Kalle Malkki

SOSIAALISESTA MEDIASTA KERÄTTY MASSADATA JA SEN ANALYTIIKKA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2018

TIIVISTELMÄ

Malkki, Kalle
Sosiaalisesta mediasta kerätty massadata ja sen analytiikka
Jyväskylä: Jyväskylän yliopisto, 2018, 94 s.
Tietojärjestelmätiede, kandidaatin tutkielma
Ohjaaja(t): Clements, Kati

Sosiaalinen media ja massadata ovat nousseet 2000-luvulla merkittäviksi tekijöiksi yhteiskunnissa ja liiketoiminnassa. Sosiaalinen media koostaa sisälleen suuren määrän käyttäjiä, jotka tuottavat myös huomattavia määriä dataa ja tämä tekee sosiaalisesta mediasta yhden suurimmista massadatan lähteistä. Tässä tutkielmassa pyritään selvittämään, minkälaista on sosiaalisen median massadata ja massadata-analytiikka. Tutkielma toteutettiin kirjallisuuskatsauksena ja kirjallisuutta valittaessa pyrittiin valitsemaan julkaisuja, jotka on julkaistu 2010 jälkeen.

Sosiaalisen median massadatan huomattiin soveltuvan hyvin massadatan määritelmiin ja sillä todettiin olevan tiettyjä sille ominaisia piirteitä struktuurin ja datan luotettavuuden osalta. Analytiikan osalta todettiin erityisesti event processing -mallilla olevan merkittävä rooli sosiaalisen median massadata-analytiikan osalta.

Asiasanat: massadata, massadata-analytiikka, sosiaalinen media,

ABSTRACT

Malkki, Kalle
Social media big data and its analytics
Jyväskylä: University of Jyväskylä, 2018, 94 p.
Information Systems, Bachelor's Thesis
Supervisor(s): Clements, Kati

Social median and big data have had a significant role in shaping the business and societies of the 21st century. Social media includes many users who produce a lot of data and this fact make social media one of the biggest sources of big data. The goal of this thesis is to find the nature of social media big data. The thesis was done as a literature review and literature from 2010 and newer were prioritized.

Social media big data was found to conform to the definition of big data and it was also noted to have some key characteristics unique to it, which concern the data structure and reliability of the data. On the social media big data analytics, it was found that event processing model is one of the key analytics models for social media big data.

Keywords: Big data, big data analytics, social media

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Tuotetun datan määrä on kasvanut huomattavasti 2000-luvulla (Gandomi, & Haider, 2015.), internetin ja erityisesti tämän mahdollistaman sosiaalisen median myötä. Tämä datan määrän kasvu on johtanut uusiin liiketoiminnallisiin mahdollisuuksiin, mutta on myös nostanut esille datan säilöntään sekä prosessointiin liittyviä ongelmia ja tämä datan määrän kasvu on myös johtanut massadatan, englanniksi ”big data”, syntyyn käsitteenä.

Massadatan konseptin katsotaan saaneen alkunsa jo 1990-luvulla, mutta alkoi vakiintua ja ilmetä tieteellisissä artikkeleissa enemmän vasta 2010-luvulla. (Gandomi, & Haider, 2015.) Massadata on ollut yksi suurimmista puheen aiheista 2010- luvulla erityisesti informaatioteknologian saralla ja se on muovannut merkittävästi yhteiskuntia sekä liiketoimintaa (Boyd & Crawford 2012).

Massadatan noustessa esille ilmiönä on myös noussut esille tarve analysoida kasvavia data määriä ja tämä on johtanut massadata-analytiikan syntyyn. Massadata-analytiikka linkittyy vahvasti business intelligence -käsitteeseen ja se koostuu analyyttisistä menetelmistä jotka käyvät läpi suuria datajoukkoja (Fan, Lau, & Zhao, 2015) (Russom, 2011).

Myös sosiaalinen median on kasvattanut suosiotaan merkittävästi 2000-luvulla ja sen kasvun on mahdollistanut web- ja mobiiliteknologioiden kehittyminen tänä aikana, jotka ovat yhdessä luoneet ympäristön jossa käyttäjät voiva luoda ja jakaa sisältö, sekä kommunikoida keskenään (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011) (Xiang, & Gretzel, 2010). Sosiaalinen median käsitti sisälleen 2017 yli kaksi miljardia käyttäjää.

Sosiaalisen median on suuri osa webympäristöä ja täten se on myös yksi suurimmista massadatan lähteistä ja muovaa myös täten merkittävästi massadatan rakennetta (Bello-Orgaz, Jung, & Camacho, 2016). Täten tässä tutkielmassa pyritään tarkemmin selvittämään, minkälaista on juuri sosiaalisen median massadata ja minkälaiset analyyttiset mallit parhaiten soveltuvat sosiaalisen median massadata-analytiikkaan. Tämän pohjalta tutkielman tutkimuskysymykseksi muodostui: ”Mitkä analyyttiset mallit soveltuvat parhaiten sosiaalisen median massadatan analytiikkaan?”.

Tutkielma toteutettiin kirjallisuus katsauksena, jonka aineisto pyrittiin rajoittamaan informaatioteknologian alan tieteellisiin julkaisuihin. Lisäksi Julkaisut pyrittiin rajoittamaan 2010-luvulle ja ylöspäin erityisesti massadatan osalta, jota on suurimmalta osin käsitelty tieteellisissä julkaisuissa vasta 2010-luvulla. Rajoituksesta jouduttiin kuitenkin ajoittain poikkeamaan, rajoitusta vastaavien relevanttien lähteiden puutteen takia. Lähteitä haettiin seuraavista tietokannoista, Google Scholar, Scopus, ProQuest, IEEE explorer ja Academic Search Elite.

Rakenteeltaan tutkielma etenee seuraavasti, luvussa kaksi käydään läpi sosiaalisen median käsite, tarjotaan teoreettinen viitekehys sosiaalisen median seitsemän kulmakiven muodossa ja tutkitaan tarkemmin sosiaalisen median sisältöä. Luvussa kolme käydään läpi massadata ja massadata-analytiikka käsitteinä. Neljännessä luvussa esitellään tutkielman tulokset, jotka koostuvat lukujen kaksi ja kolme tietojen yhdistelystä. Viidennessä luvussa on tutkielman yhteenveto.

2 SOSIAALINEN MEDIA

Sosiaalinen media on suhteellisen löyhä käsite, mutta sen usein nähdään olevan yhdistelmä mobiili- ja webpohjaisia teknologioita, jotka kokonaisuudessaan luovat vuorovaikutteisen alustan, jolla eri käyttäjät voivat kommunikoida, luoda sisältöä tai jakaa sisältöä keskenään (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011) (Xiang, & Gretzel, 2010). 2017 arvioitiin sosiaalisen median käyttäjiä kokonaisuudessaan olevan n. 2.46 miljardia ja vuoteen 2021 mennessä määrän arvioidaan kasvavaan yli kolmeen miljardiin (Statista, 2018). Sosiaalisen median kehittyminen verkkoympäristössä voidaan katsoa olevan luontainen kehityksen suunta verkkoympäristössä, kun otetaan huomioon internetin alkuaika, jolloin sen tarkoituksena oli vapaasti jakaa informaatiota käyttäjien kesken ja juuri tämä toiminto on yksi sosiaalisen median keskeisimpiä tekijöitä (Kaplan, & Haenlein, 2010).

