Aihemallinnus journalismin työkaluna:

menetelmä ja mahdollisuudet

Eleonora Tellervo Alariesto

Haaga-Helia ammattikorkeakoulu
Medianomi amk
Opinnäytetyö
2022

Tiivistelmä

Tekijä(t)
Alariesto, Eleonora Tellervo

Tutkinto
Medianomi

Raportin/Opinnäytetyön nimi
Aihemallinnus journalismin työkaluna: menetelmä ja mahdollisuudet

Sivu- ja liitesivumäärä
35

Nykyään dataa on ja sitä syntyy jatkuvasti valtavia määriä eri organisaatioiden ja ihmisten toimesta – erityisesti tekstimuotoista dataa tarkastelemalla journalistit voivat löytää tietoa ja juttuaiheita. Datajournalismi on erikoistunut datan hyödyntämiseen journalistisissa tarkoituksissa.

Tässä opinnäytetyössä esitellään algoritmiin perustuva ohjaamaton koneoppimismenetelmä, aihemallinnus. Menetelmää voi käyttää suurten tekstiaineistojen tarkasteluun. Tutkimuksen pyrkimyksenä on selostaa menetelmän tausta ja käyttö sekä pohtia sen käyttömahdollisuuksia journalismin työkaluna.

Aihemallinnus havaitsee aineistosta piileviä rakenteita, joita kutsutaan aiheiksi. Aihemallinnus tarkastelee aineistoa sen tilastollisten ominaisuuksien kautta, havainnoiden aineiston dokumenteissa esiintyvien sanojen esiintymistiheyttä. Tuloksenaan aihemallinnus esittää sanalistoja eli ’aiheita’ sekä aiheiden jakautumia aineiston dokumenteissa. Aiheet ovat yhdessä esiintyvien sanojen joukkoja, jotka paljastavat, mistä aineistossa on kyse. Menetelmää voidaan käyttää ikään kuin esikatseluikkunana suureen aineistoon.

Aihemallinnus on teknisesti vaativa menetelmä ja sen käyttö edellyttää perehtyneisyyttä. Aihemallinnuksen tekijän on dokumentoitava työnsä vaiheet tarkasti ja kyettävä arvioimaan aihemallinnuksen laadukkuutta. Aihemallinnuksen tulokset ovat määrällistä tietoa ja niistä tulkintojen tekeminen vaatii menetelmän kautta saavutettujen havaintojen yhdistämistä laadulliseen analyysiin.

Aihemallinnusta on jo käytetty muun muassa puolueohjelmien, uutisartikkelien, twiittien ja laululyriikoiden tarkasteluun. Menetelmä on otettu ahkeraan käyttöön esimerkiksi yhteiskuntatieteissä, politiikan tutkimuksessa, lääketieteessä, psykologiassa ja informaatiotutkimuksessa.

Journalismissa aihemallinnusta voidaan käyttää valtavaan tekstiaineistoon tutustumisessa, tiedon haussa, aineistojen vertailussa ja menetelmän tuloksia voidaan visualisoida erilaisia muuttujia hyödyntäen. Aihemallinnuksen tuloksista ei ehkä yksinään riitä jutuksi, mutta niitä hyödyntäen journalisti voi tehdä sellaisia havaintoja, joita lähiluvussa ei ehkä huomattaisi ja kysyä journalistisesti mielenkiintoisia kysymyksiä.

Asiasanat
journalismi, aihemallinnus, tekoäly, tiedonhankinta, datajournalismi, data

Sisällys

1 Johdanto

Journalismi on kulkenut käsi kädessä teknologian kanssa koko sen historian ajan. Erityisesti julkaisemiseen liittyvä teknologia ja sen eri muodot ovat vaikuttaneet journalismin kehittymiseen. Nyt alalla pohditaan, miten hyödyntää tekoälyä ja automaatiota siten, että ne palvelisivat parhaiten journalistisia päämääriä.

Datajournalismin määritelmää on tapailtu vuosia, mutta kuten journalismi, on myös data moniselitteinen asia. Datajournalismin lähteenä voi toimia esimerkiksi offshore-pankista vuodetut asiakirjat, vuosikymmenten aikana radiossa soitettujen top10 -kappaleiden lyriikat tai internetin eri sivustojen sekä ihmisten ja organisaatioiden käyttämien teknisten välineiden tuottama data. Journalistin valmiudet käsitellä tuota dataa ovat ensiarvoisen tärkeitä. Osaavissa käsissä miltei mikä tahansa data voi muotoutua uutiseksi tai jutuksi.

