillä ei ole ostoskoriin siirrettyjä tuotteita, ja eikä siten ostoja. Asiakashankinnan kannalta nämä asiakkaat ovat kiinnostavampia kuin ostoja ja ostoskoriin tuotteita siirtäneet vahvasti sitoutuneet asiakkaat, sillä markkinoinnilla voidaan vaikuttaa näiden asiakkaiden ostoaikeisiin.

Asiakkaiden sessioiden yhteenlaskeminen ja *ostoskori*-muuttujan skaalaaminen TF-IDF:llä (muunnos 5.2) muodosti viisi klusteria, joilla oli ostoskoriin siirrettyjä tuotteita ja ostoja, mutta sessioiden kestot ja tuotteet vaihtelivat alhaisesta korkeaan. Näin ollen sessioiden yhteenlaskeminen ei eritellyt klustereita sitoutumisasteen mukaisesti, mutta toi esille ostoskorien ja ostojen merkityksen sitoutumisessa. Sessioiden lukumäärän huomioimisessa korkean keston ja vähäisen tuotteiden katselumäärän mukaan klustereista ei voi päätellä vahvasti sitoutunutta klusteria. Toisaalta vahvasti sitoutuneeksi voidaan nähdä klusteri, jolla session kesto ja tuotteiden määrät ovat vähäisiä, mutta jossa on eniten ostoja ja ostoskoriin siirrettyjä tuotteita. Nämä asiakkaat tietävät mitä haluavat, jolloin asiakashankintavaiheessa he eivät ole kiinnostavia. Selkeää klusteria, jossa olisi asiakkaita, jotka eivät osaa tehdä ostopäätöstä ei löytynyt, sillä klusterissa, jossa on eniten sessioita ja katseltujen tuotteiden määriä on myös ostoskoriin siirrettyjä tuotteita ja ostoja.

Tarkastellessa sitoutumista tuotteen ja keston avulla, niin selkeitä sitoutumisen asteita ei ole tuloksien pohjalta havaittavissa. Huomioitaessa ostoskorit ja ostot, joita muissa tutkimuksissa ei oltu huomioitu, sitoutumisasteita on mahdollista havaita. Tätä tukee myös erottelevien muuttujien kuvaajat, sillä ostoskoriin siirrettyjen eri tuotteiden määrät erottelevat eniten ja katseltujen eri tuotteiden määrät vähiten.

7.2 Asiakkuuden kehittäminen

Doornin ym. (2010) mukaan sitoutunut asiakas tuottaa aktiivisesti sisältöä asiakasyhteisölleen. Lisäksi Malthousen ym. (2013) mukaan sisällön tykkääminen tai jakaminen ovat heikkoa sitoutumista ja sisällön luominen puolestaan ilmentää vahvaa sitoutumista. Sitoutumisasteittain ryhmitellyissä käyttäjäryhmissä vahvasti sitoutuneet käyttäjät olivat ainoat, jotka loivat julkaisuja yhteisölle. Heikosti sitoutuneet käyttäjät olivat vain tykänneet julkaisuista tai kommenteista. Näiden asteiden välissä olevien käyttäjien sitoutumistoiminta muodostui tykkäyksien lisäksi kommenteista.

Heikosti sitoutuneita käyttäjiä oli eniten, varsinkin vain julkaisuista tykänneitä käyttäjiä oli melkein kolmasosa kaikista käyttäjistä. Julkaisujen tykkäyksiä voidaan pitää heikoimpana sitoutumistoimintona vähäisemmän vuorovaikutuksen vuoksi. Vuorovaikutusmuuttujien pois jättäminen lisäsi vahvasti sitoutuneiden ryhmiä muodostaen niistä toiseksi suurimman ryhmän. Kaiken kaikkiaan dataan tehtiin kolme muunnosta poistamalla datasta juuri vuorovaikutusmuuttujat ja eniten sitoutumistoimintaa tekevät käyttäjät. Eniten sitoutumistoimintaa tekevien käyttäjien poistamisella ei ollut vaikutusta sitoutumisasteihin.

Doornin ym. (2010) mukaan sitoutuminen vaikuttaa käyttäjän oman toiminnan lisäksi toisten käyttäjien toimintaan. Sitoutunut käyttäjä on hyvä yhteistyökumppani, jolla saa yhteyden toisiin käyttäjiin. Näin ollen kehittämisen näkökulmasta kiinnostavimpia ryhmiä ovat juuri vahvasti sitoutuneet ryhmät. Keskivertoisesti ja heikosti sitoutuneet käyttäjät tykkäävät ja kommentoivat näiden julkaisuja, jolloin vahvasti sitoutuneiden käyttäjien kautta on mahdollista vaikuttaa heikommin sitoutuneisiin. Kuvioiden 19 ja 23 perusteella vahvasti sitoutuneissa on yksi eniten seurattu ryhmä, joiden julkaisujen reaktioiden määrät ovat suuria; varsinkin tykkäykset ja kommentit. Tämän ryhmän käyttäjiä voidaankin pitää hyvinä yhteistyökumppaneina.

Jokaisessa sitoutumisasteessa oli tunteitaan ja kokemuksiaan monipuolisesti ilmaiseva käyttäjäryhmä, koska heillä oli eniten erilaisia reaktioita julkaisuihin. Näillä käyttäjillä voi olla toimintaa ja palveluita kehittäviä ehdotuksia, jolloin näiden käyttäjäryhmien tarkempi tutkiminen sekä yhteistyö heidän kanssaan olisi hyödyllistä. Toisaalta Sashin (2012) mukaan nämä ryhmät voidaan nähdä sitoutuneina, koska käyttäjä on sitoutunut, kun hänellä on vahva

emotionaalinen side yhteisöön.