Sosiaalinen media on myös suonut yrityksille hyvän alustan markkinoinnille. Sosiaalisten median kanavien avulla voidaan tavoittaa suhteellisen helposti useita potentiaalisia asiakkaita, joten suurin osa nykypäivän yrityksistä toteuttaa markkinointia vaihtelevalla laajuudella sosiaalisessa mediassa (Saravanakumar, & SuganthaLakshmi, 2012). Sosiaalisessa mediassa markkinointi voi tapahtua monella eri tavalla, esimerkiksi word-of-mouth markkinointi on hyvin yleistä sosiaalisen median eri foorumeilla ja koostuu usein kuluttajien tuotearvioista (Mangold, & Faulds, 2009) (Li, & Du, 2011). Lisäksi sosiaalisessa mediassa voidaan pyrkiä ylläpitämään kuluttajien bränditietoisuutta ja brändiin sitoutumista, markkinoimalla ja tuottamalla omaa sisältöä sosiaalisessa mediassa. (Hoffman, & Fodor, 2010).

2.1 Sosiaalisen median seitsemän kulmakiveä

Kietzmann, ym. (2011) artikkelissa ”Social media? Get serious! Understanding the functional building blocks of social media”, määritellään sosiaalisen median seitsemän kulmakiven avulla, jotka ovat identiteetti, keskustelu, jakaminen, läsnäolo, suhde, maine ja ryhmät. Kirjoittajat toteavat, että kulmakivet eivät ole toisiaan poissulkevia ja että kaikkien palasia ei tarvitse olla läsnä, jotta sivusto saavuttaisi sosiaalisen median määritelmän (Kietzmann, Hermkens, McCarthy, & Silvestre, 2011.). Viitekehystä on myöhemmin käytetty sosiaalista mediaa koskevien tutkimusten pohjana (Kietzmann, Silvestre, McCarthy, & Pitt, 2012). Seuraavaksi käydään lyhyesti läpi, kyseessä olevan teoreettisen viitekehyksen kulmakivet, sen mukaan miten Kietzmann, ym. (2011), ovat ne artikkelissaan määritelleet.

Taulukko 1 Sosiaalisen median seitsemän kulmakiveä (Kietzmann, Silvestre, McCarthy, & Pitt, 2012).

Kulmakivi	Määritelmä
Identiteetti	Käyttäjän itsensä esilletuomisen laajuus
Keskustelu	Käyttäjien keskinäisen keskustelun laajuus
Jakaminen	Jaetun sisällön määrä sosiaalisen median kanavan sisällä
Läsnäolo	Kuinka paljon käyttäjät tietävät muita käyttäjiä olevan
Suhde	Käyttäjien välisten suhteiden laajuus
Maine	Käyttäjien sisällön ja muiden käyttäjien sosiaalisenstatuksen tiedostamisen laajuus
Ryhmät	Ryhmien muodostamisen, joko pakonedessä tai käyttäjien omasta toimesta, laajuus.

Identiteetillä viitataan yksilöiden henkilökohtaisten tietojen, kuten nimen ja iän paljastamisen määrään. Tämä usein tapahtuu tiedostetusti ja tiedostamattomasti ja jaettu informaatio on usein subjektiivisten ajatusten ja mielipiteiden muodossa. Identiteetti osuuden sosiaalisessa mediassa voidaan katsoa olevan yksi merkittävimmistä asioista joita yritykset voivat hyödyntää, sillä käyttäjien jakamia tietoja voidaan hyödyntää monella tapaa yritysten toiminnassa (Kietzmann, ym. 2011.).

Keskustelu edustaa viitekehyksessä käyttäjien välisen kommunikoinnin laajuutta, eli kuinka laajalti sosiaalisen median kanava tarjoaa käyttäjille mahdollisuuksia kommunikoida yksilöinä tai ryhmissä (Kietzmann, ym. 2011.).

Jakamisella viitataan käyttäjien jakaman sisällön määrään. Jakaminen tarjoaa yritykselle mahdollisuuden saada suuren joukon sosiaalisen median käyttäjiä jakamaan yrityksen tuottamaa sisältöä, kun sisältö kohdennetaan oikealle sisällölle (Kietzmann, ym. 2011.).

Läsnäololla sosiaalisessa mediassa viitataan siihen, että kuinka hyvin käyttäjät tiedostavat muiden käyttäjien olemassaolon, eli siis kuinka hyvin käyttäjä on tietoinen muiden käyttäjien sijainnista suhteessa itseensä todellisessa ja virtuaalisessa maailmassa. Yritysten näkökulmasta tämä tarkoittaa sitä, että kuinka hyvin he voivat tietyn kanavan kautta tavoittaa yksilöitä (Kietzmann, ym. 2011.).

Suhde tarkoittaa viitekehyksessä käyttäjien välisiä yhteisiä tekijöitä, jotka ovat usein käyttäjien välisen interaktion kulmakiviä. Tähän sisältyy myös käyttäjien välisten yhteyksien muodot jotka määrittelevät sen, miten kommunikaatio tapahtuu ja kuinka sisältöä jaetaan (Kietzmann, ym. 2011.).

Maineella viitataan käyttäjien tiedostamaan sosiaaliseen asemaan, sosiaalisessa mediassa. Maineen merkitys vaihtelee sosiaalisen median kanavien välillä, mutta useimmiten maineessa on kyse luottamuksesta (Kietzmann, ym. 2011.).

Ryhmillä tarkoitetaan käyttäjien ryhmien muodostus mahdollisuuksia sosiaalisen median kanavan sisällä. Viitekehyksessä tunnistetaan kaksi isoa ryhmätyyppiä. Ensimmäisessä, käyttäjät muodostava ryhmiä omien kontaktiensa pohjalta ja toisessa ryhmät ovat enemmänkin tosielämän kerhojen vastineita.

2.2 Sosiaalisen median sisältö

Iso osa sosiaalisen median sisällöstä on käyttäjien tuottamaa ja sitä tuotetaan useissa eri formaateissa kuten videona, kuvina, tekstinä tai äänenä. Tämä käyttäjien tuottama sisältö vaihtelee laadullisesti hyvälaatuisen, väärinkäytön ja spämmin välillä (Agichtein, ym., 2008.). Tämä laadun vaihtelevuus sisällössä, johtaa sosiaalisen median sisällön luotettavuuden laskuun. Sosiaalisessa mediassa voidaan tahallisesti tai vahingossa jakaa vääristeltyä tai täysin väärää tietoa ja tämä tekee tehokkaasta sisällön etsimisestä vaikeaa (Ivanov, Vajda, Lee, & Ebrahimi, 2012) (Bian, Liu, Zhou, Agichtein, & Zha, 2009).

Sosiaalisen median sisällölle on myös ominaista hashtagien käyttö. Hashtagit ovat sosiaalisen median sisällä käytettäviä käyttäjien määrittelemiä tageja, joilla pyritään luokittelemaan saman tyylistä sisältöä yhteen ja helposti muiden käyttäjien löydettäväksi (Tsur, & Rappoport, 2012.). Hashtageihin kuitenkin pätee sama ongelma kuin sosiaalisen median sisältöön yleisesti, eli käyttäjät voivat tarkoituksellisesti tai vahingossa väärinkäyttää tageja, tehden näin sisällön hausta tagipohjaisesta sisällönhausta epätarkempaa (Ivanov, Vajda, Lee, & Ebrahimi, 2012).

