TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

KLAUS HURME

ASIAKASVIESTIEN AUTOMAATTISEN LUOKITTELUN TUOMAT TEHOKKUUSHYÖDYT FINANSSIALAN YRITYKSEN ASIAKAS-PALVELUSSA

Diplomityö

Tarkastaja: professori Samuli Pekkola
Tarkastaja ja aihe hyväksytty
Talouden ja rakentamisen tiedekunta-neuvoston kokouksessa 9.11.2016

TIIVISTELMÄ

KLAUS HURME: Asiakasviestien automaattisen luokittelun tuomat tehokkuushyödyt finanssialan yrityksen asiakaspalvelussa
Tampereen teknillinen yliopisto
Diplomityö, 81 sivua, 2 liitesivua
Joulukuu 2016
Tietojohtamisen diplomi-insinöörin tutkinto-ohjelma
Pääaine: Informaatioanalytiikka
Tarkastaja: professori Samuli Pekkola
Ohjaaja: Jari Jussila

Avainsanat: luokitteleminen, koneoppiminen, hahmontunnistus, asiakaspalvelu, data, liiketoimintatiedon hallinta, analytiikka, virtuaaliagentti

Työn tavoitteena oli selvittää kohdeyritykselle, mitä tehokkuushyötyjä voidaan saavuttaa asiakasviestien automaattisella luokittelemisella heidän asiakaspalvelussaan. Samalla työ pohjustaa, kuinka asiakasviestejä voidaan automaattisesti luokitella. Kohdeyritys on yksi finanssialan toimijoista. Finanssialalle on tulossa useampia muutoksia lähitulevaisuudessa, mitkä pakottavat alan toimijoita hallitsemaan dataansa paremmin ja ketterämmin.

Luokitteleminen sinällään ei tuo vielä mitään hyötyjä. Hyödyt tulevat vasta, kun luokitellun datan kautta saadaan tehtyä parempia päätöksiä joko manuaalisesti tai automaattisesti. Tämän johdosta teoriaosuus ei pureudu suoraan luokittelemiseen, vaan lähtee liikkeelle siitä, kuinka datasta voidaan luoda kilpailukykyä. Toisekseen syvennytään tutkimaan minkälaisia asiakaspalveluun saapuvat viestit ovat ja kuinka niitä voidaan käsitellä automaattisesti. Lopulta päädytään koneoppimisen ja hahmontunnistuksen teorioihin, joilla automaattista luokittelua voidaan tehdä.

Työssä kerättiin tietoa kohdeyrityksestä työpajan avulla. Kerätystä tiedosta selvisi, että nykyinen asiakasviestien luokitustarkkuus ei mahdollista tunnistettuja tehokkuushyötyjä asiakaspalvelulle. Vasta tarkemman luokituksen jälkeen asiakaspalvelun tietotarpeet voidaan tyydyttää. Tietotarpeet liittyvät ratkaisukykyyn ja asiakasviestien sisällöllisiin asioihin. Tunnistettujen tietojen jalostus hyödyiksi onnistuu myös perinteisimmillä tiedonhaun menetelmillä, eikä automaattista luokittelemista välttämättä tarvita.

Automaattisen luokittelemisen hyödyt tulevat järjestelmän reaaliaikaisuudesta. Asiakkaat lähettäisivät vähemmän viestejä ja asiakaspalvelijat kykenisivät vastaamaan nopeammin, jos heillä olisi viestin lähetys- ja saapumishetkellä tarkempaa tietoa asiakasviestille tyypillisistä ominaisuuksista. Pohdinnassa päädyttiin tulokseen, että asiakaspalvelun tehokkuutta voidaan parhaiten nostaa tarjoamalla tarkasti määriteltyyn luokkaan sopiva mallivastaus automaattisesti. Tällöin virtuaaliagentit voivat hoitaa asiakaspalvelua.

ABSTRACT

KLAUS HURME: Efficiency benefits in customer service of a finance company by automation of classification of customer messages
Tampere University of Technology
Master of Science Thesis, 81 pages, 2 Appendix pages
December 2016
Master's Degree Programme in Information and Knowledge Management
Major: Information Analytics
Examiner: Professor Samuli Pekkola
Adviser: Jari Jussila

Keywords: Classification, Machine learning, Pattern recognition, Customer service, Data, Business intelligence, Analytics, Virtual agents

The objective of this research was to investigate how automatic classification of customer messages can increase effectiveness in target company's customer service. Secondly the study investigates how automatic classification can be done. The target company is an operator in finance sector. Recent trends suggest that the finance sector will undergo multiple changes in the near future, which will force the operators to handle their data better and more agile.

The classification merely does not bring any benefits. The benefits will emerge when classified data is used, either manually or automatically, for better decision making. Therefore, this study's theoretical part will not go straight to the classification. Instead, the study will first take a look on how data can be transformed to a competitive advantage. Secondly, this study will investigate customer messages and how those can be handled automatically. In the end this study will show how classification can be done automatically using machine learning and pattern recognition systems

Data for this study was gathered with a workshop method. Analyzing this data, it was recognized that the current classification precision of customer messages does not allow effectiveness benefits to appear. A more precise classification model needs to be implemented in order to fulfill the information needs of customer service. The information needs are concerning problem solving ability and content related issues. These recognized information needs can be mined with more traditional information retrieval methods and automatic classification is not necessary required.

Automatic classification system will bring benefits via real time analytics. The customer would send less messages and the customer service agents would be able to reply faster, if they had more information when the messages are sent and received. The study's discussion part came to a conclusion, that the most efficient method for improving effectiveness is to match precisely classified customer message with an answer template. If this is the case, the virtual agents could handle the customer service.

ALKUSANAT

Haluan ensimmäisenä kiittää kohdeyritystä diplomityön kirjoitusmahdollisuudesta. He antoivat minulle mahdollisuuden kirjoittaa mistä tahansa aiheesta, minkä kokisin itse mielenkiintoiseksi. Löydettyäni mielenkiintoisen aiheen, kohdeyritys tuki työn etenemistä ja osoitti, että näkökulmani aiheeseen on arvokas. Lisäksi he motivoivat, että tutkimisen jälkeen on valmiina pelikenttä, jossa oppeja pääsee hyödyntämään.

Lämmin kiitos myös Tampereen teknilliselle yliopistolle ja erityisesti ohjaajilleni Samuli Pekkolalle ja Jari Jussilalle. Saatiin yhdessä valmis työntekijä maailmalle, vaikka välillä tuntui, että opetetut asiat ovat kaukana reaalimaailmasta. Nyt hiukan varpaita yritysmaailmaan kastaneena ymmärrän paremmin, kuinka opit vastaavat yhteiskunnan tarpeisiin.

Tutkimuksen alkuvaiheilla sain ympäristöstäni kuulla kommentteja, että kokopäivätyön, diplomityön ja viiden kuukauden tavoiteaikataulun yhdistäminen ei ole toteutettavissa. Nämä kommentit ovat niitä, jotka saivat allekirjoittaneen motivoitumaan lähtemään pitkien työpäivien jälkeen koululle jatkamaan diplomityötä. Yksi parhaimmista tavoista kohottaa omaa itseluottamusta on osoittaa epäilijöiden olleen väärässä.

Seuraavaksi elämässäni aukeaa henkilökohtaisesti hankala vaihe, kun tarvitsee ruveta erikoistumaan pienelle osa-alueelle. Tätä ennen olen pystynyt pitämään oppini ja työkokemukseni rikkaina.

Jos teet kaikkea, et ole hyvä missään.

Tampereella, 20.12.2016

Klaus Hurme

SISÄLLYSLUETTELO

KUVALUETTELO

TAULUKKOLUETTELO

1. JOHDANTO

Tämän työn toimeksianto on tullut eräältä suomalaiselta finanssialan yritykseltä, jolla on suuri tahtotila muuttaa toimintaansa enemmän digitaaliseksi erityisesti sisältä päin. Yhtenä kehitysalueena ovat asiakaspalvelukanavat, joista tässä työssä tarkastellaan yrityksen verkkopalvelusta löytyvää viestintäkanavaa ja sitä kautta tulevia asiakasviestejä. Toimeksianto on tullut liiketoiminnan puolelta, joten tässä työssä ei uppouduta syvällisestä algoritmeihin tai muihin matemaattisiin laskentakaavoihin. Kohdeyrityksen toivomus työn suhteen oli, että se valmistelee organisaatiota uusiin teknologisiin menetelmiin. Heitä kiinnostaa erityisesti tietää, kuinka voidaan luokitella uusia asiakasviestejä vanhan viestidatan perusteella. Verkkopalvelun viestit ovat vain yksi asiakkaisiin liittyvä datanlähde. Muita asiakasdataa tuottavia lähteitä ovat asiakkaiden kanssa käytävät puhelinkeskustelut, sosiaalinen media, konttorit ja transaktiot. Tämän johdosta työn oppeja voidaan soveltaa myös kohdeyrityksen muihin osa-alueisiin.

Asiakkaista tulee niin valtava määrä dataa, että ihmisillä ei ole yksinkertaisesti aikaa tarkastella sen sisältöä tarkemmin. Ihmisen huomiosta on tullut organisaatioiden tärkeimpiä resursseja (Jiawei et al. 2001, s. 1). Tästä johtuen tulee löytää keinoja, joilla dataa voidaan automaattisesti analysoida, tiivistää, luokitella, sekä etsiä ajankohtaisia trendejä ja poikkeamia niistä (Lagus 2000, s. 10). Tiedonlouhinnan tieteenhaara tarjoaa työkaluja, joilla strukturoitua ja epästrukturoitua dataa voidaan käsitellä automaattisesti edellä mainittuihin tyyleihin. Pelkät työkalut eivät ole vielä riittäviä, vaan tarvitaan lisäksi toimivaa liiketoimintatiedonhallintaa muuttamaan data osaksi parempaa päätöksentekoa.

Tässä työssä tutkitaan tiedonlouhinnan yhden osa-alueen, automaattisen luokittelemisen, tuomia potentiaalisia hyötyjä. Samalla tutkitaan myös kokonaisvaltaisesti sitä, kuinka luokitteleminen on osa liiketoimintatiedon hallintaa. Aiheen kokonaisvaltainen näkemys on tarpeen, koska luokitteleminen ei yksistään tuo hyötyjä. Hyötyjä tulee vasta, kun luokittelun tuomia mahdollisuuksia käytetään. (Watson & Wixom 2007.) Ihmisen aivot ovat tottuneet suorittamaan vastaavanlaisia tehtäviä erittäin nopeasti. Kuvitellaan vaikka tilannetta, jossa tehtävänä on tunnistaa kaverisi kasvot luokkakuvasta, matkapuhelin taskustasi pelkällä kosketuksella, tai tehdä hajuaistiin perustuva päätös, tarvitseeko vaatekappale laittaa pesuun (Duda et al. 2001, s. 1). Kaikissa näissä esimerkeissä on luokittelu ja vanhan datan hyödyntäminen osana ajatusprosessia, vaikka näitä elementtejä ei aina tunnistaisi. Lisäksi näitä kaikkia esimerkkejä yhdistää, että pelkkä tehtävän suorittaminen ei tuo hyötyjä. Hyödyt tulevat vasta sen jälkeen, kun tiedon perusteella tehdään toimenpiteitä.

Työn taustalla on olettamus, että informaatioteknologia on kypsynyt tällä tieteenhaaralla niin pitkälle, että tekstin automaattinen luokittelu on mahdollista. Itse asiassa, tutkimusmotivaatio syntyi uutisista, jotka ennustavat, että lähitulevaisuudessa suuri osa asiakaspalvelusta tapahtuu virtuaaliagenttien avulla (esim. Virtanen 2016; Laaksonen 2016). Näissä uutisissa kerrotaan, että tekoäly korvaa ihmisen asiakaspalvelussa, mutta ei määritellä tarkemmin, kuinka tämä tapahtuu. Tämä työ tuo yhden näkökulman asiaan lähestymiseen, vaikka ei varsinaisesti tätä käsittele. Työ tuo kontribuutiota erityisesti uusien hyötynäkökulmien löytämisessä ja antaa kevyesti teknologiaa ymmärtävälle paremman käsityksen tekoälystä.

1.1 Kehitystyön tarve

Kohdeyrityksen toimialalle ennustetaan suuria muutoksia lähitulevaisuudessa uusien niin sanottujen fintech (Financial technology) startupien ja lainsäädännön muutosten takia. Muutosta vauhdittaa erityisesti vuoden 2018 alussa tuleva PSD2 EU-maksupalveludirektiivi, joka määrää eurooppalaiset pankit avaamaan ohjelmointirajapintoja kolmansille osapuolille, jos asiakas niin haluaa (Saarelainen 2016, ss. 22-23). OpusCapita arvioi webinaarissaan (OpusCapita 2016), että suurimmat finanssiliiketoimintaa häiritsevät tekijät tulevat lohkoketjuteknologioiden, POPPEL-direktiivin, keinoälyn ja koneoppimisen sekä mainitun PSD2-direktiivin myötä. Accenturen Satu Pulkkonen ennustaa Tivi-lehdessä (Saarelainen 2016, s. 23), että isoin vaikutus direktiiveillä tulee olemaan asiakassuhteeseen. Hän jatkaa ennustustaan, että pankeista tulee ikään kuin putkisto, jota muut osapuolet hyödyntävät ja asiakassuhteet häviävät.

Samaan aikaan fintech-startupit janoavat perinteisiltä finanssialan toimijoilta markkinaosuuksia. Esimerkiksi Tivi-lehti maalailee uhkakuvia uutisoimalla menestyneestä startup toimitusjohtajasta: ”*Mies, jota pankit pelkäävät*” (Leskinen 2016, s. 30). Lehdissä esiteltyjä startupeja yhdistää se, että kaikki pyrkivät luomaan parempaa asiakaskokemusta kuin nykyiset toimijat tarjoavat (esim. Saarelainen 2016, ss. 19-25). Edellä mainitut uutiset nostavat esille sitä uhkaa, mikä kohdeyritystä koskee. Digitalisaatio muuttaa pelikenttää ja horjuttaa markkinoilla pitkään hallinneiden yritysten asemaa. Nykypäivänä maantieteellisellä sijainnilla ei ole enää niin väliä, valtio ei pysty suojelemaan omia kotimaisia toimijoitaan, ja teknologioita voidaan kopioida nopeasti. Kilpailussa erottautumiskeinot ovat vähentyneet ja kilpailua on lähinnä siitä, kuka pystyy suorittamaan tehokkaimmin omaa liiketoimintaa ja tekemään järkevimmät päätökset. (Davenport & Harris 2007, s. 28)

Kilpailuetua on lähdetty etsimään kohdeyrityksessä monella eri tavalla, jotta asiakaskokemus ja tehokkuus parantuisi. Yksi visio kilpailuedun lähteestä on ollut vuosien varrella kerääntynyt asiakasdata. Tässä työssä kartoitetaan mitä asiakasviestidatasta voidaan saada irti teknologisten ratkaisujen avulla, jotka liittyvät tekoälyyn. Tekoälyn idea on opettaa teknologia tekemään yhtä älykkäitä päätöksiä kuin ihminen (Merriam-Webster,

2016). Jos tässä onnistuttaisiin täysin, tietotyötä tekevät ihmiset voitaisiin korvata koneilla. Näin kokonaisvaltaisessa tilanteessa ei kuitenkaan vielä olla, joten tekoälyllä mallinnetaan ihmisen kyvykkyyksiä osittain.

1.2 Tutkimustavoitteet ja -kysymykset

Tämän tutkimuksen primääritavoite on tutkia hyötyjä, mitä voitaisiin saavuttaa asiakasviestien automaattisella luokittelemisella eräässä finanssialan yrityksessä. Erityisesti näitä hyötyjä tutkitaan yrityksen asiakaspalvelun päätöksenteon kehityksen ja tehokkuuden näkökulmista. Tutkimuksen sekundäärisenä tavoitteena on kerätä mahdollisimman kokonaisvaltainen kuva aihealueesta, jotta tutkimuksen avulla pystytään tekemään rationaalisia päätöksiä ohjelmistotoimittaja- ja osaamisvalinnoissa sekä sovittamaan automaattinen luokittelu osaksi liiketoimintaa. Tutkimuksessa ovat aihealueeseen liittyvät suomen kieliset erikoissanat käännetty myös englanniksi, jotta niistä voidaan etsiä haluttaessa helpommin lisätietoa.

Tutkimus koostuu teoreettisesta ja empiirisestä osiosta. Kokonaisuudessaan työn tarkoitus on vastata päätutkimuskysymykseen. Päätutkimuskysymys on jaettu neljään aihealueeltaan kapeampaan apukysymykseen. Niiden tarkoitus on tuoda eri näkökulmia päätutkimuskysymykseen. Neljännessä apututkimuskysymyksessä on oletus, että paremmat päätökset lisäävät yrityksen tehokkuutta välillisesti tai välittömästi.

Päätutkimuskysymys:

- Mitä tehokkuushyötyjä voidaan saavuttaa asiakasviestien automaattisella luokittelemisella finanssialan yrityksen asiakaspalvelussa?

Apututkimuskysymykset

- Miten datasta voidaan saada kilpailuetua?
- Minkälaisia ovat asiakasviestit ja miten niitä voidaan käsitellä koneellisesti?
- Miten asiakasviestejä voidaan luokitella automaattisesti?
- Mitä tehokkuuteen ja päätöksentekoon liittyviä hyötyjä voidaan jalostaa asiakasviestidatasta kohdeyrityksen asiakaspalvelussa?

Teoriaosuus vastaa kolmeen ensimmäiseen alakysymykseen. Empiirisessä osiossa haetaan vastausta neljänteen alakysymykseen. Empiriaosuus perustuu asiakaspalveluun liittyvien henkilöiden vastauksiin, jotka kerättiin työpajan avulla. Pohdintaluvussa empirian tuloksia käsitellään teoriaan kytkettynä. Teorian eri apukysymykset toimivat runkona pohdintaluvulle. Yhteenvetoluvussa pohditaan päätutkimuskysymystä isommassa mittakaavassa.

1.3 Tutkimuksen rajaukset

Tutkimuksen tavoite on kartoittaa hyötyjä, jotka ovat toteutettavissa lyhyellä aikavälillä. Täten tutkimus pyrkii rajaamaan pois sellaiset mahdollisuudet, jotka eivät vaikuta olevan toteutettavissa nykyteknologialla. Tiedonlouhinnan teknologia on kehittynyt merkittävästi viime aikoina erityisesti suurten datamäärien käsittelykyvykkyyden suhteen, ja tästä johtuen useita alueen valmiita ohjelmistoja ja osaamisia löytyy markkinoilta. Tämä johtaa tilanteeseen, missä kohdeorganisaation ei tarvitse omata syvällistä tietämystä eri algoritmien ja menetelmien toiminnasta, vaan ne voidaan ostaa ulkopuolisilta alihankkijoilta. Rieuf (2016) esittää selkeän kokonaiskuvan palveluita tarjoavista yrityksistä. Lähtökohtaisesti voidaan ajatella, että dataa voidaan jalostaa ja visualisoida täsmälleen niin kuin halutaan. Tästä johtuen tärkeämpi kysymys on, että mitä datasta halutaan saada irti ja kuinka organisaation ulkopuolisilta palveluntarjoajilta voidaan rationaalisesti ostaa palveluja.

Edellisen kappaleen argumenttien perusteella tämän tutkimuksen laajuudesta jätetään pois tulevaisuuden teknologiat sekä algoritmien ja menetelmien syvällinen esitteleminen. Tutkimuksen näkökulma on pääosin keskittynyt yrityksen sisäisen toiminnan kehittämiseen, eikä tarkoituksena ole tarkastella asiaa asiakkaan näkökulmasta. Välillä tutkimuksessa asiakkaan näkökulma otetaan myös mukaan, koska asiakaspalveluprosesseissa asiakas on aina osa arvon luontiketjua. Lisäksi tähän työhön tehtiin rajaus, että kohdeyrityksen dataa ei oteta mukaan tutkielmaan. Tämän johdosta tutkimuksessa pysytään teoreettisella tasolla ja tehdään alustavaa selvitystyötä käytännön hankkeelle.

1.4 Tutkimusstrategia ja –menetelmät

Tässä aliluvussa perustellaan, minkälaisia tutkimusstrategioita ja –menetelmiä tutkimuksessa on pyritty käyttämään. On kuitenkin huomattava, että tässä tutkimuksessa ei olla täysin yksisuuntaisia käytettyjen otteiden, menetelmien ja päättelymuotojen kanssa. Saaranen-Kauppinen & Puusniekka (2009, ss. 5-6) yleisesti tiivistävät asian, että harvoin tutkimuksista voidaan sanoa yksiselitteisesti, että se edustaa jotain tiettyä valintaa. He jatkavat, että usein lopulta kyse on eri suuntauksien sekoituksista, jossa esitetyt pääsuuntaukset ovat tutkimuksen tavoitteita.

Tämä työ tulee olemaan osa liiketaloustieteitä. Olkkosen (1994) mukaan tällä tutkimusalueella merkittävimmät tieteenkäsitykset ovat hermeneutiikka ja positivismi. Hermeneutiikka painottaa tutkijan merkitystä lopputuloksien kannalta. Tämä johtuu siitä, että tieteenkäsitykseen kuuluu pitkälti tulkintaa ja kokonaisuuksien ymmärtämistä. Tulokset pohjautuvat enemmän kvalitatiiviseen aineistoon ja tutkimus ei välttämättä ole toistettavissa. Tällaisille tutkimuksille ominaista on, että tutkimuksella ei voida koskaan ymmärtää tutkittavaa ilmiötä sen laajuudeltaan ja syvyydeltään, vaan tutkimus on vain pinnan

raapimista (Töttö, 2004). Positivismi on lähellä hermeneutiikan vastakohtaa edellä mainituista näkökulmista. Sen yksi ajatuksista on, että tutkimus tulee olla tutkijasta riippumaton ja toistettavissa. Jos toinen tutkija käyttäisi samoja aineistoja ja menetelmiä, tulisi hänen päätyä vastaaviin tuloksiin. (Olkkonen 1994, s. 34-38.) Tämän työn tutkija ei koe, että tämän tutkimuksen kanssa identtisiin tuloksiin on päästävissä, koska tutkimus riippuu paljon tulkitsijasta. Tämä johtuu siitä, että tutkimus perustuu kvalitatiiviseen aineistoon, jota joudutaan tulkitsemaan suhteessa enemmän kuin kvantitatiivisia aineistoja. Nämä asiat vahvistavat, että tässä työssä käytetään pääosin hermeneutiikan tieteenkäsitystä. Lisäksi tutkimuksen luonne on eksploratiivinen, joka vahvistaa valintaa.

Saunders et al. (2009, s. 61) jakaa tutkimukset kahtia induktiiviseen ja deduktiiviseen lähestymiseen sen mukaan, minkälaista päättelyä tutkimuksessa käytetään. Deduktiivisella päättelyllä tarkoitetaan erikoisempien väitteiden johtamista yleisistä totuuksista. Sen päämääränä on testata ja vahvistaa kirjallisuudessa esitettyjä teorioita. Induktiivisella päättelyllä tarkoitetaan lähestymistapaa, jossa yleisiä väitteitä johdetaan erikoisista tunnetuista tosiasioista. Tämän lähestymistavan päämääränä on uuden teorian luominen. Nämä päättelyt ovat siis erisuuntaisia. Lisäksi tutkimus voi olla myös yhdistelmä näistä kahdesta erilaisesta lähestymistavasta, jolloin lähestymistapaa kutsutaan abduktiiviseksi päättelyksi. (Olkkonen 1994, s. 29.) Tässä tutkimuksessa tarkoituksena on luoda teoreettinen viitekehys kirjallisuuden perusteella. Luodun viitekehyksen pohjalta suunnitellaan ja määritellään aineiston keräykseen liittyvät asiat. Kerättyä aineistoa peilataan lopulta teorian löydöksiin. Tämä tutkimus noudattaa siis deduktiivista päättelyä. (Saunders 2009, s. 61.)

Hermeneuttiseen tieteenkäsitykseen kuuluvia tutkimusmenetelmiä kutsutaan yleisesti case-tutkimuksiksi (tapaustutkimus). Olkkonen (1994, s. 52) korostaa, että tämä on yleisnimi perinteisentyyppiselle hermeneuttiselle tutkimukselle, ja siitä saatetaan käyttää eri lähteissä myös nimitystä toiminta-analyyttinen-, toiminta- tai kliininen tutkimus. Vaikka käsitteet ovat eri nimisiä, on niissä paljon yhteisiä piirteitä, minkä takia niitä ei ole mielekästä erotella toisistaan. Näitä kaikkia eri tapaustutkimuksen nimityksiä yhdistää perinteisen ja kliinisen tutkimuksen piirteet. Perinteisissä tutkimuksissa korostetaan käytännönläheisyyttä, jossa käytännön tietoa systematisoidaan ja jalostetaan. Kliinisissä tutkimuksissa perehdytään asioihin ongelman ytimen ymmärtämisen kautta. (Olkkonen 1994, s. 52-63.) Saaranen-Kauppisen & Puusniekan (2009) mukaan tapaustutkimuksen yksi tärkeimmistä eduista suhteessa muihin tutkimusmenetelmiin on, että tutkimuksen tulokset ovat ymmärrettävissä helposti luodun kontekstin avulla ja täten myös hyödyt saadaan implementoitua vähemmällä vaivalla. Tämä kuvaus sopii tutkimuksen tavoitteisiin.

Muissa lähteissä tapaus- ja toimintatutkimuksen vertaaminen synonyymeiksi ei ole niin mielekästä. Tämän tutkimuksen pääpaino on antaa kohdeyritykselle muutosehdotuksia, mihin suuntaan organisaatiota tulisi kehittää tutkitulla osa-alueella. Muutosehdotukset perustuvat teoriaosuuden kirjallisuuskatselmukseen, jota peilataan kohdeyrityksen nyky-

tilaan. Tutkimuksessa keskitytään erityisesti teknologian tuomiin hyötyihin osana liiketoimintaprosesseja. Eskola & Suonranta (2008) kutsuvat tällaista tutkimusta toimintatutkimukseksi. Saaranen-Kauppinen & Puusniekka (2009, ss. 41-42) ovat samoilla linjoilla, mutta korostavat, että toimintatutkimuksissa otetaan tutkittavat aktiivisesti mukaan osaksi tutkimusta. He jatkavat, että toimintatutkimuksen tarkoituksena on johtaa muutosprosessia osana tutkimustyötä siten, että tutkittavat ja tutkija ovat aktiivisesti yhteistyössä. Näin ei ole tämän tutkimuksen yhteydessä, koska varsinainen muutosprosessin aloitus on tarkoitus aloittaa tämän työn pohjalta. Näiden argumenttien ja pohdintojen perusteella tätä työtä kuvaa parhaiten tapaustutkimus.

Tutkimusaiheesta lähdettiin muodostamaan kuvaa teorian pohjalta ilman ennakko-oletuksia. Aihepiiri oli tutkijalle aluksi melko tuntematon. Teoriaosuuden tiedonhakua suoritettiin TTY:n kirjaston tietokantapalveluiden (IEEE Xplore, Emerald Insight, Science Direct ja Scopus) avulla, ja lisäksi käytettiin Google Scholaria. Tutkimuksen edetessä tiedonhakuun käytettiin enemmän Google Scholar -palvelua, koska tutkija huomasi, että muiden tietokantojen artikkelit ovat myös mukana hakutuloksissa tätä kautta. Lisäksi palvelun tiedonhaun tekoäly on huomattavasti parempi, jolloin saadut hakutulokset ovat relevantimpia. Palvelu suosii hakutuloksissa artikkeleita, joihin on viitattu useimmiten muiden tutkijoiden toimesta. Tästä voidaan tehdä johtopäätös, että korkealle asetetut hakutulokset ovat muiden tutkijoiden toimesta todettu luotettaviksi, käyttökelpoisiksi ja helposti ymmärrettäviksi. Viittausmäärin perustuva suosiminen palvelussa aiheuttaa tosin luonnollisen ongelman, että uusimpia artikkeleita ei löydy, koska niihin ei olla ehditty viittaamaan niin paljon. Tämä asia huomioitiin käyttämällä muita mainittuja tietokantoja mukana tutkimuksessa. Hakusanoina käytettiin tutkimuskysymyksissä esiintyviä termejä suomeksi ja englanniksi. Tärkeimpiä hakusanoja olivat: tekstin luokittelu (text classification), koneoppiminen (machine learning), hahmon tunnistus (pattern recognition) yhdistettynä luvun asiasanoihin. Näiden lisäksi käytettiin paljon muita hakusanoja, koska käsitellyt aiheet ovat hyvin monitieteellisiä. Relevantit lähteet suodatettiin hakutulosten tiivistelmien perusteella. Jäljelle jääneitä dokumentteja käytettiin lähteenä, ja näiden dokumenttien lähdeluettelosta löydettiin lisää aihealueen tutkimuksia. Lopputuloksena näistä muodostui teoreettinen viitekehys aihealueelle.

Tapaustutkimuksissa voidaan käyttää tiedonkeräyksen menetelmänä esimerkiksi haastatteluja, asiakirjoja, havainnointia ja kyselyitä. Yleisesti tapaustutkimuksissa käytetään useita eri lähteitä ja tiedonkeruumenetelmiä. Tällä strategialla mahdollistetaan mahdollisimman laaja ymmärrys tutkimuksen kontekstista ja sovelletuista prosesseista. (Saunders et al. 2009, s. 146.) Mahdollisimman laajaan ymmärrykseen tutkija ajatteli sopivimmaksi menetelmäksi työpajaa. Tässä menetelmässä on etuna se, että tutkittavat voivat luoda yhteisen näkemyksen yrityksen nykytilasta ja tutkimuskysymyksistä, eivätkä saadut vastaukset ole riippuvaisia yksilöiden taustoista ja työnkuvista. Näin myös työpajasta saatu tieto on paljon valmiimpi ja yhtenäisempi tuotos suhteessa siihen, jos tiedonkeräysmenetelmäksi olisi valittu joku yksilöllinen menetelmä.

Työn teoria- ja empiriaosuus ovat jossain määrin toisistaan irrallisia. Tutkimuksen tarkoitus on tehdä näistä osioista varsinainen synteesi pohdinta ja yhteenvetoluvuissa. Tähän strategiaan päädyttiin, koska tutkija koki aihealueen suoranaisen käsittelemisen olevan liian haastava työpajan tutkittaville, jotka eivät ole niin tekniikkatietoisia. Lisäksi kohdeyritys oli tyytyväinen tähän strategiaan.

Taulukko 1: Tutkimuksessa käytetyt menetelmät

Tutkimusfilosofia	Hermeneuttinen
Tutkimuksen lähestyminen	Deduktio
Tutkimusmenetelmä	Toimintatutkimus
Tiedonkeruumenetelmät	Kirjallisuus, työpaja
Tutkimustyyppi	Kvalitatiivinen

1.5 Tutkimuksen rakenne

Työn ensimmäinen luku on tutkimuksen johdanto. Tutkimuksen teoriaosuus alkaa tämän jälkeen ja se koostuu kolmesta eri luvusta (luvut 2, 3 ja 4), jotka vastaavat järjestyksessä kysymyksiin ”miksi”, ”mitä”, ja ”miten”. Lähestymistapa aiheeseen on yleisestä yksityiskohtaisempaan näkökulmaan. Työn toisessa luvussa vastataan kysymykseen ”miksi”. Tähän kysymykseen vastataan tutkimalla, miten datasta voidaan saada kilpailuetua. Datasta voidaan saada kilpailuetua hyvin monella eri tavalla, joten tämä luku on luonteeltaan hyvin yleinen. Kolmannessa luvussa vastataan kysymykseen ”mitä”. Tämän tutkimuksen keskiössä ovat asiakasviestit ja sen ominaisuudet. Täten kirjallisuuden avulla tutkitaan, minkälaisia ovat asiakasviestit ja miten niiden sisältöä voidaan käsitellä automatiikan keinoin. Neljäs teoreettinen luku syventyy kysymykseen ”miten”. Tämän luvun tarkoitus on kertoa yksityiskohtaisesti, miten tutkimuksen keskiössä olevia asiakasviestejä voidaan luokitella automaattisesti. Luvussa syvennytään koneoppimisen ja hahmontunnistuksen keinoihin.

Tutkimuksen empiirinen osio perustuu työpajassa kerättyyn dataan. Työpajan menetelmät ja suoritus on esitelty luvussa viisi. Luvussa kuusi on esitelty työpajasta saadut tulokset. Seitsemäs luku on pohdintaluku, missä peilataan työpajassa saatuja tuloksia suhteessa käsiteltyyn teoriaan. Pohdinta on pyritty rakentamaan pääsääntöisesti samalla tavalla kuin teoriaosuus. Siinä vastataan kohdeyrityksen nykytilanteeseen suhteutettuna uudestaan kysymyksiin ”miksi”, ”mitä” ja ”miten”. Työn varsinainen osio päättyy yhteenvetolukuun (luku 9), jossa tiivistetään tutkimuksen tärkeimmät havainnot yhteen, ja poh-

ditaan, miten tutkimuksen tuloksia voidaan käyttää muissa organisaatioissa. Näiden lisäksi työn lopussa on kerättynä työssä käytetyt lähteet ja liitteet. Lähteiden ja viitteiden perusteella lukija voi syventää omaa tietämystään halutulla osa-alueella.