Eräs tapa lähestyä dataa, on tässä opinnäytetyössä esiteltävä menetelmä, aihemallinnus. Se on tietojenkäsittelytieteen alalla kehitetty laskennallinen menetelmä. Se perustuu algoritmiin, joka käy läpi esimerkiksi tekstikokoelman luoden tilastollisen kuvauksen aineiston sisältämien sanojen esiintyvyydestä. Aihemallinnus nostaa tuloksekseen aineistosta havaitsemiaan piileviä aiheita – siksi nimi. Aihemallinnuksella voidaan käsitellä tuhansia eri tekstejä kerralla, enemmänkin.

Aihemallinnus voi toimia ikään kuin esikatseluikkuna. Sen avulla voi saada käsityksen koko aineistosta, tarkastella lähemmin yksittäisiä aineiston osia tai esittää aineistosta havaittuja trendejä visualisoinnein. Menetelmää käyttämällä voi saada alustavan käsityksen aineiston sisällöstä ja sisällöllisten aiheiden jakautumisesta aineistossa.

Muuttumaton algoritmi ja koneen vakioitu tapa löytää aineistoista piileviä rakenteita ei tee aihemallinnuksesta täydellistä menetelmää. Aihemallinnusta valmistellessa sen käyttäjä joutuu tekemään valintoja, joilla on vaikutuksensa tuloksiin. Jokainen tätä opinnäytetyötä varten luettu lähde muistuttaa noiden valintojen dokumentoinnin tärkeydestä ja aihemallinnuksen tekijän tulkintavallasta tuloksien suhteen.

Journalismiin kuuluva itsekriittinen ote sopii yhteen tunnollista dokumentointia peräänkuuluttavan ja tulkinnoista tietoisen aihemallinnuksen maailmaan. Tämän opinnäytetyön aihe on hyvin tekninen, mutta kyseessä on journalismin opinnäytetyö. Tavoitteenani on esitellä aihemallinnus menetelmänä ja ilmiönä. Tulen esittelemään myös aihemallinnuksen käyttötapauksia, pohdin menetelmän sopivuutta ja ideoin sen käyttöä journalismin työkaluna.

2 Menetelmä

Tämä opinnäytetyö pyrkii esittelemään laskennallisen menetelmän, aihemallinnuksen, selvittämään menetelmän taustaa sekä kuvailemaan ja pohtimaan sen käyttömahdollisuuksia journalismissa. Työ on ennen kaikkea tutkiva, mutta siinä myös ideoidaan ja pyritään antamaan esimerkkejä aihemallinnuksen käytöstä. Opinnäytetyön aihetta lähestytään kirjallisuuskatsauksen keinoin.

Opinnäytetyötä voidaan pitää tutkivana siksi, että työssä pyritään ratkaisemaan tiedollinen ongelma: miten aihemallinnusta voisi käyttää journalismin työkaluna? Tämän lisäksi opinnäytetyössä etsitään vastausta seuraaviin kysymyksiin: missä vaiheessa jutuntekoa aihemallinnusta voidaan käyttää, minkälaisia aineistoja journalisti voisi aihemallintaa ja minkälaisiin kysymyksiin journalisti voisi aihemallinnuksen keinoin etsiä vastauksia? Jotta voitaisiin vastata edellä esitettyihin kysymyksiin, on myös selvitettävä mitä aihemallinnus on ja miten sitä käytetään.

Vastausta näihin kysymyksiin haetaan tekemällä katsaus aihetta käsittelevään kirjallisuuteen. Salmisen (2011) mukaan kirjallisuuskatsauksella pyritään rakentamaan kokonaiskuva asiakokonaisuudesta. Tämän opinnäytetyön aiheeseen liittyvän kirjallisuuden katsaus on ennen kaikkea kuvaileva. Kuvailevaa kirjallisuus katsausta voidaan luonnehtia yleiskatsaukseksi ja se pyrkii kuvaamaan tutkittavaa ilmiötä laaja-alaisesti. Kuvailevassa kirjallisuuskatsauksessa aineiston keruu ja tutkimuskysymyksen asettelu saa olla väljempää, kuin esimerkiksi systemaattisessa kirjallisuuskatsauksessa. (Salminen, 2011)

Aihemallinnus on verrattain nuori laskennallinen menetelmä, aikaisimmat maininnat siitä löytyvät vuosituhannen vaihteen ajoilta (Kherwa & Bansal, 2020). Tässä opinnäytetyössä lähteenä käytetty kirjallisuus löytyi verkkohauin ja erilaisten akateemisten julkaisuportaalien kautta. Valtaosa aihemallinnukseen liittyvistä lähteistä on vertaisarvioituja tieteellisiä julkaisuja tai artikkeleja. Lähteiden etsimisessä johtoajatuksena oli lähteen laatu ja esitellyn aihemallinnuksen soveltuvuus journalistiseen käyttöön. Journalismiin liittyvät lähteet ovat myös pääosin verkkohauin löytyneitä.