3 MASSADATA JA MASSADATA-ANALYTIIKKA

Kerätyn datan määrä on kasvanut viime vuosien saatossa huomattavasti ja tämän seurauksena on syntynyt käsite massadata ja tämän pohjalta vielä massadata-analytiikka. Massadataa ja sen analytiikkaa kuvastaa usein datan suuri määrä ja tämän suuren datan määrän analysointi ja molemmat käsitteet käydään syvemmin läpi tässä luvussa (Wu, Zhu, Wu, & Ding, 2014.) (Gandomi, & Haider, 2015.).

3.1 Massadata

Massadata on merkittävässä roolissa nykyisissä yhteiskunnissa, kulttuurissa, liiketoiminnassa sekä tieteessä, erityisesti informaatioteknologian alalla (Boyd & Crawford 2012). Massadatalla yksinkertaisimmillaan usein viitataan suuriin määriin dataa, jota on usein kerätty tarkoituksella, johonkin tarkoitukseen, esimerkiksi strategisen hyödyn saavuttamiseen liiketoiminnassa (Manyika, ym., 2011). Massadataa kertyy yrityksille ja organisaatioille verkon kautta useista lähteistä kuten eri sosiaalisen median kanavista, sähköposteista, sivustoille tehdyistä klikkauksista sekä hakukoneista. Massadata kuitenkin asettaa useita haasteita sitä hallinoiville organisaatioille ja nämä haasteet usein liittyvät massadatan säilöntää ja prosessointiin (Sagiroglu, & Sinanc, 2013.).

Boyd & Crawford (2012) argumentoivat massadatan olevan muuttumassa käsitteenä, teknologian kehityksen myötä. Prosessointi tehon kasvaessa, pystyy tavalliset kotikoneet prosessoimaan samoja data määriä mihin ennen tarvittiin supertietokoneita. Tutkijat kuitenkin noteeraavat, data määrien olevan suuria, mutta eivät näe tätä massadataa käsitteenä määrittelevänä tekijänä. Boyd ja Crawford kuvaavatkin massadataa enemmänkin teknologisena, analyyttisenä ja mytologisena ilmiönä.

> **Teknologia**: Datan prosessointia, analysointia ja keräystä varten prosessointitehon maksimointi
>
> **Analyysi**: Toistuvien kuvioiden suurista data määristä ja näiden kuvioiden pohjalta tehdyt johtopäätökset
>
> **Mytologia**: Ajatus siitä, että suuret data määrät tarjoavat korkeamman tason tietoa, joka voi tarjota näkemyksiä, jotka olivat ennen mahdottomia.

Boyd & Crawford huomauttavat massadataan liittyvän myös utopistisia ja dystopisia näkemyksiä, sekä erittäin optimistisia ajatuksia massadatan suomasta potentiaalista.

Boyd & Crawford tarjoavat hieman erilaisen kuvauksen massadatalle, kuvailemalla sitä teknologisten, analyyttisten ja mytologisten tekijöiden kautta.

Useimmissa massadataa koskevissa artikkeleissa kuitenkin massadata määritellään neljän V:n-mallin kautta, joka käydään läpi seuraavassa kappaleessa (Gandomi, & Haider, 2015).

3.1.1 Massadatan V-malli

Massadataa kuvattaessa monet tutkijat käyttävät kolmen V:n-mallia (Gandomi, & Haider, 2015). Mallissa tunnistetaan kolme kuvaavinta massadatan ominaisuutta, jotka ovat velocity (nopeus), variety (moninaisuus) ja volume (määrä) (Sagiroglu, & Sinanc, 2013).

Nopeus viittaa mallissa datan luonnin nopeuteen, tämän datan prosessoinnin nopeuteen ja siihen, että kuinka nopeasti sen suomaan tietoon tulisi reagoida. Uutta dataa syntyy koko ajan suurissa määrissä ja yhä nopeammalla vauhdilla, johtuen muun muassa, älypuhelimien suuresta käytöstä ja erilaisen digitaalisten median kanavien käytöstä, kuten sosiaalinen media. Strategisen edun saavuttamiseksi olisi täten suotavaa analysoida dataa suoraan datavirrassa, jotta strategista dataa voitaisiin tuottaa mahdollisimman kilpailukykyisesti (Gandomi, & Haider, 2015) (Sagiroglu, & Sinanc, 2013.).

Moninaisuudella v-mallissa tarkoitetaan datan lähteiden vaihtelevuutta sekä datan eri muotoja, jotka ovat strukturoitu, osittain strukturoitu ja strukturoimaton, joista kerrotaan lisää tämän kappaleen seuraavassa osiossa (Sagiroglu, & Sinanc, 2013). Uudet teknologiat sekä analyyttiset menetelmät ovat suoneet uudenlaisen datan tuottamista, kuten asiakasdemografioiden analysointia kasvojentunnistuksen avulla ja asiakaskäyttäytymisen seuraaminen verkkosivulla klikkausten avulla (Gandomi, & Haider, 2015).

Määrä yksinkertaisesti viittaa datan suureen määrään. Puhuttaessa massadatan määrästä usein puhutaan tera- tai petatavuista. Massadatan määrä kuitenkin riippuu eri tekijöistä, kuten ajasta ja datan tyypistä. Dataa kuitenkin usein tuotetaan niin paljon, että sitä kaikkea on vaikea prosessoida nykyisillä teknologioilla (Gandomi, & Haider, 2015) (Sagiroglu, & Sinanc, 2013.).

Edellä mainittujen kolmen v:n lisäksi ovat tietyt organisaatiot lisänneet kolme v:tä lisää kuvaamaan massadataa. IBM lisäsi totuudenmukaisuuden (veracity) yhdeksi massadatan kuvaavaksi aspektiksi. Tällä IBM viittaa datan epäluotettavuuteen, joka on tietyille lähteille ominaista. SAS on myöhemmin lisännyt vaihtelevuuden ja kompleksisuuden massadatan kuvaaviksi tekijöiksi, viitaten datan tuottonopeuksien variaatioon sekä siihen, että data tuotetaan usein usean lähteen tuloksena. Oracle on myös lisännyt oman ulottuvuuden malliin, joka on arvo (value). Arvolla Oracle viittaa siihen, että lähtökohtaisesti datalla ei ole suurta arvoa, mutta analytiikan avulla datasta tuotettu tieto on hyvin arvokasta (Gandomi, & Haider, 2015.).

3.1.2 Data tyypit

Massadata jaeteen usein myös sen datantyyppien, mukaan ja usein massadataa koskevassa kirjallisuudessa tunnistetaan kolme eri data tyyppiä, strukturoitu,

osittain-strukturoitu sekä strukturoimaton data ja nämä datatyypit tulevat usein monissa eri formaateissa, kuten videona, tekstinä ja kuvina (Wu, Zhu, Wu, & Ding, 2014). Nämä kolme datatyyppiä eroavat toisistaan usein rakenteellisesti datan sisältämien viitteiden mukaan, jotka auttavat koneita tunnistamaan dataa (Manyika ym, 2011, s. 33.) (Gandomi, & Haider, 2015.).

Strukturoitu data viittaa taulukoituun dataa, jota usein löytyy taulukkolaskentaohjelmista sekä relaatiotietokannosta (Manyika ym, 2011, s. 33). Strukturoitudata on usein tunnistettavissa jonkin viitteen avulla ja täten koneet pystyvät helposti järjestelemään ja analysoimaan strukturoitua dataa (Sagiroglu, & Sinanc, 2013) (Manyika ym, 2011, s. 33). Massadatan kannalta ongelmallista on kuitenkin se, että vain 5% siitä on strukturoitua dataa (Gandomi, & Haider, 2015).