2. LUOKITTELUN HYÖDYT

Luokitteleminen on ihmisen keino ymmärtää maailmaa paremmin jakamalla se yksinkertaisempiin objekteihin. Se on keino, jolla laitamme asiat järjestykseen toistensa suhteen. (Parrochia & Neuville 2013) Jotta ymmärtäisimme luokittelun hyötyjä paremmin, tulee ensin katsoa asiaa kauempaa. Tätä kautta pääsemme kiinni paremmin siihen, kuinka luokittelu voi tukea liiketoimintaa. Teidon luokitteleminen ei tuo suoranaisesti hyötyjä, vaan hyödyt tulevat vasta paremman ymmärryksen ja jalostetun tiedon käyttämisen kautta (Watson & Wixom 2007).

Jokaisella yrityksellä on erilaiset tarpeet informaation suhteen. Valmis työkaluratkaisu, joka on toiminut toisessa yrityksessä, ei tuo haluttuja tuloksia kaikissa yrityksissä. Yrityksen avoimet kysymykset, käytetty sanasto, säännöt ja lähteet määrittelevät minkälainen työkalu on toimivin. Lisäksi tulee muistaa, että liiketoiminta on se yrityksen osa, joka määrittelee mitä työkalulla halutaan saada aikaiseksi. Liian usein analytiikkahankkeisiin lähdetään teknologisesta näkökulmasta miettimättä liiketoimintahyötyjä. Ilman liiketoiminnan ohjausta tekstin automaattinen luokittelu on resurssien hukkaamista. (Markham et al. 2015.) Myös Laihonen et al. (2013, s. 67) on samoilla linjoilla toteamalla, että liiketoiminta ohjaa teknologiaa ja teknologia mahdollistaa liiketoimintaa.

Yritykset ovat huomanneet, että jokainen kontaktihetki asiakkaan kanssa on mahdollisuus parantaa asiakassuhdetta. Suuri osa näistä kontaktoinneista tapahtuu kirjoitetun tekstin välityksellä. (Zhu et al. 2009, s. 12.) Tässä luvussa tutkitaan mikä on datan rooli yleisesti osana liiketoimintaa, ja miten kirjoitettua asiakasviestidataa voidaan hyödyntää. Tämä kappale vastaa tutkimuksen apukysymykseen: *Miten datasta voidaan saada kilpailuetua?* Tähän kysymykseen vastataan neljän alaluvun avulla. Ensimmäisessä alaluvussa tutkitaan, kuinka datasta voidaan saada liiketoimintaa tukevaa ymmärrystä. Toisessa alaluvussa tutkitaan monipuolisesti, minkälaisilla hyötynäkökulmilla dataa voidaan lähestyä. Kolmannessa alaluvussa keskitytään asiakaspalvelun hyötypotentiaaliin, ja viimeisessä eli neljännessä alaluvussa esitetään, mitä hyötyjä muut organisaatiot ovat saavuttaneet datan avulla.

2.1 Datasta ymmärrykseen

Data yksistään ei tuo arvoa, vaan sitä tulee jalostaa arvoa tuottavaksi resurssiksi. Resurssipohjaisessa ajattelussa organisaation kilpailukyky määrittyy omistettavien resurssien mukaan. (Barney 1996, s. 469) Tästä näkökulmasta lyhyen aikavälin kilpailuetu voidaan muuttaa kestäväksi kilpailueduksi resursseilla, jotka ovat arvokkaita, harvinaisia, vaike-

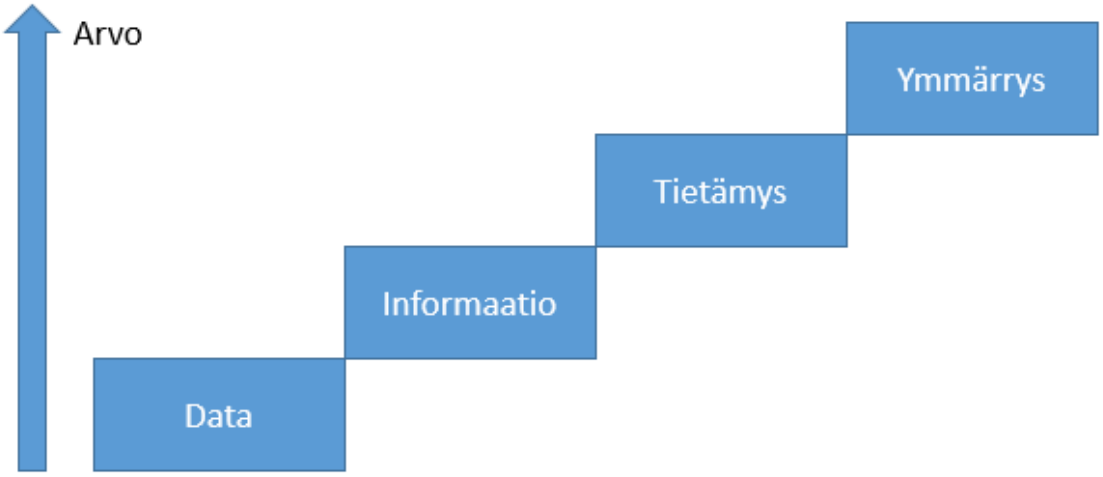
asti kopioitavissa ja myös vaikeasti korvattavissa (Laihonen et al. 2013, s. 24). Asiakasviestidata sopii edellä mainitun kilpailuedun lähteeksi, kunhan siitä saadaan jalostettua ymmärrystä liiketoiminnan tueksi. Salo (2013, s. 138) ennustaa, että tulevaisuudessa datasta tulee yksi merkityksellisimmistä ja yleisesti jopa ainoista kestävän kilpailuedun lähteistä. Hän jatkaa, että hyödynnettynä datasta saatavaa etumatkaa kilpailussa on vaikea kuroa kiinni. Liiketoimintatiedon hallinta (engl. *business intelligence*) voidaan nähdä haasteeseen vastaavana toimintana, joka tuo arvoa tiedosta. (Laihonen et al. 2013, s. 45)

Analytiikka on osa liiketoimintatiedon hallintaa, jonka avulla datasta saadaan normaalia arvokkaampaa kilpailuetua ja ennakoivampaa ymmärrystä. Normaalilla kilpailuedulla tässä viitataan perus- ja erikoisraportteihin, selvityksiin ja hälytyksiin, jotka ovat usein perinteisiä organisaation tiedon johtamisen toimia. Analytiikalla tarkoitetaan edistyneempiä toimia eli tilastollisia ja kvantitatiivisia analyysejä, selittäviä, optimoivia ja ennustavia malleja sekä toiminnan johtamista ja päätöksenteon perustamista tosiasioihin. (Davenport & Harris 2007, ss. 26-27.) Tärkeintä analytiikassa on se, että päätökset tehdään tiedon eikä mututuntuman perusteella. Lisäksi kun toiminta perustuu tietoon, voidaan päätöksentekoa automatisoida. (Davenport et al. 2010, s. 1.) Davenport (2013) mukaan analytiikka on kehittynyt jo kolmannelle tasolle eli niin sanottuun sukupolveen. Hänen mukaan ensimmäinen analytiikan sukupolvi ei hyödyntänyt massadataa (engl. *big data*), toinen sukupolvi otti massadatan mukaan, ja kolmannen sukupolven analytiikalla yritykset eivät pelkästään hyödynnä itse massadataa, vaan se on osa heidän ulkopuolelle tarjottuja palveluja ja tuotteita. Massadata on määritelty tarkemmin myöhemmin.

Watson & Wixomin (2007) mukaan liiketoimintatiedon hallinnasta ei tule hyötyjä automaattisesti, vaan se edellyttää organisaatiolta seuraavia ehtoja. Ensimmäiseksi, heidän mukaansa johdon tulee uskoa ja sitoutua asiaan, tarjota sopivia resursseja ja vaatia tietoon perustuvia päätöksiä koko organisaatiolta, jotta liiketoimintatiedon hallinta olisi onnistunutta. Tiedolla johtaminen tulee saada osaksi organisaatiokulttuuria. Toiseksi, liiketoiminnan strategian tulee olla linjassa tiedon johtamisen strategiaan. Vain strategioiden ollessa linjassa toistensa suhteen, voi liiketoimintatiedon hallinta edistää liiketoiminnan menestymistä. Kolmas ehto liittyy tiedon käyttäjiin. Käyttäjille tulee tarjota tietotarpeisiin sopivat työkalut, koulutus ja tuki. Lisäksi datan tulee olla laadukasta, jotta käyttäjät voivat luottaa siihen. Davenport et al. (2010, ss. 19-21) jakavat liiketoimintatiedon menestyksen ennakkoedellytykset dataan, organisaatioon, johtamiseen, tavoitteisiin ja analyysiin. Nämä ovat pääpiirteittäin linjassa Watson & Wixomin (2007) määrittelemiin ehtoihin. Davenport et al. (2010) lisäksi korostavat, että data on kaiken analytiikan ennakkoedellytys. Ilman dataa ei olisi analytiikkaa, ja toisaalta, ilman hyvää dataa, ei olisi hyvää analytiikkaa.

Dataa muodostuu yrityksen operatiivisesta ympäristöstä. Datan muuntautuminen arvoa tuottavaksi resurssiksi voidaan kuvailla hyvin porrasmallilla, jossa data jalostuu informaatioksi, informaatio tietämykseksi ja tämä edelleen ymmärrykseksi (Laihonen et al 2013, ss. 17-18; Ackoff 1989). Joissain lähteissä (esim. Tuomi 1999) porrasmalliin on

lisätty myös älykkyys ja viisaus, mutta liian tarkka tiedon tasojen erottelu yleisesti sekoittaa kommunikointia, eikä ole tämänkään työn kannalta mielekästä (Nürnberger et al 2009). Porrasmalli on esitetty kuvassa 1. Porrasmalli havainnollistaa ideaa, että data on vähemmän kuin informaatio, informaatio vähemmän kuin tietämys ja tietämys vähemmän kuin ymmärrys. Lisäksi malli havainnollistaa sitä perusideaa, että tarvitaan aina dataa ennen kuin informaatiota tai muita mallin jälkiportaita voi ilmaantua. Ymmärrystä on voinut tulla myös aiempien kokemuksien ja oppien yhteydessä, minkä takia yksilöillä voi olla erilaiset kyvykkyydet saada datasta ymmärrystä. Mallissa datan oletetaan olevan eristäytyneitä yksinkertaisia faktoja. Näitä faktoja syntyy organisaation tapahtumien (esim. transaktiotapahtumien) myötä. Kun nämä faktat laitetaan kontekstiin ja yhdistetään rakenteeseen, syntyy informaatiota. Informaation syntyminen edellyttää datan manipulointia, esittämistä ja tulkitsemista. Kun informaatiolle annetaan tietty tarkoitus tulkintaan, syntyy tietämystä. Tietämyksen syntyminen edellyttää informaation testailua, validointia ja luokittelua. Kun ihmiset tekevät tietämyksen perusteella päätöksiä eri vaihtoehdoista, käyttäytyminen muuttuu viisaaksi ja ymmärretään tulkittavaa ilmiötä. (Tuomi 1999, ss. 105-106; Earl 1994, s. 59)



Kuva 1: Datasta ymmärrykseen. Mukaillen Ackoff (1989) ja Laihonen et al. (2013)

Salon (2013, s. 136) mukaan olemme yhä vahvemmin siirtymässä datavetoiseen talouteen, jossa dataa hyödyntämällä voidaan tunnistaa tai jopa ennakoida markkinoiden asiakastarpeet ja vastata niihin. Hänen ennustuksessaan on kaksi olettamusta: ymmärretään mitä on tapahtunut ja ymmärretään, mitä tulee tapahtumaan. Jotta voidaan ymmärtää mitä on tapahtunut, tulee meidän omata kokonaisvaltainen ymmärrys omasta liiketoimintaympäristöstä. Pelkkä sattumanvarainen datan kerääminen ei tuo vastauksia asiakastarpeista. Dataa tulee kerätä ymmärryksen kautta, jolloin edellä esitelty malli mennään toiseen suuntaan – ymmärryksestä dataan. (Tuomi 1999, ss. 107-108.) Laihonen et al. (2013, s.

46) korostaa myös samaa asiaa esittämällä, että liiketoimintatiedon hallinnan prosessimalli alkaa tietotarpeiden määrittelyllä eikä datan keräyksellä. Tässä prosessivaiheessa selvitetään keskeiset tietotarpeet, vähennetään turhan tiedon keräämistä ja edistetään relevantin tiedon hyödyntämistä. Edistämisellä tarkoitetaan esimerkiksi missä formaatissa dataa hankitaan ja minkälaiseen rakenteeseen se tallennetaan. (Tuomi 1999, s. 107.) Laihosen et al. (2013, s. 46) mallin mukaan seuraava vaihe on tiedon hankinta. Kolmantena vaiheena on tiedon prosessointi ja analysointi. Neljäntenä tiedon jakaminen ja viimeisenä vaiheena tiedon hyödyntäminen ja palautteen saaminen. Tämän tutkimuksen kontribuutio kulkee myös porrasmallia ylhäältä alas. Toisin sanottuna tutkimus tuo ymmärrystä ensin ongelma-alueeseen ja liiketoiminnan tietotarpeisiin, ja näiden perusteella tullaan hankkimaan tulevaisuudessa dataa määrätietoisesti.

2.2 Hyötynäkökulmat

Hyödyllä tämän tutkimuksen yhteydessä tarkoitetaan hyvää tai auttavaa tulosta tai vaikutusta, joka syntyy implementoidusta ratkaisusta (Merriam-Webster 2016). Jotta tiedosta saataisiin kilpailukykyä, tulee tietoa hyödyntää tehokkaasti. Tämä edellyttää, että päätöksiä tekevät henkilöt saavat tarvitsemansa tiedon käyttökelpoisessa muodossa ja oikeaan aikaan (Laihonen et al. 2013, s. 49). Brynjolfsson et al. (2011) tutkimuksen mukaan yritykset, jotka käyttävät dataa päätöksenteon pohjana, ovat keskimäärin 5-6 % kannattavampia kuin kilpailijat. Datavetoinen päätöksenteko edellyttää, että tietoa jaetaan tehokkaasti. Kanavana voi olla määrämuotoinen tietotuote sähköpostin tai kuukausiraportin muodossa, mutta myös jäsentymättömämpi tapa voi olla yhtä toimiva. Tällaisia tapoja voivat olla esimerkiksi käytävä- tai kahvikeskustelut, puhelinkeskustelu tai epämuodollinen tapaaminen. Se miten tieto jaetaan, ei vaikuta saadun hyödyn arvoon. Sen sijaan sillä on suurta vaikutusta, jos yhtenäisiä toimintatapoja ja prosessia ei ole. Silloin tieto voi jäädä täysin jakamatta, jolloin jokainen hankkii tietonsa omista lähteistään ja tekee omat analyysinsä. Samat analyysit saatetaan tehdä useampaan kertaan ja erilaisin tuloksin, mikä ei ole tehokasta toimintaa. (Laihonen et al. 2013, ss. 48-50)

Kilpailuetua tavoitellaan datan avulla kahdella eri pelikentällä tällä hetkellä. Kilpailua on kaukonäköisesti siitä, kuka saa kerättyä dataa toiminnastaan ja toimintaympäristöstään mahdollisimman moniulotteisesti. Kerätään siis hyötypotentiaalia tulevaisuutta varten. Toisella pelikentällä keskitytään siihen, miten jo kerätystä datasta saadaan kaikki mahdollinen irti. (Salo 2013, s. 138.) Analytiikka keskittyy jälkimmäiseen ja sen toiminta voidaan erotella eri näkökulmiin, joita on esitetty kuvassa 2**.** Näkökulmat on esitetty avainkysymysten avulla riippuen siitä, minkälaisiin kysymyksiin analytiikalla haetaan vastausta. Kysymykset ovat aseteltu ajan suhteen: haetaanko hyötyjä ymmärtämällä menneisyyttä, nykyhetkeä vai tulevaisuutta? Toinen jako on tehty innovatiivisuuden suhteen: haetaanko hyötyjä jo tiedetyn informaation perusteella vai haetaanko uusia oivalluksia? Kun päätöksiä tehdään jo tiedetyn informaation perusteella, käytetään nykyistä tietoa vain tehokkaammin. (Davenport et al. 2010, s. 6.) Olennaista on myös, kuka analyyseja tekee.

Jos useampi eri henkilö tekee samoja analyyseja, voi tiedon jalostuksesta kumuloitua kallis prosessi, kuten Watson et al. (2004) osoittaa tutkimuksessaan.

Kuvan 2 ensimmäinen menneisyyden solu on perinteistä liiketoimintaraportointia eikä varsinaisesti analytiikkaa, ja vastaa kysymykseen *mitä tapahtui*. Toinen solu antaa kuvaa reaaliaikaisesta tilanteesta, johon voidaan asettaa säännöillä tiettyjä hälytyksiä, joihin voidaan reagoida. Se vastaa kysymykseen *mitä on nyt tapahtumassa*. Esimerkiksi jokin sensori voi antaa normaalista poikkeavan havaintoarvon. Kolmas solukysymys tarkoittaa yksinkertaista ekstrapolointia, jossa menneisyyden mallien perusteella arvioidaan tulevaisuutta. Tämä vastaa kysymykseen *mitä tulee tapahtumaan.* (Davenport et al. 2010, s. 6.)

Mentäessä kuvan 2 matriisissa oivallusriville, tarvitaan erilaisia työkaluja syvemmän ymmärryksen tuottamiseen. Watson & Wixomin (2004) mukaan tämä on liiketoimintatiedon hallinnalle luonnollinen kehityssuunta, kun aikaa kuluu ja organisaatiossa on halua kehittää toimintaansa. Luokitteleminen on yksi työkaluista. Muita työkaluja ovat esimerkiksi visualisointi, korrelaatiot, regressio ja ennustaminen (Lagus 2000). Oivalluksien tekeminen menneisyydestä vaatii tilastotieteellisiä aktiviteettejä ja se vastaa kysymykseen *minkä takia, ja miksi joku asia tapahtui*. Oivallukset nykyhetkeen antavat neuvoa *mitä teemme seuraavaksi* -kysymykseen. Esimerkiksi mikä finanssialan tuotetarjous kiinnostaa asiakkaita tällä hetkellä? Oivallukset tulevaisuuteen tulevat ennustamisen, optimoinnin ja simuloinnin tekniikoilla, jotta pystytään luomaan parhaimmat tulokset tulevaisuudessa. Tämä solu vastaa kysymykseen *mikä on parasta tai pahinta mitä voi tapahtua*. (Davenport et al. 2010, ss. 6-7.)

Kuva 2: Avainkysymykset hyötyihin, mukailtu Davenport et al (2010, s. 7)

Edellä mainittuihin kysymyksiin saadaan vastauksia eri analyyttisillä lähestymistavoilla. Lähestymistavat voidaan jakaa neljään Banerjee et al. (2013) mukaan: deskriptiiviseen eli kuvailevaan (engl. descriptive analytics), diagnostiiviseen eli selittävään (engl. diagnostic analytics), prediktiiviseen eli ennustavaan (engl. predictive analytics) ja preskriptiiviseen eli ohjailevaan analytiikkaan (engl. prescriptive analytics). Kuvailevan analytiikan avulla haetaan usein vastausta menneisyyden kysymyksiin. Sen tarkoitus on selittää ilmiöitä eri mittauksien avulla, jotka ilmaisevat relevantteja ulottuvuuksia datasta. Diagnosoiva analytiikka hakee usein vastausta siihen miksi jokin asia tapahtui. Tämä myös viittaa menneisyyden näkökulmaan. Prediktiivinen analytiikka pyrkii ennustamaan tulevaa vastaamalla kysymykseen *mitä todennäköisesti tulee tapahtumaan*. Ohjailevan analytiikan fokus on myös tulevassa, ja sen avulla pyritään tekemään mahdollisimman tietoperusteisia päätöksiä tulevan suhteen. Ohjaileva analytiikka on näistä edistynein ja samalla haasteellisin toteuttaa, koska se käyttää hyväksi kuvailevan, selittävän ja ennustavan analytiikan keinoja. (Banerjee et al. 2013.) Kaisler et al. (2014) toteavatkin, että usein tarvitaan eri näkökulmaisia analytiikan keinoja liiketoimintaongelmien ratkaisemiseen.

Näiden näkökulmien ymmärtäminen voi tuntua yksinkertaiselta, kun datan määrä on sen verran pieni, että ihminen pystyy hallitsemaan sitä. Ongelmat nousevat esille vasta, kun datan ominaisuudet täyttävät massadatan (engl. *big data*) määritelmät, jolloin datojen väliset suhteet katoavat (Salo 2013, s. 65; Bayer & Lanyen 2012). Alun perin massadataan yhdistettiin volyymi, vauhti ja vaihtelevuus, mutta nykyään ilmiötä kuvaileviin termeihin on lisätty myös todenmukaisuus (Salo 2013, s. 21; Fertier et al. 2016). Kaisler et al (2013) kertoo, että koko massadatan käsite lähti alun perin siitä, että dataa ei pystytty käsittelemään perinteisillä metodeilla ja työkaluilla. Tämä edelleen johti siihen, että hyötyjä ei saatu kaivettua datasta kahden eri syyn takia. Ensinnäkin, teknisesti ei pystytty rakentamaan sopivia systeemejä käsittelemään dataa tehokkaasti ja toiseksi, ei ollut työkaluja löytämään relevanttia tietoa päätöksenteon tueksi. Nyt tilanne on muuttunut kuitenkin huomattavasti massadata termin kehitysvaiheesta, ja Salon mukaan (2013, s. 52) ”lähes kaikki ovat pelissä mukana” viitaten nykyhetkeen. Hän esittelee kirjassaan useita erilaisia onnistuneita ratkaisuja Amazon, EMC, Google, HP ja IBM -yritysjäteiltä. Salon (2013) ja Russom (2011) raporteista käy ilmi, että hyvin useassa tilanteessa valtavaa laskentatehoa vaativaan käsittelyyn kannattaa harkita ulkopuolisten toimijoiden tarjoamia ratkaisuja.

Vaikka analytiikka on suuren megatrendin asemassa, sen hyötynäkökulmia tulee peilata sen aiheuttamiin kustannuksiin kuten muissakin investointilaskelmoinneissa. Ja näiden laskelmien tulee olla positiivisen puolella, jotta niitä kannattaa toteuttaa. Tätä asiaa peräänkuuluttaa Salo (2013, s. 143). Duda et al. (2001, s. 3) nostavat taas esiin kustannukset, mitkä aiheutuvat väärästä analyyttisestä johtopäätöksestä. Vääriä johtopäätöksiä tulee varsinkin silloin, kun automaatioastetta nostetaan ja ihmiset eivät ole osana asiantuntijasysteemiä. Vääriä johtopäätöksien mahdollisuutta on käsitelty enemmän luvussa 4.4.4. Jos esimerkiksi asiakasviesteistä tehdyn johtopäätöksen perusteella annetaan asiakkaalle

väärä vastaus, voi aiheutuva kustannus olla pieni. Jos taas sairaalassa analysoidaan potilaan sairaus väärin ja hän saa vääriä lääkkeitä, voi tästä aiheutuva kustannus nousta huomattavan korkeaksi. (Duda et al 2001, s. 3.)

2.3 Asiakaspalvelun hyötypotentiaali

Asiakaspalvelulla on tiettyjä ominaisuuksia mitkä voivat tehdä siitä erityisen mielenkiintoisen kohteen analytiikkahankkeille. Ensinnäkin, yleisesti voidaan sanoa, että asiakaspalvelu on ensimmäinen taho organisaatiossa, jolle raportoidaan havaitut ongelmat tai puutteet organisaation palveluissa ja tuotteissa. Asiakaspalvelijan intresseihin havaitut ongelmat eivät välttämättä kuulu, mutta ne voivat kuulua organisaation muiden osien mielenkiinnon kohteisiin. (Salo 2013, ss. 29-30.) Yksi asiakaspalvelija saa tietoonsa useasti vain yksittäisiä tapauksia, mutta yhdistelemällä useamman asiakaspalvelijan havaintoja voidaan tehdä merkittäviä löytöjä. Tällaisten ongelmien tunnistaminen voi tuoda suurta hyötyä, kun pystytään esimerkiksi estämään ongelmien laajentuminen. Bijmolt et al. (2010) esittävät tästä näkökulmasta, että asiakkaat tulisi nähdä osana yrityksen strategiaa, jossa he luovat arvoa yhdessä yrityksen kanssa (engl. *value co-creation*). He kritisoivat tutkimuksessaan, että usein yritys ja asiakkaat nähdään erillisinä toimijoina, joista vain yritys tuottaa arvoa. He jatkavat, että kun asiakkaat saadaan osaksi yhteistä arvonluontia, he ovat useammin taipuvaisia suosittelemaan yritystä ystävilleen (engl. *word-of-mouth*). Wangenheim & Bayón (2007) laskelmien mukaan word-of-mouth –markkinoinnissa onnistuva yritys nostaa 40 % asiakkaan elinkaariarvoa. Toinen asiakaspalveluun heijastuva luonteenpiirre liittyy reaaliaikaisuuteen. Sieltä voitaisiin saada reaaliaikaista tietoa tuotteiden ja palveluiden ominaisuuksista, kampanjoiden onnistumisesta sekä käyttäjien mieltymyksistä. (Salo 2013, ss. 29-30) Nämä asiat ovat lähes samoja, mitä haetaan sosiaalisten medioiden analyyseillä (esim. Nadeem 2012).

Asiakaspalvelu sisältyy hetkiin, joissa asiakas ja organisaation edustaja kohtaavat fyysisesti tai virtuaalisesti. Nämä hetket ovat ratkaisevia asiakkaan kokemalle laadulle. Palveluiden johtamisessa näitä tilanteita voidaan kutsua *totuuden hetkiksi* tai epäonnistuessaan *kurjuuden hetkiksi*. (Grönroos 1998, s. 68; Manning & Reece 2004, s. 371) Totuuden hetki on organisaation näkökulmasta tilaisuus, jossa voidaan osoittaa asiakkaalle tässä ja nyt palvelun laatu. Jos vastausta ei saada toimitettua nopeasti, asiakkaan kokema epätyytyväisyys kasvaa suhteessa odotusaikaan (Davis & Heineke 1998). Asiakkaan näkökulmasta tämän hetken epäonnistuminen voidaan korjata vain uudella totuuden hetkellä. Tämä on huomattavasti hankalampaa, joten jokainen totuuden hetki tulisi olla onnistunut. Pahimmassa tapauksessa epäonnistunut totuuden hetki voi johtaa asiakkaan menetykseen. Useassa tutkimuksessa ollaan osoitettu, että asiakkaan tyytyväisenä pitäminen on merkittävästi halvempaa kuin uusien löytäminen (katso Beaujean et al 2006). Epäonnistunutta totuuden hetkeä voidaan yrittää korjata esimerkiksi ottamalla yhteyttä asiakkaaseen tilanteen korjaamiseksi, mutta tämä on tehottomampaa suhteessa hyvin hoidettuun totuuden hetkeen. Organisaation johdolla ei ole mahdollisuutta valvoa jokaista totuuden

hetkeä, joten analytiikka voi tuoda tähän asiaan kontribuutiota. (Grönroos 1998, ss. 68-69)

Tämän tutkimuksen yhteydessä asiakkaat ovat se osapuoli, joka tekee aloitteen totuuden hetkeen. Näissä hetkissä on mahdollista viedä keskustelua myös asiakasviestin sisältämän aihealueen ulkopuolelle. Eichfeld et al. (2006) tutkimuksen mukaan jopa 25 % uusista tuloista pystytään tuottamaan finanssialalla asiakaspalvelupuheluissa, joihin lähtökohtaisesti asiakas ei ole soittanut ostaakseen uusia tuotteita. Tutkimuksen aikaan verkkoviestintä ei ole ollut niin suosittua, mutta oletetusti tämä voisi toimia myös tällä alueella. Erityisesti seuraamalla kokonaisvaltaisesti asiakaskäyttäytymistä voitaisiin ennustaa, mitä asiakas tarvitsee seuraavaksi. Esimerkiksi Khan et al. (2015) tutkimuksessa pystyttiin ennustamaan liki 90 % tarkkuudella tuleeko asiakas irtisanoutumaan. Tutkimus perustui sisäisistä tietojärjestelmistä saatuun asiakaskäyttäytymisdataan. Jos tällaista tietoa saadaan esille asiakaspalvelijoiden totuuden hetkiin, voidaan palvelun laadulla mahdollisesti vaikuttaa asiakaspoistumaan. Beaujean et al. (2006) tutkimuksessa kysyttiin asiakkailta, mikä tekee heidän totuuden hetkistä positiivisia. Yleisimmät vastaukset olivat samoja kuin edellä käsitellyt: proaktiivisten ehdotusten saaminen tarpeiden mukaan sekä se, että asiakaspalvelijat ottivat asiakkaiden historian ja tarpeet huomioon. Samassa tutkimuksessa myös selvitettiin, mikä tekijät vievät kurjuuden hetkiin. Suurimpana tekijänä oli myynti tai palvelu, joka ei vastannut tarpeisiin. Näistä johtuen on myös asiakaspalvelijoiden päätöksenteko oltava datalähtöistä, jota voidaan edesauttaa analyyttisellä järjestelmällä.

Asiakasviesteistä tai asiakaspalvelutilanteeseen muista järjestelmistä jalostettu analytiikka voi tarjota hyvin monipuolisia hyötyjä palvelun laadun eri osa-alueeseen. Grönroos (2000, s. 117) jakaa yleisesti palvelun laadun osa-alueet viiteen, jotka pätevät myös asiakaspalveluun:

1. Konkreettinen ympäristö. Tällä tarkoitetaan käytettyjä toimitilojen, laitteiden ja materiaalien miellyttävyyttä sekä asiakaspalvelijoiden ulkoista olemusta.
2. Luotettavuus. Tällä tarkoitetaan palvelun kykeneväisyyttä tarjota oikea vastaus heti ensimmäisellä kerralla täsmällisesti ja virheettömästi.
3. Reagointialttius. Tällä tarkoitetaan asiakaspalvelijoiden halukkuutta auttamaan asiakkaita. Tämä sisältää myös palvelun läpinäkyvyyden eli kerrotaan missä tilassa heidän pyytämä palvelu on ja koska he saavat vastauksensa.
4. Vakuuttavuus. Tämä tarkoittaa, että asiakkaat saadaan luottamaan organisaatioon ja tuntemaan olonsa turvalliseksi. Työntekijät ovat aina kohteliaita ja osaavia.
5. Empatia. Tällä tarkoitetaan, että yritys ymmärtää asiakkaiden näkökulman ja toimii heidän etujensa mukaisesti.

Edellä mainittujen palvelun laadun osa-alueiden onnistumista voidaan mitata SERVQUAL-menetelmällä. Tässä menetelmässä näitä viitta eri osa-aluetta kuvataan yleensä 22 erilaisella kysymyksillä, joihin vastaajat ilmoittavat, kuinka he kokevat asian arvioimalla asteikolla yhdestä seitsemään. (Grönroos 2000, s. 117) Uuden analytiikka-hankkeen implementoimista osaksi järjestelmiä voidaan testata esimerkiksi ennen ja jälkeen mittauksilla SERVQUAL-menetelmällä.

2.4 Hyödyt kirjallisuudessa

Analytiikan ja luokittelun soveltamisen kautta saatuja erilaisia hyötyjä on rajattomasti, joten kyse on enemmän luovuudesta ja innovatiivisuudesta. Jokainen organisaatio voi valjastaa menetelmät omalla tavallaan osaksi organisaatiota tuottamaan arvoa. Myös kirjallisuudessa on tutkittu monipuolisesti eri tapauksia, kuinka massadatasta on saatu arvoa ja miten siinä ollaan onnistuttu. Tähän kappaleeseen on kerätty kirjallisuudessa esiintyneitä tapauksia hyötytaulukkoon 2. Taulukkoon on kerätty automaattisella luokittelemisella saavutettuja hyötyjä, jonka pohjana on ollut suuri massadata.

Kuten aikaisemmin jo todettiin, sama ratkaisu ei sellaisenaan toimi kahdessa eri organisaatiossa. Hyötytaulukon tarkoitus ei ole tarjota valmiita ratkaisuja, joista voidaan valita mieluisin, vaan tarkoitus on antaa lukijalle ja kohdeyritykselle ideoita, joita voidaan suhteuttaa omaan organisaatioon. Tarkoituksena on myös laajentaa lukijan näkökulmaa kohti monipuolisempaa näkemystä aiheesta. Hyötytaulukko koostuu viidestä eri sarakkeesta. Ensimmäisessä sarakkeessa on ilmaistu epästrukturoitu data, josta haluttuja hyötyjä on lähdetty hakemaan. Toisessa sarakkeessa on ongelma, joka on motivoinut hyödyntämään analytiikkaa. Kolmannessa sarakkeessa on hyödyt, mitä ongelman ratkaisemisella voidaan saavuttaa. Neljännessä sarakkeessa on tutkimuslähde, jotta ratkaisusta voidaan saada lisätietoa. Huomioitavaa on, että osa sarakkeiden merkinnöistä on tutkijan omia tulkintoja ja ne eivät välttämättä täsmää todelliseen tilanteeseen.

Taulukko 2: Massadatasta jalostettuja hyötyesimerkkejä kirjallisuudessa.