7.3 Asiakkuuden säilyttäminen

Kehittämisen vaiheen heikosti sitoutuneet käyttäjäryhmät ovat kiinnostavia asiakkuuden säilyttämisen kannalta, koska heidän vähäinen toimintansa voi ennakoida poistumista. Tutkimalla heikosti sitoutuneiden sosiaalisia kontakteja vahvasti ja keskivertoisesti sitoutuneisiin huomattiin, että vahvimmin sitoutuneen ryhmän käyttäjistä hieman yli puolet oli niitä, joiden julkaisuja heikoiten sitoutunut ryhmä oli tykännyt. Tämä vahvistaa sitä, että yhteistyössä vahvasti sitoutuneiden ryhmien kanssa voidaan vaikuttaa heikosti sitoutuneisiin. Täten sosiaalisen median avulla pystytään todentamaan käyttäjät, jotka kannattaa säilyttää sosiaalisten kontaktiensa vuoksi (Malthouse ym. 2013).

Tutkielman pohjalta mielenkiintoista olisi tutkia yhteisön sosiaalista verkostoa sen Facebookin sosiaalisen median datan ja verkkoteorian avulla. Verkkorakenteen kautta saataisiin selville, millaisia sosiaalisia verkostoja yhteisössä on ja niiden keskiössä olevat henkilöt. Tämä parantaisi sitoutumisen selvittämistä asiakkuuden säilyttämisen vaiheessa, mutta tämän tutkielman rajoissa verkkoteoriaa ei ollut mahdollista soveltaa.

8 Yhteenveto

Asiakkaan sitoutumisen mittaaminen on lisääntynyt vuosien aikana ja varsinkin sosiaalisen median käytön lisääntyminen ovat muuttaneet asiakkaan sitouttamisen päätavoitteeksi yrityksille (Ryan 2017; Okazaki ym. 2015). Sitoutumisen tutkimiseen hyödynnetään yhä enemmän reaaliaikaista sivukyselydataa ja sosiaalisen median dataa, joita myös tässä tutkielmassa on käsitelty. Reaaliaikaiset datajoukot ovat suuria, joten tutkielmassa hyödynnettiin KDD-prosessia ja klusterointia sen tiedonlouhinnan metodina.

KDD-prosessin aikana datoista muodostettiin erilaisia muunnoksia, joille suoritettiin MATLAB:illa K:n keskiarvon klusterointi neliöllisellä euklidisella etäisyydellä. Optimaalisten klusterimäärien arvioimiseen käytettiin monia klusterien validointi-indeksejä, jotka Jauhiainen ja Kärkkäinen (2017) ja Hämäläinen, Jauhiainen ja Kärkkäinen (2017) olivat tutkimuksissaan todenneet suorituskyvyiltään parhaimmiksi. Klustereiden määrän tulkintaa heikensi kuitenkin sivukyselydatan suuruus, mihin pyrittiin vaikuttamaan esikäsittelyssä TF-IDF-skaalauksella. Joidenkin muunnosten kohdalla klustereiden määrät jouduttiin päättämään validointi-indeksien kuvaajien tulkinnan perusteella.

KDD-prosessin avulla saatiin uutta ja hyödyllistä tietoa asiakkaan sitoutumisesta asiakkuuden elinkaaren eri vaiheissa, mitä ei olisi saavutettu pelkästään alkuperäisten datojen tietoja tarkastelemalla. Datojen eri muunnoksien klusteroinnilla saatiin muodostettua asiakkuuden elinkaaren mukaisesti asiakasryhmiä, jotka auttoivat muodostamaan käsityksen asiakkaiden sitoutumisesta ja käyttäytymisestä. Asiakashankintavaiheessa olennaista on kiinnittää huomiota asiakasryhmiin, jotka eivät osaa tehdä ostopäästöstä. Asiakkuuden kehittämisen vaiheessa kiinnostavia asiakasryhmiä ovat vahvasti sitoutuneet ja asiakkuuden säilyttämisessä heikosti sitoutuneet asiakasryhmät.

Lähteet

Aggarwal, Charu C., ja Chandan K. Reddy. 2014. *Data Clustering : Algorithms and Applications.* Chapman / Hall/CRC.

Backiel, Aimee, Bart Baesens ja Gerda Claeskens. 2014. "Mining Telecommunication Networks to Enhance Customer Lifetime Predictions". *International Conference on Artificial Intelligence and Soft Computing. ICAISC 2014. Lecture Notes in Computer Science* 8468:15–26. doi:10.1007/978-3-319-07176-3_2.

Bahari, T. Femina, ja M. Sudheep Elayidom. 2015. "An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour". *Procedia Computer Science* 46:725–731. doi:10.1016/j.procs.2015.02.136.

Baumann, Annika, Johannes Haupt, Fabian Gebert ja Stefan Lessmann. 2017. "Changing perspectives: Using graph metrics to predict purchase probabilities". *Expert Systems With Applications:* 137–148. doi:10.1016/j.eswa.2017.10.046.

Berendt, Bettina, Bamshad Mobasher, Myra Spiliopoulou ja Jim Wiltshire. 2001. "Measuring the Accuracy of Sessionizers for Web Usage Analysis". *Workshop on Web Mining at the First SIAM International Conference on Data Mining:* 7–14. Viitattu 18. joulukuuta 2018. http://facweb.cti.depaul.edu/research/TechReports/TR01-006.pdf.