Strukturoidun datan vastakohtana on strukturoimaton data. Toisinkuin strukturoitu data strukturoimaton data ei sijaitse ennalta määritellyissä kentissä, kuten taulukkolaskentaohjelmissa, vaan on usein vapaamuotoista tekstiä, kuten kirjoissa, blogiteksteissä ja sähköposteissa, tai se voi olla myös viitteetöntä kuva tai video dataa (Manyika ym, 2011, s. 33). Strukturoimattomalta datalta puuttuu kunnollinen rakenne, joka tekee sen analysoinnista vaikeaa koneille (Gandomi, & Haider, 2015).

Strukturoidun ja strukturoimattoman datan välimuotona toimii osittain-strukturoitu data. Kuten strukturoimaton data, osittain-strukturoitu data ei mukaudu ennalta määriteltyjen kenttien mukaisesti, mutta toisinkuin strukturoimaton data, osittain-strukturoitu data sisältää viitteitä, joiden avulla koneet pystyvät helpommin analysoimaan tämän tyyppistä dataa. Osittain-strukturoidun datan viitteet eivät kuitenkaan noudata yhtä tiukkoja standardeja toisinkuin strukturoitu data. Esimerkkejä osittain-strukturoidusta datasta ovat XML ja HTML merkitty teksti (Manyika ym, 2011, s. 33.) (Gandomi, & Haider, 2015.) (Sagiroglu, & Sinanc, 2013.).

3.2 Massadata-analytiikka

Massadata-analytiikka voidaan nähdä kehittyneiden analyyttistenmenetelmien ja massadatan risteytyksenä ja on täten yksi merkittävimmistä business itelligence (BI) trendeistä 2010-luvulla. Massadata-analytiikassa analyyttiset menetelmät käyvät läpi massadata joukkoja ja sen ideana on tuottaa strategista etua yrityksille esimerkiksi markkinoinnin saralla (Fan, Lau, & Zhao, 2015) (Russom, 2011.). Ennen kuin siirrytään eteenpäin massadata-analytiikassa, on määriteltävä käsitteet, joista se koostuu. Massadata on jo tutkielmassa aiemmin määritelty, joten seuraavaksi määritellään kehittyneet analyyttiset menetelmät.

Kehittyneillä analyyttisillä menetelmillä tarkoitetaan, analyyttisten sovellutusten sarjoihin tai joukkoihin jotka auttavat optimisoimaan organisaation toimintaa tai tekemään ennusteita tulevaisuuden kehityssuunnoista. Ne linkittyvät vahvasti yrityksen tai organisaation it-infrastruktuuriin ja niiden tulisi kohdistua vahvasti yrityksen liiketoimintaan, jotta BI-strategiat voitaisiin toteuttaa madol-

lisimman onnistuneesti. Kehittyneet analyyttiset sisältävät käytännön sovelluksia kuten yrityksen tai organisaation tiedon louhinnan kyvykkyyden sekä suositusjärjestelmien luonnin (Bose, 2009.) (Russom, 2011, s. 26.).

Massadata-analytiikka on siis massadatan ja kehittyneiden analyyttisten menetelmien risteys, jossa massadata tuo datan ja kehittyneiden analyyttisten menetelmien avulla siitä pyritään tuottamaan arvokasta informaatiota, jota käytetään BI-strategioiden toteuttamisessa.

3.2.1 Datan-rinnastusmallit

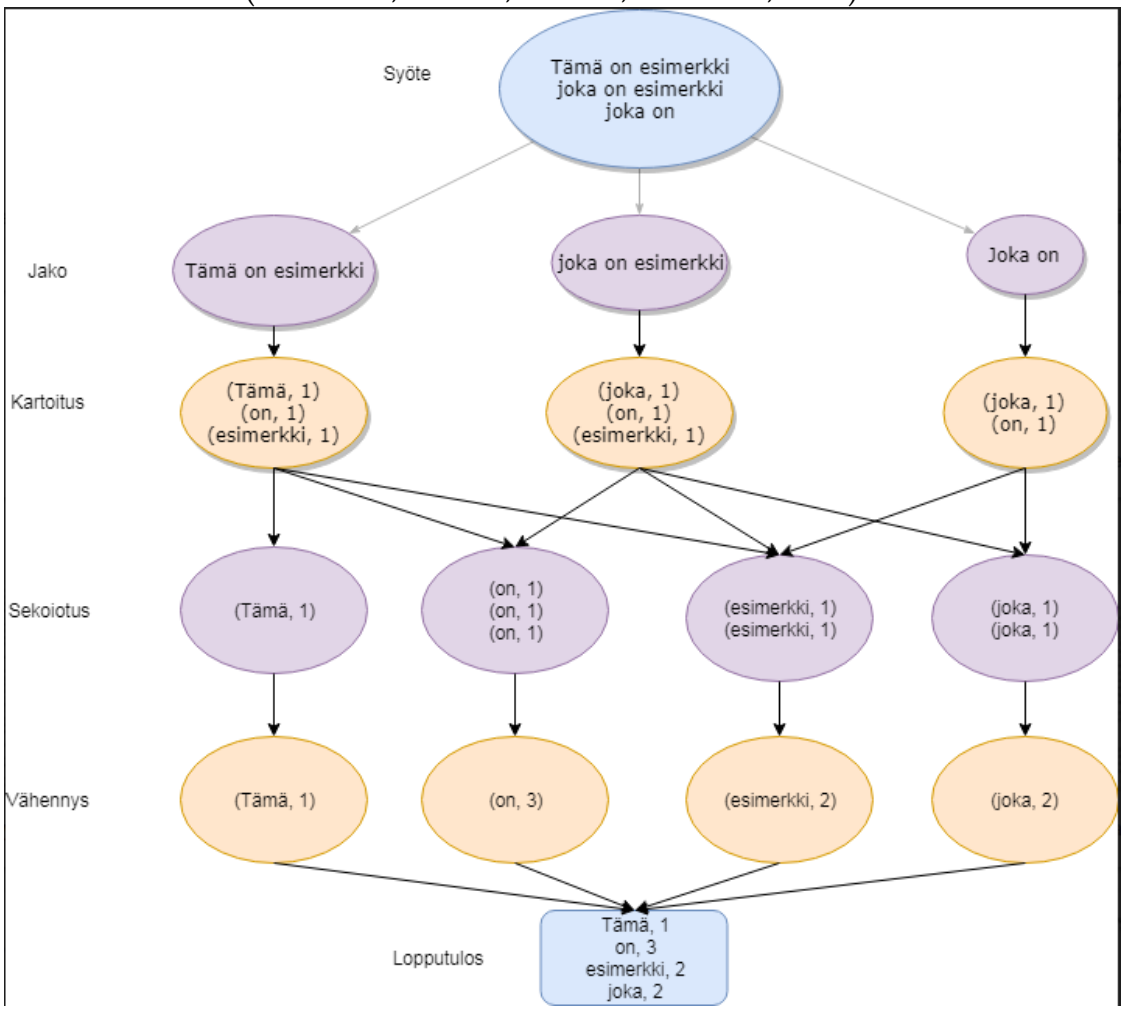
Data-analytiikkaa tyypillisesti sovelletaan datan rinnastuksessa, joka käsittää sisälleen yksittäisille datapalasille tehdyt laskennalliset toimenpiteet (Kambatla, Kollias, Kumar, & Grama, 2014). Datan rinnakkaisuus saadaan esille käyttämällä SPMD-tekniikka (Single Program Multiple Data), jossa yksittäinen operaatio sovelletaan jokaiseen datan palaseen, potentiaalisesti rinnakkaisesti (Kambatla, Kollias, Kumar, & Grama, 2014). Kambatla ym, (2014) käyvät läpi artikkelissaan, Trends in big data analytics, kolme eri kyseessä olevaan tekniikkaan liittyvää mallia, jotka ovat riippuvaisia syötteen tyypistä. Kyseiset mallit ovat batch prosessing, Bulk synchronous parallel processing ja Event processing.