Epästrukturoitu massadata	**Ongelma**	**Hyödyt**	**Lähde**
Asiakasviestit	Samat viestit kirjoitetaan uudestaan	Valmiiksi kirjoitettujen vastauksien käyttäminen	(Zhu et al. 2009, ss. 12-13)
Asiakasviestit	Viestit menevät asiakaspalvelijoille, jotka eivät osaa vastata	Osaavan asiakaspalvelijan löytäminen tehokkaammin	(Zhu et al. 2009, ss. 12-13)

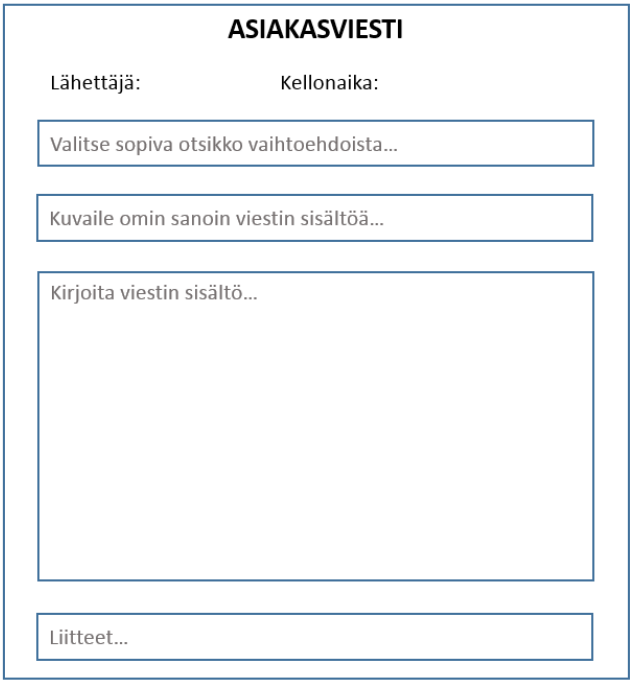
Asiakasviestit	Asiakaspalvelijat kopioivat mallivastauksia ja ohjaavat asiakkaita itsepalveluun	Automatisoitu ja sitä kautta edullisempi asiakaspalvelu	(Zhu et al. 2009, ss. 12-13)
Sähköpostit	Roskapostin käsittely vie aikaa. Sähköpostikansion hallitseminen vie aikaa	Tehokkuuden nostaminen tiedon helpomman saatavuuden avulla	(Zhu et al. 2009, ss. 11-12)
Sosiaaliset mediat	Ei tiedetä, miten ihmiset kokevat yrityksen tuotteet ja palvelut	Vastaamalla asiakastarpeisiin pystytään kehittämään tuotteita ja palveluita	(Liu 2012)
Sosiaaliset mediat	Negatiivista puheista seuraa lumipalloefektinä lisää negatiivisia puheita, jotka vahingoittavat organisaation imagoa	Pyritään kääntämään negatiivinen puhe positiiviseksi imagon parantamiseksi	(Chmiel et al 2011)
Potilashistoria ja muut sisäisest tietojärjestelmät	Osa potilaista ei oikeasti tarvitsisi niin laadukasta hoitoa terveyden puolesta	Hoidetaan niitä potilaita, jotka oikeasti hoitoa tarvitsevat	(Ray et al. 2015; Kent 2014)
Sisäiset tietojärjestelmät	Ei tarkkaa tietoa asiakaspoistumaan vaikuttavista tekijöistä	Vähennetään asiakaspoistumaa	(Khan et al. 2015; Douglas 2011)
Markkinaraportit, julkiset uutiset	Tuntematon markkina	Uusien asiakkaiden löytäminen	(Markham et al 2015)
Hallituksen miljoonat dokumentit	Kasvumahdollisuus	Uusien asiakkaiden löytäminen	(Markham et al 2015)
Sisäiset tietojärjestelmät	Ei tietoa asiakaskäyttäytymisestä	Paremman asiakasarvon tuottaminen	(Kent 2014)

Lentotiedot, sosiaalinen media	Terrorismi ja biologiset uhat lennoilla	Lentoturvallisuuden parantuminen	(Kent 2014)

3. ASIAKASVIESTI JA SEN KÄSITTELEMINEN

Asiakasviesteillä tarkoitetaan tämän tutkimuksen yhteydessä luonnollisella kielellä kirjoitettuja viestejä, joita asiakkaat lähettävät kohdeyritykselle verkkopalvelun avulla. Tässä luvussa perehdytään asiakasviestien rakenteeseen ja niiden sisältöön. Lisäksi luku tutkii yhden asiakasviestin käsittelyä koneellisesti. Yleisesti tiedonlouhinnan menetelmät perustuvat tavalla tai toisella perinteiseen tiedonhakuun, joten myös sitä käsitellään lyhyesti. Luku pyrkii vastaamaan tutkimuksen alakysymykseen *”Minkälaisia ovat asiakasviestit ja miten niitä voidaan käsitellä koneellisesti?”*

Kohdeyrityksen tapauksessa asiakasviestit koostuvat viidestä eri pääelementistä. Ensimmäinen pääelementti on viestien sisältämä vapaamuotoinen kenttä, jossa asiakas kuvaa omin sanoin kontaktin syyn. Toinen pääelementeistä on itse kuvailtu otsikon aihe, joka tarkentaa valittua yläotsikkoa. Yläotsikko valitaan listasta, jonka vaihtoehdot ovat määriteltynä ennakkoon. Tämä on kolmas pääelementti. Neljäntenä elementtinä on asiakasviestin liitetiedostot, jotka eivät ole pakollisia, joten ne eivät kaikissa viesteissä ilmene. Viides elementti on taustalla pyörivä tekninen data, johon liittyvät esimerkiksi lähetysaika, lähettäjän IP-osoite ja asiakasnumero. Pääelementtejä on havainnollistettu kuvan 3 avulla.



Kuva 3: Asiakasviesti ja sen pääelementit

3.1 Data ja sen tyypit

Datasta on eri lähteissä eri sanoilla kuvailtuja määritelmiä, mutta yleisesti kuvailut tukevat toistensa määritelmiä. Nürnberger et al. (2009) määritys datalle vaikuttaa olevan samansuuntainen, mitä muualla kirjallisuudessa tapaa (esim. Laihonen et al. 2013): data on joukko irrallaan olevia symboleita, havaintoja tai faktoja joilla ei ole syvällisempää merkitystä kuin niiden olemus. Hyvä data on ennakkoedellytys kaikelle analytiikalle. Sen tulee olla siistiä tarkkuuden ja formaatin suhteen, jotta voidaan olla hyviä analytiikassa (Davenport et al. 2010, ss. 19-23). Esimerkiksi Googlen pääekonomi on sanonut, että ”*Googlella ei ole parempia malleja, meillä on vain enemmän dataa*” viitaten siihen, että heidän analytiikka on yksi maailman parhaimmista (katso Erhardt 2015). Tarvitaan siis erittäin paljon dataa siistissä muodossa, jotta analytiikkahankkeista saadaan hyötyjä irti.

Dataa voidaan tyypitellä monesta erilaisesta näkökulmasta. Tässä kappaleessa sitä on kuvailtu kahdesta erilaisesta. Ensimmäinen vastakkainasettelun näkökulma liittyy datan dynaamisuuteen. Tässä tyyppijaottelussa data voidaan jakaa staattiseen ja virtaavaan dataan. Staattisella datalla viitataan dataan, joka pysyy muuttumattomana ja jonka käsitteleminen on siten helpompaa. Helpomman hallittavuuden takia perinteiset tiedonlouhinnan sovellukset keskittyvät tähän tyyppiin enemmän kuin virtaavaan dataan (Salo 2013, s.24). Virtaavalla datalla tarkoitetaan dataa, joka on jatkuvassa muutostilassa. Tällaisessa tapauksessa tietokanta kokee usein lisäyksiä, päivityksiä ja poistoja. Tämän työn yhteydessä vanha asiakasviestihistoria edustaa staattista dataa. Siihen kohdistuu muutoksia vain uusien asiakasviestien saapuessa. Kun uusi asiakasviesti saapuu vanhaan staattiseen dataan, voi se vaikuttaa asiakasviestin luokittelun sääntöihin, jota tässä työssä tutkitaan. Asiakasviestihistorian ollessa satojatuhansia, ellei miljoonia viestejä, on hyvin todennäköistä, että uusi asiakasviesti ei muuta luokittelusääntöjä. Tämä tarkoittaa, että kyseisen asiakasviestin piirteet ovat jo ennestään tuttuja. Jos asiakasviestejä taas on kymmeniä tai satoja, joudutaan todennäköisesti opettelemaan uusia luokittelun sääntöjä. Asiakasviestien automaattista luokittelemista ja oppivuutta on käsitelty enemmän luvussa 4. (Salo 2013, ss. 61-62.)

Virtaavan datan näkökulmasta tämän työn yhteydessä voidaan pitää yksittäistä saapuvaa asiakasviestiä, jolle voidaan suorittaa toimenpiteitä, ennen kuin se liittyy osaksi staattista asiakasviestihistoriaa. Asiakkaalle esimerkiksi voidaan ehdottaa kysymykseen vastausta jo ennen kuin hän lähettää koko viestiä. Tälle esimerkille on ennakkoedellytys, että staattisesta datasta on jo olemassa käsitys, joten asia on monimutkaisempi. Asiasta tulee entistä monimutkaisempi, kun tehtävänä on analysoida epästrukturoitua virtaavaa dataa. (Venkatesan et al. 2016) Liikkuvan datan analysointiin tarvitaan äärimmäisen tehokkaita ja tarkoitukseensa optimoituja ”*Stream Computing*” –ratkaisuja. Esimerkiksi IBM on erikoistunut näihin kahdella eri ratkaisullaan: InfoSphere Streams ja IBM PureData for Analytics. Avoimen lähdekoodin Hadoop sopii myös ongelmaan (Venkatesan et al. 2016).

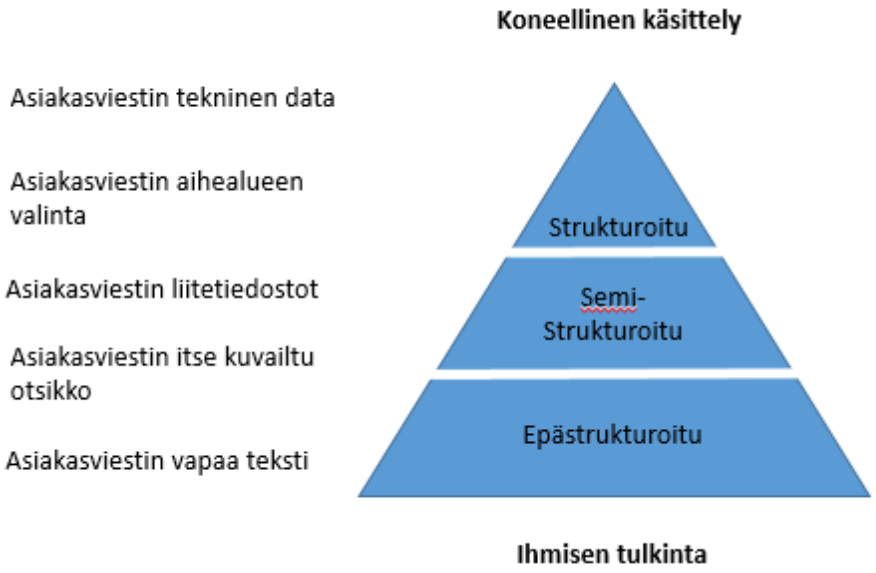
Staattiseen dataan tarvitaan taas tehokkaita tiedon keräämiseen, louhimiseen ja analysointiin soveltuvia ratkaisuja, joita IBM:llä ovat esimerkiksi IBM InfoSpehere BigInsights ja IBM Social Media Analytics. (Salo 2013, ss. 61-62)

Toinen datan tyyppiä kuvaileva näkökulma liittyy datan rakenteeseen. Datan rakenteen avulla on helppo havainnoida luonnollisen kielen luonnetta ja sen haasteita automaattiseen käsittelyyn. Datan rakenteella kuvataan tapaa, miten se on tallennettuna tietokantaan. Päämääränä on, että dataa voidaan käyttää mahdollisimman tehokkaasti. (Encyclopedia Britannica) Yleisesti datan rakenne jaetaan strukturoituun (engl. *structured data*), semi-strukturoituun (engl. *semi-structured data*) ja epästrukturoituun dataan (engl. *unstructured data*) (esim. Li et al 2008, Salo 2013). Strukturoidulla datalla tarkoitetaan dataa, jolla on rakenne. Epästrukturoitu data on taas tämän vastakohta ja viittaa dataan, jolla ei ole rakennetta. Näiden kahden ääripään väliin mahtuu dataa, jolla on osittain rakennetta ja osittain ei. Tällaista dataa kutsutaan semi-strukturoiduksi. Datojen suhdetta on verrattu taloustieteistä tutumpaan Pareton-jakaumaan 85/15, jonka mukaan 85 % datasta on luonteeltaan enemmän epästrukturoitua ja vain 15 % on strukturoitua. (Salo 2013, s. 25-26) Määrämuotoista eli strukturoitua dataa koneet osaavat käsitellä paremmin. Epästrukturoidun datan käsittelyyn tarvitaan usein ihmisen tulkintaa. (Blumberg & Atre 2003) Tästä näkökulmasta katsottuna tämä työ voidaan nähdä aiheena, kuinka epästrukturoidusta datasta saadaan strukturoitua.

Jos edellä mainittuja datarakenteita verrataan kohdeyrityksen pääelementteihin, voidaan huomata pääelementtien piirteiden muodostavan rakenteellisia eroja. Ensimmäinen pääelementti, asiakasviestin vapaamuotoinen kenttä, täyttää selvästi epästrukturoidun datan piirteet. Asiakas voi kirjoittaa kenttään mitä tahansa, eikä kentällä ole mitään rajoitteita. Asiakkaan tarkoitus on, että hän tulee ymmärretyksi asiakaspalvelijan toimesta, joka tietysti antaa tietynlaisen rajoitteen mitä hän kirjoittaa ja miten asian ilmaisee. Toisaalta järjestelmässä on jo olemassa muutamia oletuksia ja rajoituksia, minkä takia tämä elementti täyttää myös semistrukturoidun datan piirteitä. Järjestelmä olettaa, että asiakkaan varsinainen ongelma tulee tässä kohdassa ilmi, joten tämä kenttä on voitu nimetä esimerkiksi <asia>-elementiksi. Järjestelmä myös todennäköisesti rajoittaa käyttäjää kirjoittamaan asiansa tietyn merkkimäärän sallimissa rajoissa.

Toinen pääelementti, omin sanoin kuvailtu otsikko, täyttää osittain strukturoidun datan piirteet. Lähtökohtaisesti sen tarkoitus on kuvata ensimmäisen elementin sisältöä muutamilla sanoilla. Koska sanoja on vähemmän, on erilaisten otsikoiden määrä pienempi kuin erilaisten sisältöviestien. Lisäksi järjestelmä tunnistaa tämän kentän todennäköisesti <otsikko>-elementtinä. Kolmas pääelementti, aihealueen valinta ennakkoon määritellystä valikoimasta, täyttää strukturoidun datan piirteet. Järjestelmä tietää, että tässä kentässä tulee vastaus esimerkiksi <aihealue> nimiseen elementtiin, johon se antaa vastausvaihtoehdot. Kun vastausvaihtoehtoja on rajoitettu määrä, niiden kvantitatiivinen käsitteleminen on helpompaa. Neljäs pääelementti liittyy asiakasviestien liitetiedostoihin. Yleisesti kuvia ja muita multimediatiedostoja voidaan pitää semi-strukturoituna datana (Salo 2013,

s. 25). Viidennen pääelementin (esim. lähetysaika, IP-osoite) data on suunniteltu avustamaan koneellista käsittelyä. Tästä johtuen se on strukturoitua dataa. Datan rakennetta suhteessa kohdeyrityksen asiakasviesteihin on havainnollistettu kuvassa 4.



Kuva 4: Datan rakenne suhteessa asiakasviestin pääelementteihin. Pyramidi havainnollistaa datan määrää. Mukaillen Salo (2013) ja Blumberg & Atre (2003)

3.2 Viestien sisältö

Useat tutkijat sanovat, että tekstien koneellinen käsitteleminen voi olla erittäin haastavaa, jos ihmiset saavat kirjoittaa niihin mitä tahansa haluavat. Helpoin tapa ratkaista tätä ongelmaa on rajoittaa aihealuetta (engl. *domain*). (Katso Kent 2014) Viestien sisällön rajoittamista kuvaa hyvin avoimen aihealueen (engl *open domain*) ja suljetun aihealueen (engl *closed domain*) lähestymistapa (Agirre et al. 2009). Paukkeri (2012, ss. 9-10) mukaan moni luonnollista kieltä käsittelevä sovellus on suunniteltu toimimaan erityisesti tietyllä aihealueella.

Avoimessa aihealueessa ihmisen voivat kirjoittaa lähtökohtaisesti mitä tahansa ja mistä tahansa. Tällaisia paikkoja ovat esimerkiksi sosiaaliset mediat. Ongelmana näissä on, että käytetty sanasto voi olla niin laajaa, että sääntöpohjaisille koneille ei voida opettaa kaikkia sääntöjä kohtuullisen ajan puitteissa. Toinen ongelma on, että samoilla sanoilla tarkoitetaan useampia eri asioita. (McCarthy et al. 2004.) Suljetussa aihealueessa asia on toisin, joten sääntöpohjaiset koneet toimivat tällä alueella paremmin, koska käytetyssä

sanastossa on enemmän säännönmukaisuuksia (Agirre et al. 2009; Paukkeri 2012). Tällaista aihealuetta edustavat esimerkiksi kirjautumisen vaativat finanssipalvelut, joiden takana lähetetään viesti. Tällöin kirjautunut henkilö on suuremmalla todennäköisyydellä tietoinen jo sanastosta, mitä finanssipalvelussa käytetään. Hän voi yksilöllisesti viitata esimerkiksi tiettyyn finanssituotteeseen palvelun nimellä tai käyttää aihepiirille tunnusomaista sanastoa. Kun samaa sanastoa käytetään avoimen aihepiirin yhteydessä, sen merkitystä ei välttämättä tunnisteta. Esimerkiksi sana *nostaa* finanssialalla yleisesti viittaa tilanteeseen, jossa asiakas saa rahaa pankkikortin avulla automaatista. Avoimessa aihealueessa sillä yleisemmin viitataan muihin asioihin.

Viestien sisältöä voidaan kuvata metadatan avulla. Metadatalla tarkoitetaan tietoa tiedosta. Tämä tieto voi olla käyttäjille sellaista, jota he eivät havaitse. Toisaalta se voi olla hyvinkin näkyvissä, kuten asiakasviestin otsikko. Esimerkiksi asiakasviestin kolmas pääelementti eli otsikon valitseminen määritellystä listasta, on vain tietoa tiedosta. Se vain tiivistää viestin sisällön eikä tuo mitään uutta tietoa viestin kannalta. Tällaisissa tapauksissa tieto on suunnattu lähtökohtaisesti koneiden käsittelyä varten. Toisena metadataesimerkkinä, joka ei ole palvelun käyttäjille näkyvissä on edellisessä luvussa käytetyt <aihealue> ja <otsikko>-elementit. Niiden tarkoitus on ilmaista datan rakennetta HTML-merkkauksella ja täten palvella järjestelmän toimivuutta. Tässä tapauksessa ensisijaisesti selaimen toimivuutta.

3.2.1 Viestien sisällön koneellinen ymmärtäminen

Viestien sisältö rakentuu hyvin subjektiivisesti sen perusteella, kuka sen kirjoittaa, mistä hän on kotoisin ja miten hän kokee asian olevan. Näitä asioita tulee ottaa huomioon, kun koitetaan siirtää ymmärrystä koneelle tekstin pohjalta. Koneellista ymmärrystä tutkii erityisesti tietokonelingvistiikan (engl. *Computational linguistics*) ja luonnollisen kielen käsittelyn (engl *Natural language processing*, NLP) tieteenhaarat. (Paukkeri 2012, ss. 7-11.) Luonnollisella kielellä viitataan kommunikointitapaan, jota ihmiset käyttävät toistensa kanssa ymmärtääkseen toisiaan. Keskustelu voi tapahtua tekstin tai puheen muodossa, eikä mikään taho ole rajoittamassa kommunikointitapaa. Pääasia on, että viestin vastaanottaja ymmärtää, mitä viestin lähettäjä haluaa kertoa. Ongelmat ja rajoitteet syntyvät vasta, kun teknisen järjestelmän tulisi ymmärtää opittujen sääntöjen perusteella, mitä ihminen haluaa viestiä. Nämä säännöt tulee ihmisen opettaa järjestelmälle.

Luonnolliselle kielelle on ominaista, että sitä ei voida koskaan mallintaa tarkoiksi säännöiksi, koska kieli on jatkuvan muutoksen alla. Ihmiset jatkuvasti venyttävät ja taivuttelevat sanoja sekä lauserakenteita uusiin muotoihin, joita kone ei aikaisemmin ole tiennyt. Lisäksi sanoja lainataan jatkuvasti toisista kielistä ja kokonaan uusia sanoja keksitään. Vaikka kone ei kaikkea tiedä, voimme tilastotieteen ja todennäköisyysteorian avulla löytää todennäköisesti merkityksellisimmät osat tekstistä. (Manning & Schütze 1999, ss. 3-4) Näiden avulla mahdollistamme paremman koneen ja ihmisen välisten ymmärryksen.

Ymmärryksen kanssa tulee myös ongelmia, kun luonnollisessa kielessä ilmaantuu kirjoitusvirheitä. Nämä mainitut ongelmat liittyvät vain kielen esiintymismuotoon, joka ei yksistään ratkaise kommunikaatio-ongelmia. Chowdhuryn (2003) mukaan kielen esiintymismuodon lisäksi luonnollisen kielen ymmärtäminen vaatii ajatusprosessin ymmärtämistä ja maailmantietämystä.

Ajatusprosessin ymmärtämiseen ei riitä, että seuraa pelkästään yksittäisiä sanoja tai lauseita, vaan pitää katsoa laajempaa esiintymiskontekstia. Toisin sanoen tarvitsee ymmärtää sanojen ja lauseiden semantiikkaa (engl. *semantics*). Semantiikan loogiseen rakentamiseen tarvitaan kirjoittajalta ja vastaanottajalta yhtenäistä kieli- (engl. *grammar*) ja lauseoppia (engl. *syntax*). Lisäksi asiaa hankaloittaa yksittäisten sanojen tai useammista sanoista koostuvien termien monimerkityksellisyys ja synonyymisyys. Samat sanat voivat tarkoittaa eri asioita (esimerkiksi kuusi). Toisaalta eri sanat voivat tarkoittaa samaa asiaa (kännykkä – matkapuhelin). (Ikonomakis et al 2005)

Kolmantena ymmärtämisen edellytyksenä on maailmantietämys. Jokainen ihminen tietää, että koira ja kissa edustavat eläinlajeja, mutta tietokone ei tätä lähtökohtaisesti ymmärrä. Tekstiä käsitteleville systeemeille teksti edustaa vain merkityksetöntä syötettä eikä maailmantietämystä ole. Sanojen oikeiden merkitysten opettaminen on lähes mahdotonta. Näille systeemeille sanat ’ei’ ja ’hei’ ovat lähes yhtä samanlaisia kuin esimerkiksi sanat ’kissa’ ja kissat’. Näille sanoille ihminen osaa automaattisesti luoda suuria assosiaatioeroja, mutta koneelle ne ovat lähtökohtaisesti vain uniikkeja erilaisia tekstisyötteitä. (Paukkeri 2012, s. 15-16)

Näiden kolmen ongelmakentän ratkaisuja tutkii erityisesti luonnollisen kielen käsittelyn tieteenhaara, jonka menetelmiä on käsitelty tekstin automaattisen luokittelemisen prosessimallin luvussa 4.4.1. Tieteenhaaran käytännön sovellukset ovat tosin keskittyneet ratkomaan englannin kielen ongelmia ja parantamaan tämän kielen koneellista ymmärrystä. Tämä on ymmärrettävää, kun huomioi englannin kielen aseman globaalisti. Sen sijaan muiden marginaalisempien kielten, kuten suomen kielen, ongelmiin ei ole kehitetty niin paljon sovelluksia. Englannin kielen ylivoimaisuuden osoittaa W3C-organisaation arviot internetissä käytetyistä kielistä (W3Techs 2016). Laskennan mukaan internetin kaikista sivuista on 52,8 % englanniksi, 6,4 % venäjäksi, 5,6 % saksaksi, 4,9 % espanjaksi ja 4 % ranskaksi. Lisäksi huomion arvoista on, että maailman väkirikkaimman maan Kiinan (Kiinan kielet yleisesti) prosenttiosuus on vain 2,0 %, toiseksi suurimman eli Intian (hindin) osuus 0,1 % ja suomen kielen osuus 0,3 %. Tämä laskenta kuvaa tilannetta erinomaisesti, koska internetin dokumentteja käytetään usein osana sovelluksien kehittämistä (esim. Joachims 1996). Onhan internet maailman suurin julkisesti saatavilla oleva dokumenttivarasto, jonka aineisto on valmiiksi elektronissa formaatissa. Mainituille kielille, joita ei käytetä niin paljon, on pyritty rakentamaan kieli- ja aihealueriippumattomia sovelluksia, jotka perustuvat lähtökohtaisesti vain tilastotieteisiin ja todennäköisyysteoriaan. Näiden toimivuutta tukee koneoppiminen, jonka perusajatus on oppia iteraatiokier-

rosten kautta merkityksellisiä osia tuntematta välttämättä kieltä ollenkaan. Tällaiset menetelmät tulevat tarpeen maailmassa, jossa käytetään jopa 7000 erilaista kieltä. (Paukkeri 2012, s. 94) Näistä kielistä jopa 2000 on Koskenniemi et al. (2012) mukaan kuolemassa sukupuuttoon, koska ne eivät täytä digiajan vaatimuksia.

3.2.2 Suomen kielen erityishaasteet

Tilastotieteiden ja luonnollisen kielen käsittelyn tutkimukset ovat usein keskittyneet englannin kieleen. Tästä johtuen myös tutkimuksissa esitetyt, testatut ja kehitetyt menetelmät, kuten sanaluokittimet ja ontologiat, eivät ota kantaa, kuinka ne toimivat suomen kielellä. Korenius et al. (2004) toteaa, että englannin kieli on suhteellisen helppo prosessoitavaksi verrattuna suomeen. Hän perustelee, että suomen kielessä on rikas taivutusjärjestelmä ja arvioiden mukaan substantiiveilla on jopa 2000, adjektiiveilla 6000 ja verbeillä yli 12000 erilaista taivutusmuotoa, joka tuo suuren teknisen haasteen. Taivutusmuotojen lisäksi sanoja johdetaan ja yhdistellään paljon suomen kielessä. Sanakirjoissa hakusanoista perussanoja (esim. kirja) on noin 10-15 %, johdoksia (esim. kirjasto) 20-30 % ja yhdyssanoja (esim. kirjastokortti) 60-70 %. Tähän kun lisätään, että suomen sanajärjestys on hyvin vapaa, voidaan todeta, että suomen kielen erityispiirteet ovat haasteellisia kieliteknologioiden kannalta. (Koskenniemi et al. 2012)

Vaikutuksia osoittaa hyvin Koehn et al. (2003) tutkimus, jossa analysoitiin Euroopan Parlamentin aineistoa, joka on käännetty 11 eri kielelle. Tutkimuksen mukaan kaikilla eri kielillä on sama informaatio. Tutkimuksen tulokset on esitetty taulukossa [2]. Taulukosta voidaan huomata, että suomen kieli on hyvin tiivistä sanamäärien suhteen. Esimerkiksi ranskan kielisessä dokumentissa sanoja oli 326 000 kun suomen kielessä sanoja oli 203 000. Huomattavaa on, että ranskan kielessä uniikkeja sanoja oli 16 400, kun suomen kielessä niitä oli vastaavasti 37 000. Tämä johtuu erityisesti edellä mainituista asioista.

Taulukko 3: Eri kielien vertailua sanamäärien suhteen. Arvot perustuvat eri kielillä kirjoitettuun dokumenttiin, joka sisältää saman informaation. Tekstissä viitatut arvot lihavoitu. (mukailtu Koehn, 2003)

Kieli	Sanojen määrä	Uniikkien sanojen määrä	Uniikkien sanojen määrä suhteessa kaikkiin sanoihin
ranska	**326 k**	16,4 k	5,0 %
kreikka	322 k	23,0 k	7,1 %
espanja	309 k	18,4 k	6,0 %

portugali	303 k	18,1 k	6,0 %
englanti	299 k	12,6 k	**4,2 %**
hollanti	299 k	17,6 k	5,9 %
italia	291 k	18,3 k	6,3 %
saksa	274 k	22,9 k	8,4 %
tanska	272 k	20,4 k	7,5 %
ruotsi	268 k	21,9 k	8,2 %
Suomi	203 k	37,0 k	**18,2 %**

Taulukosta [2] voidaan huomata suomen kielen haasteellisuus ja englannin kielen helppous uniikkien sanojen suhteen, joka edelleen korreloi, kuinka helppoa kieltä on käsitellä koneellisesti. Mitä vähemmän uniikkeja sanoja on, sitä ”strukturoidumpi” kieli on. Suomen kielessä melkein joka viides sana on sellainen, joka ei ole aikaisemmin esiintynyt tekstissä. Englannin kielessä vastaava luku on noin joka 25. Huomioiden tämä tutkimus ja se, että englanti on internetin käytetyin kieli, ymmärrämme miksi sovelluskehittäjät ovat kiinnostuneet erityisesti kehittämään englantia tukevaa kieliteknologiaa. Arppe (2008) kuvaa tätä asiaa hyvin: *”Kun esimerkiksi englantia varten pystyy kehittämään yksinkertaisen kielenkäsittelyohjelmiston kuten oikolukijan käytännössä listaamalla ja kompressoimalla yleisimmät satatuhatta sanaa, suomen kohdalla pitäisi samaa tekniikkaa noudattaen listata, jos ei satoja niin vähintään kymmeniä miljoonia eri sanamuotoja, jotta vastaava oikolukija olisi yhtä kattava.”*

Huolimatta tehtävän haasteellisuudesta, moni yritys on lähtenyt kehittämään suomen kieltä tukevaa kieliteknologiaa 1980-luvulta alkaen kuten Lingsoft Oy, Kielikone Oy, Connexor Oy ja Gurusoft. Myös joitain vapaasti käytettäviä sovelluksia löytyy kuten Turku BioNLP Groupin kehittämät työkalut. Osa edellä mainituista yrityksistä perustuu onnistuneisiin yliopistotutkimuksiin, jotka ovat johtaneet liiketoimintaan. (Arppe 2008) Suomen kieltä koskevia tutkimuksia löytyy keskitetysti Turku BioNLP Groupin kotisivuilta (Turku BioNLP Group 2016). Suurin osa tosin sivuilla olevista tutkimuksista ei ole saatavilla tämän tutkimuksen resursseilla.

Muutamia relevantteja tutkimuksia on saatavilla, joista Korenius et al. (2004) tutkimus osoittaa, että suomen kieltä voidaan hyvin automatisoida luokittelutehtävien kannalta. Heidän tutkimuksessaan 84 567 erilaisen termin dokumenttikokoelma saatiin supistettua 1500 merkitykselliseen termiin lemmaus ja PCA-menetelmillä (luvussa 4.4 lisää menetelmistä). Samoilla linjoilla on myös Ginter et al. (2013) kertoen, että heidän testaama

morfologinen analyysi pystyi palauttamaan sanat perusmuotoon 88-98 % tarkkuudella. Nämä tulokset ovat lupaavia. Luvussa 4.4 on perehdytty yksityiskohtaisemmin, mitä näillä tuloksilla tarkoitetaan. Suomen kieltä voidaan käsitellä automaattisesti myös kieliriippumattomilla sovelluksilla, joita tutkii esimerkiksi Paukkeri (2012) väitöskirjassaan. Hän kertoo tutkimuksessaan tuloksista niukasti, mutta esimerkiksi Solorio et al. (2004) tutkimuksen mukaan myös kieliriippumattomilla menetelmillä päästään hyviin tuloksiin.

3.3 Tiedonhaku

Tekstin luokitteleminen koneellisesti on nykyään tiedonhaun, koneoppimisen ja hahmontunnistuksen aikaansaannos (Sebastiani 2002). Hänen mukaan tiedonhaun tekniikoita käytetään kolmessa eri tekstiluokittelun vaiheessa, joita esitellään luvussa 4:

1. Luokittelussa luodaan sanahakemistoja opetteludatan perusteella.
2. Tiedonhaun tekniikoita hyödynnetään piirteiden laskennassa.
3. Tiedonhaun menetelmillä arvioidaan luokittelun toimivuutta.

Tiedonhaulla (engl. *information retrieval*) tarkoitetaan prosessia, jonka tavoitteena on tietotarpeiden tyydytys (Alkula 2000, s. 24**)**. Prosessi koostuu tiedon esittämisen, etsimisen ja tallentamisen vaiheista silloin, kun ihminen esittää tiedonhakutehtävän (Ingwersen 2002, s. 49). Tiedonhaku voidaan suorittaa täsmäävillä sanoilla (engl. *keyword searching*) tietokantaa vastaan, ja saada vastauksena kyselyyn täsmääviä dokumentteja tai niiden osia. Tämä on näkökulma, mihin yleisesti tutut internetin hakukoneet perustuvat. Monimutkaisempi tiedonhaku liittyy tekoälyn tarjoamiin teknologisiin mahdollisuuksiin ja se ottaa huomioon, että ihmisellä ei välttämättä ole selkeää käsitystä omista tietotarpeistaan. (Encyclopedia Britannica; Alkula 2000, s. 25.) Tekoälyn tarkoituksena on tulkita ihmisen kyselyitä mahdollisimman älykkäästi, jotta kyselyihin tulisi mahdollisimman kiinnostavat vastaukset. Tässä näkökulmassa otetaan huomioon muun muassa luvussa 3.2.1 esitettyjä ongelmia liittyen viestisisältöjen koneelliseen ymmärtämiseen. Myös näitä menetelmiä hyödynnetään moderneissa hakukoneissa. Esimerkiksi Googlen hakukone ehdottaa käyttäjälle hakusanoja jo kyselyn kirjoitusvaiheessa aavistellen, mistä käyttäjä voisi olla kiinnostunut. Samalla se myös huomioi oletettavia kirjoitusvirheitä tarjoamalla oikein kirjoitettuja hakukyselyitä.