Berry, Michael J., Gordon S. Linoff ja Michael J. A. Berry. 2004. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management.* 2. painos. John Wiley & Sons, Incorporated.

Bhatia, Sumit, Jingxuan Li, Wei Peng ja Tong Sun. 2014. "Monitoring and Analyzing Customer Feedback Through Social Media Platforms for Identifying and Remedying Customer Problems". *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining:* 1147–1154. doi:10.1145/2492517.2500287.

Bijmolt, Tammo H. A., Peter S. H. Leeflang, Frank Block, Maik Eisenbeiss, Bruce G. S. Hardie, Aurelie Lemmens ja Peter Saffert. 2010. “Analytics for Customer Engagement”. *Journal of Service Research* 13 (3): 341–356. doi:10.1177/1094670510375603.

Bowman, Douglas, ja Das Narayandas. 2001. “Managing Customer-Initiated Contacts with Manufacturers: The Impact on Share of Category Requirements and Word-of-Mouth Behavior”. *Journal of Marketing Research* 38 (3): 281–297. doi:10.1509/jmkr.38.3.281.18863.

Bramer, Max. 2013. *Principles of Data Mining*. 2. painos. Springer. Viitattu 30. lokakuuta 2018. http://www.springer.com/series/7592.

Bucklin, Randolph E., ja Catarina Sismeiro. 2009. “Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing”. *Journal of Interactive Marketing* 23 (1): 35–48. Viitattu 24. heinäkuuta 2018. https://doi.org/10.1016/j.intmar.2008.10.004.

Calinski, T., ja J. Harabasz. 1974. “A dendrite method for cluster analysis”. *Commun. Stat. Theory Methods* 3:1–27.

“Classification: ROC Curve and AUC”. 2019. Viitattu 20. helmikuuta 2019. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

“CRM (customer relationship management)”. 2018. Viitattu 24. syyskuuta 2018. https://searchcrm.techtarget.com/definition/CRM.

Culbert, Ben, Bin Fu, James Brownlow, Charles Chu, Qinxue Meng ja Guandong Xu. 2018. “Customer Churn Prediction in Superannuation: A Sequential Pattern Mining Approach”. *Databases Theory and Applications. ADC 2018. Lecture Notes in Computer Science* 10837:123–134. doi:10.1007/978-3-319-92013-9_10.

Cvijikj, Irena Pletikosa, ja Florian Michahelles. 2013. “Online engagement factors on Facebook brand pages”. *Social Network Analysis and Mining* 3 (4): 843–861. doi:10.1007/s13278-013-0098-8.

Dasgupta, Tirthankar, Lipika Dey ja Ishan Verma. 2016. “Fuzzy Multi-label Classification of Customer Complaint Logs Under Noisy Environment”. *International Joint Conference on Rough Sets, Lecture Notes in Computer Science* 9920:376–385. doi:10.1007/978-3-319-47160-0_34.

Davies, David L., ja Donald W. Bouldin. 1979. “A Cluster Separation Measure”. *IEEE Transaction on Pattern Analysis and Machine Intelligence* PAMI-1 (2): 224–227.

Desgraupes, Bernard. 2013. “ClusterCrit: Clustering Indices”. *R Package Version 1.2.3.*

Doorn, Jenny van, Katherine N. Lemon, Vikas Mittal, Stephan Nass, Doreen Pick, Peter Pirner ja Peter C. Verhoef. 2010. “Customer Engagement Behavior: Theoretical Foundations and Research Directions”. *Journal of Service Research* 13 (3): 253–266. doi:10.1177/1094670510375599.

Du, Juan, Biying Tan, Feida Zhu ja Ee-Peng Lim. 2013. “Social Listening for Customer Acquisition”. *International Conference on Social Informatics:* 75–80. doi:10.1007/978-3-319-03260-3_7.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander ja Xiaowei Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. *KDD, 96:* 226–231. Viitattu 10. marraskuuta 2018. http://dl.acm.org/citation.cfm?id=3001460.3001507.

Estivill-Castro, Vladimir. 2002. “Why So Many Clustering Algorithms: A Position Paper”. *SIGKDD Explor. Newsl.* 4 (1): 65–75. doi:10.1145/568574.568575.

Estivill-Castro, Vladimir, ja Jianhu Yang. 2000. “Fast and Robust General Purpose Clustering Algorithms”. *PRICAI 2000 Topics in Artificial Intelligence:* 208–218. doi:10.1007/3-540-44533-1_24.

Everitt, Brian S., Sabine Landau, Morven Leese, Daniel Stahl, Dr Sabine Landau ja Dr Morven Leese. 2010. *Cluster Analysis.* 5. painos. John Wiley & Sons, Inc.

Faed, Alireza, Omar K. Hussain ja Elizabeth Chang. 2014. “A methodology to map customer complaints and measure customer satisfaction and loyalty”. *Service Oriented Computing and Applications* 8 (1): 33–53. doi:10.1007/s11761-013-0142-6.

Fayyad, Usama, Gregory Piatetsky-Shapiro ja Padhraic Smyth. 1996a. “From Data Mining to Knowledge Discovery in Databases”. *Artificial Intelligence Magazine* 17 (3): 37–54. doi:10.1609/aimag.v17i3.1230.

———. 1996b. “Knowledge Discovery and Data Mining: Towards a Unifying Framework”. *KDD, 96:* 82–88. Viitattu 30. lokakuuta 2018. http://dl.acm.org/citation.cfm?id=3001460.3001477.

———. 1996c. “The KDD Process for Extracting Useful Knowledge from Volumes of Data”. *Communications of the ACM* 39 (11): 27–34. doi:10.1145/240455.240464.