Taulukko 2 Datan-rinnastusmallit (Kambatla, Kollias, Kumar, & Grama, 2014)

Datan-rinnastusmalli	**Määrittely**
Batch prosessing	Suurien data-aineistojen prosessointiin tarkoitettu malli, jossa data pyritään jakamaan osiin, jotta se olisi helpommin ja tehokkaammin prosessoitavissa.
Bulk synchronous parallel processing	Toimintaperiaatteeltaan samanlainen kuin batch prosessing -malli, mutta soveltuu paremmin iteratiivisten algoritmien toimintaan, koska dataa ei tarvitse ladata kullakin prosessointi kierroksella datavarastosta toisinkuin batch prosessing -mallissa.
Event processing	Data prosessoidaan suoraan datavirrasta, eikä datavarastosta ja dataa ei myöskää voida jaotella joukkoihin. Data joko prosessoidaan yksittäisesti

	tai tapahtuma syötevirrasta pyritään tunnistamaan tapahtuma malleja.

Batch prosessing mallissa käsitellään suuria data-aineistoja, jossa moneen data osaan kohdistuvia operaatioita pystytään ryhmittelemään tehokkaasti. Ensimmäinen merkittävä Batch prosessing mallin paradigma oli Googlen MapReduce-ohjelmointimalli (Kuvio 1) (Kambatla, Kollias, Kumar, & Grama, 2014.). MapReduce-mallissa dataan kohdistuva laskenta automaattisesti rinnakkaisettaan suurten kone ryppäiden läpi, se myös käsittelee koneiden virheitä ja ajoittaa koneiden välistä kommunikaatiota. Laskenta mallissa suoritetaan map- ja reduce-komennoilla. Map-komennolla kartoitetaan MapReduce datakirjastoja yhteen ja reduce-komennolla data rajoitetaan relevanttiin dataan, jonka relevanssi on riippuvainen käyttäjän syöttämästä avainsyötteistä (Dean, & Ghemawat, 2008.). MapReduce mahdollistaa esimerkiksi tehokkaan data-analytiikan sekä tiedonlouhinnan (Kambatla, Kollias, Kumar, & Grama, 2014).



KUVIO 1 MapReduce malli (Bello-Orgaz, Jung, & Camacho, 2016)

Bulk synchronous parallel processing –malli, lyhennettynä BSP, on vastaus batch prosessing –mallin ongelmaan, jossa iteratiiviset algoritmit operoivat saman syötteen parissa kunkin iteraation aikana. Batch prosessing –mallia käytettäessä data tulisi kunkin iteraation alussa ladata uudelleen datavarastosta ja kyseinen operaatio on hyvin resurssi-intensiivistä koneille. BSP:ssä laskenta etenee iteroituna, eli kunkin iteraation aikana, samanaikaiset laskennat toteutetaan rinnastetusti ja tätä seuraa synkronointi, jossa eri toimenpiteet kommunikoivat keskenään tarpeen mukaan. Kukin laskenta operaatio käy läpi dataa paikallisessa solmussa ja syöte voidaan pitää välimuistissa eri tasoilla, eikä sitä tarvitse ladata uudelleen. BSP-malliin pohjautuvia ohjelmia käytetään esimerkiksi sosiaalisen median analysointiin (Kambatla, Kollias, Kumar, & Grama, 2014.).

Event processig –malli ideana ”tapahtumien” prosessointi suoraan datavirrasta. Toisin kuin aiemmissa malleissa, aikarajoitteiden takia dataa ei voida jakaa joukkoihin, eikä sitä varsinaisesti tallenneta, vaan data pyritään prosessoimaan sitä mukaa, kun se tulee. Tyypillisiä tapahtumapohjaisia sovelluksia ovat tapausten, prosessointi, valvonta ja tapahtumista ilmoittaminen ja näihin sovelluksiin usein liittyy monimutkainen tapahtumien prosessointi ja tapahtumavirran syötteenä luku sekä joko yksittäisten datan palojen itsenäinen prosessointi tai monimutkaisten kuvioiden tunnistaminen tapahtumista. Aikarajoitteiden tiukentuessa tämänlaiset mallit yleistyvät ja näitä malleja käytetään paljon sosiaalisessa mediassa, kun halutaan esittää käyttäjille sopivimpia mainoksia (Kambatla, Kollias, Kumar, & Grama, 2014.).

3.2.2 Analytiikan hyödyntäminen

Analytiikan koetaan olevan merkittävä strateginen etu liiketoiminnassa, erityisesti markkinoinnin osalta (LaValle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011) (Fan, Lau, & Zhao, 2015). LaValle, Lesser, Shockley, Hopkins & Kruschwitz (2011) tuovat esille artikkelissa, Big Data, Analytics and the Path From Insights to Value, kolme eri analytiikan hyödynnystasoa, tavoitteellinen, kokenut ja muuttunut. Nämä eri tason toimijat eroavat toisistaan sen mukaan, kuinka laajasti analytiikkaa hyödynnetään toiminnassa.

Tavoitteellisen tason toimijat keskittyvät usein pääosin toimintojen ja prosessien automaation ja tehokkuuden kehittämiseen ja näiden toimenpiteiden tarkoituksena on kulujen leikkaaminen. Tavoitteellisilla toimijoilla on yleensä puutteita työkalujen sekä kyvykkäiden työntekijöiden suhteen, jotka auttaisivat yritystä paremmin keräämään, ymmärtämään ja toimimaan analytiikan pohjalta. Tavoitteelliset toimijat usein käyttävät analytiikkaan pelkästään toiminnan perusteluun (LaValle ym,2011.).

Kokeneentason toimijat ovat usein saaneet kokemusta analytiikasta, onnistuneiden toimenpiteiden seurauksena tavoitteellisen tason aikana. Tämän seurauksena kokeneentason toimijat eivät näe analytiikkaa pelkästään kulujen leikkaamisen työkaluna, vaan enemmänkin toimintaa ohjaavana tekijänä. Tällaisilla organisaatioilla on jo kehitteillä tehokkaampia tapoja kerätä, ymmärtää ja toimia analytiikan pohjalta (LaValle ym,2011.).

Muuttuneentason organisaatioilla on jo huomattavaa kokemusta analytiikasta laajalla skaalalla. Nämä toimijat näkevät analytiikan kilpailullisena etuna. Muuntuneentason toimijat eivät enää keskity niinkään toimintojen kulujen leikkaamiseen ja keskittyvät enemmän asiakkaiden tuottavuuteen ja pyrkivät tekemään kohdennettuja sijoituksia markkinarakoihin. Analytiikka tällä tasolla nähdään toimintaa kuvaavana tekijänä (LaValle ym,2011.).

Kuten aiemmin mainittiin, markkinointi on yksi merkittävimmistä massadata-analytiikan hyödyntämisen kohteista. Erityisesti sosiaalisten medioiden tuottama data on mahdollistanut yrityksille tehokkaamman toiminnan markkinoinnin saralla. Kerättyä dataa voidaan hyödyntää markkinoinnissa, asiakastietojen keräämisessä, tuotekehityksessä, myynninedistämisessä, hinnoittelussa sekä sijainti kohtaisessa mainonnassa (Fan, Lau, & Zhao, 2015.).