Haut kohdistuvat tietokantaan, joka koostu tietueista. Tekstitietokannoissa tietueita kutsutaan myös dokumenteiksi, jotka edelleen rakenteellisesti voidaan jakaa eri kenttiin. Esimerkiksi tämän luvun alussa esiteltyjä pääelementit voidaan nähdä yhden asiakasviestidokumentin kenttinä, ja kenttiä voidaan merkitä esimerkiksi <otsikko>-rakenteella, kuten luvussa 3.1 tehtiin. Useat hakukoneet mahdollistavat tiedonhaun eri tasoille ja eri kenttiin. Googlen kuvahaulla voidaan esimerkiksi hakea kissa nimistä <img>-kenttää, ja tämän työn yhteydessä useasti käytetty IEEE Xplore -elektroninen tietokanta mahdollistaa haun 25 erilaiseen dokumentin tietokenttään. Nämä IEEE Xploren -kentät sisältävät lähinnä

metadataa varsinaisesta dokumentin sisällöstä. Pienissä ja erityisesti strukturoiduissa tietokannoissa haku voi olla jouhevaa käymällä kaikki metadatakenttien sisältämät merkkijonot lävitse. Asia on toisin, jos haetaan kaikkien dokumenttien kaikista sanoista tiettyä hakulauseketta. Tällaisessa tilanteessa ei ole mielekästä käyttäjän kannalta odottaa dokumentti toisen perään, kun tiettyä hakulauseketta etsitään.

Tarvitaan tehokkuutta nostavia työkaluja. Käänteistiedostot eli hakemistot (engl. *index*) nostavat hakujen tehokkuutta. Hakemistojen idea on koota dokumenteissa esiintyneet sanat aakkosjärjestykseen, ja kirjata sanan perään dokumentit, missä sana on esiintynyt. Laajoissa tietokannoissa käytetään lisäksi käänteistiedostojen hakemistoa eli niin sanottua sanakirjatiedostoa (engl. *dictionary file*) tehostamaan hakua. (Järvelin 1995, s. 98; Alkula 2000, s. 23-24.) Myös asiakasviestien automaattisen luokittelun yhteydessä (luku 4.4.1) esitetyt esikäsittelyn menetelmät ovat sovellettavissa tehokkuuden nostamiseen.

4. ASIAKASVIESTIEN AUTOMAATTINEN LUOKITTELU

Edellisen luvun yhtä asiakasviestiä voidaan pitää dokumenttina, joka koostuu tekstistä ja siihen yhdistetystä metatiedosta kuten otsikosta ja lähetysajasta. Tässä luvussa tutkitaan, kuinka näistä alkioista voidaan luokitella loogisia kokonaisuuksia. Luku lähestyy aihetta eri käsitteiden määritelmien avulla ja esittelemällä automaattisen luokittelun prosessimallin. Luku vastaa tutkimuskysymykseen: *Miten asiakasviestejä voidaan luokitella automaattisesti?* Selvennettäköön vielä, että automaattisella luokittelemisella tarkoitetaan automatiikan hyödyntämistä osana luokitteluprosessia, eikä täysin autonomista systeemiä. Tässä luvussa syvennytään ohjatun oppimisen järjestelmiin kohdeyrityksen ongelmakuvailun takia.

Tämän kappaleen tarkoitus on tutkia monipuolisesti mitä eri menetelmiä voidaan käyttää automaattisen luokittelujärjestelmän rakentamiseen, joten suurta arvoa ei anneta menetelmien yksityiskohtaiselle esittelemiselle. Niistä löytää lisätietoa kappaleissa käytetyistä lähdemateriaaleista.

4.1 Määritelmä ja lähitermit

Kroeze et al. (2002, s. 254) määrittelee luokittelemisen perinteiseksi älykkääksi toiminnaksi, jossa monimutkaista ilmiötä jaetaan pienempiin ja johdonmukaisiin osiin, luokkiin, rakenteisiin tai yksiköihin, jotta ilmiöstä saataisiin helpommin ymmärrettävä ja ohjattava. Scott & Marshall (2009) määrittelevät luokittelemisen prosessinäkökulmasta, jossa mittaaminen on avaintekijä. Mittaamalla monimerkityksellisten asioiden ominaisuuksia voidaan löytää yhteisiä tekijöitä, joiden perusteella asiat jaetaan edelleen ryhmiksi tai luokiksi. Zhu et al. (2009, ss. 5-6) julkaisu kuvaa luokittelemista ihmisen toiminnan kautta. Kun jotain uutta asiaa näytetään ihmiselle, hän intuitiivisesti sovittaa asiaa omaan tietämykseensä sen perusteella, mitä tuttuja elementtejä hän tunnistaa. Manning & Schütze (1999, s. 612) määrittelevät luokittelemisen tehtäväksi, jossa universaaleja objekteja järjestetään kahteen tai useampaan luokkaan tai kategoriaan. Luokat voivat olla joko binäärisiä, jatkuvia tai kategorisia, ja yksittäinen alkio voi kuulua yhteen (engl. *single label*) tai useampaan luokkaan (engl. *multilabel*) (Sebastiani 2002).

Samoihin määritelmiin perustuu myös automaattinen luokitteleminen, jota voidaan kutsua myös koneelliseksi luokittelemiseksi. Jotta luokitteleminen voi tapahtua automaattisesti, täytyy systeemille joko opettaa tietämystä aihekokonaisuudesta (engl. *supervised learning)* tai systeemi hankkii tietämyksensä omiin algoritmeihin perustuen (engl. *unsupervised learning) (*Zhu et al. 2009, s. 5). Näiden ääripäiden väliin mahtuu myös vahvistettu oppiminen (engl. *semi-supervised learning*), joka hyödyntää molempien

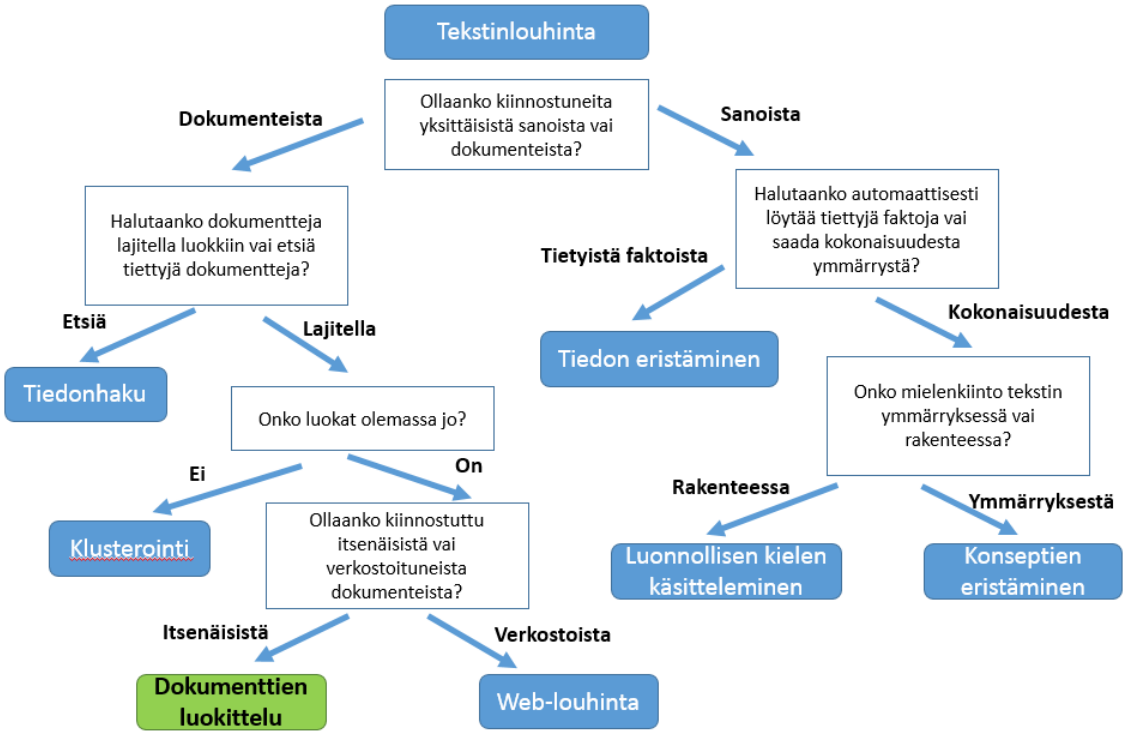
edellä mainittujen menetelmien oppeja (Zhu 2005). Kokonaisuudessaan automaattinen luokitteleminen on hyvin monitieteellinen aihealue ja siihen törmää useissa eri lähdekirjallisuuksissa. Kirjallisuuden kirjoa nostaa se, että automaattisella luokittelemisella voidaan tarkoittaa eri asioita. Tällä voidaan tarkoittaa automaattista luokittelemista ennakkoon määriteltyihin luokkiin, mikä on tämän tutkimuksen päänäkökulma. Toisekseen tällä voidaan tarkoittaa eri luokkien löytämistä. Kolmannekseen tarkoitus voi olla löytää luokkia ja sijoittaa tekstit näihin automaattisesti, mikä tunnetaan paremmin klusterointina. Neljännekseen voidaan tarkoittaa mitä tahansa toimintaa, jossa tekstiä laitetaan ryhmiin käyttäen esimerkiksi edellisten luokitusten tunnuspiirteitä. (Sebastini 2002) Automaattista luokittelemista on tutkittu useissa eri tutkimusalueiden kirjallisuudessa, koska eri alueilla käytetään samoja yleiskäyttöisiä algoritmeja, työkaluja ja menetelmiä. Luokittelemiseen viitataan ainakin seuraavissa tiedekirjallisuuksissa:

- Tiedonlouhinta (Aggarwal & Zhai 2012)
- Tietämyksen muodostaminen (Norton 1999)
- Tekstinlouhinta (Aggarwal & Zhai 2012)
- Koneoppiminen (esim Sebastiani 2002)
- Hahmontunnistus (esim Webb et al 2011)
- Luonnollisen kielen käsitteleminen (esim. Manning & Schütze 1999, s. 575-607)
- Tekoäly (Barr et al. 1989)
- Analytiikka, erityisesti ennustava analytiikka (esim. Hammond & Varde 2013)

Moni edellä mainituista termeistä ovat aihealueiltaan päällekkäisiä. Jos tarkastelee termien määrittelyjä tieteellisissä julkaisuissa, on eroja vaikea huomata. Jotta lukija pystyy hahmottamaan kokonaiskuvan, on seuraavaksi yritetty ehkä hiukan keinotekoisesti eritellä termejä hierarkiseen ja rinnakkaiseen järjestykseen. Esitetyt selitykset yksinkertaistavat asiaa ja niissä on painotettu erityisesti sanojen semantiikasta johtuvia näkökulmaeroja. Termien välisiä suhteita on lisäksi visualisoitu kuvassa 5.

Tiedonlouhinta (engl. *Data mining*) painottaa tiedon suurta määrää, josta voidaan löytää mielenkiintoisia ja käyttökelpoisia suhteita ja hahmoja. Sen tavoite on pyrkiä etsimään datasta monipuolisesti erilaisia löytöjä ja sen takia sitä kutsutaan myös englanniksi *Knowledge Discovery in Database* eli tietämyksen muodostamiseksi. Nämä termit ovat siis lähellä toistensa synonyymeja. (Encyclopedia Britannica 2016.) Luokitteleminen on vain yksi tiedonlouhinnan menetelmä. Laguksen (2000) mukaan muita tiedonlouhinnan menetelmiä ovat tekstien yleiskuvan ja tiivistelmien luominen, trendien ja muutosten tunnistaminen ajan kuluessa, riippuvuuksien ja odottamattomien datayhteyksien löytäminen, päätösten tukeminen sekä datan visualisointi. Tiedonlouhinnan yksi työkaluista on tekstinlouhinta (engl. *text mining*), jota voidaan kutsua myös dokumenttien louhinnaksi tai tekstianalytiikaksi. Se soveltaa tiedonlouhinnan tekniikoita pääasiassa epästrukturoituun tekstidataan. (Dörre et al. 1999.) Tyypillisimpiä sovelluskohteita voi tarkastella kuvasta 5.

Tutkimuksen pääpaino on dokumenttien louhinnassa, joka on merkattu kuvaan vihreällä. Tiedonlouhintaa voidaan myös tehdä manuaalisesti, mutta se ei aina ole mielekästä, käyttötarkoituksesta riippuen. Tästä johtuen koneita opetetaan toimimaan kuten ihminen, joten niille opetetaan tekoälyä (engl. a*rtificial intelligence*). Jotta kone voisi toimia läheskään kuten älykäs ihminen, tulee sen ymmärtää ihmisten kommunikointia (luonnollisen kielen käsitteleminen), oppia omasta toiminnastaan (koneoppiminen) sekä tunnistaa säännöllisyyksiä ja hahmoja (hahmontunnistus). Luonnollisen kielen (engl. *natural language processing*) ymmärtäminen painottaa ihmisen ja tietokoneen välistä ymmärrystä. Jotta ymmärrys voisi tapahtua, tulee koneen generoida ja ymmärtää ihmisten käyttämää kommunikaatiota. (Cammack et al. 2006.) Koneoppimisessa (engl. *machine learning*) suurin merkitys on sillä, että käytetty tietokoneohjelma oppii kokemuksensa perusteella. Oppimisella tarkoitetaan, että ohjelmalla on kyky muuttaa omaa toimintaansa uuden saadun informaation perusteella. (Miquel Porta 2014; Encyclopedia Britannica.) Hahmontunnistuksessa (engl. *pattern recognition*) pääpaino on sillä, että pystytään löytämään erilaisia sääntöjä, hahmoja tai kaavoja (engl. *patterns)* käytetystä data-aineistosta tiettyjä ominaisuuksia mittaamalla (Encyclopedia Britannica, Butterfield & Ngondi 2016).



Kuva 5: Tekstinlouhinnan osa-alueet esitettynä päätöspuun avulla. Mukaillen Miner et al. (2012, s. 33)

Kuvan 5 esitystapa päätöspuun ja kysymysten avulla on yksi tapa esitellä automaattisen luokittelun suhdetta muihin menetelmiin. Erilaisia havainnollistavia kuvaajia löytyy useampia ja niiden sisällön asettelu riippuu lähestymistavasta (vertaa esim Miner et al. 2012, s. 38). Osa kuvaajista on ristiriitaisia toistensa suhteen. Esimerkiksi edellisessä kappa-

leessa määriteltiin, että luonnollisen kielen käsittelyä tarvitaan dokumenttien luokitteluun. Tässä puumallissa luokittelu ja luonnollisen kielen käsitteleminen soveltuvat täysin eri menetelmiin. Kroeze et al. (2002) kritisoi aihealueen termien monimuotoisuutta tutkimuksessaan. He muun muassa esittelevät kahdeksan erilaista määritelmää tekstinlouhinnalle.

4.2 Koneoppiminen ja hahmontunnistus

Asiakasviestien luokittelemista voidaan lähestyä monesta edellä esitetyltä abstraktiotasolta ja näkökulmasta, kuten luvussa 4.1 osoitettiin. Tämän työn näkökulmasta mielenkiintoinen kysymys on erityisesti ”*kuinka uusia asiakasviestejä voidaan luokitella vanhan viestidatan perusteella*?” Hotho et al. (2005) mukaan oli menetelmä mikä tahansa, tekstin luokitteleminen alkaa aina opetusdatan avulla, josta luodaan luokittelumalli. Toisin sanottuna luokittelu voi syntyä vain, kun on olemassa tietty konteksti mihin uuden alkion piirteitä peilataan. Täten asetettu kysymys voidaan muotoilla uudelleen ”*kuinka uusia asiakasviestejä voidaan luokitella?”* Samalla vanha viestidata luo sen kontekstin, johon uuden alkion tulee sijoittua. Koska halutaan luokitella uusia asiakasviestejä suhteessa vanhojen viestien luokkiin, on kyse ohjatun oppimisen menetelmistä. Tästä johtuen ohjatun oppimisen prosessimallia käsitellään tarkemmin luvussa 4.4. Tätä prosessia voidaan tehdä myös manuaalisesti, mutta tekstin luokitteleminen voi olla erittäin työläs tehtävä ajan ja vaivan suhteen. Tehokkaammin tätä voidaan tehdä automaattisesti koneoppimisen ja hahmontunnistuksen menetelmillä, jolloin manuaalityöntekijät voivat keskittyä haastavampiin tehtäviin. (Miner et al. 2002, ss. 882-884.)

Oppimisprosessi perustuu lähtöaineistoon eli korpukseen (engl. *corpus*), jonka perusteella luokittelumallit rakennetaan (Bird et al. 2015). Lähtöaineiston rajallisuuden vuoksi tulosten tarkkuus ei koskaan saavuta täydellisyyttä. Ääretöntä dataa ei voi olla olemassa. Varsinkin kun luokittelun kohteena on jokin reaalimaailman asia, ei jatkuvia arvoja voida sensoreilla mallintaa tarkasti. Tekstinluokittelun yhteydessä tätä ongelmaa ei tosin ole, koska teksti on jo lähtökohtaisesti diskreetissä tilassa. Lisäksi käytetyn kielen sanojen lukumäärä on rajallinen. Tekstin luokittelun yhteydessä yleisempi ongelma liittyy datan puhtauteen. Datan puhtauteen vaikuttaa usein kirjoitusvirheet, joita voi olla suhteellisen paljon, kuten Carvalho & Curto (2014) osoittavat tutkimuksessaan. Toisaalta täydellisyyttä ei myöskään haluta saavuttaa, koska se usein vaatii tarkoitukseen sopimattoman pitkän laskentatehokkuuden ja –ajan. Tästä johtuen tulee luokitteluun valita tapauskohtaisesti sopivimmat menetelmät, jotka vastaavat tehtävään riittävän hyvin. Käytännön sovelluksissa algoritmien valinta on aina tasapainoilua tarkkuuden ja laskenta-ajan välillä. (Lagus 2000; Pilaszy 2005.) Käytäntö on myös osoittanut, että usein ongelma ei ole puutteellinen määrä dataa, vaan että datan perusteella ei pystytä vastaamaan liiketoiminnan ongelmiin (Miner et al. 2002). Pilaszy (2005) tiivistää kolme yleisintä käytännön ongelmaa liittyen koneoppimisen ja hahmontunnistuksen sovelluksiin tekstin luokittelemisessa:

- Sanojen sijaintia ei yleensä huomioida, joten oleellinen osa kontekstia voi jäädä huomioimatta
- Uudenlaisia dokumentteja tulee järjestelmän tuotantovaiheessa, joita ei oltu osattu huomioida kehityksessä
- Erilaisten sanojen lukumäärä nousee dokumenttien lukumäärän suhteen, eikä kaikkia sanoja voida huomioida kehitysdatan perusteella

Kirjallisuudessa koneoppimisen ja hahmontunnistuksen menetelmien kehitykseen useimmiten käytetty korpus on Reuters-kokoelma. Tämä kokoelma sisältää uutisia, jotka on luokiteltu niiden sisällön mukaan. Esimerkiksi Reuters-21578 -kokoelma sisältää 21578 uutisdokumenttia, jotka on jaetultu 114 eri luokkaan (Gao et al. 2014). Tutkimuksissa on eri menetelmillä saatu erilaisia tuloksia oikeiden luokitusten suhteen, mutta yleisesti Reuters-korpuksen kanssa ohjatun koneoppimisen luokittelutarkkuus on yli 80 % (esim Khan & Qamar 2015; Gao et al. 2014; Joachims 1998). Tarkkuutta on käsitelty prosessimallin yhteydessä luvussa 4.4.4. Yli 80 % tarkkuudesta voidaan tehdä johtopäätös, että menetelmät voivat mahdollistaa liiketoimintaa tukevia asioita.

Koneoppimista ja hahmontunnistusta ei tunnu olevan mielekästä erotella toisistaan. Esimerkiksi hahmontunnistukseen syventyvässä kirjassa *"Statistical Pattern Recognition*" (Webb et al. 2011, s. 1-2) todetaan, että koko kirjan nimeksi olisi sopinut *koneoppiminen*. Lisäksi samassa kirjassa ilmaistaan, että kaikki sen kirjan asiat voidaan tiivistää yhdellä sanalla *luokitteleminen*. Useat aihetta käsittelevät kirjat on myös nimetty koneoppimisen ja hahmontunnistukseen liittyvillä otsikoilla kuten esimerkiksi Anzai (2012) – *Pattern Recognition & Machine Learning* tai Fu (1968) – *Sequential methods in pattern recognition and machine learning*. Tämä johtuu osiltaan siitä, että nämä asiat toimivat aina parina tämän tyyppisen tekoälyn luonnissa. Anzai (2012, ss. ix-x) mukaan mitään aiheeseen liittyvää algoritmia ei voida tunnistaa selväksi kuuluvan vain toiseen osa-alueeseen. Näiden tekijöiden johdosta tuntuu luontevalta, että tässä tutkimuksessa termejä ei erotella toisistaan.

Koneoppimisen ja hahmontunnistuksen menetelmien kannalta on tärkeätä tiedostaa millä tavalla kone oppii. Oppimisella tässä yhteydessä viitataan siihen, että kone pääsee kerta toisensa jälkeen parempiin tuloksiin. Tämä tapahtuu algoritmin avulla, jossa yritys-erehdys paria hyödynnetään. Jos virhe havaitaan, sen syntyminen analysoidaan ja pyritään siihen, että jatkossa vastaavanlaista virhettä ei synny. Analysoinnin voi tehdä ihminen tai kone. Koska yhä useampi päätös voidaan nykyään tehdä automaattisesti, tulee yritysten miettiä, mitkä päätökset voidaan jättää koneille ja mitkä vaativat ihmisen asiantuntijuutta (Davenport & Harris 2007, s. 191).

4.3 Oppimisen muunnelmat

Oppimisen muunnelmia on kolmea erilaista: ohjattu, ohjaamaton ja vahvistettu oppiminen (esim. Duda et al. 2001). Ohjatulla oppimisella tarkoitetaan menetelmää, jossa ulkopuolinen kouluttaja opettaa algoritmia saavuttamaan halutun lopputuloksen kehitysdatasta. Käytetty kehitysdata sisältää valmiit luokitukset (engl. *labeled data*). Ohjatussa oppimisessa idea on etsiä algoritmeja, jotka mahdollisimman tehokkaasti ja tarkasti pystyvät luomaan yleisiä hypoteeseja, joiden perusteella uudet tuntemattomat alkiot asetetaan osaksi luokkajärjestelmää. Toisin sanottuna tarkoituksena on luoda tietty luokittelumalli, jonka jokaiselle luokalle koneen tehtävä on etsiä ennustettavia piirteitä. Tämä edellyttää, että tuntemattomissa alkioissa tulee olla jotain järjestelmälle tuttuja piirteitä. (Kotsiantis et al. 2007) Tällaisessa tilanteessa siis tiedetään jo, minkälaisia luokkanimityksiä tuntemattomille alkioille voidaan antaa. Kun koneelle annetaan tämän jälkeen uusi tuntematon syöte, osataan syötteen piirteiden avulla automaattisesti löytää sille tietty luokka. Thorsten (1997) mukaan ohjatun oppimisen menetelmät ovat tehokkaimmillaan silloin, kun luokat ovat binäärisiä ja jokaisen tekstidokumentin kohdalla tehdään päätös, kuuluuko se tiettyyn luokkaan vai ei. Ohjatun oppimisen menetelmiä kutsutaan kirjallisuudessa myös luokittelemiseksi (engl. *classification*) ja erottelemiseksi (engl. *discrimination*) (Webb et all 2011, s. 7).

Ohjaamattomassa oppimisessa ei harjoitteludatalle anneta minkäänlaisia vastinarvoja (engl. *unlabeled data*) eikä menetelmässä käytetä ulkopuolista kouluttajaa. Tätä kutsutaan kirjallisuudessa myös klusteroinniksi (engl. *clustering*). Tavoite on löytää luonnollisia ryhmiä tai samanlaisia esimerkkitapauksia datan seasta. Erityisesti tässä menetelmässä arvoa tuovat tuntemattomat, mutta hyötykäyttöiset luokat, joita voidaan kutsua myös löydöksiksi. (Jain et al. 1999.) Löydökset koostuvat alkioista, jotka ovat toistensa kanssa samanlaisia tiettyjen piirteiden valossa. Yleisesti alkioiden samanlaisuus määritellään etäisyysfunktiolla. Paukerin (2012, s. 2) mukaan varsinkin datan määrän kasvaessa näitä löydöksiä voi olla hankala löytää ja se kuluttaa suhteettoman paljon aikaa. Tutkija myös kertoo, että ohjaamattomien oppimisen menetelmiä käytetään usein ilmiön mallintamisessa, eikä niinkään eri hypoteesien osoittamisessa. Myös Duda et al. (2001, ss. 517-518) ovat samoilla linjoilla Paukerin kanssa. Heidän mielestään ohjaamattoman oppimisen keinoja kannattaa käyttää ainakin viidessä perustapauksessa, kun datan määrä on suuri:

- **Luokittelijoiden suunnittelussa**. Datan määrän ollessa erittäin suuri voi luokittelupiirteiden etsiminen olla hyvin kallista. Tällöin ohjaamatonta oppimista kannattaa hyödyntää suunnitteluvaiheessa.
- **Luokkien löytämisessä**. Suuresta datamassasta voi olla hankala löytää luokkia, joten käytetään ohjaamatonta oppimista luokkien löytämisessä. Tämän jälkeen nimetään järjestelmän löytämät luokat.

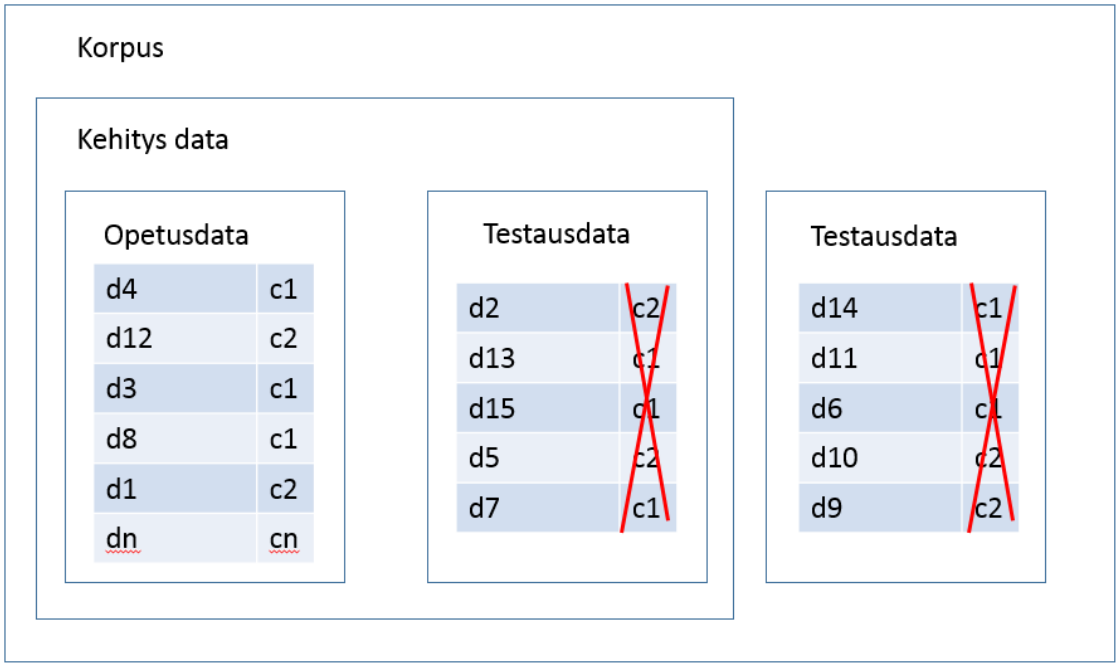
- **Muuttuvien luokkien tunnistaminen**. Jotkut luokat voivat muuttua ajan kuluessa, joten luokkien piirteiden uudelleenmääritys on tehokkaampaa automatisoinnin keinoin.
- **Piirteiden etsinnässä**. Ohjaamattoman oppimisen järjestelmä voi löytää automaattisesti sellaisia piirteitä, jotka muuten voisivat jäädä huomaamatta tai olisivat vaikeasti löydettävissä.
- **Datan ymmärtäminen**. Datan tutkimisen aikaisessa vaiheessa voi olla vaikea hahmottaa koko datan laajuutta. Ohjamaaton oppiminen voi automaattisesti helpottaa datan ymmärtämistä visualisoimalla ja laskennalla.

Yksi oppimisen muunnelma on vahvistusoppiminen (engl. *reinforcement learning*). Tässä muunnelmassa hyödynnetään sekä ohjatun että ohjaamattoman oppimisen piirteitä. Tästä menetelmästä on olemassa useampia variaatioita, mutta yleisesti niitä voidaan luonnehtia seuraavilla kuvailuilla. Opetusdata tarjotaan oppivalle järjestelmälle ulkopuolisen opettajan toimesta. Tämän jälkeen opettaja tarkkailee ja mittaa, kuinka hyvin järjestelmä suoriutuu toiminnan eri vaiheista. Opettajan tarkoitus ei ole kertoa mitä toimintoja tehdään seuraavaksi, vaan ennemmin kertoa, mikä yksittäinen toiminto toi halutun lopputuloksen. Ideana on jatkuva parantaminen yrityksen ja erehdyksen kautta, jossa asiantuntija toimii tuomarina. (Barto & Sutton 1997; Kotsiantis et al. 2007.) Toinen oppimisen välimallimuunnelma on hetkellinen erotteluoppiminen (engl. *temporal difference learning*). Tämä muunnelma on vähän vastaavanlainen kuin vahvistusoppiminen, mutta opettaja analysoi vasta jälkikäteen, mikä järjestelmän vaihe tuotti huonon tuloksen. Tämän johdosta muunnelmasta käytetään myös nimitystä myöhäistetty vahvistusoppiminen (engl. *delayed reinforcement learning*). (Tesauro 1995.) Tätä menetelmää on käytetty erityisesti peleihin liittyvissä sovelluksissa, joissa pyrkimys on opettaa kone voittamaan ihminen (esim. Tesauro 1995; Schraudolph et al. 1994).

4.4 Ohjatun oppimisen prosessimalli

Automaattiseen luokittelemiseen löytyy useita erilaisia prosessimalleja riippuen siitä, mihin luokittelemisen ongelmaan haetaan vastausta, millä oppimisen muunnelmalla luokittelua tehdään ja mitä dataa on saatavilla. Tämän tutkimuksen yhteydessä ongelma on, kuinka uusia asiakasviestejä voidaan luokitella vanhan viestidatan perusteella, mikä sisältää jo valmiin luokituksen. Tällaiseen ongelmaan soveltuu ohjatun oppimisen prosessimalli, jossa kyse on yleistämisestä (engl. *generalization*) (Kotsiantis et al. 2007). Yleistetään, että uusi ennestään tuntematon asiakasviesti saa saman luokituksen kuin lähtöaineistossa lähes vastaavanlaiset asiakasviestit ovat saaneet. Tämä edellyttää, että lähtöaineisto kattaa kaikki mahdolliset asiakasviestien variaatiot. Kun lähtöaineisto on saatu täyttämään mahdollisimman hyvin edellinen ehto, jaetaan se dokumenttien suhteen sattumanvaraisesti kahteen osaan. Osaan jolla ennustava malli kehitetään ja osaan, jolla mallia testataan. On tärkeä huomioida, että mallin testausta varten käytetty osa myös sisältää

oikeat luokitukset, mutta niitä ei paljasteta koneelle. Oikeiden luokitusten avulla voidaan evaluoida mallin toimivuus. Lisäksi Bird et al. (2014) suosittelee, että kehitykseen käytetty aineisto jaetaan edelleen kahtia: varsinaiseen opetteludataan ja kehityksen testausdataan, jonka avulla tehdään virheanalyyseja. Optimaalinen korpuksen jakaminen on esitetty kuvassa 6.



Kuva 6: Optimaalinen korpuksen jakaminen Bird et al (2015) mukaan, missä d=dokumentti, c=luokka. Testausdatoissa luokat ovat piilotettu koneilta.