Ennen itse opinnäytetyön aiheen eli aihemallinnuksen esittelyä luodaan yhteys journalismiin ja dataan. Näitä aiheita lähestytään myös kirjallisuuden kautta, narratiivisella ja toimituksella otteella. Narratiivista kirjallisuuskatsausta Salminen (2011) luonnehtii metodisesti keveimmäksi kirjallisuuskatsauksen muodoksi. Toimituksellista katsausta hän kuvaa napakaksi ja julkaisun teemaa tukevaksi. (Salminen, 2011)

Työ etenee seuraavasti: tietoperustan alussa esitellään ensin keskeiset käsitteet journalismi, data ja datajournalismi, sekä selvitetään hieman datajournalistista jutuntekoa. Tämän jälkeen luvussa

3.2 edetään esittelemään aihemallinnuksen taustaa tietojenkäsittelyä, tekoälyä, koneoppimista ja luonnollisen kielen käsittelyä kuvaillen. Seuraavaksi päästään itse aihemallinnuksen pariin. Luvussa 3.3 kuvaillaan aihemallinnus menetelmänä ja avataan sen käyttöä sekä kuvaillaan esimerkein, miten menetelmää on aiemmin käytetty.

Luvussa 4 yhdistetään aihemallinnus ja journalismi sekä pohditaan menetelmän mahdollisia käyttökohteita. Luvussa vastataan tutkimuksen pääkysymykseen: miten aihemallinnusta voisi käyttää journalismin työkaluna? Tämän lisäksi luvussa pohditaan esimerkkien ja ideoinnin kautta sitä, missä vaiheessa jutuntekoa aihemallinnusta voidaan käyttää, minkälaisia aineistoja journalisti voisi aihemallintaa ja minkälaisiin kysymyksiin journalisti voisi aihemallinnuksen keinoin etsiä vastauksia?

Luvussa 5 tehdään yhteenvetoa ja pohditaan opinnäytetyön tuloksia. Luvussa myös pyritään arvioimaan opinnäytetyön tekoprosessia ja tekijän oppimista.

3 Tietoperusta

Jotta voitaisiin pohtia aihemallinnuksen käyttöä journalismissa, on ensin otettava selvää, mitä itse journalismi on. Datajournalismi vaatii myös avaamista käsitteenä ja journalismin alalajina. Myös data on määriteltävä ja sen ominaisuuksia kuvailtava, sillä aihemallinnus on yksinkertaisimmillaan datan käsittelyä ja tarkastelua.

Ennen aihemallinnusta on esiteltävä sen juuret, jotka ovat tietojenkäsittelyssä. Aihemallinnus perustuu tekoälyyn, ohjaamattomaan koneoppimiseen ja luonnollisen kielen käsittelyyn ja myös nämä käsitteet on avattava. Tämän jälkeen selostetaan aihemallinnuksen toimintaa sekä käyttöä ja selitetään siihen liittyviä käsitteitä. Viimeiseksi tietoperustassa esitellään muutamia esimerkkejä aihemallinnuksen käytöstä.

3.1 Journalismi ja data

3.1.1 Journalismi

”Journalismi on yhteiskunnallista tiedonvälitystä” toteaa Jaakkola oppikirjassaan Hyvä journalismi (2013). Jaakkolan mukaan journalismin pyrkimys on varmistaa, että yhteiskunnallisesti merkityksellinen tieto välittyy nopeasti ja luotettavasti. Journalistisen sisällön tuotantoa ohjaavat ammatilliset normit ja käytännöt. Journalismia voidaan pitää myös ammatillisena arvojärjestelmänä, jonka arvoihin lukeutuu: ajantasaisuus, todenmukaisuus, puolueettomuus, itsenäisyys, edustavuus ja vastuullisuus. (Jaakkola, 2013)