Finto, suomalainen sanasto- ja ontologiapalvelu. 2019. “Tietotermit”. Viitattu 12. toukokuuta. http://urn.fi/URN:NBN:fi:au:tt:t6.

Goldenberg, Jacob, Barak Libai, Sarit Moldovan ja Eitan Muller. 2007. “The NPV of bad news”. *Journal of Marketing* 24 (3): 186–200. doi:10.1016/j.ijresmar.2007.02.003.

Guelman, Leo, Montserrat Guillen ja Ana M. Perez-Marin. 2012. “Random Forests for Uplift Modeling: An Insurance Customer Retention Case”. *Modeling and Simulation in Engineering, Economics and Management. MS 2012. Lecture Notes in Business Information Processing* 115:123–133. doi:10.1007/978-3-642-30433-0_13.

Hand, David, Heikki Mannila ja Padhraic Smyth. 2001. *Principles of Data Mining.* The MIT Press.

He, Wu, Shenghua Zha ja Ling Li. 2013. “Social media competitive analysis and text mining: A case study in the pizza industry”. *International Journal of Information Management* 33 (3): 464–472. doi:10.1016/j.ijinfomgt.2013.01.001.

Hämäläinen, Joonas, Susanne Jauhiainen ja Tommi Kärkkäinen. 2017. “Comparison of Internal Clustering Validation Indices for Prototype-Based Clustering”. *Algorithms* 10 (3): 1–14. doi:10.3390/a10030105.

Jain, A. K., M. N. Murty ja P. J. Flynn. 1999. “Data Clustering: A Review”. *ACM Comput. Surv.* 31 (3): 264–323. doi:10.1145/331499.331504.

Jauhiainen, Susanne. 2017. *Knowledge discovery from physical activity.* Pro gradu -tutkielmat. University of Jyväskylä. Viitattu 5. helmikuuta 2018. http://urn.fi/URN:NBN:fi:jyu-201705302561.

Jauhiainen, Susanne, ja Tommi Kärkkäinen. 2017. "A Simple Cluster Validation Index with Maximal Coverage". *In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESAINN 2017)* 26–28:293–298. Viitattu 1. tammikuuta 2019. http://www.i6doc.com/en/.

Jones, Karen Sparck. 1972. "A Statistical interpretation of term specificity and its application in retrieval". *Journal of Documentation* 28 (1): 11–21. doi:10.1108/eb026526.

Kamakura, W., C.F. Mela, A. Ansari ym. 2005. "Choice Models and Customer Relationship Management". *Marketing Letters* 16 (3–4): 279–291. doi:10.1007/s11002-005-5892-2.

Kantardzic, Mehmed. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms.* 2. painos. John Wiley & Sons, Inc. Viitattu 29. huhtikuuta 2018. https://ieeexplore.ieee.org/servlet/opac?bknumber=6105606.

Kazienko, Przemyslaw, Piotr Brodka ja Dymitr Ruta. 2009. "The Influence of Customer Churn and Acquisition on Value Dynamics of Social Neighbourhoods". *Visioning and Engineering the Knowledge Society. A Web Science Perspective.WSKS 2009. Lecture Notes in Computer Science* 5736:491–500. doi:10.1007/s11761-013-0142-6.

Kunz, Werner, Lerzan Aksoy, Yakov Bart, Kristina Heinonen, Sertan Kabadayi, Francisco Villarroel Ordenes, Marianna Sigala, David Diaz ja Babis Theodoulidis. 2017. "Customer engagement in a Big Data world". *Journal of Services Marketing* 31 (2): 161–171. doi:10.1108/JSM-10-2016-0352.

Lehtinen, Jarmo R. 2004. *Asiakkuuksien aktiivinen johtaminen.* Helsinki: Edita Publishing Oy.

Liu, Haibin, ja Vlado Keselj. 2006. "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users future requests". *Data & Knowledge Engineering:* 304–330. doi:10.1016/j.datak.2006.06.001.

Luhn, H. P. 1957. “A Statistical Approach to Mechanized Encoding and Searching of Literary Information”. *IBM Journal:* 309–317. Viitattu 3. huhtikuuta 2019. http://web.stanford.edu/class/linguist289/luhn57.pdf.

Malthouse, Edward C., Michael Haenlein, Bernd Skiera, Egbert Wege ja Michael Zhang. 2013. “Managing Customer Relationships in the Social Media Era: Introducing the Social CRM House”. *Journal of Interactive Marketing* 27:270–280. doi:10.1016/j.intmar.2013.09.008.

Manning, C.D., P. Raghavan ja H. Schutze. 2009. *Scoring, term weighting, and the vector space model.* Cambridge University Press. Viitattu 3. huhtikuuta 2019. https://nlp.stanford.edu/IR-book/pdf/06vect.pdf.

Montgomery, Alan L., Shibo Li, Kannan Srinivasan ja John C. Liechty. 2004. “Modeling Online Browsing and Path Analysis Using Clickstream Data”. *Marketing Science* 23 (4): 579–595. doi:10.1287/mksc.1040.0073.

Montgomery, Douglas C., Elizabeth A. Peck, G. Geoffrey Vining ja G Geoffrey. Vining. 2015. *Introduction to Linear Regression Analysis.* John Wiley & Sons, Incorporated.

Moro, Sergio, Paulo Rita, Mercedes Rozano ja Bernardo Vala. 2016. “Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach”. *Journal of Business Research* 69 (9): 3341–3351. doi:10.1016/j.jbusres.2016.02.010.