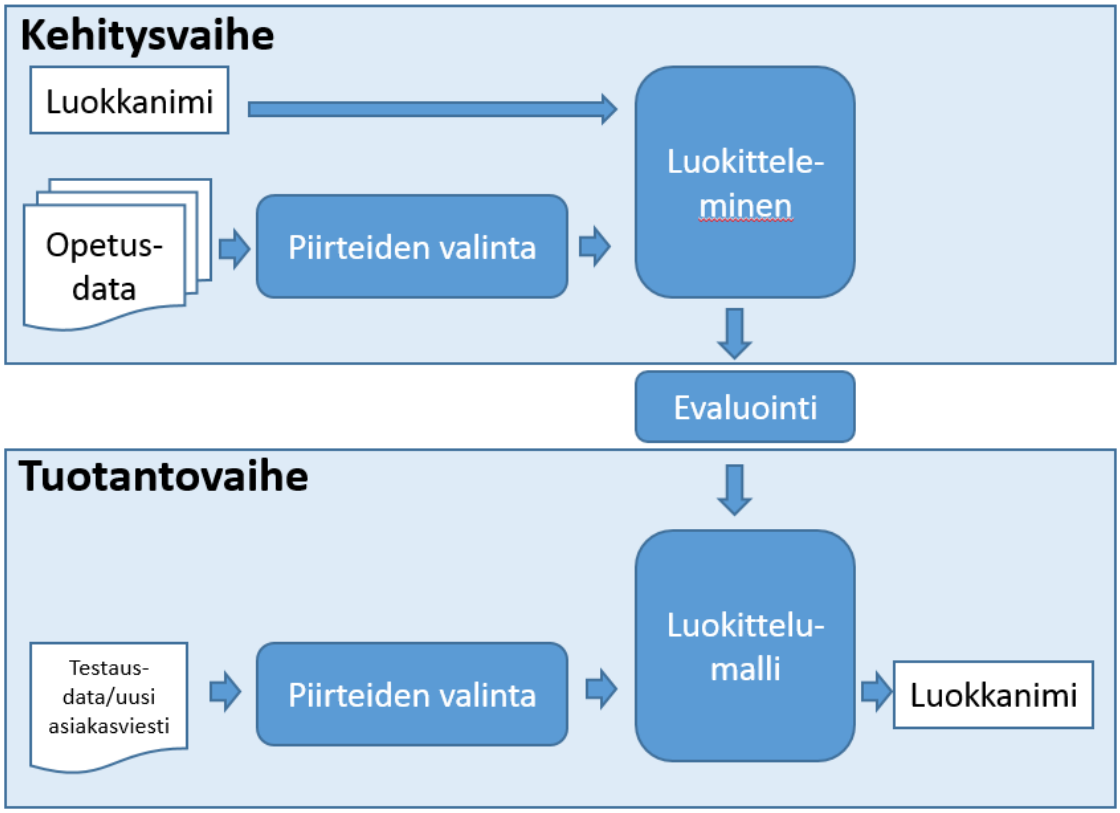
Alkuaineiston määrittelyn jälkeen alkaa varsinaisen luokittelumallin rakennus. Luokittelumallin rakentamisesta löytyy kirjallisuudessa erilaisia prosessivariaatioita ja nimityksiä (esim. Xia et al. 2016; Zhu et al 2009, s. 25; Kurbatow 2015; Ghalehtakia et al. 2014; Duda et al. 2001, ss. 14-16; Shah & Patel 2016; Sebastiani 2002), mutta niistä löytyy tiettyjä säännönalaisuuksia. Nämä säännönalaisuudet voidaan jakaa viiteen eri kohtaan, joita käsitellään tämän luvun aliluvuissa. Aliluvuissa esitellään myös yleisimpiä kohdan menetelmiä syventymättä yksityiskohtiin. Suurimmat vaikutteet esiteltäviin vaiheisiin tulevat Rana et al. (2014) tutkimuksesta, koska sen tutkimusongelma vastaa hyvin tämän tutkimuksen ongelmaan.

1. Esikäsitteleminen (aliluku 4.4.1)
2. Piirteiden valinta (aliluku 4.4.2)
3. Luokitteleminen (aliluku 4.4.3)
4. Mallin evaluointi (aliluku 4.4.4)
5. Uusien viestien luokitteleminen (aliluku 4.4.4)

Eri tutkimuksissa on annettu painoarvoa eri kohdille, lähestytty hiukan erilaisista näkökulmista ja käytetty erilaisia nimityksiä. Esimerkiksi Xia et al. (2016) duplikoi vaiheet

koskemaan varsinaista tekstiä ja sen metadataa. Ideana heillä on saavuttaa parempia tuloksia *Stackoverflow*-verkkopalvelun kysymystekstien luokittelemisessa valitsemalla piirteitä kahdelta eri tasolta: otsikko ja sisältötasolta. Kurbatow (2015) yhdistää piirteiden valinnan ja luokittelemisen yhdeksi vaiheeksi. Shah & Patel (2016) taas jakaa piirteiden valinnan piirteiden valinnaksi ja erottelemiseksi (engl. *feature selection and extraction*). Zhu et al. (2009, s. 25) kutsuvat vaiheita hiukan eri nimillä. He kutsuvat piirteiden valintaa konseptien mallinnukseksi (engl. *concept modeler*) ja luokittelemista semanttiseksi koneeksi (engl. *semantic engine*). Sebastiani (2002) käyttää piirteiden valinnassa termejä dokumenttien indexointi (engl. *document indexing*) ja ulottuvuuksien vähentäminen (engl. *dimensionality reduction*). Paukkeri (2012) esittelee työssään kieliriippumattomia sovelluksia, jotka eroavat suuresti tässä luvussa esitellystä mallista.

Kokonaisvaltaisesti prosessimallia voidaan havainnollistaa kuvassa 7 esitetyllä tavalla. Siinä on eroteltuna tekstinluokittelun kehitys- ja tuotantovaihe. Kehitysvaihe alkaa opetusdatan esikäsittelyllä. Tämän jälkeen datasta etsitään ja valitaan tiettyjä piirteitä, joilla eri luokat voidaan tunnistaa. Samalla myös määritellään millä tavalla dokumentteja halutaan esitettävän (esim. vektoreina). Kun jokainen dokumentti on esitetty samalla tavalla, voidaan käyttää luokittelemisen algoritmeja laittamaan dokumentit suhteelliseen järjestykseen. Samalla kun suhteellinen järjestys määritetään, annetaan myös dokumenteille halutut luokkanimet. Näin ollaan saatu rakennettua luokittelumalli, jonka toimivuutta testataan ja evaluoidaan. Evaluointi suoritetaan hyödyntämällä tuotantoon siirrettävää järjestelmää ja aiemmin määriteltyä testausdataa. Jos tulokset ovat riittävällä tarkkuudella suhteessa tutkimusongelmaan, voidaan siirtyä varsinaiseen tuotanto- eli ennustusvaiheeseen. Ennustusvaiheessa tulee uusi tuntematon dokumentti järjestelmään. Sille tehdään samat esikäsittelemisen ja piirteiden laskennan toimenpiteet, jonka jälkeen sitä sovitetaan aiemmin luotuun malliin. Mallin perusteella annetaan dokumentille luokitus. (Bird et al. 2015; Rana et al. 2014.)



Kuva 7: Koneoppimisen ja hahmontunnistusjärjestelmän kehitys ohjatun oppimisen keinoin. Mukaillen Bird et al. (2015) ja Rana et al. (2014).

4.4.1 Esikäsitteleminen

Esikäsittelemisellä voidaan yleisesti tarkoittaa useaa erilaista tapaa valmistella dataa koneelliselle käsittelylle. Sitä voidaan siivota (engl. *data cleaning*), integroida muun datan kanssa (engl. *data integration*), muuttaa toisenlaiseksi (engl. *data transformation*) tai sitä voidaan vähentää (engl. *data reduction*). Näitä menetelmiä tarvitaan datan laadun nostamiseksi. Toisaalta on huomioitava, että nämä käsittelytavat muuttavat dataa lopullisesti, jolloin paluuta alkutilanteeseen ei enää ole (Lagus 2000). Datan siivouksella tarkoitetaan korjaavia toimenpiteitä, jotka saattavat johtua puuttuvista, epämääräisistä ja epäjohdonmukaisista arvoista. Data integraatiota tarvitaan sellaisissa tapauksissa, kun halutaan yhdistellä useampia eri datan lähteitä. Datan muuntamisella tarkoitetaan arvojen muuntamista paremmin ymmärrettävään formaattiin. Datan vähennyksellä tarkoitetaan toimenpiteitä, joilla vähennetään epärelevanttia dataa tehokkaamman laskennan vuoksi. Tuloksien tulisi silti säilyä samana. (Jiawei & Kamber 2001, s 108-116.)

Edellisen kappaleen menetelmät ovat yleispäteviä korkealla tasolla. Tekstimuotoisen datan esikäsittelyyn on kehitelty oma tieteenhaaransa, ja siihen sopivat työkalut. Tätä tieteenhaaraa kutsutaan luonnollisen kielen käsittelyksi (engl. *natural language processing*,

NLP). NLP-tekniikoita käytetään esikäsittelyn yhteydessä sanojen määrän vähentämiseksi. Esimerkiksi Carvalho & Curto (2014) tutkivat potilastietojärjestelmää luonnollisen kielen käsittelyn näkökulmasta, joka sisälsi yhteensä 156 miljoonaa sanaa. Järjestelmässä oli uniikkeja sanoja 260 180 joista vain 31 527 (~12 %) oli tunnettuja sanoja. Kaikki muut sanat olivat tuntemattomia käytetyille sanahakemistoille ja suurin osa vaihtelusta johtui kirjoitusvirheistä. Tutkimuksessa korostetaan esikäsittelyn tärkeyttä, jossa poistetaan kirjoitusvirheitä ja turhia taivutusmuotoja. Toinen asia mitä tutkimus korostaa on, että esikäsittelyn menetelmät toimivuus ei ole ennalta arvattavissa, joten optimaalisin luokittelumalli saavutetaan useiden iteraatiokierroksien avulla. Lisäksi tutkimuksessa huomautetaan, että esikäsittelyvaiheessa esiintyvät virheet monistuvat myöhempiin prosessivaiheisiin.

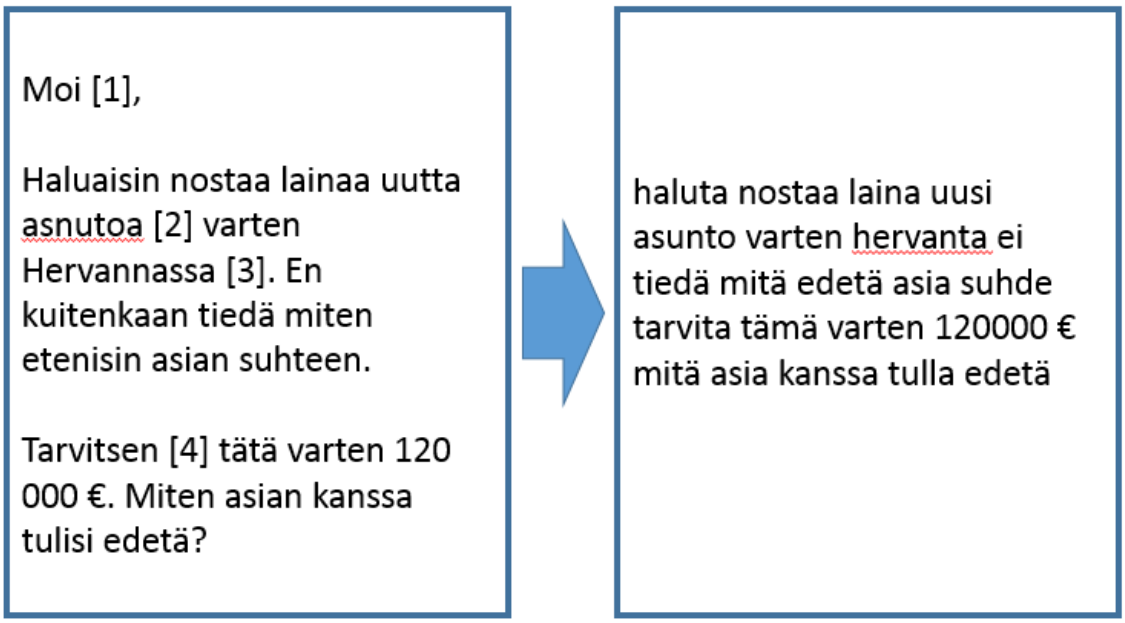
Kappaleessa 3.2.2 osoitettiin suomen kielen monimuotoisuutta erityisesti sanojen suhteen. Tämän johdosta NLP-menetelmien tehokas käyttö on toimivalle suomen kieliselle luokittelumallille välttämätön. Zhu. et al (2009, s. 123) ja Ghalehtaki et al. (2014) tutkimuksista on yhdistelty seuraava lista NLP-menetelmistä:

- Kielen tunnistaminen (engl. *language identification*)
- Merkkien normalisointi (engl. *normalization*)
- Turhien sanojen poistaminen (engl. *stopword removal*)
- Morfologiset analyysit (engl. *morphological analysis)*
- Kirjoitusvirheiden korjaus (engl. *spelling correction*)

Suomessa on käytössä suomen ja ruotsin kielet virallisina äidinkielinä. Lisäksi ulkomaalaiset saattavat käyttää englantia kommunikointikielenään. Yleisesti NLP-järjestelmät olettavat, että kaikki prosessoitavat dokumentit ovat kirjoitettu samalla kielellä (katso Lui & Baldwin 2012). Kielen tunnistaminen on kuitenkin ensimmäinen askel reaalimaailman datan prosessoinnissa. Tämän menetelmän tekeminen onnistuu vertailemalla tiedetyn kielen korpusta suhteessa kohdedataan. (Lui & Baldwin 2012.) Tällä menetelmällä vältetään, että yksittäinen ”vääränkielinen” dokumentti ei tuo suuria poikkeamia rakennettuun malliin.

Merkkien normalisoinnilla tarkoitetaan ylimääräisten sanavariaatioiden poistamista esimerkiksi muuttamalla kaikki isot kirjaimet pieniksi (Ghalehtaki et al. 2014). Tässä saatetaan menettää merkitystä yleis- ja erisnimien suhteen (vrt. laakso ja sukunimi Laakso). Toinen tapa parantaa merkityksellisten sanojen arvoa on poistaa turhat sanat. Tällä viitataan arvottomien sanojen poistamiseen, jotka esiintyvät lähes jokaisessa dokumentissa. Suomen kielessä viisi yleisintä sanaa ovat ”ja, on, ei, että, hän”, joiden arvo luokittelun kannalta on vähäinen (Verbix 2016). Kolmas tapa vähentää sanavariaatiota on palauttaa sanat perusmuotoon. Tällaista menetelmää kutsutaan morfologiseksi analyysiksi. Yksi morfologinen menetelmä on stemmaus (engl. *stemming*), jossa poistetaan sanojen alkuja loppuliitteitä. Toinen lähellä stemmausta on lemmaus (engl. *lemmatization*), jonka tarkoitus on palauttaa termien muodot sanakirjoista löytyviin perusmuotoihin. Usein tähän

käytetään hakemistona hyödyksi sanakirjaa. Tutkimuksen mukaan lemmaus toimii paremmin suomen kielen kanssa kuin stemmaus. (Korenius et al. 2004.) Kettusen & Baskayan (2011) tutkimuksen mukaan suomen kieleen on olemassa useampia vaihtoehtoja, joilla sanoja voidaan onnistuneesti palauttaa perusmuotoonsa. Kurbatowin (2015) tutkimuksen mukaan erityisesti sanojen palauttaminen perusmuotoonsa tuo huomattavaa tarkkuutta luokittelun kannalta. Menetelmien käyttöä on havainnollistettu kuvassa 8. Kuvassa havainnollistetaan turhan sanan poistamista, kirjoitusvirheen korjaamista ja tekstin normalisointia. Lisäksi sanat on muutettu perusmuotoon.



Kuva 8: Yksinkertaistettu esimerkki tekstin esikäsittelemisestä. [1] sana poistetaan, koska se on turha. [2] sana korjataan, koska siinä on kirjoitusvirhe. [3] sana normalisoidaan muuttamalla isot kirjaimet pieniksi. [4] sana muutetaan perusmuotoon morfologisella analyysillä.

Näitä menetelmiä käytettäessä on hyvä muistaa, että samalla kun asiakasviestien sanavariaatiot vähenevät, sanoista voi tippua merkityksellisiä osia pois. Merkitykset tippuvat entisestään yleisellä *bag-of-words* lähestymisellä, missä sanojen järjestyksellä ja suhteilla ei ole merkitystä. Näiden asioiden huomioimatta jättäminen on tutkimusten mukaan vähemmän merkityksellinen asia luokittelun kannalta (Hotho et al. 2005). Hyödyt tulevat siitä, että sanavariaatiot vähenevät ja samaa tarkoittavat sanat voidaan analysoida yhden hakutermin avulla. (Katso Rana et al. 2014.) Kirjoitusvirheiden korjauksen menetelmä ei lähtökohtaisesti muuta sanojen merkitystä. Myös tähän voidaan käyttää stemmauksen ja lemmauksen keinoja. Kirjoitusvirheiden poistaminen on tärkeätä, koska Carvalho & Curto (2014) osoittavat tutkimuksessaan niiden olevan erittäin yleisiä luonnollisen kielen korpuksissa.

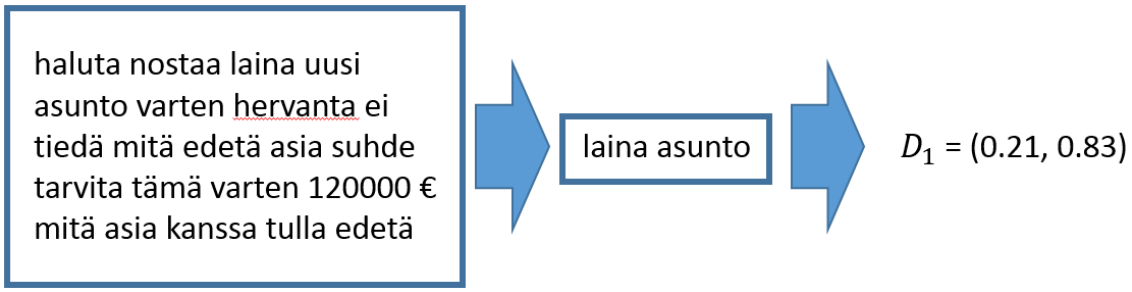
4.4.2 Piirteiden valinta

Piirteiden valinnalla (engl. *feature selection*) tarkoitetaan oikeiden luokitteluominaisuuksien löytämistä kaikista mahdollisista vaihtoehdoista (Lagus 2000). Scott & Marshallin (2009) mukaan koko luokittelemisen prosessissa on kyse mittaamisesta. Tästä näkökulmasta katsottuna tässä kappaleessa käsitellään mittareiden valintaa. Jos käytetään sensoreita kuvaamaan reaalimaailmaa, voi erilaisia signaaleja olla satoja tai jopa tuhansia kuvaamaan objektin piirteitä. Kaikkien signaalien käsitteleminen ei ole tarpeen, sillä ne eivät välttämättä kuvaa objektia luokittelun kannalta mitenkään. On siis valittava oikeat ja mahdollisimman tehokkaat piirteet kuvaamaan objektia. (Talonen 2015, s 61-62.) Myös monet muut tutkijat ovat tästä samaa mieltä toteamalla, että piirteiden valinta on luokittelun suurimpia ongelmia (Shah & Patel 2016; Rana et al. 2014; Sebastiani 2002; Yang & Pedersen 1997). Toisaalta Throstenin (1997) tutkimuksessa osoitettiin, että piirteiden karsiminen ei ole edes välttämätöntä. Tämä tutkimus suoritettiin 9947 piirteellä. Luku on suhteellisen alhainen joihinkin käytännön kohteisiin verrattuna.

Yksi toimiva tapa on muuttaa dokumentit tässä vaiheessa sanojen kokoelmiksi (engl. *bag of words)* (Rana et al 2014). Muutoksen jälkeen sanojen järjestykset ja rakenteet katoavat, ja sanat vastaavat täysin dokumentin piirteistä. Kuten edellisessä alaluvussa todettiin, sanojen järjestyksellä ja lauserakenteilla ei ole suurta merkitystä. Jos yksinkertaistetaan asiaa, voi jo yksi termi tai sana olla tarkka ja tehokas piirre kategorisoinnille. Esimerkiksi jos dokumentista löytyy termi ”*Tarja Halonen*”, voidaan dokumentti kategorisoida politiikkaan liittyväksi. Tällaisille sanoille tulee antaa enemmän painoarvoa dokumenttien luokittelemisessa. (Yang & Pedersen 1997.) Dokumenteissa voi olla muitakin piirteitä kuin sanojen tarkastelu, mutta niiden yhteys luokituksen valintaan on todennäköisesti heikko. Esimerkiksi asiakasviestistä voidaan laskea sanojen ja lauseiden määrää, mutta ne tuskin vaikuttavat asiakasviestin luokkaan.

Jos sanat ilmaisevat tekstidokumenttien piirteitä, voidaan sanoja poistamalla vähentää piirteiden määrää. Tätä voidaan kutsua myös termien indeksoinniksi (Hotho et al. 2015). Menetelmiä kutsutaan aggressiivisiksi, jos piirteitä (sanoja) poistetaan paljon. Esimerkiksi Yang & Pedersenin (1997) tutkimuksessa uniikkeja sanoja poistettiin jopa 98 % ja silti luokittelun tarkkuus pysyi samana. Koska myös esikäsittelyvaiheessa poistetaan sanoja, on piirteiden valinta jossain määrin myös esikäsittelemistä tai toisinpäin. Piirteitä valitaan luokitteluennustusten tarkkuuden parantamiseksi ja epärelevantin tiedon poistamiseksi. Lisäksi parannetaan oppivuuden tehokkuutta vähentämällä laskennallista ja muistin tarvetta, vähennetään tulevan datan keräyksen hintaa ja myös, jotta datamalleista saadaan yksinkertaisempia ja sitä kautta ymmärrettävämpiä. (Webb et al. 2011, s. 435) Gao et al. (2014) osoittaa tutkimuksessa, että 100 piirrettä tekstidokumenttien yhteydessä voi olla riittävä määrä. Tätä suurempi määrä ei välttämättä paranna luokittelun tarkkuutta merkittävästi. Tutkimuksessa tosin käytettiin tuhatta dokumenttia ja kahta eri luokkaa.

Piirteiden valinnan yhteydessä tulee määritellä, miten ja millä perusteilla dokumentit indeksoidaan. Rana. et al. (2014) mukaan tämä on luokittelun kannalta tärkein vaihe. Sebastiani (2002) mukaan tyypillisin indeksointitapa on esittää dokumentit vektoreina, missä vektorin eri ulottuvuudet kuvaavat tiettyjen sanojen tai termien esiintyvyyttä dokumentissa. Näiden sanojen ja termien tulee olla mahdollisimman merkityksellisiä kuvaamaan dokumenttia. Tekstin merkityksellisten osien löytämisellä tarkoitetaan ison tekstin jakamista pieniin avaimiin (engl. *tokenization*). Avaimiin ja segmentteihin jakamisella tarkoitetaan sellaisten merkkijonojen löytämistä, jotka ovat tilastollisesti merkittäviä luokittelupäätöksen tekemiseksi. Avaimet, jotka esiintyvät lähes kaikissa tai vain muutamissa dokumenteissa eivät tuo lisäinformaatiota luokittelun näkökulmasta. (Shah & Patel 2016.) Piirteiden valintaa ja dokumentin representaation muutosta on havainnollistettu kuvalla 9. Yleisesti vektorit ovat moniulotteisempia, mutta kuvassa on pyritty yksinkertaistamaan asiaa.



Kuva 9: Yksinkertaistettu esimerkki merkityksellisten piirteiden valinnasta ja dokumentin representaatiosta kaksiulotteisen (kahden piirteen) vektorin avulla. Vektorissa piirteiden voimakkuutta kuvattu arvoilla.

Piirteiden valinnassa käytetään yleensä kahta eri tapaa, joita voidaan käyttää yksittäin tai yhdessä (Khalid et al. 2014). Yksinkertaistaen asiaa, ensimmäisessä vaiheessa otetaan huomioon koko data ja käytetään statistiikkaa hyödyksi mallintamaan kaikista korpuksen sanoista tärkeimmät. Menetelmiä on paljon, mutta kirjallisuudesta saa kuvan, että suosituimmat menetelmät tähän vaiheeseen ovat: *document frequency*, *information gain* ja *Chi square*-algoritmit (Yang & Pedersen 1997; Xiong 2014; Shah & Patel 2016). Toisessa vaiheessa parannetaan ensimmäisen vaiheen lopputulosta (engl. *feature extraction*) analysoimalla, mitkä termit kuvaavat luokkaa parhaiten. Tähän yleisin tapa vaikuttaa olevan *principle component analysis* (PCA) (Khalid et al. 2014; Shah & Patel 2016). Menetelmät esitellään seuraavaksi lyhyesti.

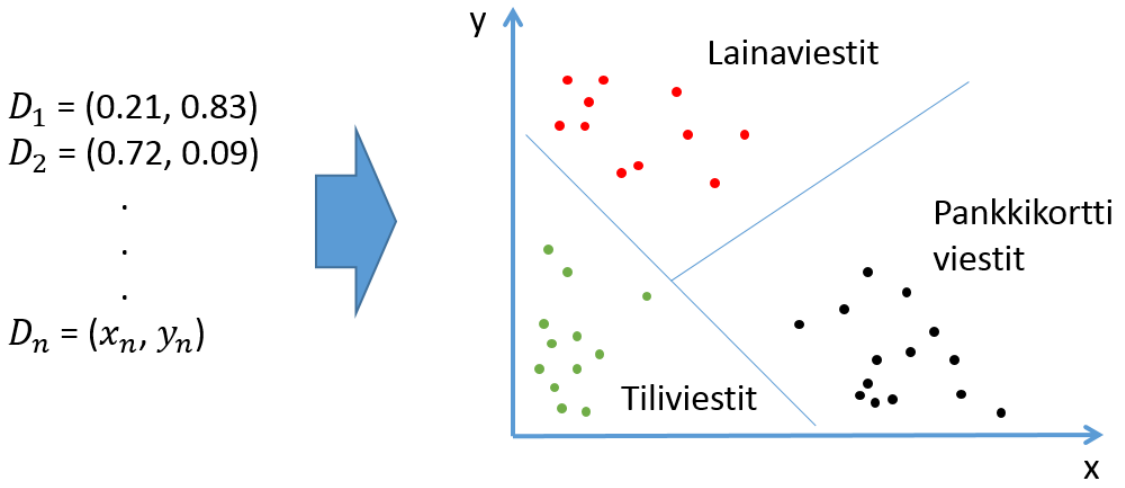
Document frequency ilmaisee, kuinka monta kertaa tietty termi esiintyy dokumentissa. Jokaiselle termille lasketaan esiintyvyysluku ja jos tämä luku jää alle määritellyn raja-arvon, termi poistetaan. Tämä menetelmä skaalautuu hyvin isoihin korpuksiin (Shah & Patel 2016). *Information gain* ilmaisee, kuinka paljon jokaisella uniikilla termillä on informaatioarvoa luokituksen kannalta. Laskenta tapahtuu sen perusteella, esiintyykö termi dokumentissa vai ei. Tässä menetelmässä poistetaan sanastosta ne termit, joiden informaatioarvo on alle määritetyn raja-arvon. (Yang & Pedersen 1997.) *Chi square*-algoritmit

testaavat luokan ja termin itsenäisyyttä. Menetelmään voi tutustua lähdemateriaalissa lisää (esim. Yang & Pedersen 1997). *Principle component analysis* idea on vähentää vektorin ulottuvuuksien eli dokumentin piirteiden määrää analysoimalla, mitkä ovat tärkeimpiä komponentteja luokittelun kannalta. Kone laskee automaattisesti pääkomponentit tarkempaan analyysiin. (Shah & Patel 2016.)

Vektorien muodostuksessa ja piirteiden valinnassa tulee huomioida, että niitä ei tule harjoittaa liikaa opetusdatan kanssa. Tällaisessa tapauksessa saatetaan päätyä tilanteeseen, jossa kehitetty luokkajärjestelmä toimii erinomaisesti opetusdatan kanssa, mutta järjestelmä ei toimi enää tuotantovaiheessa uudenlaisten dokumenttien kanssa. On tärkeätä muistaa, että ohjatun oppimisen perusperiaate on yleistäminen ja se, että opetusdata ei ole koskaan täydellinen. Tästä johtuen piirteiden valinnassa tulee huomioida vähän löysemmin mukaan tekijöitä, jotka mahdollistavat tulevaisuudessa luokittelun. (Sebastiani 2002.)

4.4.3 Luokitteleminen

Tällä vaiheella tarkoitetaan parhaimpia tuloksia antavan luokittelumallin löytämistä kaikista mahdollisista malleista (Lagus 2000). Prosessivaiheen ideana on, että edellisessä luvussa syntyneet dokumenttien representaatiot (esim. vektorit) luodaan malliksi, mihin uudet tuntemattomat dokumentit voidaan asettaa. Samalla tässä vaiheessa annetaan luokille nimet. Tietyllä luokalla tulee olla tietyt luokkasäännöt, minkälaisia piirteitä eli sanoja se hyväksyy luokkaansa. (Rana et al. 2014) Luokittelemisen toimivuutta on tutkittu erilaisissa tarkoituksissa kuten uutisotsikoiden (Drury et al. 2011; Liu et al. 2012) ja sosiaalisen median viesteissä (Dilrukshi et al. 2013). Kaksiulotteisten dokumenttien sijoittumista eri luokkiin on havainnollistettu kuvassa 10. Aikaisemmin esitetty asiakasviesti (D_1) on sijoitettu lainaviestit-alueelle sen piirrearvojen perusteella. Näiden eri dokumenttipisteiden avulla voidaan laskea tietyt luokkarajat, joita käytetään myöhemmin tuotantovaiheessa.



Kuva 10: Esimerkki kaksiulotteisten asiakasviestien luokittelemisesta.

Luokitteluja voi tehdä monella erilaisella tavalla ja algoritmilla. Luokittelun algoritmien määrä on suuri, mutta kirjallisuudessa suosituimmat menetelmät tuntuvat olevan *K-lähimmän naapurin, Naiivi Bayes, Support Vector Machines, neuraalien verkkojen* ja *päätöspuiden algoritmit* (Rana et al. 2014).

K-lähimmän naapurin (engl, *K-Nearest Neighbor*, *KNN*) algoritmin idea hyvin yksinkertainen. Uuden alkion luokka määrittyy sen mukaan, mikä teksti on ollut harjoitteludatassa tätä lähimpänä. Uusi alkio saa saman luokituksen kuin lähin naapurinsa vektoriavaruudessa. Menetelmän perustavalaatuinen idea on, että jos uudelle alkiolle löytyy täsmälleen identtinen alkio harjoitteludatassa, se saa saman luokan. Jos identtistä ei löydy, oletetaan, että lähimmän alkion luokka on oikea. (Manning & Schütze 1999, s. 604.) Menetelmä soveltuu parhaiten ongelmiin, missä on käytössä useampia eri luokkia. Menetelmän heikkous on sen raskas laskenta, joka näkyy käyttäjälle hitautena. (Shah & Patel 2016.)

Naiivin Bayeksen idea on verrata uutta dokumenttivektoria kaikkien luokkien määrittelyihin yksi kerrallaan. Lopputulemana syntyy todennäköisyystaulukko sisältäen yhtä monta saraketta kuin luokkia on. Näistä valitaan todennäköisin luokka. Naiivin Bayeksen menetelmän etu on sen helppokäyttöisyys ja ymmärrettävyys, joten sen oppiminen ja käyttäminen ovat nopeaa. Varjopuolena menetelmä ei tuota niin tarkkoja tuloksia, mihin muilla menetelmillä voitaisiin päästä. (Chakrabarti 2003.) Toinen menetelmän heikkous on, että se olettaa luokkien olevan itsenäisiä, vaikka todellisuudessa ne voivat olla hyvin samanlaisia (Shah & Patel 2016).

Tutkimusten mukaan (esim. Pilaszy 2005; Chakrabarti et al, 2003) mukaan tukivektorikone (engl. *support vector machine*) on yksi tehokkaimmista oppivista algoritmeista tekstin kategorisoinnille. Se soveltuu parhaiten kahden luokkanimen aineistoille, mutta toimii myös hyvin useammalla luokalla. Pilaszy (2005) kommentoi hiukan ristiriitaisesti, että tukivektorikone on erityisen hyvä siksi, koska sen avulla voidaan ottaa huomioon rajaton määrä piirteitä ja silti laskentatehoa ei tarvita suhteettoman paljon. Ongelma on, että mitä

enemmän luokkia on, sitä monimutkaisemmaksi tulevat sen ymmärtäminen ja implementointi. Tämä on myös menetelmän heikkous. (Shah & Patel 2016.) Sen toiminta perustuu vektorilaskentaan ja sen avulla muodostettuun turva-alueeseen (engl. *hyperplane*). Ideana on, että tämä turva-alue maksimoidaan mahdollisimman suureksi, jotta luokittelu pystyisi erottelemaan helposti syötedokumentit eri luokkiin. Turva-alueen toisella puolella olevat pisteet kuuluvat toiseen luokkaan ja toisella toiseen. (Pilaszy 2005)

Neuroverkot (engl. *neural networks*) nimi tulee siitä, että niiden toiminta pyrkii jäljittelemään ihmisen aivojen toimintaa (sana neuro viittaa aivoihin). Neuroverkot koostuvat silmukoista (kutsutaan neuroneiksi), jotka ovat kytkettynä toisiinsa. Jokainen neuroni on yksi tiedonkäsittely-yksikkö. Syöteneuronit esittävät eri termejä, tulosneuronit eri luokkia ja kytkennät kuvaavat eri termien ja luokkien riippuvuussuhteita. Menetelmä perustuu siihen, että alkutilanteessa dokumentissa esiintyvät neuronit aktivoituvat, jolloin muodostuu syy-seuraussuhteiden sarja, joka johtaa tekstidokumentin luokitukseen. (Sebastiani 2002.) Lisäksi menetelmä hyödyntää neuronien keskinäisiä yhteyksiä tallentaakseen opittua informaatiota. Menetelmä sopii hyvin soveltaviin ja monimutkaisiin ongelmiin, jotka ovat luonnostaan epälineaarisia. Sen etu on myös monipuolisessa oppivuudessa, jota se hyödyntää uusilla toistokerroilla. (Haykin 1998.)