Mark Deuze tunnistaa artikkelissaan ”Mitä on journalismi?” viisi olennaista elementtiä, joista journalismin voidaan katsoa koostuvan. Ensimmäisenä hän toteaa journalismi olevan julkista palvelua – se on informaation keräystä, jäsennystä ja jakamista yleisölle sekä vallan vahtikoirana toimimista. (Deuze, 2005)

Toiseksi elementiksi Deuze nimeää objektiivisuuden. Journalisti pyrkii puolueettomuuteen, neutraaliuteen ja reiluuteen. Kolmas elementeistä on autonomia. Vapaus ja itsenäisyys tarkoittavat sitä, ettei journalisti luovuta päätöksentekovaltaa työstään toimituksen ulkopuolelle tai anna esimerkiksi painostuksen vaikuttaa työnsä sisältöön. (Deuze, 2005)

Neljäs elementti journalismissa on ajankohtaisuus ja välittömyys. Ne ovat asioita, jotka liittyvät vahvasti myös uutisen määritelmään. Eettisyys on viides elementti, jonka Deuze nimeää. Sillä hän viittaa journalistin kykyyn arvioida omaa toimintansa oikeutusta ja pätevyyttä. (Deuze, 2005)

3.1.2 Data

Tietotekniikan sanakirja määrittelee datan olevan 1) tietokoneen käyttämästä ja käsittelemästä tiedosta käytetty nimitys, 2) asian säännönmukainen esitys viestittävässä muodossa joko ihmisen muistissa tai keinotekoisessa tietovarastossa. (Jaakohuhta, 2011)

Cambridge Dictionary määrittelee datan informaatioksi, erityisesti faktoiksi tai numeroiksi, jota kerätään tutkittavaksi, arvioitavaksi ja päätöksenteon tueksi, tai elektronisessa muodossa olevaksi informaatioksi, jonka voi käsitellä ja tallentaa tietokoneella. (Cambridge Dictionary, haettu 26.1.2022)

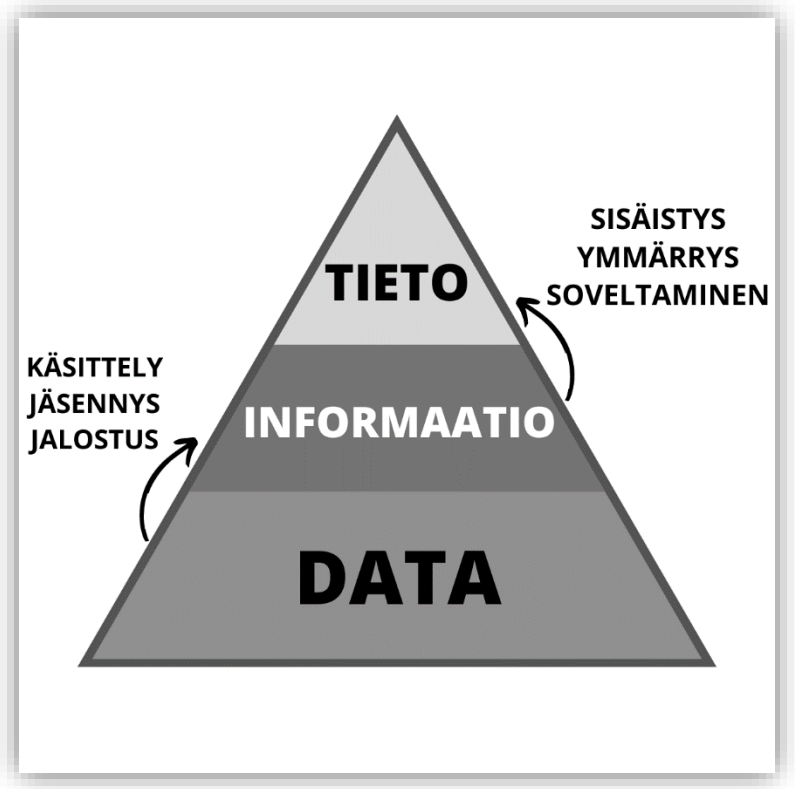
Merriam-Webster sanakirja antaa sanalle data kolme määritelmää. Ensimmäiset kaksi mukailevat edellä esitettyjä määritelmiä. Kolmannen mukaan data on anturilaitteen tai elimen tuottamaa tietoa, joka sisältää sekä hyödyllistä että epäolennaista tai tarpeetonta tietoa ja jota on käsiteltävä mielekkääksi. (Merriam-Webster, haettu 26.1.2022)