Mostafa, Mohamed M. 2013. “More than words: Social networks text mining for consumer brand sentiments”. *Expert Systems with Applications* 40 (10): 4241–4251. doi:10.1016/j.eswa.2013.01.019.

Mäntyneva, Mikko. 2003. *Asiakkuudenhallinta.* 2. painos. Vantaa: WSOY.

Netzer, Oded, James M. Lattin ja V. Srinivasan. 2008. “A Hidden Markov Model of Customer Relationship Dynamics”. *Marketing Science* 27 (2): 185–204. doi:10.1287/mksc.1070.0294.

Okazaki, Shintaro, Ana M. Diaz-Martin, Mercedes Rozano ja Hector David Menindez-Benito. 2015. “Using Twitter to engage with customers: a data mining approach”. *Internet Research* 35 (3): 416–434. doi:10.1108/IntR-11-2013-0249.

Ordenes, Francisco Villarroel, Babis Theodoulidis, Jamie Burton, Thorsten Gruber ja Mohamed. 2014. “Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach”. *Journal of Service Research* 17 (3): 278–295. doi:10.1177/1094670514524625.

Pakhira, Malay K., Sanghamitra Bandyopadhyay ja Ujjwal Maulik. 2004. “Validity index for crisp and fuzzy clusters”. *Pattern Recognition* 37 (3): 487–501. doi:10.1016/j.patcog.2003.06.005.

Pansari, Anita, ja V. Kumar. 2017. “Customer engagement: the construct, antecedents, and consequences”. *The Journal of the Academy of Marketing Science* 45 (3): 294–311. doi:10.1007/s11747-016-0485-6.

Piatetsky-Shapiro, Gregory. 1990. “Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop”. *Artificial Intelligence Magazine* 11 (5): 68–70. doi:10.1609/aimag.v11i4.873.

Potdar, Vidyasagar, Sujata Joshi, Rahul Harish, Richard Baskerville ja Pornpit Wongthongtham. 2018. “A process model for identifying online customer engagement patterns on Facebook brand pages”. *Information Technology & People* 31 (2): 595–614. doi:10.1108/ITP-02-2017-0035.

Pyyhtiä, Tomi, Seppo Roponen, Mikko Seppä, Teemu Relander, Raino Vastamäki, Janne Korpi, Mark Filenius, Kati Sulin ja Jani Engberg. 2013. *Digin mitalla. Verkkomarkkinoinnin ja -myynnin mittaamisen käsikirja.* Helsinki: Mainostajien Liitto.

Raphaeli, Orit, Anat Goldstein ja Lior Fink. 2017. “Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach”. *Electronic Commerce Research and Applications* 26:1–12. doi:10.1016/j.elerap.2017.09.003.

Ray, Siddheswar, ja Rose H. Turi. 2000. “Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation”. *In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques.* Viitattu 3. tammikuuta 2019. https://pdfs.semanticscholar.org/0ec1/32fce9971d1e0e670e650b58176dc7bf36da.pdf.

Rendon, Erendira, Itzel Abundez, Alejandra Arizmendi ja Elvia M. Quiroz. 2011. “Internal versus External cluster validation indexes”. *International journal of computers and communications* 5 (1): 27–34. Viitattu 1. tammikuuta 2019. https://pdfs.semanticscholar.org/2054/29d63883b041436fdf2be8170a1f98fa90da.pdf.

Retailrocket. 2018. “Retailrocket recommender system dataset”. Lisenssi CC BY-NC-SA 4.0, https://creativecommons.org/licenses/by-nc-sa/4.0/. Viitattu 12. marraskuuta. https://www.kaggle.com/retailrocket/ecommerce-dataset#events.csv.

———. 2019. “Cheltenham’s Facebook Group”. Lisenssi GNU General Public License, version 2, http://www.gnu.org/licenses/old-licenses/gpl-2.0.en.html. Viitattu 20. helmikuuta. https://www.kaggle.com/mchirico/cheltenham-s-facebook-group/version/1/home.

Riaz, Sumbal, Mehvish Fatima, M. Kamran ja M. Wasif Nisar. 2017. “Opinion mining on large scale data using sentiment analysis and k-means clustering”. *Cluster Comput:* 1–16. doi:10.1007/s10586-017-1077-z.

Roberts, Christopher, ja Frank Alpert. 2010. “Total customer engagement: designing and aligning key strategic elements to achieve growth”. *Journal of Product & Brand Management* 19 (3): 198–209. doi:10.1108/10610421011046175.

Roehm, Michelle L., ja Michael K. Brady. 2007. “Consumer Responses to Performance Failures by High-Equity Brands”. *Journal of Consumer Research* 34 (4): 537–545. doi:10.1086/520075.

Rousseeuw, Peter J. 1986. “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”. *Journal of Computational and Applied Mathematics* 20:53–65. doi:10.1016/0377-0427(87)90125-7.

Ryan, Damian. 2017. *Understanding Digital Marketing: Marketing strategies for engaging the digital generation.* 4. painos. New York: Kogan Page Ltd.

Saarela, Mirka, ja Tommi Kärkkäinen. 2015. "Analysing Student Performance using Sparse Data of Core Bachelor Courses". *Journal of educational data mining:* 3–32. Viitattu 22. tammikuuta 2019. https://jyx.jyu.fi/handle/123456789/46677.

Sashi, C.M. 2012. "Customer engagement, buyer-seller relationships, and social media". *Management Decision* 50 (2): 253–272. doi:10.1108/00251741211203551.