Lisäksi luokittelua voidaan tehdä päätöspuiden (engl. *decision trees*) avulla, kuten esimerkiksi kuvassa 5 havainnollistettiin. Tässä menetelmässä luokittelumalli muodostetaan hierarkkiseksi rakenteeksi, joka koostuu solmuista ja lehdistä. Solmut kuvaavat tiettyjä ehtoja, jotka joko täyttyvät, täyttyvät osittain tai eivät täyty. Lehdet kuvaavat luokittelun tapauksessa päätöksiä eli luokkia. Päätöspuun hierarkiaa jatketaan niin kauan, kunnes ehdoista muodostuva alipuu ei sisällä uusia luokkia, arvioitavat ominaisuudet loppuvat tai puun maksimi syvyys saavutetaan. Dokumenttien luokittelu tapahtuu siirtymällä ylhäältä alaspäin ja tarkastelemalla, kuinka se täyttää ehdot. (Manning & Schütze 1999, ss. 576-578.) Menetelmän vahvuus on sen yksinkertaisuus, mutta menetelmällä sorrutaan usein ylisovittamaan luokkia (Shah & Patel 2016).

4.4.4 Luokittelun evaluointi

Kuten aiemmin on esitetty, lähtöaineiston rajallisuuden ja monen muun syyn takia luokittelujärjestelmien mallit voivat olla virheellisiä. Tämän vuoksi on tärkeätä käyttää evaluointia osana luokittelun prosessia, jotta opitaan löytämään parhaimmin tehtävästä suoriutuvat algoritmit. Ohjatun koneoppimisen järjestelmän kehitys päättyy evaluointivaiheeseen, jonka perusteella saadaan palautetta rakennetun systeemin toimivuudesta. Palautteen perusteella pyritään kehittämään luokittelujärjestelmän toimivuutta kehittämällä koko prosessia tai jotain sen vaihetta. Järjestelmän toimivuutta testataan testausdatalla, joka on erotettu alkuvaiheessa korpuksesta. Tätä testausdataa ei olla käytetty osana järjestelmän kehitystä. Tarkoituksena on verrata, antaako rakennettu malli samat luokitukset kuin korpuksessa oli todellisuudessa. (Manning & Schütze 1999, ss. 576-578.)

Luokittelua voidaan evaluoida erilaisilla mittareilla. Tekstinluokittelun yhteydessä käytetyimmät mittarit ovat tarkkuus (engl. *precision*) ja saanti (engl. *recall*), joiden laskennassa käytetään hyödyksi totuustaulua. Totuustaulu on esitetty kuvassa 11. Totuustaululla verrataan asiantuntijoiden antamia luokituksia eli niin sanottuja todellisia luokituksia suhteessa automaattiseen luokittelujärjestelmän tarjoamiin. TP (true positive) kuvaa oikeiden luokkanimen saaneiden dokumenttien määrää. TN (true negative) kuvaa dokumenttien määrää, jotka ovat todellisuudessa ja järjestelmän mielestä väärin luokiteltu Ne ovat myös oikein luokiteltu. FP (false positive) ja FN (false negative) tilanteissa järjestelmä on toiminut väärin. (Sebastiani 2002; Rana et al. 2014.)

Luokkanimi		Todellinen luokitus (asiantuntijoiden)	
		KYLLÄ	EI
Luokittelu-Järjestelmän luokitus	KYLLÄ	TP	FP
	EI	FN	TN

Kuva 11: Järjestelmän toimivuuden laskeminen totuustaulun avulla. Mukaillen Sebastiani (2002) ja Manning & Schütze (1999, ss. 576-577)

Hyvä saanti ilmaisee tietotarpeiden täyttymistä, kun tarkkuudella pyritään siihen, että eitoivottujen luokitusten määrä on mahdollisimman matala. Saannin laskenta on esitettynä kaavassa [1] ja tarkkuuden kaavassa [2]. Kaavasta saadaan murtolukuja, jotka usein käännetään tarkkuutta ilmaisevaksi prosentiksi, kuten aikaisemmin on esitetty. Mitä suurempi murtoluku, sitä toimivampi luokittelujärjestelmä. (Alkula 2000; Sebastiani 2002.)

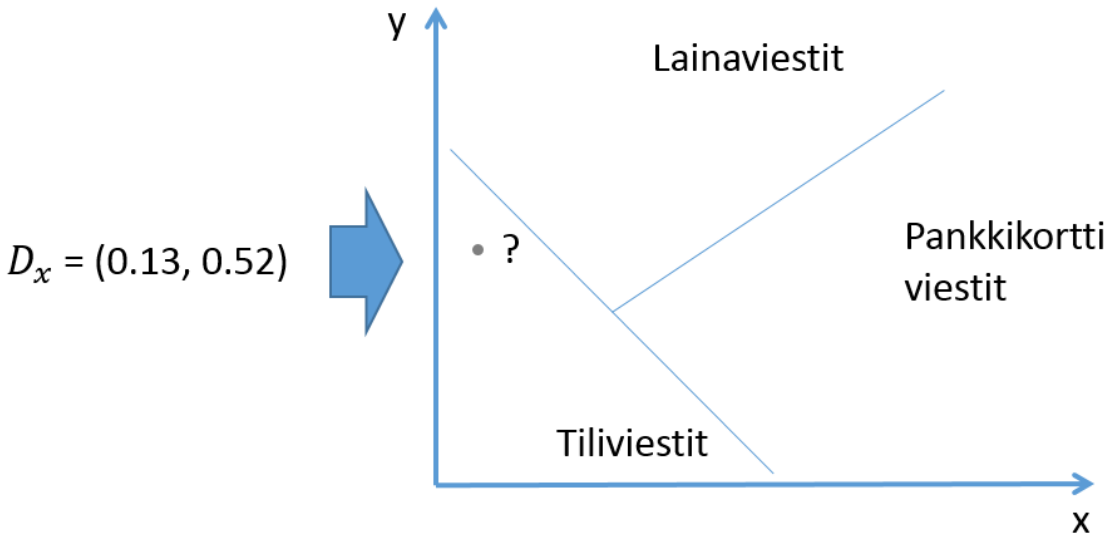
$$Saanti = \frac{TP}{TP+FN} \quad [1]$$

$$Tarkkuus = \frac{TP}{TP+FP} \quad [2]$$

Edelliset kaavat ilmaisevat yhden luokkanimen eli lokaalin luokituksen toimivuutta. Järjestelmän globaali toimivuus saadaan laskettua summaamalla eri luokkien lokaalit tarkkuus- tai saantiluvut yhteen, ja jakamalla nämä luokkanimien määrällä. Parhaimman globaalin toimivuuden löytäminen vaatii useita toistoja prosessissa. Mitä enemmän luokkia

on, sitä hankalampi on löytää globaalisti optimaalinen huippukohta. Tämä johtuu siitä, että luokkamalleissa piirteet ovat riippuvuussuhteissa toistensa kanssa, jolloin tietyn luokan optimoiminen voi vaikuttaa toisen luokan lokaaliin toimivuuteen negatiivisesti. (Manning & Schütze 1999, ss. 576-577.)

Liiketoiminta ja käyttötarkoitus määrittelevät, mitkä ovat riittäviä tarkkuuksia globaaleille ja lokaaleille saanti- ja tarkkuusluvuille. Esimerkiksi jos järjestelmä pystyy ennustamaan 95 % tarkkuudella, että saapunut asiakasviesti liittyy lainoihin, voi tämä olla optimaalinen osaksi asiakaspalvelua, vaikka globaali tarkkuus olisi vain 30 %. Joka tapauksessa, kun käyttötarkoitukseen optimaalinen toimivuus on saavutettu, voidaan järjestelmä siirtää tuotantovaiheeseen. Tässä vaiheessa järjestelmään saapuu uusi tuntematon asiakasviesti, jonka järjestelmä käsittelee vastaavilla algoritmeilla kuin kehitys- ja testivaiheissa. Viesti esikäsitellään, piirteet valitaan ja asiakasviesti saa mahdollisesti oikean luokituksen. Kuvassa 12 on havainnollistettu tuntemattoman dokumentin x saapumista järjestelmään tuotantovaiheessa.



Kuva 12: Esimerkki luokittelumallista tuotannossa. Tuntemattoman dokumentin ennustettu luokka piirteiden valossa on "tiliviestit".

Dokumentin sanojen perusteella on laskettu piirteiden arvoiksi 0.13 (x) ja 0,52 (y). Näillä arvoilla dokumentti sijoittuu mallissa *tiliviestit* alueelle. Sijoittuminen kuvaa luokkaa, johon dokumentti todennäköisesti sijoittuu. Sijoittumiselle on laskettu aikaisemmin tarkkuusluku kuvaamaan, kuinka todennäköisesti viesti on todellisuudessa aiheeseen liittyvä.

5. TUTKIMUKSEN MENETELMÄT JA SUORITUS

Tämän tutkimuksen empiirisessä osiossa haetaan vastausta tutkimuksen viimeiseen apukysymykseen: *”Mitä tehokkuuteen ja päätöksentekoon liittyviä hyötyjä voidaan jalostaa asiakasviestidatasta kohdeyrityksen asiakaspalvelussa*?*”* Hyötyselvitysten tekeminen tulisi olla jokaisen aihealueen projektin lähtökohta (Markham et al. 2015). Tässä kappaleessa esitellään empiirisen tutkimuksen lähtökohdat, käytetyt menetelmät ja kuinka tutkimus suoritettiin.

Asiakasviestejä halutaan luokitella esimerkiksi sen takia, koska tietyn luokan viestit jakavat samoja ominaisuuksia (Fu 1968, ss. 1-2). Ominaisuuksia voivat olla esimerkiksi keskimääräinen palveluaika, vastauksissa käytetty mallipohja tai aihealueen asiantuntija. Tutkimuksen empiirisessä osiossa pyritään selvittämään haluttuja luokkaominaisuuksia.

5.1 Tutkimuksen menetelmät

Luvussa 4 esiteltyyn prosessimalliin ei voida siirtyä suoraan käytännön ongelmissa. Prosessimalli olettaa, että menetelmän tavoitteet ovat selvät ja kehitykseen käytettävä data on määritelty. Markham et al. (2015) muotoilee asian hyvin toteamalla, että yrityksille tulee olla selvää ennen implementointia, miten näitä työkaluja käytetään ja mitä niiltä odotetaan. Toisin sanottuna syöte (engl. *input*) ja tulos (engl. *output*) tulee olla selkeästi tunnistettu sekä resurssit, roolit ja vastuut jaettu. Tähän näkökulmaan tähtää Webb et al. (2011, ss. 4-5) julkaisussa esitelty kokonaisvaltaisempi malli, jolla voidaan lähteä ratkaisemaan koneoppimisen ja hahmontunnistuksen ongelmia tyhjältä pöydältä. Malli koostuu seitsemästä eri vaiheesta:

1. **Ongelman muotoilu**: tavoite on saada selvä ymmärrys tulevasta tutkimuksesta ja suunnitella tulevat vaiheet. Samalla myös määritellään haluttuja tuloksia esimerkiksi tutkimuskysymysten tai hypoteesien avulla (Markham et al. 2015).
2. **Datan suunnittelu ja kerääminen**: Määritellään tunnettuja ja potentiaalisia datan lähteitä (Markham et al. 2015). Tehdään mittauksia tietyille muuttujille ja kirjataan ylös yksityiskohdat, kuinka data aiotaan kerätä.
3. **Alustava datan arviointi**: tarkastellaan dataa, lasketaan datan pääpiirteitä, tutustutaan aineiston tärkeimpään sanastoon, etsitään sääntöjä ja luodaan diagrammeja datasta, jotta saadaan parempi ymmärrys tutkimuksen kohteesta.
4. **Piirteiden valinta ja vähentäminen**: Valitaan mitattavat muuttujat, jotka ovat sopivia tehtävään. Käytetään NLP-menetelmiä datan yksinkertaistamiseksi (Markham et al. 2015).

5. **Löytöjen tekeminen ohjaamattoman oppimisen avulla**: Tämä vaihe voidaan nähdä tutkivana vaiheena, joka voi tuoda menestyksellisiä löytöjä tutkimuksen kannalta. Toisaalta, tämä voi olla datan esikäsittelyn vaihe ohjatun luokittelun ongelmaan.
6. **Luokitteleminen**: Rakennetaan luokittelumalli, johon uuden datan tulee sijoittua.
7. **Tulosten arviointi**: Testataan ja evaluoidaan luokittelemisen toimivuutta totuustaulun avulla. Tässä vaiheessa asiantuntijoiden tulee arvioida, onko dataa tarpeeksi ja onko se luotettavaa vastaamaan liiketoiminnan ongelmiin (Markham et al. 2015).

Tämän tutkimuksen tavoite on valmistella organisaatiota uusiin teknologisiin menetelmiin. Valmistelulla tarkoitetaan tässä yhteydessä mallin vaiheita yksi ja kaksi. Vaiheesta kolme alkaen mallissa mennään käytännön tekemisen puolelle, joka jää tämän tutkimuksen laajuuden ulkopuolelle. Näitä kaikkia vaiheita on käsitelty tämän tutkimuksen teoriaosuudessa (erityisesti luvussa 4.4), jotta tämän tutkimuksen laajuuden ulkopuolelle jäävät vaiheet voidaan toteuttaa myöhemmin. Esitelty prosessimalli antaa suoraviivaiset askeleet etenemiselle, jos kohdeyritys haluaa tämän tutkimuksen pohjalta lähteä tavoittelemaan hyötyjä.

Kahteen ensimmäiseen prosessivaiheeseen haettiin vastausta mahdollisimman monipuolisesti osallistavan työpajan avulla. Pavelin et al. (2014) määrittelee osallistavien työpajojen menetelmän olevan joukko strukturoituja tapahtumia ryhmälle, jonka osallistujien tarkoitus on etsiä ongelmaa ja sen ratkaisua tiettynä ajan hetkenä tietyssä paikassa. Erityisesti tämä menetelmä rohkaisee luovaan ajatteluun ja voi tuottaa nopeita ideoita ja ratkaisuja. Eroa perinteisiin tapaamisiin syntyy siitä, että menetelmän tarkoitus on stimuloida luovuuteen yhteisen toiminnan kautta. Menetelmä sopii esimerkiksi hyvin ongelmien ratkaisemiseen sekä prioriteettien, strategioiden ja visioiden päättämiseen. (Pavelin et al. 2014.) Osallistava työpaja nimettiin tarkemmin *tulevaisuusverstaaksi*, koska sen määritelmissä tuntui olevan kaivattuja lisäyksiä, mitkä puuttuivat osallistavien työpajojen määritelmistä. Tulevaisuusverstaan perusidea on löytää ratkaisuja liittyen tulevaisuuteen, jotka saattaisivat jäädä huomioimatta erityisesti yksilöhaastatteluilla. Sen tarkoituksena on, että verstaaseen osallistuvat henkilöt täydentävät toistensa ideoita, ja saavat muiden ideoista lisää omia ideoita. Tämä mahdollistaa niin sanottujen heikkojen signaalien löytämisen, mikä on yksi tärkeimmistä tavoitteista tulevaisuusverstaissa. Heikoilla signaaleilla tarkoitetaan oivallusta, josta voi tulevaisuudessa tulla jotain merkittävää. (Vainio 2009.)

Kiimamaa (2003) esittää, että tulevaisuusverstaan toteutus voidaan jakaa neljään eri vaiheeseen. Ensimmäinen vaihe on valmisteleva, jossa tarkoituksena on johdattaa tutkittavat ongelman ääreen ja esitellä verstaan tavoitteet. Toisessa vaiheessa tutkittavat keskittyvät itse ongelmaan ja sen epäkohtiin. Näkemykset kirjataan ylös paperille. Kolmannessa vaiheessa tarvitaan luovuutta, kun aletaan kerätä ratkaisuja edellisessä kohdassa ilmenneisiin

ongelmiin. Ratkaisuja ei kritisoida tässä vaiheessa. Neljännessä vaiheessa evaluoidaan luotuja ratkaisuja. Pavelin et al. (2014) lisää, että toteutuksessa on todella tärkeätä kirjata ylös kaikki tutkimuksesta saadut artefaktit eli tulokset, jotta ne muistetaan jatkossa.

Tulevaisuusverstaan kysymystenasettelussa päädyttiin kahteen erilaiseen malliin. Kysymyksiä esitettiin sekä yksilöiden että laajemmin organisaation näkökulmista. Yksilönäkökulman kysymyksiin päätettiin ottaa hyvin täsmälliset kysymykset liittyen aiheeseen. Toisaalta taas kysymykset eivät saaneet olla liian täsmällisiä, koska silloin ei olisi saavutettu haluttuja tuloksia. Esimerkiksi varsinaisen tutkimuskysymyksen esittäminen ei tutkijan hypoteesin mukaan olisi ollut antoisa: *”Mitä tehokkuushyötyjä voidaan saavuttaa asiakasviestien automaattisella luokittelemisella finanssialan yrityksen asiakaspalvelussa?”*. Organisaatiotason näkökulmasta aihetta lähestyttiin sovelletulla gap-analyysillä. Gap-analyysilla tarkoitetaan menetelmää, jossa esitetään tutkittaville nykytila ja haluttu lopputila (Business Dictionary 2016). Tutkittavat pohtivat tämän jälkeen, mitä ongelmia ratkaisemalla tähän tilanteeseen päästään.

5.2 Tulevaisuusverstaan toteutus

Verstaan toteutukseen käytettäviin resursseihin tutkija sai tietynlaiset reunaehdot kohdeyritykseltä. Kohdeyrityksen ohjaaja määritteli, ketkä tilaisuuteen kutsutaan ja missä tilaisuus järjestetään. Tutkija itse sai määritellä kuinka kauan tilaisuus kestää, mutta tähän vaikutti kohdeyrityksessä vallitseva organisaatiokulttuuri. Lähtökohtaisesti tutkija on havainnoinut, että kohdeyrityksessä kokoukset kestävät keskimäärin tunnin, jonka jälkeen on jo seuraavan aihepiirin kokouksia. Lisäksi jokaisella osallistuvalla henkilöllä on lopulta oikeus päättää itsenäisesti, aikooko osallistua kokoukseen. Vaikka Pavel et al. (2014) määrittelevät, että työpajojen tulisi kestää minimissään 2-3 tuntia, ja parhaimmat tulokset saadaan kokonaisessa päivässä, ei tutkija kokenut tätä tapahtuman markkinoinnin kannalta järkeväksi. Ottaen kaikki tekijät huomioon, tutkija päätti työpajan keston olevan optimaalinen ja tarpeiden mukainen, kun se kestää puolitoista tuntia.

Verstaaseen kutsuttiin kohdeyrityksen työntekijöitä, jotka tietävät asiakaspalvelun päätöksentekoon ja tehokkuuteen liittyvistä ongelmista. Käytännössä työn tilaaja kutsui tilaisuuteen useampia tiimipäälliköitä, joilla oli kokemusta myös varsinaisesta asiakaspalvelutyöstä, muutaman prosessikehitysasiantuntijan, asiakaspalvelijoiden työvuorovastaavan ja asiakaskokemuksen parantamiseen keskittyvän henkilön. Edellä mainitut toimihenkilöt liittyivät suoranaisesti asiakaspalvelutyöhön. Tämän lisäksi tilaisuuteen osallistui kaksi yksikön päällikköä, joiden mielenkiinnon kohteisiin kuuluivat myös asiakaspalvelun kehittämiseen liittyvät toimenpiteet, vaikka heidän toiminta-alueensa on laajempi. Verstaaseen osallistui yhteensä yksitoista henkilöä, joista yksikön päälliköt ehtivät olemaan vain puoli tuntia paikalla. Tämä osallistujamäärä oli riittävä huomiomaan aihetta laaja-alaisesti.

Verstaan tulokset kerättiin kahdella eri tavalla. Ensisijaisesti tärkeimmät ajatukset kerättiin tutkijan muodostamaan posteriin. Tämä posteri toimi koko verstaan keskiössä. Posteriin kerättiin tutkimukseen osallistuneiden vastauksia post-it –lappujen avulla. Toisekseen posti-it lapuille tiivistettyjen vastauksien perusteluja kerättiin nauhoittamalla koko tulevaisuusverstaan äänimaailma. Tämän tarkoitus tutkimuksen kannalta oli varmistaa, että posteriin kerätyt näkökulmat on ymmärretty oikein. Posteriin oli tarkoitus kerätä niukasti muutamilla sanoilla ydinasioita, joita henkilöt perustelivat keskustelemalla. Epäselvyystilanteita varten nauhoite on saatavilla tutkijalta 31.1.2017 asti, jonka jälkeen se poistetaan lopullisesti. Tällaisen epäformaalin tilaisuuden nauhoitteita ei ole tarkoituksenmukaista säilyttää.

Tutkija koki, että aihealueen suoranainen käsitteleminen liiketoiminnan kanssa on hiukan haastavaa teknisillä termeillä, joita tämän tutkimuksen teoriaosuudessa on käytetty. Tämän takia tulevaisuusverstaassa käytettiin mahdollisimman arkikielistä termistöä ja ymmärrettäviä ajatusmalleja. Tämä antoi tutkijalle vastuun peilata saatuja tuloksia teoriaan. Tämä päätös perustui tutkijan oletukseen, että näin päästään parhaimpiin tuloksiin. Verstaan agenda eteni kuvan 13 mukaisesti, joka on sovellettu erityisesti aiemmin esitellyistä Kiimamaan (2013) ja Webb et al. (2011, ss. 4-5) viitekehyksistä. Näiden kahden eri aihepiirien viitekehykset yhdistettiin sopivaksi kokonaisuudeksi.

Kuva 13: Työpajan agenda. Sovellettu Kiimamaa (2003) ja Webb et al. (2011, ss. 4-5)

Kuvassa 13 esitetyssä mallissa ensimmäisessä vaiheessa esivalmisteltiin verstaaseen osallistuneet henkilöt johdattelemalla heidät aihepiiriin, kertomalla miten aihepiiri liittyy kohdeyritykseen ja esittelemällä tutkimuksen tavoitteet. Toisessa vaiheessa poikettiin Kiimamaa (2003) esittämästä mallista, ja haettiin tietoa, jolla voitaisiin parantaa tulevaisuusverstaaseen osallistuneiden ja asiakaspalvelijoiden päätöksentekoa. Ajatuksena oli, että paremmille päätöksillä voidaan vaikuttaa suoraan tai epäsuoraan asiakaspalvelun tehokkuuteen. Lisäksi tämän vaiheen tarkoitus oli aktivoida verstaaseen osallistuneet henkilöt ajattelemaan oikeita asioita täsmällisillä kysymyksillä. Tämä vaihe toteutettiin esittämällä kaksi kysymystä, joihin heidän tuli vastata yksittäin:

1. Jos saisit tietää ihan mitä vaan vuoden sisällä tulleista asiakasviesteistä, mikä olisi arvokkainta tietoa osana päätöksentekoasi?
2. Jos asiakaspalvelija saisi tietää ihan mitä vaan vuoden sisällä tulleista asiakasviesteistä, mikä tieto toisi mielestäsi eniten arvoa hänen päätöksentekoon?

Kolmannessa vaiheessa syvennyttiin tutkimuksen tavoitteisiin laajemmalla tasolla sovelletulla gap-analyysin menetelmällä. Tutkittavat asetettiin kahteen eri ryhmään etsimään juurisyitä, minkä takia tulevaisuusverstaassa annettuihin tavoitteeseen on hankala päästä. Molemmille ryhmille annettiin oma kysymyksensä liittyen tavoitteisiin, johon heidän tuli vastata. Kysymykset ryhmille olivat:

1. Miksi asiakkaat lähettävät meille niin paljon viestejä?
2. Miksi asiakaspalvelijat ovat niin hitaita vastaamaan asiakasviesteihin?

Tämän jälkeen neljännessä vaiheessa suoritettiin löytyneiden ongelmien evaluointi. Tarkoituksena oli löytää ongelma, minkä ratkaiseminen tuo suurimman liiketoimintahyödyn. Tämä eroaa Kiimamaa (2003) mallista siten, että siinä evaluoidaan ratkaisuja, kun tässä tapauksessa evaluoinnin kohteena ovat ongelmat. Evaluointi tehtiin siten, että jokaisen henkilön piti antaa kolme pistettä ratkaisulle, joka tuo suurimman hyödyn. Kaksi pistettä ratkaisulle, joka tuo toisiksi suurimman ja yksi piste sille, joka tuo kolmanneksi suurimman hyödyn. Pisteitä annettiin vain oman ryhmän löytämille ongelmille. Täten aiemmin esitetyille kahdelle eri kysymykselle löytyi molemmille kolme suurinta ongelmaa. Viimeisessä eli viidennessä vaiheessa pureuduttiin löydettyihin suurimpiin ongelmiin. Aiemmin luodut ryhmät miettivät, mitä tietoa tarvitaan tulevaisuudessa, jotta ongelma ratkeaa.

Taulukko 4: Empiirisen osion toteutus

Tutkimusmenetelmä	Tulevaisuusverstas
Tutkimukseen osallistuneiden määrä	9 henkilöä oli koko verstaan ajan. 2 henkilöä poikkesi 30 minuutiksi.

Osallistuneiden suhde tutkimusalueeseen	Asiakaspalvelun tiimipäälliköitä 5 henkilöä, Asiakaspalvelun prosessikehitysasiantuntijoita 2 henkilöä, asiakaskokemuksesta vastaava, voimavarasuunnittelija, yksikönpäälliköitä 2 henkilöä
Tutkimukseen käytetty aika	1,5 tuntia

6. TUTKIMUSTULOKSET

Tässä luvussa esitetään luvussa 5 kuvaillun tulevaisuusverstaan tulokset. Tulevaisuusverstaassa haettiin vastausta avustavaan tutkimuskysymykseen: *”Mitä tehokkuuteen ja päätöksentekoon liittyviä hyötyjä voidaan jalostaa asiakasviestidatasta kohdeyrityksessä?”*. Luvussa 5 on esitelty perustelut käytetyille kysymyksille ja menetelmille. Tutkimuksen tulokset kerättiin post-it lapuilla posteriin, joten tässä kappaleessa on esiteltynä tulokset lyhytsanaisesti. Pohdintaluvussa on käyty tuloksia lävitse teoriaan suhteutettuna.

6.1 Päätöksenteon parantaminen

Päätöksentekoa tutkivia kysymyksiä oli kaksi kappaletta. Näiden kysymysten tarkoitus oli selvittää, minkä tuntemattoman tiedon tutkittavat kokevat arvokkaimmaksi yksilötasolla. Ensimmäisen kysymys oli: *”Jos saisit tietää ihan mitä vain vuoden sisällä tulleista asiakasviesteistä, mikä olisi arvokkainta tietoa osana päätöksentekoasi?”* Tämän kysymyksen tulokset ovat esiteltynä taulukossa 5.

Taulukko 5: Ensimmäisen yksilökysymyksen vastaukset: "Jos saisit tietää ihan mitä vain vuoden sisällä tulleista asiakasviesteistä, mikä olisi arvokkainta tietoa osana päätöksentekoasi?"

Vastaus
Sisältö. Missä kaikkialla viesti on kiertänyt ja kuka sen lopulta on hoitanut → kenelle viesti olisi heti kuulunut ohjautua
Ratkaisuaste → Kuinka monta viestiä on voitu hoitaa ja asia ratkaista suoraan lisäarvoa tuottaen
Kuinka paljon pystymme hoitamaan kuntoon ensimmäisen kontaktikäsittelyn aikana?
Ratkaisukyky. Sisältääkö viesti riittävät tiedot asian hoitamiseen? Olisiko asiakas voinut hoitaa asian itse verkkopalvelussa.
Ratkaisukyky. Yhteydenoton aihe
Kuinka monessa viestissä on ohitettu myynninpaikat?
Pääaihealueet. Millä asialla asiakkaat meitä lähestyvät?
Mikä oli primäärikontaktin syy?

Miksi asiakas ei pystynyt hoitamaan asiaa verkossa itse? Mitkä asiat nousee useimmin esiin?

Näistä vastauksista esiin pistää kaksi usein mainittua teemaa. Ensimmäinen teema liittyy ratkaisukykyyn. Jopa puolet tutkittavista koki, että ratkaisukykyyn liittyvä data olisi arvokkainta tietoa osana heidän päätöksentekoaan. Toinen puolikas oli kiinnostunut sisältöön liittyvistä asioista. Tarkemmilla sisältöanalyyseilla voidaan tuottaa heille arvokkainta tietoa.

Toinen kysymys oli: *”Jos asiakaspalvelija saisi tietää ihan mitä vaan vuoden sisällä tulleista asiakasviesteistä, mikä tieto toisi mielestäsi eniten arvoa hänen päätöksentekoon?”*. Kysymyksen vastaukset on esitetty taulukossa 6.

Taulukko 6: Toisen yksilökysymyksen vastaukset: "Jos asiakaspalvelija saisi tietää ihan mitä vaan vuoden sisällä tulleista asiakasviesteistä, mikä tieto toisi mielestäsi eniten arvoa hänen päätöksentekoon?"

Vastaus
Mitä asiaa kysytään useimmiten?
Mitkä asiat asiakas kokee tärkeimmiksi tuottamassamme palvelussa?
Miten asiakas haluaa hänelle vastattavan ja mikä siinä on tärkeintä?
Mikä on asiakkaalle tärkeää?
Mitä osaamista ja ohjeistusta tarvitsen usein / harvoin?
Auttoiko annettu ratkaisu asiakasta kerralla?
Kuinka paljon vuoden aikana on mennyt ohi lisäarvon tuottamisen paikkoja?
Ratkaisukyky. Kuinka monessa tilanteessa asiakasneuvoja itse on tai ei ole voinut auttaa asiakasta ja tuottaa lisäarvoa? Miksi ei?
Ratkaisukyky. Miten asiakkaan ongelma on saatu kerralla kuntoon? Onko asiakkaan vielä pitänyt palata asian tiimoilta uudestaan?

Myös tässä kysymyksessä pistää samat teemat esiin kuin aiemmassa. Ratkaisukykyyn ja sisältöön liittyvät asiat toisivat huomattavaa parannusta asiakaspalvelijoiden päätöksentekoon. Tämän lisäksi asiakasarvoon liittyviä vastauksia tuli muutama.

6.2 Ongelmien löytäminen ja evaluointi

Tässä kappaleessa on pureuduttu ryhmänä annettuihin kysymyksiin. Tarkoitus oli kollektiivisesti selvittää, minkälaisia ongelmia organisaatiossa koettiin olevan. Ryhmä jaettiin kahtia ja ensimmäiselle ryhmälle annettiin kysymys: *”Miksi asiakkaat lähettävät meille niin paljon viestejä?”*. Ryhmä sisälsi neljä henkilöä. Ongelmien löytymisen jälkeen niille annettiin pisteitä sillä perusteella, minkä ongelman ratkaiseminen heidän mielestään tuo suurimman liiketoimintahyödyn. Pisteytyksen idea oli priorisoida suurimmat ongelmat. Kysymyksen vastaukset ja niille annetut pisteet on koottu taulukkoon 7.

Taulukko 7: Ensimmäisen ryhmäkysymyksen vastaukset: "Miksi asiakkaat lähettävät meille niin paljon viestejä?"

Ongelma	**Pisteet evaluoinnista**
Asiakas kyselee keskeneräisistä asioista, koska ei tiedä koska häntä palvellaan	*10*
Asiakas ei voi hoitaa asiaansa verkkopalvelussa	*8*
Meiltä lähtevän virheellisen informaation vuoksi viestiketjut laajenevat	*6*
Helppo kanava ikävien asioiden hoitoon	-
Halpa kanava	-
Osa asiakkaista on jatkuvassa yhteydessä kohdeyritykseen	-
Viestikanava käytössä 24/7	-
Ei jonotusta	-

Pisteiden jakautumisesta voidaan huomata, että ryhmässä vallitsi yhdenmielisyys ongelmien priorisoinnin suhteen. Ryhmän yksilöiden tuli jakaa kolmelle eri kohdalle pisteitä. Kaikki pisteet menivät kolmelle ongelmalle. Ne ongelmat, jotka eivät pisteitä saaneet, eivät ole tutkittavien mielestä tärkeimpiä kehityskohteita. Tutkittavat kokivat, että erityisesti kolmen syyn takia asiakkaat lähettävät kohdeyritykselle niin paljon viestejä. Nämä suurimmat ongelmat ovat järjestyksessään:

1. Asiakas kyselee keskeneräisistä asioista, koska ei tiedä koska häntä palvellaan.
2. Asiakas ei voi hoitaa asiaansa verkkopalvelussa.

3. Meiltä lähtevän virheellisen informaation vuoksi viestiketjut laajenevat.

Toiselle ryhmälle annettiin kysymys: *”Miksi asiakaspalvelijat ovat niin hitaita vastaamaan asiakasviesteihin?”*. Ryhmä koostui viidestä eri henkilöstä. Tämän ryhmän vastaukset ja niille annetut pisteet on koottu taulukkoon **x.**

Taulukko 8: Toisen ryhmäkysymyksen vastaukset: "Miksi asiakaspalvelijat ovat niin hitaita vastaamaan asiakasviesteihin?"