Rydenfelt, Haapanen, ja Lehtiniemi toteavat, että ”datalla tarkoitetaan yleisesti raaka-ainetta, jota syntyy, kun maailmaa abstrahoidaan mittareiden tai kategorioiden avulla. Data on ikään kuin rakennuspalikoita informaation ja tiedon tuottamiseksi.” He mainitsevat artikkelissaan Dataa näkyvissä (2021) myös hiljattain kirjallisuuteen ilmaantuneen käsitteen datafikaatio. Sillä tarkoitetaan laadullisten, ihmisiin ja sosiaaliseen maailmaan liittyvien ilmiöiden muuttamista mitattavaan muotoon. (Rydenfelt, Haapanen & Lehtiniemi, 2021)

Datafikaatio on informaatioteknologian ominaisuus, mutta se on myös hyödyllistä nähdä yhteiskunnallisena ilmiönä. Kaikki käyttämämme teknologia on ihmisen laatimaa ja siten sen tuottama data heijastaa teknologian kehitysprosessissa tehtyjen valintojen arvoja. (Rydenfelt, Haapanen & Lehtiniemi, 2021)

Dataa on siis monenlaista ja määritelmiä sille on useita. Yhteistä datan eri muodoille ja määritelmille on se, että sitä pidetään tiedon yksikkönä, joka kuvaa jotakin asiaa, ilmiötä tai tapahtumaa. Mittaushetken lämpötila, sanat ja niiden järjestys teoksessa Pikku prinssi sekä twiitti metatietoineen ovat kaikki omilla tavoillaan dataa.

Dataa voidaan pitää tiedon esiasteena. Se on jäsentymätöntä ja jalostettavissa informaatioksi. Data muodostaa pohjan pyramidille, jolla Uskali ja Kuutti havainnollistavat datan, informaation ja tiedon välistä suhdetta. Käsittely, jäsennys ja jalostaminen muuntavat datan informaatioksi. Tämä muodostaa pyramidin toisen tason. Pyramidin ylin, kolmas kerros on tieto. Tieto on sisäistettyä informaatiota, ymmärrystä ja tietotaitoa. (Uskali & Kuutti, 2016)



Kuva 1 Tiedon pyramidi kuvaa datan, informaation ja tiedon suhdetta (Uskali & Kuutti, 2016, 56)

3.1.3 Datajournalismi

Datajournalismia voidaan pitää journalismin alalajina tai journalistisena toimintamallina. Uskalin ja Kuutin teoksessa Datajournalismin työkäytännöt datajournalismi kuvataan työprosessina, jossa hyödynnetään tietotekniikan mahdollistamia uusia tapoja tuottaa entistä parempia journalistisia lopputuotteita. (Uskali, & Kuutti, 2016)

Jaakkola taas määrittelee datajournalismin olevan ”juttujen tuottamista viranomaisten avoimesti verkossa olevia tietoaineistoja yhdistelemällä, suodattamalla ja analysoimalla. Jutut voivat olla esimerkiksi visualisointeja tai interaktiivisia verkkosovelluksia, --.” Tässä määritelmässä mukaan tulevat verkko julkaisualustana ja avoin data. (Jaakkola, 2013)

Nieminen ehdottaa pro-gradussaan, että ”datajournalismi on journalistinen prosessi, jossa tuotetaan datan avulla uutta tietoa verkkoon tai verkko edellä ja johon liittyy usein visuaalisia ja interaktiivisia elementtejä.” Toisaalta hän myös toteaa, että datajournalistinen tuotos voi olla perinteinen tekstijuttu kuvien kera. (Nieminen, 2015)

Vaikka datajournalismi tuntuu tuoreelta ja internetiin sidotulta ilmiöltä, ovat journalistit, tutkijat ja mielipidevaikuttajat käyttäneet dataa hyväkseen jo ennen tietokoneiden aikaa. Sairaanhoitaja

Florence Nightingale käytti 1850-luvulla tilastodataa ansiokkaasti laatiessaan visualisoinnin brittisotilaiden kuolinsyistä Kriminsodassa. Piirakkakaavio-esitys osoitti selkeästi, että huono hygienia ja taudit harvensivat brittisotilaiden rivejä enemmän kuin taistelut. (Rogers, The Guardian, 2010. Uskali & Kuutti, 2016)

Visualisoinnissa datan kautta tuotettu informaatio järjestellään helpommin käsitettävään muotoon. Visualisointi voi myös auttaa itse journalistia hahmottamaan aineistosta teemoja ja vastauksia kaipaavia kysymyksiä. Pääasia visualisoinnissa on luoda esitys data-analyysin tuloksista ja edesauttaa valtavasta ja/tai vaikeaselkoisesta aineistosta löytyneiden havaintojen hahmottamista. (Uskali, & Kuutti, 2016)