Schultz, Carsten D. 2017. "Proposing to your fans: Which brand post characteristics drive consumer engagement activities on social media brand pages?" *Electronic Commerce Research and Applications* 26 (11-12): 23–34. doi:10.1016/j.elerap.2017.09.005.

Su, Qiang, ja Lu Chen. 2015. "A method for discovering clusters of e-commerce interest patterns using click-stream data". *Electronic Commerce Research and Applications* 14 (1): 1–13. doi:10.1016/j.elerap.2014.10.002.

Subramania, Kumar, Alexander Velkov, Irene Ntoutsi, Peer Kröger ja Hans-Peter Kriegel. 2011. "Density-based community detection in social networks". *2011 IEEE 5th International Conference on Internet Multimedia Systems Architecture and Application:* 1–8. doi:10.1109/IMSAA.2011.6156357.

"Termiharava". 2019. Viitattu 27. helmikuuta 2019. http://www.terminfo.fi/sisalto/termiharava-148.html.

Trusov, Michael, Randolph E. Bucklin ja Koen Pauwels. 2009. "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site". *Journal of Marketing* 73 (5): 90–102. doi:10.2139/ssrn.1129351.

Vanderveld, Ali, Addhyan Pandey, Angela Han ja Rajesh Parekh. 2016. "An Engagement-Based Customer Lifetime Value System for E-commerce". *Proceeding KDD 16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining:* 293–302. doi:10.1145/2939672.2939693.

Wangenheim, Florian v., ja Tomas Bayon. 2007. “The chain from customer satisfaction via word-of-mouth referrals to new customer acquisition”. *Academy of Marketing Science* 35:233–249. doi:10.1007/s11747-007-0037-1.

Wartiainen, Pekka, ja Tommi Kärkkäinen. 2015. “Hierarchical, prototype-based clustering of multiple time series with missing values”. *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning:* 95–100. Viitattu 24. tammikuuta 2019. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-91.pdf.

Venkatesan, Rajkumar. 2017. “Executing on a customer engagement strategy”. *The Journal of the Academy of Marketing Science* 45 (3): 289–293. doi:10.1007/s11747-016-0513-6.

Verhoef, Peter C., Jenny van Doorn ja Matilda Dorotic. 2007. “Customer Value Management: An Overview”. *Journal of Research and Management,* numero 2: 51–68. doi:10.1177/1094670510375603.

Verhoef, Peter C., Werner J. Reinartz ja Manfred Krafft. 2010. “Customer Engagement as a New Perspective in Customer Management”. *Journal of Service Research* 13 (3): 247–252. doi:10.1177/1094670510375461.

Villanueva, Julian, Shijin Yoo ja Dominique M. Hanssens. 2008. “The Impact of Marketing-Induced Versus Word-of-Mouth Customer Acquisition on Customer Equity Growth”. *Journal of Marketing Research* 45 (1): 48–59. doi:10.1509/jmkr.45.1.48.

Vo, Nhi N. Y., Shaowu Liu, James Brownlow, Charles Chu, Ben Culbert ja Guandong Xu. 2018. “Client Churn Prediction with Call Log Analysis”. *Database Systems for Advanced Applications* 10828:752–763. doi:10.1007/978-3-319-91458-9_47.

Xu, Rui, ja Don Wunsch. 2009. *Clustering*. New Jersey: John Wiley & Sons, Inc. Viitattu 6. marraskuuta 2018. https://ieeexplore.ieee.org/servlet/opac?bknumber=5236612.

Zadeh, Amir Hassan, ja Ramesh Sharda. 2014. “Modeling brand post popularity dynamics in online social networks”. *Decision Support Systems* 65:59–68. doi:10.1016/j.dss.2014.05.003.

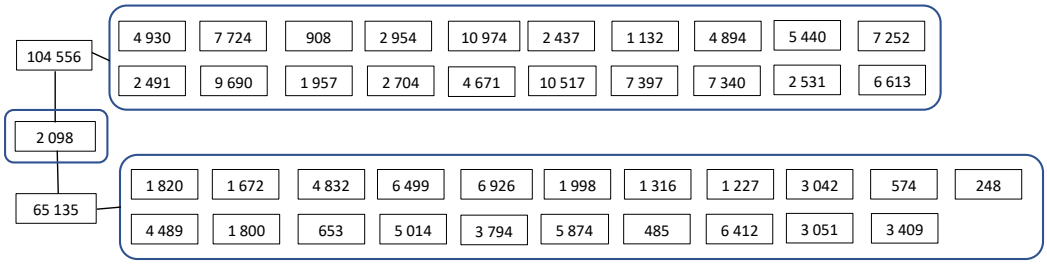
Zaki, Mohammed J., ja Wagner Meira Jr. 2014. *Data Mining and Analysis: Fundamental Concepts and Algorithms.* Cambridge University Press, toukokuu.

Zhao, Qinpei, ja Pasi Fränti. 2014. “WB-index: A sum-of-squares based index for cluster validity”. *Data & Knowledge Engineering* 92:77–89. doi:10.1016/j.datak.2014.07.008.

Äyrämö, Sami. 2006. *Knowledge mining using robust clustering.* Väitöskirja 63. University of Jyväskylä. Viitattu 30. lokakuuta 2018. http://urn.fi/URN:ISBN:951-39-2655-9.