Ongelma	**Pisteet evaluoinnista**
Järjestelmien moninaisuus vaikeuttaa tiedon löytämistä	*10*
Organisaation toimintatavat, -mallit ja räätälöinti luovat vaihtelua	*7*
Jälkitöiden vaatimukset	*6*
Epäoptimaalinen työvirranohjaus suhteessa saapuviin asiakasviesteihin	*3*
Viestin sisältö edellyttää yhteydenpitoa muihin sidosryhmiin tai tukeen	*3*
Samojen tietojen tallentaminen useaan paikkaan	*1*
Asiakasviestien moninaisuus	-
Työajan käyttö oikeaan asiaan	-
Motivaatio, asenne ja työssäjaksaminen	-
Epätasainen työvirta	-
Toimintamallien noudattamatta jättäminen	-
Keskeytykset	-
Asiakas ei osaa käyttää palvelua, jolloin he voivat antaa väärän otsikkotarkenteen	-

Puutteet osaamisessa	-

Tässä ryhmässä ei oltu niin yhdenmielisiä ongelmien priorisoinnin suhteen kuin ensimmäisen ryhmän evaluoinnissa. Lisäksi toinen ryhmä kehitteli useampia vastausvaihtoehtoja. Ongelmat, jotka eivät pisteitä saaneet, eivät ole tutkittavien mielestä tärkeimpiä kehityskohteita. Kolme tärkeintä asiaa, minkä takia asiakaspalvelijat eivät pysty toimimaan tehokkaasti, ovat:

1. Järjestelmien moninaisuus vaikeuttaa tiedon löytämistä.
2. Organisaation toimintatavat, -mallit ja räätälöinti luovat vaihtelua.
3. Jälkitöiden vaatimukset

Näiden ongelmien ratkaiseminen tuo suurimman tehokkuuteen liittyvän liiketoimintahyödyn. Lisäksi osa tutkittavista koki seuraavat asiat tehokkuutta heikentäviksi: epäoptimaalinen työvirranohjaus suhteessa saapuviin asiakasviesteihin, viestin sisältö edellyttää yhteydenpitoa muihin sidosryhmiin tai tukeen ja samojen tietojen tallentaminen useaan paikkaan. Näihin mainittuihin ongelmiin ei lähdetä etsimään ratkaisua tämän tutkimuksen yhteydessä.

6.3 Ratkaisujen etsiminen

Edellisessä kappaleessa esiteltiin tapa, jolla löydettiin kolme merkityksellisintä ongelmaa molemmista tehokkuuteen vaikuttavista kysymyksistä. Tämän jälkeen mietittiin, millä tiedolla ongelmaa pystytään ratkaisemaan. Taulukoissa 9, 10 ja 11 on esitetty kolme merkityksellisintä tiedonlähdettä, jolla pystytään vaikuttamaan kysymykseen: *”Miksi asiakkaat lähettävät meille niin paljon viestejä?”*.

Taulukko 9: Ensimmäisen ryhmäkysymyksen (asiakkaiden kysymysten suuri määrä) suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Asiakas kyselee keskeneräisistä asioista, koska ei tiedä koska häntä palvellaan (esim. lainaprosessi)
Tilannetiedon välittäminen asiakkaalle
Asiakkaalle tieto prosessin alussa sen odotetuista aikatauluista
Pitäydytään luvatussa aikataulussa. Ilmoitetaan asiakkaalle jos luvatussa aikataulussa ei pysytä

Tutkittavat kokivat, että asiakkaat eivät kysele niin paljoa keskeneräisistä asioista, jos heille toimitetaan tilannetietoa prosessin aikana sekä prosessin alussa. Lisäksi pelkän tilannetiedon toimittaminen ei ole riittävä, vaan kohdeyrityksen tulee pysyä aikataulussaan.

Taulukko 10: Ensimmäisen ryhmäkysymyksen (asiakkaiden kysymysten suuri määrä) toisiksi suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Asiakas ei voi hoitaa asiaansa verkkopalvelussa
Uusi verkkopalvelu
Automatisoinnin laajempi hyödyntäminen esim. ohjelmistorobotiikalla
Verkkopalvelun hakukoneen parantaminen

Tutkittavat kokivat, että asiakas voisi hoitaa asiansa verkkopalvelussa, jos kohdeyrityksellä olisi parempi ja monipuolisempi verkkopalvelu. Toisaalta he myös kokivat, että nykyisessä verkkopalvelussa hakukone ei ole toimiva. Tämän parantaminen vähentäisi asiakkaiden kysymyksiä. Kolmannekseen he vastasivat laajemmin, että automatisoinnin hyödyntäminen esimerkiksi ohjelmistorobotiikan avulla vähentäisi kysymysten määrää. Tällä tutkittavat viittasivat, että asiakkaat kyselevät keskeneräisten asioiden perään, koska tietokannoissa tiedot eivät päivity reaaliajassa. Ohjelmistorobotiikalla tieto saataisiin virtaamaan tehokkaammin järjestelmästä toiseen.

Taulukko 11: Ensimmäisen ryhmäkysymyksen (asiakkaiden kysymysten suuri määrä) kolmanneksi suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Meiltä lähtevän virheellisen informaation vuoksi viestiketjut laajenee
Mallivastausten kuntoon laittaminen
Virheiden tarkempi tilastointi, ja tähän perustuva systemaattinen korjaus sekä osaamisen kehittäminen

Kohdeyritykseltä lähtee virheellistä informaatiota asiakkaille, koska mallivastaukset eivät ole kohdallaan. Toisekseen, virheitä ei tilastoida tarpeeksi systemaattisesti. Tämä johtaa siihen, että kohdeyrityksessä ei olla tietoisia mitä virheitä asiakaspalvelijat tekevät usein. Näin ei osata puuttua ja korjata suurimpia ongelmia.

Seuraaviin kolmeen taulukkoon 12, 13 ja 14 on kerätty tietoa, millä pystyttäisiin vastaamaan kysymykseen: *”Miksi asiakaspalvelijat ovat niin hitaita vastaamaan asiakasviesteihin?”*.

Taulukko 12: Toisen ryhmäkysymyksen (asiakaspalvelijoiden hitaus) suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Järjestelmien moninaisuus vaikeuttaa tiedon löytymistä
Järjestelmien yhtenäistäminen
Robotiikka
Tuplanäytöt

Tutkittavat kokivat, että asiakaspalvelijat löytäisivät tiedon nopeammin, jos järjestelmät olisivat yhtenäisiä. Tällä hetkellä eri järjestelmissä tieto tulee hankkia eri tavalla riippuen kohdejärjestelmästä. Lisäksi joidenkin järjestelmien tiedot eivät ole keskenään linjassa. Järjestelmistä saataisiin enemmän yhtenäisiä, jos ohjelmistorobotiikalla puututtaisiin järjestelmien synkronointiin toistensa kanssa. Myös tuplanäytöt toisivat asiakaspalvelijoille enemmän tietoa näkymiin yhdellä vilkaisulla.

Taulukko 13: Toisen ryhmäkysymyksen (asiakaspalvelijoiden hitaus) toisiksi suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Organisaation toimintatavat, - mallit ja räätälöinti
Yhtenäinen tapa toimia organisaatiossa (yhtenäinen hinnasto ja toimintamallit). ”Olemme me, ei se ja tuo” -ajattelu

Tutkimuksessa paljastu, että organisaatiossa on liikaa vaihtelua yksiköiden toimintatapojen ja –mallien välillä. Tämä asia voitaisiin ratkaista yhtenäistämällä organisaation toimintaa hinnastojen ja toimintamallien avulla. Tutkittavat kaipasivat enemmän yhteisöllisyyttä organisaatiossa.

Taulukko 14: Toisen ryhmäkysymyksen (asiakaspalvelijoiden hitaus) kolmanneksi suurin ongelma ja sen ratkaisuehdotukset.

Ongelma: Jälkitöiden vaatimukset
Vähennetään tietovaatimuksia, koska tietoa ei hyödynnetä
Laskutukseen liittyvät tiedot automaattisesti asiakastiedoista yhdistämällä
Toimintatapojen leanaus

Asiakaspalvelijat toimisivat nopeammin, jos heillä ei olisi niin paljon vaatimuksia tiedon

tallentamisen suhteen. Tutkittavat kokivat, että osa jälkitöistä on turhia, jolloin vaatimuksia voitaisiin vähentää. He myös totesivat, että automatisoinnilla voitaisiin vähentää jälkitöitä erityisesti laskutukseen liittyvissä asioissa. Kolmannekseen he mainitsivat, että toimintatavoissa on ylimääräisiä asioita, joita voitaisiin kehittää lean-periaatteilla.

7. POHDINTA

Tässä luvussa koostetaan empiriaosuuden pohjalta, mitä tehokkuutta nostavia hyötymahdollisuuksia on tunnistettavissa kohdeyrityksessä. Näitä hyötymahdollisuuksia sovitetaan tämän tutkimuksen teoriaosuuden tuloksiin ja pohditaan, kuinka ne olisivat toteutettavissa. Näin ollen tässä luvussa luodaan kokonaisuudesta synteesi, jossa annetaan potentiaalisia kehitysehdotuksia kohdeyritykselle.

Empiriaosuudessa kerättiin tehokkuuteen liittyviä ongelmia sekä niihin vastaavia ratkaisuehdotuksia. Tässä luvussa ei varsinaisesti tartuta enää itse ongelmiin, vaan pohditaan, kuinka ratkaisuehdotuksia voidaan viedä käytäntöön. Näillä ratkaisuehdotuksilla pystyttäisiin parantamaan päätöksentekoa ja asiakaspalvelijoiden tehokkuutta, sekä vähentämään asiakkaiden kontaktimääriä. Ratkaisuehdotuksia on analysoitu ja koostettu taulukkoon, joka löytyy liitteestä 1. Analyysissä on pohdittu mitä ennakkoedellytyksiä tarvitaan luokittelun näkökulmasta, jotta ratkaisua voidaan hyödyntää. Analyysissä eri keinojen määrä pyrittiin pitämään mahdollisimman vähäisenä, jotta aihetta olisi helpompi käsitellä.

Tuloksia käsitellään ja peilataan teoriaan kolmivaiheisesti. Ensimmäisessä vaiheessa pohditaan, miten asiakasviestidatasta voidaan saada kilpailuetua kohdeyrityksessä. Toisessa osiossa pohditaan, kuinka asiakasviestejä tulisi käsitellä automaattisesti. Kolmannessa osiossa annetaan suositus kohdeyritykselle kuinka he voivat lähteä rakentamaan automaattista luokittelujärjestelmää. Nämä kysymykset yhdessä vastaavat siihen, kuinka kohdeyritys voi saada tehokkuutta ja päätöksentekoa tukevia hyötyjä asiakaspalvelussa.

7.1 Asiakasviestidatasta kilpailuetua

Tulevaisuusverstaassa saadun datan perusteella vaikuttaa siltä, että kohdeyrityksen liiketoimintatiedon hallinnassa on monipuolisesti kehitettävää. Monipuolisuudella tässä yhteydessä tarkoitetaan, että löydettyjen ratkaisuehdotusten toteuttamiseen riittää perinteisemmät tavat, joihin ei tarvita automaattista luokittelemista. Saatuja ratkaisuehdotuksia voidaan tavoitella muilla tavoilla, kuten järjestelmä- ja laitehankinnoilla, yleisellä automatisointiasteen nostamisella, muutosjohtamisella, asiakasarvon tutkimisella ja lean-periaatteiden hyödyntämisellä. Tässä luvussa pohditaan saaduista ratkaisuista erityisesti niitä, joihin voidaan hyödyntää asiakasviestidataa.

Saatujen ratkaisuideoiden analyysistä esiin pistää teema, että moni toisistaan riippumaton taho haluaisi samoja asioita osaksi omaa, asiakkaan tai asiakaspalvelijoiden päätöksentekoa. Lisäksi tiedetään, että tietotarpeisiin vastaavaa dataa on olemassa, mutta sitä ei jalosteta sopivaan muotoon tai sitä ei jaeta tehokkaasti ympäri organisaation. Tällä hetkellä yrityksen arvokasta, harvinaista ja vaikeasti korvattavissa olevaa dataa ei saada muutettua

kestäväksi kilpailukyvyksi (Laihonen et al. 2013, s.24). Vaikuttaa jopa siltä, että tällä hetkellä datasta ei saada jalostettua tehokkaasti lyhyen aikavälin kilpailuetua perus- ja erikoisraporttien tai selvityksien avulla (Davenport & Harriksen 2007, ss. 26-27). Koska kohdeyritys on kiinnostunut aiheesta ja tuki tämänkin tutkimuksen toteutusta, suunta on oikea.

Erityisesti liitteen 1 analyysissä paljastuu, että viestidatasta haluttaisiin tietää ratkaisukykyyn ja sisältöön liittyviä asioita, jotka kuuluvat Davenport et al (2010, ss. 6-7) normaalin raportoinnin piiriin. Ne ensisijaisesti vastaavat kysymykseen *mitä tapahtui*. Ratkaisukykyasioissa halutaan tiedonkeräyksen perusteella tietää, kuinka nopeasti viestiketju ollaan saatu hoidettua, kuinka monella viestillä asia saatiin ratkaistua sekä kuka asiakasta miellyttävän vastauksen on tarjonnut. Käytännössä voidaan ajatella, että nämä asiat saadaan laskettua kolmen muuttujan avulla. Tarkastelemalla viestiketjun ensimmäisen ja viimeisen viestin aikaleimoja saadaan tietää, kuinka nopeasti asiakkaan ongelma on ratkaistu. Voidaan myös olettaa, että viestiketjun eri viestien määrä kertoo siitä, kuinka monella viestillä asia saatiin hoidettua. Jos viestien määrä on esimerkiksi kaksi kappaletta, viestiketju saatiin hoidettua asiakaspalvelijan ensimmäisellä vastauksella. Kolmannekseen, voidaan olettaa, että viimeisimmän viestin lähettänyt asiakaspalvelija on ratkaissut ongelman. Tutkija ei tiedä miten asiakasviestit ovat todellisuudessa tallennettuna tietovarastoon, mutta oletetaan niiden lähestyvän relaatiotaulukon 15 mukaista rakennetta.

Taulukko 15: Oletettu viestien rakenne kohdeyrityksen relaatiotietokannoissa.

Viestiketju-dokumentti	**Ensimmäinen viesti**	**Viestin otsikko**	**Viestien määrä ketjussa**	**Viimeisin asiakaspalvelija**	**Aikaleimaus**
D_1	Viesti1	Lainat	5	Jani	25.11.06-28.11.06
D_2	Viesti2	Kortit	2	Sami	26.11.06-27.11.06
D_3	Viesti3	Lainat	3	Jani	26.11.06-28.11.06
D_4	Viesti4	Lainat	2	Sami	26.11.06-26.11.06

Kuten aiemmin tässä työssä on esitetty (luku 3), viestejä luokitellaan jo asiakkaan toimesta. Tämä on esitetty relaatiotaulukossa 15 sarakkeessa ”viestin otsikko”. Ongelma on,

että viestien otsikot ovat aihepiireiltään niin laajoja, että niiden perusteella saatu statistiikka ei välttämättä tuo arvoa kuin suuremmissa organisaatiota koskevissa päätöksissä. Mitä pienemmäksi luokat muodostetaan, sitä pienempiin päätöksiin statistiikkaa voidaan hyödyntää. Tämä edellyttää myös, että statistiikka tuodaan näkyviin tiimi- ja yksilötasolle (Laihonen et al. 2013, s. 49). Luokittelua voidaan tehdä tarkemmaksi manuaalisesti tai hyödyntämällä automatiikkaa. Esimerkiksi ”lainat” luokituksen saaneita viestejä voidaan tarkemmin luokitella ”uuden lainan hakeminen” tai ”vanhan lainan maksaminen” vaihtoehtoihin. Ajatellaan, että D_1 ja D_3 viestit ovat molemmat ”uuden lainan hakeminen” –tarkennuksen omaavia viestejä. Tällöin voidaan vanhan viestihistorian perusteella laskea ratkaisukykyyn liittyviä attribuutteja taulukon 16 mukaisesti.

Taulukko 16: Ratkaisukykyyn liittyvien asioiden selvittäminen oletetun viestirakenteen perusteella.

Viestin sisältötarkennus	Keskimääräinen vastausaika (päivää)	Keskimääräinen viestiketjun koko	Aihealueen asiantuntija
Uuden lainan hakeminen	(4 + 3)/2 = 3,5 päivää	(5+3)/2 = 4 viestiä	Jani

Toisekseen, tutkittavia kiinnosti omista ja organisaation päätöksentekoa koskevista näkökulmista erityisesti viestien sisältöä koskevat tiedot. Kiinnostavia asioita olivat esimerkiksi millä asialla asiakkaat lähestyvät kohdeyritystä, olisiko asiakas voinut tehdä tämän asian itsepalveluna, miten voidaan parantaa verkkopalvelua ja sen hakukonetta, kuinka monessa viestissä on ohitettu myynninpaikat ja miten mallivastaukset saadaan laitettua kuntoon. Näihin kysymyksiin saadaan osittain vastauksia analysoimalla tarkemmin luokiteltuja asiakasviestejä. Asiaa havainnollistaa relaatiotaulukko 17. Jo pelkästään tietämällä tarkkojen aihealueiden kappalemäärät, voidaan priorisoida kehityskohteita. Taulukossa on havainnollistettu, että ”uuden lainan hakeminen” –viestit ovat yleisiä suhteessa muihin aiheisiin, ja niihin vastaamiseen menee suhteellisen kauan. Tästä voidaan tehdä johtopäätöksiä sekä analyyseja useista edellä mainituista näkökulmista.

Taulukko 17: Kaivattu uusi viestirakenne, jossa oikeat tiedot ovat helposti saatavilla.

Viestin sisältötarkennus	Kappalemäärä viimeisen viikon aikana	Luokalle analysoituja ominaisuuksia	Keskimääräinen vastausaika (päivää)	Keskimääräinen viestiketjun koko	Aihealueen asiantuntija

Uuden lainan hakeminen	2 (D_1, D_3)	Myynnillinen, vaikea itsepalvelu, asiakas usein kysyy toisessa viestissä hinnastoa	3,5	4	Jani
Vanhan lainan maksaminen	1 (D_4)	Helppo itsepalvelu	1	2	Sami

Tarkemmissa sisältöanalyyseissä voidaan esimerkiksi selvittää, mitä asiakas usein kysyy kohdeyrityksen lähettämän vastauksen jälkeen. Tätä asiaa voidaan hyödyntää lisäämällä proaktiivisesti useasti kysytty asia jo osaksi ensimmäistä viestiä, jolloin mahdollisesti pystytään välttämään asiakkaan toinen kontakti. Toisekseen, priorisoitu asia voidaan nostaa paremmin näkyville osana verkkopalvelua ja sen hakukonetta. Kolmannekseen, voidaan miettiä, miten asiakkaat saadaan tekemään kohdennettu asia useammin itsepalveluna. Neljännekseen, tiettyyn luokkaan kuuluvat viestit voidaan analysoida enemmän myynnillisiksi kuin toiset. Tällä tavalla voidaan esimerkiksi kouluttaa, että asiakaspalvelijat tunnistavat myynnin paikat paremmin.

7.2 Asiakasviestien automaattinen käsitteleminen

Edellä mainitut asiat voidaan toteuttaa tekemällä satunnaisotanta massadatasta ja asettamalla se tarkempiin luokkiin käyttämällä asiantuntijoita. Tällä perusteella voidaan statistiikan perusoppien mukaisesti yleistää, että koko massadata noudattaa likimäärin samaa jakaumaa. Tämä on kuitenkin tehoton, epätarkka, aikasidonnainen ja epätarkoituksenmukainen keino, sillä avuksi voidaan ottaa esimerkiksi koneoppimisen ja hahmontunnistuksen menetelmiä. Nämä menetelmät pystyvät reaaliaikaisesti käsittelemään koko massadataa hyödyntäen automatiikkaa. (Miner et al. 2002, ss. 882-884.) Tämän tutkimuksen teoriaosuuden valossa teknologian hyödyntäminen vaikuttaa erityisen mahdolliselta, koska asiakasviestit koskevat suljettua aihepiiriä eli finanssialaa (Agirre et al. 2009). Tämän lisäksi asiakkaat luovat suljetun aihepiirin viesteille entistä suljetummat viestiaiheet valitsemalla viesteille otsikot annetuista vaihtoehdoista. Täten voidaan olettaa, että käytetty sanasto ei ole niin monipuolista tietyn otsikon alla, ja toisiinsa liittyvät dokumentit ovat helpompi tunnistaa. On tosin hyvä huomioida, että asiakas voi antaa virheellisen otsikon, joka aiheuttaa poikkeavuuksia. Näiden takia data tulee siivota ennen menetelmän implementointia.

Tutkimusta tehdessä oli oletus, että kohdeyrityksellä olisi tarve erityisesti ohjatun koneoppimisen ja hahmontunnistuksen menetelmille. Tästä johtuen teoriaosuudessa painopiste oli ohjatulla oppimisella. Ohjattu oppiminen tarvitsee kuitenkin tietyn luokkanimityksen, johon asiakasviestit sovitetaan. Datan keräyksen jälkeen vaikuttaa siltä, että kohdeyrityksellä on erityisesti tietotarpeita tarkempien luokkanimitysten suhteen. Tämä asettaa tutkimuksen oletuksen virheelliseksi. Vasta tarkempien luokitusten jälkeen saadaan vastattua laadukkaammin tutkittavien ratkaisuehdotuksiin. Tarkempien luokitusten muodostamiseen voidaan soveltaa ohjaamattoman oppimisen eli klusteroinnin menetelmiä (Duda et al. 2001, ss. 517-518). Näillä menetelmillä voidaan löytää piirteiden valossa toisiaan lähellä olevat dokumentit, jotka todennäköisesti kuuluvat samaan luokkaan. Toisiaan lähellä oleville dokumenteille annetaan myöhemmin tarkentavia sisältöotsikoita niiden asiasisällön perusteella. Vasta kun ollaan luotu tarkemmat sisältöotsikot, voidaan implementoida ohjatun oppimisen järjestelmä tuomaan tehokkuushyötyjä. (Kotsiantis et al. 2007.)

Tunnistettuihin asiakaspalvelijoiden tehokkuuden pullonkauloihin on hankala tuoda kontribuutiota koneoppimisen ja hahmontunnistuksen menetelmillä. Suurimmat tunnistetut pullonkaulat olivat järjestelmien moninaisuus, organisaation erilaiset toimintatavat ja -mallit sekä räätälöinti ja jälkitöiden vaatimukset. Näistä asioista erityisesti organisaation erilaiset toimintatavat antavat haasteen automatiikan käytölle, jos eri asiakasryhmiä palvellaan eri toimintatavoilla. Mitä enemmän vaihtelua on datan rakenteessa ja sisällössä, sitä enemmän tarvitaan ihmisen tulkintaa (Blumberg & Atre 2003).

Toisaalta, asiakasapalvelijoiden tehokkuuteen voidaan vaikuttaa muilla keinoilla, jota ei datan keräyksen yhteydessä tunnistettu. Tehokkuutta saadaan oivaltavalla analytiikalla, kun menetelmät analysoivat jokaisen saapuvan viestin suhteessa viestihistoriaan, ja teemme tämän perusteella parempia päätöksiä tulevaan liittyen (Davenport et al. 2010, ss. 6-7). Tällainen tilanne vaatii automaattisen luokittelujärjestelmän implementointia osaksi asiakaspalvelukanavaa. Lisäksi se vaatii, että tietyille luokille on analysoitu tietotarpeisiin vastaavia ominaisuuksia. Hyvässä tilanteessa asiakaspalvelijalla on saapuneen asiakasviestin yhteydessä useita uusia työkaluja, millä asiakkaalle voidaan tuottaa entistä parempi asiakaspalvelutilanne. Asiakas olisi erittäin tyytyväinen, jos asiakaspalvelijalla olisi tietoa hänen todellisista tarpeistaan (Beujean et al. 2006). Jos oivaltava analytiikka saadaan onnistumaan, asiakaspalvelijalla voi olla seuraavia tässä työssä relevanteiksi tunnistettuja työkaluja käytössä:

- Viestiluokan asiantuntija (ratkaisukyvyn parantuminen)
- Ongelmaluokkaan sopiva mallipohja (ratkaisukyvyn parantuminen)
- Mitä asiakas oletettavasti kysyy seuraavassa kysymyksessään tässä ongelmaluokassa (ratkaisukyvyn parantuminen)
- Tietoa asiakaskäyttäytymisestä (lisäarvon tuottaminen)

Samalla järjestelmälogiikalla voidaan myös vähentää asiakkaiden kontaktimäärää. Tutkittavat tunnistivat, että asiakkaat ottavat vähemmän kontaktia, jos he saavat tietoa keskeneräisistä asioista, kohdeyritys ei lähetä virheellistä informaatiota ja jos asiakkaat pystyisivät hoitamaan asiansa itsepalveluna. Ensinnäkin, keskeneräisten asioiden kyseleminen voidaan ehkäistä antamalla vanhan viestidatan perusteella odotettuja arvoja kuinka kauan tietyn viestiluokan prosessointi tulee kestämään. Tämä informaatio voidaan toimittaa jo asiakkaan viestin lähetyksen yhteydessä. Tässä tilanteessa asiakas ei todennäköisesti kyselisi ennen odotettua palveluaikaa, koska häntä palvellaan. Toisekseen, järjestelmä voi tunnistaa viestissä haluttuja asioita, jolloin järjestelmä voi tuoda haluttuja tietoja saatavilla. Tämä vähentää virheellisen informaation määrää. Kolmannekseen, tilastoimalla useasti kysyttyjä asioita pystytään kehittämään verkkopalvelua toimivammaksi asiakkaan näkökulmasta. Useasti kysyttyihin asioihin voidaan esimerkiksi luoda toimivammat itsepalvelukanavat.

Toisaalta, vielä paremmassa tapauksessa tekoälyä omaava järjestelmä voi tunnistaa jo ennen kuin asiakas on lähettänyt viestin, että minkä tarkan luokituksen viesti on saamassa. Tällaisessa tapauksessa järjestelmä voi ehdottaa asiakkaalle aiheeseen sopivaa mallivastausta sisältäen informaatiota, linkkejä ja itsepalveluehdotuksia. Jos asiakkaan tietotarpeet saadaan tyydytettyä ennen viestin lähetystä, voidaan koko asiakaskontakti mahdollisesti estää. Mikä olisikaan kaikkien osapuolien mielestä hienompaa kuin se, että asiakkaat tyydyttäisivät tietotarpeensa ennen kuin ehtivät edes viestiä lähettämään?

7.3 Automaattisen luokittelujärjestelmän rakentaminen

Tässä työssä on pyritty antamaan valmiudet erityisesti Webb et al. (2011, ss. 4-5) ohjatun koneoppimisen kehitysmallin vaiheisiin yksi ja kaksi. Ensimmäinen vaihe on ongelman muotoilu. Ongelma on, että asiakkaalle ei toimiteta tarpeeksi tietoa viestin lähetyksen yhteydessä ja toisekseen, asiakaspalvelijalla ei ole tarpeeksi tietoa vastatessaan asiakasviestiin. Nämä vaikuttavat negatiivisesti asiakaspalvelun tehokkuuteen kahta eri kautta. Ensinnäkin asiakas kysyy lisäkysymyksiä aiheeseen liittyen, koska hän ei tiedä milloin häntä tullaan palvelemaan. Lisäksi asiakaspalvelijalla menee aikaa selvitellä asiakasviestiin liittyviä asioita, jotka hyvin suurella todennäköisyydellä on jo tiedossa muualla organisaatiossa. Tämä tieto ei vain ole helposti saatavilla ja se joudutaan kaivamaan joka kerta uudestaan. Ongelmaan ratkaisu on reaaliaikainen analytiikka, joka tarjoaa viestin luokan perusteella tietoa sen tyypillisistä ominaisuuksista.

Toinen ongelma on, että nykyiset asiakasviestien luokitukset eivät ole tarpeeksi tarkkoja. Jotta voisimme hyödyntää edellä mainituin keinoin reaaliaikaista analytiikkajärjestelmää, kohdeyritys tarvitsee ensin tarkemmat luokat asiakasviesteille. Vasta tarkempien luokitusten jälkeen kohdeyritys voi alkaa rakentaa automaattista luokittelujärjestelmää. Uuden tarkemman luokkajärjestelmän rakentamisessa voidaan hyödyntää vanhan viestidatan klusterointia tai aloittaa puhtaalta pöydältä (Duda et al. 2001, ss. 517-518). Jos aloitetaan

puhtaalta pöydältä, voidaan hyödyntää asiakkaiden ja asiakaspalvelijoiden asiantuntemusta osana luokkajärjestelmän rakennusta. Toisin sanoen, annetaan asiakkaalle mahdollisuus tarkentaa hyvin yksityiskohtaisesti ennakkoon määritellyistä luokista, mitä viesti koskee. Tällä hetkellä asiakkaiden tulee määritellä liian korkealla tasolla viestin aihe. Puuttuvat ja virheelliset luokitukset korjattaisiin asiakaspalvelijan toimesta.

Webb et al. (2011, ss. 4-5) mallissa toinen vaihe on datan suunnittelu ja kerääminen. Ensinnäkin aiemmin mainituille uusille luokille tulee laskea ja selvittää tiettyjä ominaisuuksia. Näitä ominaisuuksia on esitelty edellisessä alaluvussa. Kun uusi luokittelumalli on saatu implementoitua, ruvetaan keräämään ohjatun koneoppimisen ja hahmontunnistuksen järjestelmälle kehitysdataa. Tämän työn tutkija suosittelee, että korpusaineistoja kerätään yhtä monta kuin uudessa luokkajärjestelmässä on ylätason luokkia. Tällä pystytään luomaan hyvin rajattu sanasto.

Mallin seuraavat vaiheet järjestyksessä ovat: alustava datan arviointi, piirteiden valinta ja vähentäminen, luokitteleminen sekä evaluointi (Webb et al. 2011, ss. 4-5). Näihin vaiheisiin tämän työn tutkija suosittelee, että kohdeyritys hankkii ulkopuolista osaamista sellaiselta taholta, joka on suorittanut aiemmin vastaavanlaisia tehtäviä suomen kielisellä aineistolla. Kirjallisuudesta ei selviä yhdenmielisesti suomen kielelle sopivia parhaita käytäntöjä.

7.4 Pohdinnan yhteenveto

Tämän luvun pohdintoihin on lähdetty teoreettinen näkökulma edellä. Käytännössä tekoälyn implementointi osaksi kohdeyrityksen järjestelmiä voi olla hankalaa. Hankaluuksia tuottaa erityisesti viestien tietovarasto, jota ei välttämättä ole suunniteltu näitä asioita silmällä pitäen. Toinen työn teoriaosuudessa pohdittu hankaluus liittyy suomen kieleen. Suomen kieli on haastava koneoppimisen ja hahmontunnistuksen järjestelmille (Korenius et al. 2004). Tutkimusten perusteella tällaisen järjestelmän rakentaminen on kuitenkin mahdollista. Tämä tutkimus tuo näkökulman siihen, miten tulevaisuudessa voidaan entistä taitavammin hallita omaa dataa. Järjestelmien kehittäminen on ajankohtaista ja pakollista, koska EU:n tuomat direktiivit tulevat vaatimaan ketterää datanhallintaa sakkorangaistuksen uhalla (Nisén 2016). Perustuen tulevaisuusverstaassa saatuihin tietoihin, kohdeyritys ei joko täytä vielä ketteriä vaatimuksia tai sitten organisaatiossa ei jaeta asiakasviesteihin liittyvää dataa tehokkaasti. Tämä perustuu tutkijan vaikutelmaan, jonka mukaan osa ratkaisuista tulisi olla helposti saatavilla perinteisimmillä tiedonhaun tehtävillä.

Salon (2013, s. 138) mukaan kilpailuetua tavoitellaan datan avulla tällä hetkellä kahdella eri pelikentällä. Ensimmäisellä pelikentällä kilpaillaan kuka saa kerättyä dataa mahdollisimman paljon. Kohdeyritys vaikuttaa olevan onnistunut keräämään dataa toiminnastaan hyvin laaja-alaisesti. Sen sijaan toisella pelikentällä kilpaillaan, kuinka jo kerätystä datasta saadaan kaikki mahdollinen irti. Vaikuttaa siltä, että kohdeyrityksen tulisi kiinnittää

huomiota enemmän tähän osioon. Hyvä kehityssuunnitelma on käyttää tämän työn oppeja osana tulevaisuuden päätöksentekoa.

Tutkimuksessa kerättiin dataa tulevaisuusverstaan avulla. Pavelinin (2014) suosittelee tulevaisuusverstaaseen käytettäväksi kokonaista päivää, mutta tässä työssä käytettiin 1,5 tuntia. Vaikka aika käytettiin tehokkaasti, oli se liian vähäinen näin laajaan aiheeseen, josta tutkittavien etukäteinen tietämys oli niukka. Pidemmällä ajalla oltaisiin päästy paremmin tutkimusaiheeseen kiinni, joka edelleen olisi johtanut täsmällisempiin vastauksiin. Nyt vastauksien sovittaminen kontekstiin jäi pitkälti tutkijalle. Tämä ei ole ihanteellista tutkimuksessa, joka pyrkii olemaan mahdollisimman objektiivinen. Tulevaisuusverstas oli siitä onnistunut, että se sai osallistujien suhteen hyvin poikkileikkaavan otoksen asiakaspalveluun liittyvistä tahoista. Tällä otoksella ei kuitenkaan päästy tutkimaan asiakaspalvelun ulkopuolisia näkökulmia, kuinka hyötyjä voidaan saavuttaa muualla finanssialan organisaatiossa. Onnistunutta oli myös se, että kaikilla tutkittavilla tuntui olevan samanlaiset ajatukset tulevaisuusverstaassa esitetyistä kysymyksistä. Tämä kertoo joko siitä, että tunnistettiin todelliset ja usein esiintyvät ongelmat, tai siitä, että asiat on tunnistettu jo ennen tätä tutkimusta ja niistä on keskusteltu. Yleisesti käytettiin samanlaisia termejä ongelmien ilmaisemiseen, mikä viittaa jälkimmäiseen hypoteesiin. Verstaassa jokainen kirjoitti vastauksia post-it lapuille ennen kuin toi mielipiteensä kysymyksiin ilmi. Tämä poistaa mahdollisuuden, että joku osallistujista olisi vaikuttanut yhteisön mielipiteeseen liikaa.