3.1.4 Datajournalistinen juttu

Datajournalistisen jutun lähtökohtana voi olla itse aihe tai data-aineisto, jota tutkimalla jutun aihe muodostuu. Aiheen ollessa lähtökohtana journalistin on määriteltävä, millaista dataa aiheesta on olemassa, millaista dataa tarvitaan ja mistä se hankitaan. Tämän jälkeen on arvioitava hankitun datan laatua ja riittävyyttä. Data-aineiston ollessa juttuaiheen inspiraatio, on journalistin löydettävä datan rinnalle myös muita lähteitä. Pelkät numerot ja tilastot eivät riitä tekemään jutusta kiinnostavaa. (Uskali & Kuutti, 2016)

Uskali ja Kuutti nimeävät yhdeksi datajournalismin tavoitteista vaikuttamisen saatavilla olevan viranomaistiedon laajuuteen ja yksityiskohtaisuuteen. Avoimella datalla tarkoitetaan viranomaisen omasta aloitteesta julkisesti tarjottua data-aineistoa. Tietopyynnöin hankittu data voi kuitenkin olla journalistisesti kiinnostavampaa, koska avoimesti tarjottuun dataan viranomainen on valinnut sen mitä halutaan näyttää. Tietopyynnössään journalisti määrittelee itse tarvitsemansa datan ja muodon, jossa sen haluaa. Tietopyynnön saatuaan viranomaisen on otettava kantaa pyydetyn aineiston julkisuuteen ja luovutuksen tekniseen toteuttamiseen. Julkisuuslain (1999) mukaan viranomaisen on luovutettava aineistot ja tiedot pääsääntöisesti pyydetyllä tavalla. (Uskali & Kuutti, 2016)

Datajournalistiseen työprosessiin kuuluu olennaisesti käsillä olevan datan arvioiminen sen laadun ja luotettavuuden osalta. Data ei synny itsestään, vaan sitä kerätään ja tuotetaan. Sen taustalla on aina joku toimija, jolla on motiivinsa datan kartuttamiseen. On tärkeää tietää, kuka puhuu datan kautta, eli miksi se on kerätty ja mistä se on peräisin. Datan runsaus ja käyttökelpoisuus ei saa johtaa sen kritiikittömään käyttöön. (Uskali & Kuutti, 2016)

Jotta dataa voi käyttää jutussa lähteenä, on journalistin hahmotettava, miten kyseistä aineistoa on syytä ja mahdollista tarkastella. Datan käsittely ja analysointi vaatii journalistilta ymmärrystä

erilaisten analyysimenetelmien käyttökeloisuudesta suhteessa aineistoon. Journalistin on kyettävä esittämään aineistolleen mielekkäitä kysymyksiä. (Uskali & Kuutti, 2016)

3.1.5 Tietokoneavusteinen journalismi ja tietojenkäsittelyjournalismi

Journalismin tutkija Coddington hakee artikkelissaan määritelmiä data-, tietojenkäsittely- ja tietokoneavusteiselle journalismille. Hän toteaa kaikkien näiden olevan journalismia, joka hyödyntää määrällisiä menetelmiä. (Coddington, 2015)

1990-luvulla käsitteeksi noussut tietokoneavusteinen journalismi (eng. computer-assisted reporting, CAR) on vahvasti kytköksissä tilastotieteeseen ja yhteiskuntatieteeseen, sekä niiden tiedonkeruu- ja analyysimenetelmiin. Termin käyttö ei kuitenkaan ole enää mielekästä, edes itse käsitteen luojan Philip Meyerin mukaan, koska tietokoneesta on tullut olennainen osa kaikkea journalistista työtä. (Coddington, 2015)

Tietojenkäsittelyjournalismi (eng. computational journalism) on termeistä tuorein. Salmela on määritellyt tietojenkäsittelyjournalismin ”yhdistelmäksi algoritmeja, dataa ja tietoa yhteiskuntatieteistä.” Salmela ehdottaa maisterintutkielmassaan tietojenkäsittelyjournalismia pidettäväksi niin sanottuna sateenvarjokäsitteenä. (Salmela, 2021)