Liitteet

A Asiakashankinnan muunnoksen 1.1 hierarkkinen prototyyppipohjainen klusterointi

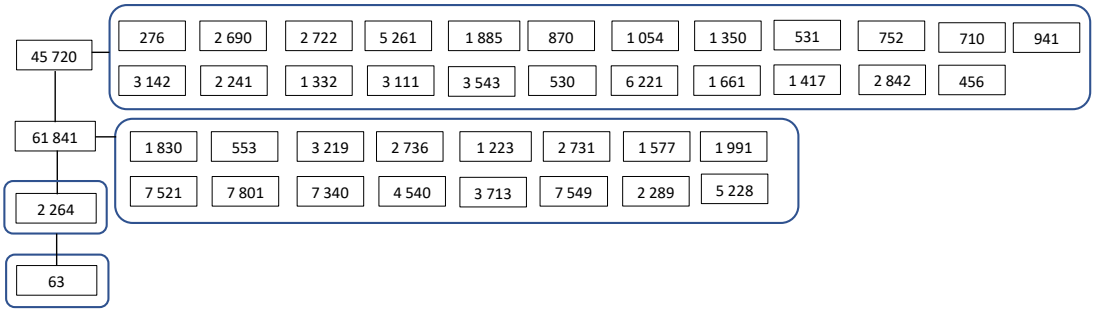


Kuvio 26: Muunnoksen 1.1 hierarkkisen prototyyppipohjaisen klusteroinnin tulos

Taulukko 24: Muunnoksen 1.1 klustereiden metadata (hierarkkinen prototyyppipohjainen)

	Koko	Asiakkaita	Kesto M_d [mm:ss]	Tuote M_d	Kori M_d	Korit >2	Koreja 1	Ostoja 1	Koreja 2, Ostoja 1
Data	171789	109888	6:46	2	0	3%	10%	3%	0.22%
C1	2098	1839	14:17	5	4	100%	0%	3%	3.15%
C2	4930	4633	6:48	2	0	3%	10%	3%	0.32%
C3	2491	2419	6:45	2	0	3%	10%	3%	0.24%
C4	7724	7159	6:54	2	0	3%	10%	3%	0.26%
C5	9690	8839	6:47	2	0	3%	10%	4%	0.23%
C6	908	893	7:01	2	0	3%	7%	3%	0.44%
C7	1957	1899	6:43	2	0	2%	9%	3%	0.10%
C8	2954	2670	6:35	2	0	3%	9%	3%	0.30%
C9	2704	2621	6:32	2	0	3%	10%	3%	0.15%
C10	10974	10001	6:49	2	0	3%	10%	3%	0.22%
C11	4671	4399	6:48	2	0	3%	10%	4%	0.21%
C12	2437	2348	7:08	2	0	3%	10%	4%	0.29%
C13	10517	9609	6:33	2	0	3%	10%	3%	0.21%
C14	1132	1110	6:57	2	0	4%	11%	4%	0.18%
C15	7397	6822	6:40	2	0	3%	10%	4%	0.19%
C16	4894	4580	6:58	2	0	3%	10%	4%	0.31%
C17	7340	6756	6:50	2	0	3%	10%	3%	0.12%
C18	2531	2428	6:53	2	0	3%	11%	4%	0.24%
C19	5440	5043	6:39	2	0	3%	9%	3%	0.24%
C20	7252	6754	6:52	2	0	3%	10%	3%	0.25%
C21	6613	6155	6:38	2	0	3%	9%	3%	0.15%
C22	1820	1731	6:33	2	0	2%	10%	3%	0.11%
C23	4489	4163	6:48	2	0	3%	10%	4%	0.22%
C24	1672	1607	6:26	2	0	3%	11%	3%	0.06%
C25	1800	1689	6:35	2	0	3%	10%	3%	0.22%
C26	4832	4496	6:59	2	0	3%	10%	3%	0.14%
C27	653	638	6:22	3	0	4%	9%	3%	0.15%
C28	6499	5897	6:47	2	0	3%	9%	4%	0.26%
C29	5014	4639	6:47	2	0	3%	11%	3%	0.14%
C30	6926	6278	6:38	2	0	3%	10%	3%	0.26%
C31	3794	3543	6:45	2	0	3%	10%	4%	0.29%
C32	1998	1892	6:46	2	0	3%	11%	4%	0.45%
C33	5874	5373	6:29	2	0	3%	10%	3%	0.12%
C34	1316	1268	7:03	2	0	4%	10%	4%	0.30%
C35	485	477	7:12	3	0	4%	12%	5%	0.21%
C36	1227	1199	6:49	2	0	3%	12%	3%	0.08%
C37	6412	5791	6:55	2	0	3%	9%	3%	0.23%
C38	3042	2864	6:46	2	0	3%	10%	4%	0.16%
C39	3051	2868	7:14	2	0	3%	10%	4%	0.26%
C40	574	545	6:33	2	0	4%	8%	4%	0.17%
C41	3409	3212	6:36	2	0	3%	10%	3%	0.18%
C42	248	247	6:10	2	0	2%	8%	3%	0.00%

B Asiakashankinnan muunnoksen 3.1 hierarkkinen prototyyppipohjainen klusterointi



Kuvio 27: Muunnoksen 1.1 hierarkkisen prototyyppipohjaisen klusteroinnin tulos

Taulukko 25: Muunnoksen 3.1 klustereiden metadata (hierarkkinen prototyyppipohjainen)