8. YHTEENVETO

Tämä tutkimus pyrki vastaamaan päätutkimuskysymykseen ”*Mitä tehokkuushyötyjä voidaan saavuttaa asiakasviestien automaattisella luokittelemisella finanssialan yrityksen asiakaspalvelussa?*”. Tutkimuskysymykseen vastattaessa selvisi, että monet kohdeyrityksen tunnistetut hyödyt ovat saavutettavissa myös muilla keinoin, kuin hyödyntämällä automaattista luokittelemista. Tutkijalle jäi sellainen vaikutelma, että tutkittavia kiinnostaneet tietotarpeet ovat lopulta tyydytettävissä hyvin yksinkertaisilla laskelmilla. Ongelma on, että asiakasviestejä on niin paljon, että näiden yksinkertaisten laskelmien tekemiseen tarvitaan niin paljon laskentatehoa, että perinteiset ohjelmistot, saati sitten ihmisen käsittelykyky, eivät näitä pysty tekemään. Tarvitaan ohjelmistoja, jotka on suunniteltu erityisesti ottamaan huomioon massadatan ominaisuudet. Kirjallisuuskatselmuksen perusteella näihin soveltuvia ratkaisuja löytyy jo osana arkipäiväistä ympäristöämme. Tämä osoittaa, että ratkaisuun voidaan hyödyntää valmista ratkaisua.

Tässä tutkimuksessa löydetyt tietotarpeet vaikuttavat olevan lopulta hyvin yksinkertaisia ja monia eri toimijoita kiinnostavia asioita. Voiko olla, että joku organisaatiossa tietää näihin vastaukset? On mahdollista, että kyseessä on liiketoimintatiedon hallinnan ongelma, jossa oikeata tietoa ei saada oikealla henkilölle oikeaan aikaan. Tällaisia ongelmia voi syntyä organisaatioissa siilomaisen toiminnan myötä. Lähtökohtaisesti jokaisessa organisaatiossa kaikkien työntekijöiden tulisi miettiä, voiko joku kollega tarvita samaa tietoa. Erityisesti tahojen, jotka toimivat keskitettyjen tietovarastojen haltijoina, tulisi pohtia tätä asiaa. Tällainen toiminta edellyttää organisaatioissa tiedolla johtamisen kulttuuria, joka tutkimuksien mukaan lähtee erityisesti sitoutuneesta yritysjohdosta, joka huomioi asian jo yritysstrategiassa.

Tiedon keräyksessä selvisi, että asiakasviestidatasta voidaan jalostaa välittömiä ja välillisiä hyötyjä. Välittömillä hyödyillä tarkoitetaan asioita, mihin vastaus selviää tutkimalla asiakasviestien massadataa. Esimerkiksi asiakaspalvelussa käytettyjä mallivastauksia voidaan laittaa kuntoon tutkimalla asiakasviesteistä, mitkä asiat tietyissä kysymyksissä jäävät asiakkaille epäselviksi. Välillisillä ratkaisuilla tarkoitetaan asioita, mihin asiakasviestien massadatasta voidaan saada viitteitä. Esimerkiksi uuden verkkopalvelun ja siihen liittyvien itsepalvelukanavien kehityksessä on tärkeätä saada tietoa, mitä asiaa asiakkaat eivät ole löytäneet nykyisestä verkkopalvelusta. Tämä ilmenee tutkimalla asiakasviestejä, joita asiakkaat kysyvät, koska eivät ole löytäneet helpompaa tapaa asian hoitamiseen. Tutkimuksessa myös selvisi, että asiakasviestidatan suurin liiketoimintahyöty on sidottuna ratkaisukykyyn ja sisältöön liittyvissä asioissa. Nämä tiedot ovat toistaiseksi piilossa, mutta saavutettavissa analysoimalla dataa tarkemmin.

Tutkimuksen yhteydessä tutkija havainnoi asiakaspalvelijoiden työtä. Yksinkertaistaen työ voidaan mallintaa seuraavaan prosessikulkuun. Suluissa on ilmaistu sama asia tämän

työn näkökulmasta. Mitä kokemattomampi asiakaspalvelija oli, sitä useammin hän käytti mallivastauksia osana toimintaansa.

1. Ymmärrä asiakkaan ongelma (Pohdi aiemman kokemuksesi perusteella onko viestissä sellaisia piirteitä, joilla voit luokitella sen tiettyyn ongelmaluokkaan.)

2. Etsi ongelmaan sopiva mallivastaus (Lähde ratkaisemaan kokemuksesi perusteella luokan tehtävää samalla tavalla kuin aikaisemmin.)

3. Räätälöi vastaus asiakkaalle sopivaksi (Käytä asiakkaaseen liittyviä muuttujia hyväksi.)

Loppujen lopuksi kyse on siis saapuvien asiakasviestien kategorisoimisesta tiettyyn ongelmaluokkaan ja luokalle sopivan mallivastauksen käyttämisestä. Mitä tarkemmin asiakasviestit ovat luokiteltu, sitä sopivampi mallivastaus voidaan tarjota. Herää siis kysymys, että tarvitseeko tulevaisuudessa ihmistä tämän prosessin tekemiseen vai voiko tekoälyä omaava virtuaalinen työntekijä hoitaa prosessin meidän puolestamme? Tätä asiaa ei tarvitse lähteä implementoimaan suoraan ottamalla virtuaaliagentteja käyttöön, vaan asiaa voidaan lähestyä asteittain vähentäen epäonnistuneiden vastauksien riskiä. Aluksi voidaan tarjota asiakaspalvelijalle vastausehdotuksia tekoälyä omaavasta järjestelmästä. Tällöin asiakaspalvelija validoi ne ennen lähetystä asiakkaalle, ja samalla opettaa vahvistetun koneoppimisen keinoin koneelle oliko vastaus hyväksyttävä.

Tutkimuksessa huomattiin, että asiakasviestien automaattinen luokitteleminen lähtee aina tarkoituksenmukaisten luokkien rakentamisesta. Vasta tämän operoinnin jälkeen voidaan hyödyntää automatiikkaa asiakasviestien luokitteluun. Uusinta ja hienointa tekoälyä omaava järjestelmä ei itsenäisesti voi rakentaa tarkoituksenmukaisia luokkia liiketoiminnalle. Tähän työhön lähdettiin harhaanjohtavalla oletuksella, että tekoäly ja nykyiset luokitukset ovat riittäviä. Tämän tyyppisissä projekteissa tullaan onnistumaan, kun muistetaan, että liiketoiminta ohjaa teknologiaa ja teknologia mahdollistaa liiketoimintaa. Kun lopulta luokat on saatu kuntoon, asiakasviestien automaattinen luokittelu tuo suurimmat hyödyt sen reaaliaikaisen data-analyysin myötä. Reaaliaikaisella data-analyysillä voidaan oivaltaa, mitä viestin toiselta osapuolelta voidaan odottaa.

Kun data on saatu otettua haltuun ja automaattinen luokittelujärjestelmä implementoitua osaksi asiakaspalveluväylää, ei olla kaukana koko paradigman muutoksesta, missä perinteinen asiakaspalveluväylä muuttuu itsepalveluväyläksi. Itsepalveluväylän toisessa päässä toimii reaaliaikaisia vastauksia tarjoava virtuaaliagentti. Tämän työn aihe itseasiassa rakentui virtuaaliagentteihin perehtymällä. Perehtymisen tuloksena valikoitui luokittelemisen näkökulma, joka antaa helpon ymmärryksen monimutkaisiin järjestelmiin. Matkalla kohti virtuaalisia agentteja voidaan saavuttaa myös muita hyötyjä. Tämän työn kontribuutio tuo selkeyttä siihen, mitä muita hyötyjä voidaan saavuttaa. Vaikka tutkimuksessa käytettiin finanssialan yritystä, ovat löydetyt hyödyt enemmän asiakaspalvelulle ominaisia, eivätkä finanssialalle.

Tämä tutkimus valmistelee ja kannustaa organisaatiota hyödyntämään automaattisen luokittelemisen menetelmiä. Tämän työn jälkeen olisi luonnollista jatkaa tutkimuksia minkälaisiin luokkiin organisaatiot haluavat asettaa epästruktuuria dataa ja millä piirteillä viestit voidaan tehokkaimmin tunnistaa tietyn luokkanimityksen omaavaksi. Lisäksi kaivataan yleistä tutkimusta liittyen suomen kielen koneelliseen käsittelyyn. Erilaisia algoritmeja on kirjallisuudessa esitetty paljon, mutta suomen kieleen kohdistuneita tutkimuksia on valitettavan vähän.

LÄHTEET

Ackoff, R. 1989. From data to wisdom. Journal of Applied Systems Analysis, vol. 16, pp. 3–9

Aggarwal, C. C., & Zhai, C. 2012. Mining text data. Springer Science & Business Media.

Agirre, E., De Lacalle, O., Fellbaum, C., Marchetti, A., Toral, A., Vossen, P. 2009. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. sivut 123-128. Association for Computational Linguistics.

Anzai, Y. (2012). Pattern Recognition & Machine Learning. Elsevier.

Alkula, R. 2000. Merkkijonoista suomen kielen sanoiksi. Väitöskirja. Tampereen yliopisto.

Arppe, A. 2008. Ei yhtä ainoaa polkua - Suomalaisia kokemuksia matkalla kielliteknologisesta tutkimuksesta liiketoimintaan. Verkkoblogi. Viitattu 18.11.2016. Saatavilla: https://kitwiki.csc.fi/twiki/bin/view/FiLT/ArppeFi

Banerjee, A., Bandyopadhyay, T., Acharya, P. 2013. Data analytics: Hyped up aspirations or true potential. Vikalpa, 38(4), 1-11.

Barney, J. 1996. The resource-based theory of the firm. Organization science, 7(5), 469-469.

Barr, A., Cohen, P., Feigenbaum, E. 1989. The handbook of Artificial Intelligence, volume IV.

Barto, A. & Sutton, R. 1997. Introduction to Reinforcement Learning. MIT Press

Beaujean, M., Davidson, J., Madge, S. 2006. The “moment of truth” in customer service. Mckinsey Quarterly, 1, 62-73.

Beyer, M. A., & Laney, D. 2012. The importance of ‘big data’: a definition. Stamford, CT: Gartner, 2014-2018.

Bijmolt, T., Leeflang, P., Block, F., Eisenbeiss, M., Hardie, B., Lemmens, A., Saffert, P. 2010. Analytics for customer engagement. Journal of Service Research 13.3: 341-356.

Bird, S., Klein, E., Loper, E. 2009. Natural language processing with Python. Chapter 6. O'Reilly Media, Inc.

Blumberg, R., & Atre, S. 2003. The problem with unstructured data. DM REVIEW, 13(42-49), 62.

Brynjolfsson, E., Hitt, L. Kim, H. 2011. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?

Business dictionary. 2016. Verkkotietosanakirja. Viitattu 16.12.2016. Saatavilla: http://www.businessdictionary.com/

Butterfield, A. & Ngondi, G. 2016. A Dictionary of Computer Science (7th edition). Oxford University Press.

Cammack, R., Atwood, T., Campbell, P, Parish, H., Smith, A., Vella, F., Stirling, J. 2006. Oxford Dictionary of Biochemistry and Molecular Biology (2nd edition). Oxford University Press.

Carvalho, J. & Curto, S. 2014. Fuzzy preprocessing of medical text annotations of intensive care units patients. In Norbert Wiener in the 21st Century (21CW), 2014 IEEE Conference on (pp. 1-7). IEEE

Chmiel, A., Sienkiewicz, J., Thelwall, M., Paltoglou, G., Buckley, K., Kappas, A., Hołyst, J. 2011. Collective emotions online and their influence on community life. *PloS one*, *6*(7), e22207.

Chowdhury, G. 2003. Natural language processing. Annual review of information science and technology, 37(1), 51-89.

Davenport & Harris 2007. Analysoi ja voita: kilpailun uusi tiede. Talentum

Davenport, T., Harris, J. Morison, R. 2010. Analytics at work: Smarter decisions, better results. Harvard Business Press.

Davenport, T. 2013. Analytics 3.0. Harvard Business Review, 91(12), 64-72.

Davis, M., Heineke, J. 1998. How disconfirmation, perception and actual waiting times impact customer satisfaction. International Journal of Service Industry Management. Vol. 9 Iss: 1, pp.64 – 73

Dilrukshi, I., De Zoysa, K. Caldera, A. 2013. Twitter news classification using SVM. In Computer Science & Education (ICCSE), 2013 8th International Conference on (pp. 287-291). IEEE.

Duda, R., Hart, P., Stork, D. 2001. Pattern Classification, 2nd Edition. 680 sivua. ISBN: 978-0-471-05669-0

Douglas, D. 2011. Watson Predictive Analytics – Classification Example (Panel 5). Springer-Verlag Berlin Heidelberg.

Drury, B., Torgo, L., Almeida, J. 2011. Classifying news stories to estimate the direction of a stock market index. In 6th Iberian Conference on Information Systems and Technologies (CISTI 2011) (pp. 1-4). IEEE.

Dörre, J., Gerstl, P., Seiffert, R. 1999. Text mining: finding nuggets in mountains of textual data. In U. Fayyad, S. Chaudhuri, and D. Madigan (eds.): Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data 123 mining, pp. 398–401. ACM Press.

Earl, M. 1994. Knowledge as strategy: reflections on Skandia International and Shorko Films. Strategic Information Syststems: A European Perspective. Chichester, UK: John Wiley and Sons, ss. 53-69.

Eichfeld, A., Morse, T., Scott, K. 2006. Using call centers to boost revenue. McKinsey Quarterly, May. http://www.alessandrosanto.com/Mck2.pdf

Encyclopedia Britannica. Verkkotietosanakirja. Viitattu 10.10.2016. Saatavilla: https://global.britannica.com/

Erhardt, J. 2015. Machine Learning vs Predictive Analytics. Verkkoblogi. Viitattu 15.11.2016. Saatavilla: http://www.wise.io/blog/machine-learning-vs-predictive-analytics

Fertier, A., Montarnal, A., Barthe-Delanoë, A., Truptil, S., Bénaben, F. 2016. Adoption of big data in crisis management toward a better support in decision-making. ISCRAM 2016 Conference.

Fu, K. 1968. Sequential methods in pattern recognition and machine learning (Vol. 52). Academic press.

Gao, Z., Xu, Y., Meng, F., Qi, F., Lin, Z. 2014. Improved information gain-based feature selection for text categorization. In Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2014 4th International Conference on (pp. 1-5). IEEE.

Ghalehtaki, R., Khotanlou, H., Esmaeilpour, M. 2014. Evaluating preprocessing by turing machine in text categorization. In Intelligent Systems (ICIS), 2014 Iranian Conference on (pp. 1-6). IEEE.

Ginter, F., Nyblom, J., Laippala, V., Kohonen, S., Haverinen, K., Vihjanen, S., & Salakoski, T. 2013. Building a Large Automatically Parsed Corpus of Finnish. In Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013); May 22-24; 2013; Oslo University; Norway. NEALT Proceedings Series 16 (No. 085, pp. 291-300). Linköping University Electronic Press.

Grönroos, C. 1998. Nyt kilpaillaan palveluilla. 4. uppl. Helsinki, Porvoo, Juva: WSOY. 338 s. ISBN 951-0-22816-8.

Grönroos, C. 2001. Palveluiden johtaminen ja markkinointi (alkuteos: Service Management and Marketing). A customer relationship management approach 2000, käännös Maarit Tillman. WSOY: Helsinki.

Hammond, K. & Varde, A. 2013. Cloud Based Predictive Analytics: Text Classification, Recommender Systems and Decision Support. 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX. pp. 607-612.

Haykin, S. & Network, N. 2004. A comprehensive foundation. Neural Networks, 2(2004).

Hotho, A., Nürnberger, A., Paaß, G. 2005. A brief survey of text mining. In Ldv Forum (Vol. 20, No. 1, pp. 19-62).

Ikonomakis, M., Kotsiantis, S., Tampakas, V. 2005. Text classification using machine learning techniques. WSEAS transactions on computers, 4(8), 966-974.

Ingwersen, P. 1992. Information Retrieval Interaction:(click on title for download access via'Documents'). Taylor Graham.

Jain, A., Murty, M., Flynn, P. 1999. Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

Jiawei, H., & Kamber, M. 2001. Data mining: concepts and techniques. San Francisco, CA, itd: Morgan Kaufmann, 5.

Joachims, T., Freitag, D., Mitchell, T. 1996. Webwatcher: A tour guide for the world wide web. In IJCAI (1) (pp. 770-777).

Järvelin, K. 1995. Tekstitiedonhaku tietokannoista: Johdatus periaatteisiin ja menetelmiin. Espoo: Suomen ATK-kustannus. 273 s. ISBN 951-762-297-X

Kaisler, S., Armour, F., Espinosa, J., Money, W. 2013. Big data: issues and challenges moving forward. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 995-1004). IEEE.

Kent, E. 2014. Text analytics–techniques, language and opportunity. Business Information Review, 31(1), 50-53. Dipparefet

Kettunen, K., & Baskaya, F. 2011. Stemming Finnish for information retrieval–comparison of an old and a new rule-based stemmer. In Proceedings of the 5th Language & Technology Conference (LTC 2011), Poznan (pp. 476-480).

Khan, K. & Qamar, U. 2015. Improved Single-Label Text Categorization by Instance Filtration. In Complex, Intelligent, and Software Intensive Systems (CISIS), 2015 Ninth International Conference on (pp. 28-35). IEEE.

Khan, M., Manoj, J. Singh, A., Blumenstock, J. 2015. Behavioral Modeling for Churn Prediction: Early Indicatiors and Accurate Predictors of Custom Defection and Loyalty. Big Data - IEEE International Congress.

Khalid, S., Khalid, T., Nasreen, S. 2014. A Survey of Feature Selection and Feature Extraction Tehniques in Machine Learning. Science and Information Conference 2014.

Kiimamaa, J., Epäilys, N. Kuikka, N. Muilu, T., Mäntysalo, R., Onkalo, P., Reinikainen, K. 2003. Tulevaisuusverstas - ongelmia, ideoita ja toteutuksen suunnittelua yhdessä. Asu Kylässä. Kokemuksia asukaskeskeisestä kylien suunnittelusta PohjoisPohjanmaalla. Nordea Tiedonantoja 2/2003. 11-17

Klösgen, W., & Zytkow, J. 2002. Handbook of data mining and knowledge discovery. Oxford University Press, Inc.

Koehn, P., Och, F., Marcu, D. 2003. Statistical phrase-based translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48-54). Association for Computational Linguistics.

Korenius, T., Laurikkala, J., Järvelin, K., Juhola, M. 2004. Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 625-633). ACM.

Koskenniemi, K., Rehm, G., Uszkoreit, H. 2012. The Finnish language in the digital age. Springer.

Kotsiantis, S., Zaharakis, I., Pintelas, P. 2007. Supervised machine learning: A review of classification techniques.

Kroeze, J., Matthee, M., Bothma, T. 2003. Differentiating data-and text-mining terminology. South African Institute for Computer Scientists and Information Technologists. ss. 93-101

Kurbatow, A. 2015. The research of text preprocessing effect on text documents classification efficiency. In " Stability and Control Processes" in Memory of VI Zubov (SCP), 2015 International Conference (pp. 653-655). IEEE.

Laaksonen, P. 2016. Miksi Tekoäly voi tarjota parempaa palvelua kuin ihminen? Salesforce verkkoblogi. Viitattu 15.12.2016. Saatavilla: https://www.salesforce.com/fi/blog/2016/12/Miksi-tekoaly-voi-tarjota-parempaa-palvelua-kuin-ihminen.html

Lagus, K. 2000. Text Mining with the WEBSOM. Helsinki University of Technology.

Laihonen, H., Hannula, M., Helander, N., Ilvonen, I., Jussila, J., Kukko, M., Kärkkäinen, H., Lönnqvist, A., Myllärniemi, J., Pekkola, S., Virtanen, P., Vuori, V., Yliniemi, T. 2013. Tietojohtaminen. Tampereen teknillinen yliopisto – Tiedonhallinnan ja logistiikan laitos

Leskinen, S. 2016. Mies, jota pankit pelkäävät. Tivi-lehti, helmikuu 2016. Talentum media.

Li, G., Ooi, B., Feng, J., Wang, J., Zhou, L. 2008. EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 903-914). ACM.

Liu, B. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

Liu, X., Rujia, G., Liufu, S. 2012. Internet news headlines classification method based on the n-gram language model. In Computer Science and Information Processing (CSIP), 2012 International Conference on (pp. 826-828). IEEE.

Lui, M., & Baldwin, T. 2012. An off-the-shelf language identification tool. In Proceedings of the ACL 2012 system demonstrations (pp. 25-30). Association for Computational Linguistics.

Manning, G. & Reece, B. 2004. Selling today: creating customer value, New Jersey: Pearson Education.

Manning, C. & Schütze, H. 1999. Foundations of statistical natural language processing (Vol. 999). Cambridge: MIT press.

Markham, S., Kowolenko, M., Michaelis, T. 2015. Unstructured text analytics to support new product development decisions. Research-Technology Management, 58(2), 30-39.

McCarthy, D., Koeling, R., Weeds, J., Carroll, J. 2004. Finding predominant word senses in untagged text. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (p. 279). Association for Computational Linguistics.

Merriam-Webster. 2016. Dictionary and Thesaurus. Viitattu 26.10. Saatavilla: https://www.merriam-webster.com/

Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., Nisbet, R. 2012. Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

Nadeem, D. 2012. Social customer relationship management (SCRM): how connecting social analytics to business analytics enhances customer care and loyalty?. International journal of business and social science 3.21.

Nisén, A. 2016. Lainsäädännön kehitys maksamisen alueella. Finanssivalvonta. Suomen Pankki.

Nürnberger, A., Seising, R., Wenzel, C. 2009. On the Fuzzy Interrelationships of Data, Information, Knowledge and Wisdom.

Norton, M. 1999. Knowledge discovery in databases. Library Trends, 48(1), 9-21.

OpusCapita. 2016. Finance 4.0. Webinaari. Järjestetty 20.10.2016

Paukkeri, M. 2012. Language-and domain-independent text mining. Aalto väitöskirja.

Pavelin, K., Pundir, S., Cham, J. 2014. Ten simple rules for running interactive workshops. *PLoS Comput Biol*, *10*(2), e1003485.

Pilaszy, I. 2005. Text categorization and support vector machines. In The proceedings of the 6th international symposium of Hungarian researchers on computational intelligence.

Porta, M. 2014. A dictionary of epidemiology. Oxford University Press

Rana, M., Khalid, S., Akbar, M. 2014. News classification based on their headlines: A review. In Multi-Topic Conference (INMIC), 2014 IEEE 17th International (pp. 211-216). IEEE.

Ray, B., Aphinyanaphongs, Y., Heffron, S., 2015. Text Classification-Based Automatic Recruitment of Patients for Clinical Trials: A Silver Standards-Based Case Study. International Conference on Healthcare Informatics, Dallas, TX, 2015, pp. 28-33.

Rieuf, E. 2016. The current state of machine intelligence 3.0. Blogikirjoitus. Data Science Central – The online resource for big data practioners.

Russom, P. 2011. Big data analytics. TDWI Best Practices Report, Fourth Quarter, 1-35.

Saaranen-Kauppinen, A. & Puusniekka, A. 2009. KvaliMOTV – Menetelmäopetuksen-tietovaranto. Tampere: Yhteiskuntatieteellinen verkkotietoarkisto. Viitattu 8.9.2016. Saatavissa: http://www.fsd.uta.fi/fi/julkaisut/motv_pdf/KvaliMOTV.pdf

Saarelainen, A. 2016. Fintech murentaa pankkeja. Tivi-lehti, syyskuu 2016. Talentum media.

Salo, I. 2013. Big data - tiedon vallankumous. Jyväskylä: Docendo Oy.

Saunders, M., Lewis, P. Thornhill, A. 2009. Research methods for business students, 5th edition ed., Pearson Education Ltd, Essex, UK, 614p.

Schraudolph, N., Dayan, P., Sejnowski, T. 1994. Temporal difference learning of position evaluation in the game of Go. Advances in Neural Information Processing Systems, 817-817.

Scott, J. & Marshall, G. 2009. A Dictionary of Sociology (3 rev. ed.). Oxford University Press.

Sebastiani, F. 2002. Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Shah, F., Patel, V. 2016. A Review on Feature Selection and Feature Extraction for Text Classification. IEEE WiSPNET 2016 conference.

Solorio, T., Pérez-Coutino, M., Montes-y-Gémez, M., Villasenor-Pineda, L., López-López, A. 2004. A language independent method for question classification. In Proceedings of the 20th international conference on Computational Linguistics (p. 1374). Association for Computational Linguistics.

Tesauro, G. 1995. Temporal difference learning and TD-Gammon. Communications of the ACM, 38(3), 58-68.

Thorsten J. 1997. Text categorization with support vector machines: learning with many relevant features, Proc. of ECML-98, 10th European Conference on Machine Learning, Springer Verlag, Heidelberg, DE, 1998, pp. 137-142

Tuomi, I. 1999. Data is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. Journal of Management Information Systems, Vol. 16 No. 3. ss 103-117. Published by: M.E. Sharpe

Turku BioNLP Group. 2016. Keskitetty tietokanta. Viitattu 15.12.2016. Saatavilla: http://bionlp.utu.fi/publications.html

Töttö, P. 2004. Syvällistä ja pinnallista. Teoria, empiria ja kausaalisuus sosiaalitutkimuksessa. Tampere: Vastapaino.

Upton, G. & Cook, I. 2014. A dictionary of Statistics. OUP Oxford. 488 s.

Venkatesan, N., Kim, E., Shin, D. 2016. PoN: Open source solution for real-time data analysis. In Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016 Third International Conference on (pp. 313-318). IEEE.

Verbix. 2000. The 10000 most frequently used words in Finnish. Verkkowiki. Viitattu 2.12.2016. Saatavilla: http://wiki.verbix.com/Documents/WordfrequencyFi

Virtanen, J. 2016. Pelkäätkö tekoälyjen vievän työsi? Tällä alalla olet jo nyt uhattuna. Tivi-verkkolehti. Viitattu 15.12.2016. Saatavilla: http://www.tivi.fi/Kaikki_uutiset/pelkaatko-tekoalyjen-vievan-tyosi-talla-alalla-olet-jo-nyt-uhattuna-6606431

Wangenheim, F. & Bayón , T. 2007. The Chain from Customer Satisfaction via Word-of-mouth Referrals to New Customer Acquisition. Journal of the Academy of Marketing Science, 35 (2), 233-249.

Watson, H., Wixom, B., Goodhue, D. 2004. Data warehousing: The 3M experience. *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance, 202*.

Watson, H., & Wixom, B. H. 2007. The current state of business intelligence. Computer, 40(9), 96-99.

Webb, A., Copsey, K., Cawley, G. 2011. Wiley Finance: Statistical Pattern Recognition (3). Hoboken, GB: Wiley.

W3Techs. 2016. Usage of content languages for websites. Web Technology Surveys. Viitattu 19.10.2016. Saatavilla: https://w3techs.com/technologies/overview/content_language/all

Xia, X., Lo, D., Correa, D., Sureka, A., Shihab, E. 2016. It Takes Two to Tango: Deleted Stack Overflow Question Prediction with Text and Meta Features.

Xiong, W. 2014. A Better Indicator for Genre Classification: Topic Word or Surface Text Feature. 2014 International Conference on Information Science, Electronics and Electrical Engineering.

Yang, Y., & Pedersen, J. 1997. A comparative study on feature selection in text categorization. In ICML (Vol. 97, pp. 412-420).

Zhu, W., Barron, S., Gallotti, M., Gupta, V., Wang, X., Magdalen, J., Singer, J. 2009. IBM Classification Module: Make It Work for you. IBM Redbooks.

Zhu, X. 2005. Semi-supervised learning literature survey.

LIITTE 1A: VASTAUSTEN ANALYSOINTI

Tässä on esiteltynä työpajan vastausten analyysi. Analyysissä pohdittiin, miten viestiä tulee luokitella, jotta tietotarpeet saadaan tyydytettyä. Lisäksi on pohdittu, mikä luokan ominaisuus tulee selvittää ratkaisuehdotusta varten. Kaikkiin asioihin ei pystytä syventymään automaattisella luokittelemisella. Nämä kohdat ovat merkattu punaisella ja niitä on analysoitu, millä muulla keinolla asiaan voitaisiin vaikuttaa. Suuren taulukkorakenteen takia liite on tuotu kuvana työhön ja jaettu A ja B osioihin, jotta ne saadaan visuaalisesti sopimaan osaksi tätä kokonaisuutta.

Taulukko	Ratkaisu	Automaattisen luokittelun kohde	Luokan nimi	Luokan ominaisuus	Muu tapa
A1	Sisältö. Missä kaikkialla viesti on kiertänyt ja kuka sen lopulta on hoitanut --> Kenelle viesti olisi heti kuulunut ohjautua	Sisältö	Sisältöotsikko	Ratkaisukyky	
A1	Ratkaisuaste --> Kuinka monta viestiä on voitu hoitaa ja asia ratkaista suoraan lisäarvoa tuottaen	Sisältö	Sisältöotsikko	Ratkaisukyky	
A1	Kuinka paljon pystymme hoitamaan kuntoon ensimmäisen kontaktikäsittelyn aikana?	Sisältö	Sisältöotsikko	Ratkaisukyky	
A1	Ratkaisukyky. Sisältääkö viesti riittävät tiedot asian hoitamiseen? Olisiko asiakas voinut hoitaa asian itse verkkopalvelussa.	Sisältö	Sisältöotsikko	Ratkaisukyky, Sisältöotsikko analyysi	
A1	Ratkaisukyky. Yhtedenotonaihe	Sisältö	Sisältöotsikko	Ratkaisukyky	
A1	Kuinka monessa viestissä on ohitettu myynninpaikat	Sisältö	Sisältöotsikko	Sisältöotsikko analyysi	
A1	Pääaihealueet. Millä asialla asiakkaat meitä lähestyvät?	Sisältö	Sisältöotsikko	Kysytyimmät sisältöotsikot	
A1	Mikä oli primäärikontaktin syy?	Sisältö	Sisältöotsikko	Sisältöotsikko analyysi	
A1	Miksi asiakas ei pystynyt hoitamaan asiaa verkossa itse? Mitkä asiat nousee useimmin esiin?	Sisältö	Sisältöotsikko	Sisältöotsikko analyysi, Kysytyimmät sisältöotsikot	
A2	Mitä asiaa kysytään useimmiten?	Sisältö	Sisältöotsikko	Kysytyimmät sisältöotsikot	
A2	Mitkä asiat asiakas kokee tärkeimmiksi tuottamassamme palvelussa?				Asiakasarvon tutkiminen
A2	Miten asiakas haluaa hänelle vastattavan ja mikä siinä on tärkeintä?				Asiakasarvon tutkiminen
A2	Mikä on asiakkaalle tärkeää?				Asiakasarvon tutkiminen
A2	Mitä osaamista ja ohjeistusta tarvitsen usein / harvoin?	Sisältö	Sisältöotsikko	Ratkaisukyky (oma)	
A2	Auttoiko annettu ratkaisu asiakasta kerralla?	Sisältö	Sisältöotsikko	Ratkaisukyky	
A2	Kuinka paljon vuoden aikana on mennyt ohi lisäarvon tuottamisen paikkoja? (myyntinäkökulma)	Sisältö	Sisältöotsikko	Sisältöotsikko analyysi	
A2	Ratkaisukyky! Kuinka monessa tilanteessa asiakasneuvoja itse on / ei ole voinut auttaa asiakasta ja tuottaa lisäarvoa pankille? Miksi ei?	Sisältö	Sisältöotsikko	Ratkaisukyky	
A2	Ratkaisukyky, miten asiakkaan ongelma on saatu kerralla kuntoon. Onko asiakkaan vielä pitänyt palata asian tiimoilta uudestaan	Sisältö	Sisältöotsikko	Ratkaisukyky	
B1a	Tilannetiedon välittäminen asiakkaalle	Sisältö	Sisältöotsikko	Ratkaisukyky	
B1a	Asiakkaalle tieto prosessin alussa sen odotetuista aikatauluista	Sisältö	Sisältöotsikko	Ratkaisukyky	
B1a	Pitäydytään luvatussa aikataulussa. Ilmoitetaan asiakkaalle jos luvatussa aikataulussa ei pysytä	Sisältö	Sisältöotsikko	Ratkaisukyky	

LIITE 1B: VASTAUSTEN ANALYSOINTI

B1b	Uusi verkkopalvelu	Sisältö	Sisältöotsikko	Kysytyimmät sisältöotsikot	
B1b	Automatisaation laajempi hyödnytäminen esim ohjelmistorobotiikalla				Automatisointiasteen nostaminen
B1b	Verkkopalvelun hakukoneen parantaminen	Sisältö	Sisältöotsikko	Kysytyimmät sisältöotsikot	
B1c	Mallivastaukset kuntoon laittaminen	Sisältö	Sisältöotsikko	Sisältöotsikko analyysi, Mallivastaus, ratkaisukyky	
B1c	Virheiden tarkempi tilastointi, ja tähän perustuva systemaattinen korjaus sekä osaamisen kehittäminen	Sisältö	Sisältöotsikko	Ratkaisukyky	
B2a	Järjestelmien yhtenäistäminen				Muutosjohtaminen, lean-periaatteet
B2a	Ohjelmistorobotiikka				Automatisointiasteen nostaminen
B2a	Tuplanäytöt				Laitehankinnat
B2b	Toimintatapojen yhteinäistäminen organisaatiossa				Muutosjohtaminen, lean-periaatteet
B2c	Vähennetään tietovaatimuksia, koska tietoa ei hyödynnetä				Muutosjohtaminen, lean-periaatteet
B2c	Laskutukseen liittyvät tiedot automaattisesti asiakastiedoista yhdistämällä				Automatisointiasteen nostaminen
B2c	Toimintatapojen leanaus				Muutosjohtaminen, lean-periaatteet