Asiakkaista kerättyjen tietojen perusteella pystytään toteuttamaan asiakas segmentointia sekä asiakas profilointia. Data-analytiikkaan pohjattuna pystytään asiakkaat jakamaan ryhmiin, jotka perustuvat näiden asiakkaiden jakamiin intresseihin ja preferensseihin. Massadata asettaa kuitenkin haasteita, erityisesti one-to-one –markkinoinnille, jossa pyritään segmentoinnin sijaan luomaan yksittäiselle asiakkaalle profiilia, johon mainonta perustuu, sillä analysoitavan datan määrä on suuri ja tuloksia tulisi saada mahdollisimman nopeasti (Fan, Lau, & Zhao, 2015.).

Data-analytiikan avulla kyetään myös tuottamaan tietoa kuluttajien tuotetyytyväisyydestä. Asiakkaiden tuotetyytyväisyydestä voidaan saada tietoa tekstipohjaisesti, kuluttajien tekemien tuotekuvausten pohjalta, mutta on myös yleistä louhia samaa tietoa kuvista. Tuotetun tiedon pohjalta pystytään tukemaan ja tehostamaan tuotetyytyväisyyden hallintaa (Fan, Lau, & Zhao, 2015.).

Data-analytiikan pohjalta voidaan myös kehittää myynninedistämisenanalytiikkaa sekä suositusjärjestelmiä. Datan pohjalta voidaan tehdä oletuksia kuluttajien reaktioista eri markkinointikampanjoihin eri ajankohtina, eri tuotteisiin ja demografioista riippuen (Fan, Lau, & Zhao, 2015.).

Hinnoittelu strategioiden luonti on myös mahdollista datananalytiikan avulla. Kerätystä datasta louhitun tiedon perustella voidaan tehdä oletuksia tiettyjen tuotteiden tulevasta kysynnästä ja täten hinnoittelua voidaan laatia kysynnän mukaisesti (Fan, Lau, & Zhao, 2015.).

Data-analytiikka mahdollistaa myös sijaintipohjaisen mainonnan. Sijaintipohjaisella mainonnalla voidaan helposti kohdata asiakkaiden tarpeita ja on tärkeä tekijä personoidussa markkinoinnissa. Pyrkimyksenä tässä mainonta tyypissä on näyttää kuluttajalle mainoksia, kuluttajan sen hetkisen sijainnin mukaan ja kuluttajan oletetun tulevan sijainnin mukaan. Tämän mainonta tavan on mahdollistanut mobiiliteknologioiden kehittyminen, minkä ansiosta kyetään helposti seuraamaan kuluttajan sijaintia reaaliajassa (Fan, Lau, & Zhao, 2015.).

4 SOSIAALISEN MEDIAN MASSADATA JA DATA-ANALYTIIKKA

Kuten aiemmin mainittiin, massadataa kerätään useasta eri lähteistä ja yksi merkittävimmistä massadatan lähteistä on sosiaalinen media. Dataa kertyy useista eri sosiaalisen median kanavista, kuten youtubesta, facebookista ja twitteristä. Kyseinen data tulee myös useissa eri formaateissa kuten videona kuvina ja tekstinä, ja pitää myös sisällään eri asioita koskevaa informaatioita, kuten demografista tietoa, tietoa harrastuksista sekä tietoa terveydestä (Bello-Orgaz, Jung, & Camacho, 2016.).

Tässä luvussa käydään läpi tutkimuksen pohjalta tehdyt löydöt ja johtopäätökset koskien sosiaalisen median ja massadatan suhdetta, eli minkälaista sosiaaline median data on massadatan kontekstissa, kuinka data-analytiikka toimii sosiaalisen median datan kanssa ja kuinka dataa voidaan hyödyntää.

4.1 Sosiaalisen median data

Sosiaalisen median massadata voidaan nähdä malliesimerkkinä massadatasta, joka lähes suoraan soveltuu massadatan V-malliin. Velocity eli datan määrän kasvunopeus (Sagiroglu, & Sinanc, 2013), ilmenee sosiaalisen median massadatassa dataa tuottavien käyttäjien kasvussa. Käyttäjien määrä on ollut viime vuosina huomattavassa kasvussa ja käyttäjien määrän on arvioitu ylittävän kolmen miljardin rajan vuoteen 2020 mennessä (Statista, 2018). Variety, eli datan vaihtelevuus (Sagiroglu, & Sinanc, 2013), ilmenee sosiaalisessa mediassa, ilmenee datan eri lähteiden, eli sosiaalisen median eri kanavien määrällä sekä datan struktuurin vaihtelevuudessa, että dataformaattien vaihtelevuudessa (Bello-Orgaz, Jung, & Camacho, 2016). Volume, eli datan määrän, voidaan myös olettaa olevan merkittävä osa sosiaalisen median massadataa (Sagiroglu, & Sinanc, 2013), sillä sen on todettu olevan yksi merkittävimmistä massadatan lähteistä ja datan määrän voidaan olettaa olevan myös suuri pelkästään käyttäjien määrän perusteella (Bello-Orgaz, Jung, & Camacho, 2016) (Statista, 2018).

Sosiaalisen median massadatan rooli malliesimerkkinä massadatasta, tarkoittaa myös sitä, että siinä ilmenee myös massadatan olennaisimmat ongelmat, joista osa voi myös hieman korostua sosiaalisen median kontekstissa. Datan säilöntä ja prosessointi ovat huomattavia ongelmia sosiaalisen median massadatassa tuotetun datan määrän takia (Bello-Orgaz, Jung, & Camacho, 2016). Sosiaalisen median massadatan ongelmien erityisominaisuutena voidaan kuitenkin nähdä sen korostunut epäluotettavuus. Massadatalle on ominaista usein suhteellinen epäluotettavuus (Gandomi, & Haider, 2015), jonka voidaan katsoa korostuvan sosiaalisen median massadatassa datan käyttäjälähtöisyyden takia (Bian, Liu, Zhou, Agichtein, & Zha, 2009). Sosiaalisessa mediassa käyttäjät saattavat tahallisesti tai vahingossa jakaa väärää tietoa (Ivanov, Vajda, Lee, & Ebrahimi, 2012),

joka voi laajemmin ilmetessään vääristää kerättyä dataa huomattavasti, laskien näin kerätyn datan luotettavuutta.

Kuten massadatan määrittelyssä todettiin, suurin osa massadatasta on strukturoimatonta (Gandomi, & Haider, 2015) ja kun huomioidaan sosiaalisen median rooli merkittävänä massadatan lähteenä (Bello-Orgaz, Jung, & Camacho, 2016.), voidaan suurimman osan sosiaalisen median datasta olevan strukturoimatonta.

Sosiaalisen median data sisältää kuitenkin myös strukturoitua ja osittain strukturoitua dataa käyttäjätietojen muodossa. Useat sosiaalisen median kanavat vaativat käyttäjätietojen täyttämistä ennen käyttöä ja esimerkiksi facebook vaatii käyttäjän nimen, sukunimen, sähköpostiosoitteen, puhelinnumeron ja syntymäajan. Näille tiedoille on usein ennalta määritellyt syöttökentät, joten niihin syötetty data on helposti strukturoitavissa kenttäspesifien viitteiden avulla, jolloin luodusta datasta tulee vähintään osittain strukturoitua (Gandomi, & Haider, 2015). Tämä typpisen datan määrä voi kuitenkin vaihdella eri sosiaalisen median kanavien välillä, sillä Kietzmann, ym. (2011) sosiaalisen median seitsemän kulmakiven -mallin mukaan käyttäjien oman identiteetin esille tuonti vaihtelee eri sosiaalisen median kanavien välillä ja tämän voidaan olettaa koskevan myös eri sosiaalisen median kanavien käyttöä vaativia käyttäjätietoja.