	Koko	Asiakkaita	Kesto M_d [[t]:mm:ss]	Tuote M_d	Kori M_d	Korit > 2	Koreja 1	Ostoja 1	Koreja 2, Ostoja 1
Data	109888	109888	9:07	3	0	5%	13%	5%	0.74%
C1	276	276	9:14	3	0	7%	11%	4%	0.72%
C2	3124	3124	9:06	3	0	4%	12%	5%	0.67%
C3	2690	2690	9:13	3	0	4%	13%	5%	0.71%
C4	2441	2441	9:46	3	0	5%	13%	5%	0.61%
C5	2722	2722	9:20	3	0	5%	13%	5%	0.66%
C6	1332	1332	8:28	3	0	5%	13%	6%	1.13%
C7	5261	5261	8:41	3	0	5%	13%	5%	0.84%
C8	3111	3111	8:53	3	0	5%	13%	5%	0.64%
C9	1885	1885	9:00	3	0	5%	13%	6%	0.74%
C10	3543	3543	8:58	3	0	5%	13%	5%	0.68%
C11	870	870	9:16	3	0	6%	12%	6%	1.38%
C12	530	530	9:00	3	0	4%	10%	5%	0.75%
C13	1054	1054	8:47	3	0	5%	13%	5%	0.38%
C14	6221	6221	8:56	3	0	5%	13%	5%	0.55%
C15	1350	1350	9:14	3	0	5%	12%	5%	0.67%
C16	1661	1661	9:52	3	0	5%	12%	5%	1.02%
C17	531	531	9:44	3	0	6%	11%	5%	1.13%
C18	1417	1417	8:42	3	0	5%	12%	5%	0.71%
C19	752	752	9:05	3	0	6%	10%	4%	1.33%
C20	2842	2842	9:33	3	0	4%	13%	6%	0.49%
C21	710	710	9:58	3	0	4%	10%	5%	0.28%
C22	456	456	9:55	3	0	4%	12%	6%	0.88%
C23	941	941	8:58	4	0	5%	13%	6%	0.53%
C24	1830	1830	8:43	3	0	5%	13%	5%	0.66%
C25	7521	7521	8:52	3	0	5%	13%	6%	0.73%
C26	553	553	8:31	3	0	5%	11%	4%	0.36%
C27	7801	7801	9:01	3	0	5%	13%	6%	0.83%
C28	3219	3219	8:56	3	0	5%	12%	5%	0.68%
C29	7340	7340	9:14	3	0	5%	12%	5%	0.69%
C30	2736	2736	9:37	3	0	5%	14%	6%	0.77%
C31	4540	4540	9:02	3	0	5%	13%	5%	0.59%
C32	1223	1223	9:09	3	0	5%	10%	5%	0.98%
C33	3713	3713	9:22	3	0	5%	13%	5%	0.70%
C34	2731	2731	9:34	3	0	5%	13%	5%	0.51%
C35	7549	7549	9:11	3	0	5%	13%	6%	0.79%
C36	1577	1577	9:57	3	0	5%	12%	6%	0.70%
C37	2289	2289	8:58	3	0	4%	14%	5%	0.61%
C38	1991	1991	9:10	3	0	4%	13%	6%	0.75%
C39	5228	5228	9:00	3	0	5%	12%	5%	0.71%
C40	2264	2264	23:43	8	4	100%	0%	9%	9.36%
C41	63	63	1:30:16	42	25	100%	0%	5%	4.76%

C Asiakashankinnan muunnoksen 4.1 klusterit

Taulukko 26: Muunnoksen 4.1 klustereiden metadata

	Data	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
Koko	109888	20187	7	14867	16807	28092	53	7	57	3705	7	215	6182	1293	19	17864	537
Asiakkaita	109888	20 187	1	14867	16807	28092	53	5	57	3705	4	215	6182	293	19	17864	537
Kesto M_d [[t]:mm:ss]	9:07	14:12	83:04:13	1:14	12:31	4:06	8:07:30	5:29:54	1:25:53	13:32	36:42:06	3:06:59	40:17	1:23:53	19:42:27	19:09	40:19
M_o	1:13	9:20	83:04:13	0:58	8:10	2:58	26:47	1:16:34	3:59	12:02	25:57:42	21:33	32:08	52:29	9:27:08	9:23	10:42
Max	83:04:13	30:00	83:04:13	2:58	30:00	8:05	19:32:47	11:58:44	6:31:28	1:19:16	46:10:13	8:27:12	2:10:31	5:03:11	30:51:48	59:33	4:01:41
Tuote M_d	3	4	1244	2	1	3	144	149	41	4	737.5	53	10	22	380	5	15
M_o	2	3	1244	2	1	3	115	83	20	2	587	42	8	13	257	4	8
Max	1 864	61	1 244	12	3	26	842	292	150	39	1350	355	125	136	1864	42	152
Korit M_d	0	0	1	0	0	0	0	90	24	2	1.5	0	0	0	0	0	8
M_o	0	0	1	0	0	0	0	64	20	2	0	0	0	0	0	0	7
Max	118	1	1	3	1	1	23	118	52	6	4	14	4	7	7	1	19
Sessioiden lkm M_d	1	1	318	1	1	1	35	18	7	1	159.5	16	3	7	88	2	3
Max	318	1	318	3	2	4	63	37	24	5	217	27	6	12	118	4	14
Korit 1	13%	8%	14%	8%	28%	7%	15%	0%	0%	0%	14%	14%	22%	19%	16%	17%	0%
Korit > 2	5%	0%	0%	1%	0%	0%	19%	71%	100%	100%	29%	26%	8%	18%	21%	0%	100%
Ostoja 1	5%	2%	0%	1%	14%	1%	6%	0%	4%	14%	14%	13%	12%	13%	16%	7%	7%
Korit 2 ja Ostoja 1	0.74%	0%	0%	0.04%	0%	0%	3.77%	0.0%	3.51%	14.33%	14.29%	7.44%	2.41%	4.87%	5.26%	0%	7.45%