Salmela toteaa, että tietotekniikka on laajentanut journalistisen työn edellytyksiä ja mahdollisia lopputuloksia. Journalismin lopputulos voi olla uutisen sijaan esimerkiksi sovellus. Laskennallinen ajattelu ja ohjelmointitaitojen hankkiminen ovat Salmelan mukaan asioita, jotka usein liitetään tietojenkäsittelyjournalismiin. (Salmela, 2021)

Eroja näiden käsitteiden välillä Coddington etsii vertaamalla niiden suhdetta ammatillisuuteen, avoimuuteen sekä käsitykseen tiedosta ja yleisöstä. Tietokoneavusteisen journalismin hän näkee vahvasti tukeutuvan journalistisiin periaatteisiin ja ammatin perinteisiin. Data- ja tietojenkäsittelyjournalismissa sen sijaan työprosessi on usein moniammatillinen, sillä mukaan otetaan esimerkiksi ohjelmoijia ja graafikkoja. (Coddington, 2015)

Avoimuus on olennainen osa datajournalismin eetosta. Tietojenkäsittelyjournalismissa avoimuutta kaventaa algoritmien ja ohjelmistojen toiminnan selittämiseen liittyvät rajoitukset. Tietokoneavusteisen journalismin perinteessä datan tai työprosessin kuvauksen tuomista osaksi journalistista lopputuotetta ei pidetä olennaisena. (Coddington, 2015)

Kunkin journalismin tyypin suhteesta tietoon Coddington vertaa tiedonhankinnan ja analyysin menetelmiä, sekä käytetyn datan kokoa. Tietokoneavusteisen journalismin alkuaikoina dataa ei ollut samalla tavalla tarjolla kuin nykyään, joten sitä luotiin itse esimerkiksi kyselytutkimuksilla.

Hypoteesin testaaminen, otanta ja tilastoanalyysi olivat yleisesti käytettyjä menetelmiä. Data- ja tietojenkäsittelyjournalismissa dataa harvemmin tuotetaan itse, vaan painopiste on olemassa olevan datan löytämisessä, hankkimisessa ja analyysissa. Molemmat journalismin tyypit nojaavat valtavien data-aineistojen analyysista nousseisiin havaintoihin. (Coddington, 2015)

Tietojenkäsittely- ja datajournalismissa yleisöön suhtaudutaan aktiivisena toimijana. Yleisö voidaan valjastaa datan käsittelyyn ja tiedon tuotantoon. Molemmat journalismin tyypit tarjoavat yleensä datan yleisön tarkasteltavaksi, jotta siitä voisi tehdä havaintoja ja johtopäätöksiä itse. Tietokoneavusteissa journalismissa yleisö nähdään passiivisempana. Tiukasti kytköksissä tutkivaan journalismiin olevana se pyrkii tuottamaan journalistisen lopputuotteen, jonka informoimana yleisö voi vaatia muutosta nykytilaan. (Coddington, 2015)

3.1.6 Journalistin atk- ja datataidot

Journalismi on kulkenut koko historiansa käsi kädessä teknologian kanssa. Se on vaikuttanut journalismiin mahdollistamalla uusia julkaisukanavia sekä tapoja tavoittaa yleisöä. Teknologian kehityksen myötä journalismi on saanut uusia työkaluja sekä raaka-aineita jutuntekoprosessia varten. (Veglis & Pomportsis, 2014)

Journalistin on tänä päivänä pystyttävä käyttämään monia erilaisia työkaluja ja palveluja. Tarvittavat taidot voidaan jakaa neljään luokkaan: perustaidot, verkkojulkaisutaidot, Web 2.0 -taidot ja verkkolähetystaidot. Perustaidot sisältävät perinteisten tekstinkäsittely- ja esitysohjelmistojen, taulukkolaskentaohjelmistojen ja tietokantojen käyttöön liittyviä taitoja. (Veglis & Pomportsis, 2014)

Verkkojulkaisutaidoiksi Veglis ja Pomportsis katsovat julkaisujärjestelmäosaamisen ja perustaidot verkkosivuston HTML- ja CSS-koodaamisessa. Web 2.0 -taidot sisältävät sosiaalisen median ja blogin ylläpitämiseen tarvittavat taidot. Verkkolähetystaidot taas sisältävät esimerkiksi podcastin tai videocastin tekemiseen ja julkaisemiseen liittyvät taidot. (Veglis & Pomportsis, 2014)

Erityisesti datajournalismin lähteenä toimiva data asettaa myös osaamisvaatimuksia journalistille. Lewis, McAdams ja Stalph toteavat laskutaidon ja kuvailevien perustilastojen ymmärryksen olevan olennaista osaamista datan kanssa työskentelevälle journalistille. (Lewis, McAdams, & Stalph, 2020)