Sosiaalisen median dataa voidaan myös mahdollisesti strukturoida käyttämällä sosiaalisen median sisältöä koskevia hashtageja. Hashtageja käytetään sosiaalisessa mediassa saman tyyppisten sisältöjen luokitteluun, jotta kyseinen sisältö olisi käyttäjille helposti löydettävissä (Tsur, & Rappoport, 2012), täten tämän, jo olemassa olevan systeemin kääntäminen datan strukturointiin, kävisi hyvin luonnollisesti. Tässä tapauksessa kyseessä oleva data strukturoitaisiin käyttämällä datan sisältämää hashtagia viitteenä ja täten datan sisältämä informaatio olisi ennalta tiedettävissä hashtagin perusteella. Tällä tavalla strukturoitu data olisi todennäköisesti korkeintaan osittain strukturoitua dataa, sillä käytetyt viitteet ovat standardeiltaan erittäin löyhiä (Gandomi, & Haider, 2015). Hashtagien käyttö datan strukturoinnissa ei kuitenkaan ole täysin ongelmatonta, sillä hashtagit ovat käyttäjien määrittelemiä ja voivat täten olla hyvinkin epätarkkoja (Ivanov, Vajda, Lee, & Ebrahimi, 2012). Hashtagit ovat kuitenkin käyttäjien määrittelemiä (Tufekci, 2014) ja niitä voidaan tahallisesti tai vahingossa väärinkäyttää, ja tämä voi datan strukturoinnissa johtaa siihen, että käytetyn viitteen alla oleva sisältö ei vastaa viitettä. Tämä ei kuitenkaan poissulje hashtageja strukturoinnin välineenä, mutta edellyttää varovaisuutta niitä käytettäessä, tässä tarkoituksessa.

4.2 Sosiaalisen median massadatan data-analytiikka

Kuten massadataa yleisesti, tulee sosiaalisen median massadataa myöskin analysoida, jotta siitä saataisiin arvoa ja strategista hyötyä (Fan, Lau, & Zhao, 2015) (Gandomi, & Haider, 2015). Seuraavaksi tässä luvussa käydään läpi datan-rinnastus-mallien soveltuvuutta sosiaalisen median massadatan analysointiin.

Batch prosessing -mallia käytetään usein suurten datamäärien läpikäyntii, joten erityisesti batch prosessing -mallin MapReduce-paradigman voitaisiin katsoa soveltuvan hyvin sosiaalisen median tuottamiin tekstipohjaisiin sisältöihin (Bello-Orgaz, Jung, & Camacho, 2016) (Kambatla, Kollias, Kumar, & Grama, 2014). MapReduce mahdollistaisin tekstin läpikännin tiettyjen avainsanojen avulla ja mahdollistaisi nopean tekstinsisällön arvioinnin. Batch prosessing mallissa data kuitenkin haetaan muistista, eikä suoraan datavirrasta, joten datan reaaliaikainen analysointi on tällä mallilla vaikeaa (Dean, & Ghemawat, 2008.).

Kuten massadataa käsittelevässä luvussa mainittiin Bulk synchronous parallel processing –malli, luo skaalautuvuus etuja Batch prosessing -malliin suhteutettuna, kun käytetään iteratiivisia algoritmeja. Todennäköisesti juuri näiden skaalautuvuusetujen vuoksi Bulk synchronous parallel processing –mallin eri paradigmat ovat kasvavissa määrin käytössä eri sosiaalisen median kanavien data-analytiikassa. Vaikka Bulk synchronous parallel processing –malli, tarjoaa merkittäviä skaalautuvuus etuja batch prosessing -malliin suhteutettuna, ei se mahdollista datan reaaliaikaista analysointia, koska data edelleenkin haetaan muistista (Kambatla, Kollias, Kumar, & Grama, 2014.).

Ratkaisu aiemmin mainittuun reaaliaikaiseen datan analysointiin on event processing -malli. Event processing -malli soveltuu hyvin käyttäjien tekemiin klikkausiin ja tilapäivityksiin reaaliaikaiseen analysointiin sosiaalisessa mediassa, sillä data haetaan suoraan datavirrasta. Tämän tyyppinen analysointi mahdollistaa sellaisten mainosten näyttämisen käyttäjille, jotka vastaavat käyttäjän sen hetkisiä tarpeita. Juuri tästä syystä voidaan event processing -mallin katsoa soveltuvan erittäin hyvin sosiaalisen median data-analysointiin, ainakin mainonnan osalta (Kambatla, Kollias, Kumar, & Grama, 2014.).

4.3 Sosiaalisen median datan hyödyntäminen

Sosiaalisen median massadatan ja massadata-analytiikan lisäksi on tärkeää käydä läpi sosiaalisen median data-analytiikan pohjalta tuotetun tiedon hyödyntämistä, jotta sosiaalisen median massadataa voidaan ymmärtää kokonaisuudessaan.

Sosiaalisen median datan hyödyntämisen voidaan katsoa olevan tärkeää erityisesti markkinoijille, sillä sosiaalisen median datan voi pitää sisällään merkittävää strategista tietoa yrityksen bränditietoisuudesta, kilpailijoiden toimista, sekä mahdollisia ratkaisuja jolla voidaan saavuttaa etua suhteessa kilpailijoihin (He, Zha, & Li, 2013). Lisäksi sosiaalisesta mediasta pystytään saamaan selville kuluttajien mielipiteitä tuotteista ja datan pohjalta voidaan myös pyrkiä luomaan uusia tuotteita ja palveluita yritykselle (He, Zha, & Li, 2013).

Seuraavaksi käydään läpi sosiaalisen median massadatan hyödyntämistä läpi Majid, ym. (2013) artikkelin, A context-aware personalized travel recommendation system based on geotagged social media data mining, kautta. Majid, ym. (2013) käyvät artikkelissaan läpi sijaintimerkityn sosiaalisen median sisällön

käyttämistä, lomakohde suositusjärjestelmien luonnissa. Tässä mallissa sosiaalisen median sisältö käydään läpi ja pyritään selvittämään käyttäjän suhtautuminen lomakohteeseen ja sijaintitietoihin pohjaten pyritään suosittelemaan saman tapaisia lomakohteita (Majid, ym. 2013.). Edelle kuvailtua mallia voidaan mahdollisesti, hieman muokaten, soveltaa yleisesti, suositusjärjestelmien luonnissa sosiaalisen median massadatan pohjalta.

5 YHTEENVETO

Tutkielmassa käytiin läpi sosiaalisen median massadataa ja massadata-analytiikka, yleisemmällä tasolla. Tehdyn tutkimuksen pohjalta todettiin sosiaalisen median massadatan täyttävän massadatan määritelmän kolmen V:n -mallin kautta. Huomattavaa oli myös sosiaalisen median massadatan struktuurissa, joka suurimmalta osin on samanlaista kuin massadatan struktuuri yleisesti, mutta sisältää potentiaalisia vaihtoehtoisia strukturointi mahdollisuuksia hashtagien muodossa. Lisäksi sosiaalisen median massadatan todettiin olevan epäluotettavampaa, sillä se on suurin osin käyttäjien tuottamaa.

Sosiaalisen median massadata-analytiikan osalta todettiin Bulk synchronous parallel processing –mallin soveltuvan parhaiten sosiaalisen median analysointiin, mutta huomautettiin MarReduce ohjelmistojen hyödyllisyys tekstipohjaisten aineistojen analysoinnissa. Event processing -mallin tärkeyttä myöskin korostettiin, sillä se tarjoaa mahdollisuuden reaaliaikaiseen analysointiin datavirtojen analysoinnin kautta.

Sosiaalisen median massadatan hyödyntämisen osalta todettiin sillä olevan merkittävä vaikutus markkinoinnin saralla, sekä strategisen edun luojana. Lisäksi käytiin myös läpi sosiaalisen median dataan perustuvan suositusjärjestelmän luontia.

Tulosten pohjalta voidaan todeta sosiaalisen median massadatalla olevan merkittävä rooli massadatassa yleisesti. Sosiaalinen media on yksi merkittävimmistä yksittäisistä massadatan lähteistä ja on täten merkittävä tekijä massadatan muovaajana. Tulokset auttavat paremmin ymmärtämään sosiaalisen median massadataa kokonaisuutena, mutta eivät kuitenkaan tarjoa ratkaisuja sosiaalisen median massadatan ongelmiin, kuten korostuneeseen epäluotettavuuteen. Lisäksi hashtagien rooli datan strukturoinnissa on tällä hetkellä vain spekuloinnin tasolla ja vaatii tulevaisuudessa tarkempaa tutkimusta, jotta niiden käytettävyyttä tässä tarkoituksessa voidaan paremmin arvioida.

Kandidaatin tutkielman rajoitteiden vuoksi aiheeseen ei voitu kovin syvälle mennä, täten sosiaalisen median massadataa päädyttiin tutkimaan yleisemmällä tasolla, jotta tutkielma ei venyisi rajojen ulkopuolelle ja jotta pystyttäisiin paremmin ylläpitämään tutkielman fokus. Aihe oli kuitenkin hieman liian laaja kandidaatin tutkielmaksi, joka vaikeutti huomattavasti fokuksen ylläpitoa tutkielman kirjoittamisen aikana. Tulokset jäivätkin todennäköisesti myös tästä syystä hyvin pintapuolisiksi.

Tutkimus tulosten pohjalta nousi esille myös useita jatkotutkimus aiheita. Tutkielmassa nostettiin esille hashtagien rooli datan strukturoinnissa ja kuten aiemmin mainittiin tätä tulisi tulevaisuudessa enemmän tutkia. Tämä vaatisi kunnollisen aineiston keräämistä sosiaalisesta mediasta, jossa käytettäisiin hashtageja datan viitteinä ja tämän jälkeen pyrittäisiin tarkastelemaan viitteiden luotettavuutta. Lisäksi tutkielmassa esille nousi sosiaalisen median datan korostunut epäluotettavuus. Jatkotutkimusaiheiksi tämän osalta ehdotetaan datan

epäluotettavuuden laajuuden tutkimista, sekä mahdollisien ratkaisujen löytämistä epäluotettavan datan poissulkemiseen.

LÄHTEET

Agichtein, E., Castillo, C., Donato, D., Gionis, A. & Mishne, G. (2008). *Finding high-quality content in social media*. In Proceedings of the 2008 international conference on web search and data mining (pp. 183-194). ACM.

Bello-Orgaz, G., Jung, J. J. & Camacho, D. (2016). *Social big data: Recent achievements and new challenges*. Information Fusion, 28, 45-59.

Bian, J., Liu, Y., Zhou, D., Agichtein, E. & Zha, H. (2009). *Learning to recognize reliable users and content in social media with coupled mutual reinforcement*. In Proceedings of the 18th international conference on World wide web (pp. 51-60). ACM.

Bose, R. (2009). *Advanced analytics: opportunities and challenges*. Industrial Management & Data Systems, 109(2), 155-172.

Boyd, D. & Crawford, K. (2012). *Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon*. Information, communication & society, 15(5), 662-679.

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., ... & Sampath, D. (2010, September). *The YouTube video recommendation system*. In Proceedings of the fourth ACM conference on Recommender systems (pp. 293-296). ACM.

Dean, J. & Ghemawat, S. (2008). *MapReduce: simplified data processing on large clusters*. Communications of the ACM, 51(1), 107-113.

Fan, S., Lau, R. Y. & Zhao, J. L. (2015). *Demystifying big data analytics for business intelligence through the lens of marketing mix*. Big Data Research, 2(1), 28-32.

Gandomi, A. & Haider, M. (2015). *Beyond the hype: Big data concepts, methods, and analytics*. International Journal of Information Management, 35(2), 137-144.

He, W., Zha, S. & Li, L. (2013). *Social media competitive analysis and text mining: A case study in the pizza industry*. International Journal of Information Management, 33(3), 464-472.

Hoffman, D. L. & Fodor, M. (2010). *Can you measure the ROI of your social media marketing?*. MIT Sloan Management Review, 52(1), 41.

Hoyer, W. D. & Brown, S. P. (1990). *Effects of brand awareness on choice for a common, repeat-purchase product*. Journal of consumer research, 17(2), 141-148.

Ivanov, I., Vajda, P., Lee, J. S. & Ebrahimi, T. (2012). *In tags we trust: Trust modeling in social tagging of multimedia content*. IEEE Signal Processing Magazine, 29(2), 98-107.

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014). *Trends in big data analytics*. Journal of Parallel and Distributed Computing, 74(7), 2561-2573.

Kaplan, A. M. & Haenlein, M. (2010). *Users of the world, unite! The challenges and opportunities of Social Media.* Business horizons, 53(1), 59-68.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P. & Silvestre, B. S. (2011). *Social media? Get serious! Understanding the functional building blocks of social media*. Business horizons, 54(3), 241-251.

Kietzmann, J. H., Silvestre, B. S., McCarthy, I. P. & Pitt, L. F. (2012). *Unpacking the social media phenomenon: towards a research agenda*. Journal of Public Affairs, 12(2), 109-119.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S. & Kruschwitz, N. (2011). *Big data, analytics and the path from insights to value*. MIT sloan management review, 52(2), 21.

Li, F. & Du, T. C. (2011). *Who is talking? An ontology-based opinion leader identification framework for word-of-mouth marketing in online social blogs*. Decision Support Systems, 51(1), 190-197.

Majid, A., Chen, L., Chen, G., Mirza, H. T., Hussain, I. & Woodward, J. (2013). *A context-aware personalized travel recommendation system based on geotagged social media data mining*. International Journal of Geographical Information Science, 27(4), 662-684.

Mangold, W. G. & Faulds, D. J. (2009). *Social media: The new hybrid element of the promotion mix*. Business horizons, 52(4), 357-365.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Russom, P. (2011). *Big data analytics*. TDWI best practices report, fourth quarter, 19, 40.

Sagiroglu, S. & Sinanc, D. (2013). *Big data: A review*. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42-47). IEEE.

Saravanakumar, M. & SuganthaLakshmi, T. (2012). *Social media marketing*. Life Science Journal, 9(4), 4444-4451.

Statista (24.1.2018) Number of social media users worldwide from 2010 to 2021 (in billions), Haettu osoitteesta https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Tsur, O. & Rappoport, A. (2012, February). *What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities*. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 643-652). ACM.

Tufekci, Z. (2014). *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*. ICWSM, 14, 505-514.

Wu, X., Zhu, X., Wu, G. Q. & Ding, W. (2014). *Data mining with big data*. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

Xiang, Z. & Gretzel, U. (2010). *Role of social media in online travel information search*. Tourism management, 31(2), 179-188.