Uskali ja Kuutti kirjoittavat datamaisesta mielenlaadusta. Tällä he tarkoittavat journalistin kykyä hahmottaa tarvitsemansa datan sisällöllisiä ja muodollisia vaatimuksia sekä arvioida datan laatua ja riittävyyttä kriittisesti. Tämän lisäksi journalistin on osattava esittää kysymyksensä datalle tietokoneen välityksellä. Datamainen mielenlaatu helpottaa journalistia löytämään mitattavia ja määrällisiä piirteitä selvitettävästä aiheesta. (Uskali & Kuutti, 2016)

Datajournalismin eri työvaiheissa tarvitaan erilaisia taitoja. Tiedonhankintaa varten on tiedettävä, mistä ja millaista dataa löytyy ja miten se saadaan omaan käyttöön. Datan puhdistamiseen voidaan käyttää ohjelmistoa, mutta journalisti käyttää myös omaa älyään oikeinkirjoituksen, välimerkkien ja esitystavan standardoimiseksi tarkkojen laskelmien saamiseksi. Datan analysointi voi tapahtua taulukkolaskentaohjelmalla, ohjelmointikielen, tietokantaohjelmiston tai visualisoinnin avulla. Datan esittäminen journalistisessa lopputuotteessa voi tapahtua kuvaajilla ja interaktiivisin toiminnoin. (Lewis et al. 2020)

Lewis ja kumppanit listaavat kolme aihealuetta, joita datan käyttöön valmentavan journalismikoulutuksen tulisi sisältää: laskutaito ja tilastojen ymmärrys, datan esittäminen, sekä datan etiikka. Lewis ja kumppanit suosittelevat myös neljättä aihetta, laskennallista ajattelua, otettavaksi osaksi opetusohjelmia. (Lewis et al. 2020)

Laskutaito ja yleinen tilastojen ymmärrys auttavat journalistia tekemään johtopäätöksiä ja välttämään virheitä tulkinnoissa. Datan esittämiseen kuuluu kyky kirjoittaa selkeästi datasta ja siihen liittyvistä epävarmuustekijöistä, kuten virhemarginaalista. Datan esittämistaitoihin liittyy myös tarkkaavainen visualisoinnin tulkinta ja arviointi, jotta esimerkiksi eroja korostavat katkaistut akselit tai väestön kasvun ja inflaation huomiotta jättävät esitykset havaitaan. (Lewis et al. 2020)

Datan etiikkaan sisältyy ennen kaikkea datan läpinäkyvyys – mistä data on peräisin, miten sitä on käsitelty ja jaetaanko se yleisön kanssa? Tämän lisäksi dataan liittyy kysymyksiä yksityisyyden suojasta. Perinteisten viranomaisten keräämien tilastojen lisäksi dataa voidaan kerätä verkosta, esimerkiksi verkkosivuilta. Tällöin dataan saattaa päätyä sellaista tietoa, jota verkkosivun ylläpitäjä ei ole tarkoittanut kerättäväksi. (Uskali & Kuutti, 2016. Lewis et al. 2020)

Laskennallinen ajattelu (engl. computational thinking) on tietokoneen ajattelutavan matkimista. Käytännössä se tarkoittaa käsillä olevan ongelman purkamista yksittäisiksi toistettaviksi toiminnoiksi. Laskennallinen ajattelu voi edesauttaa journalistia kommunikoimaan ohjelmoijan kanssa. Lewis ja kumppanit rohkaisevat laskennallisen ajattelutaidon kehittämiseen ja ohjelmoinnin ymmärtämiseen, vaikka itse ohjelmointia he eivät katso olennaiseksi osaksi journalistin taitorepertuaaria. (Lewis et al. 2020)

3.2 Tietojenkäsittely

Tietojenkäsittelytiede (engl. computer science) on tieteen ala, joka tutkii tietokonetta ja sen käyttöä. Tietokone on laite, joka rakentuu fyysisen laitteiston (engl. hardware) ja ohjelmiston (engl. software) kokonaisuudesta ja jolla voidaan tallentaa, prosessoida, esittää tai siirtää tietoa. Tietojenkäsittelytiede sisältää tutkimuksen koskien algoritmeja ja tietorakenteita, tietokone- ja

tietoverkkosuunnittelua, data- ja informaatioprosessien mallinnusta, sekä tekoälyä. (Encyclopedia Britannica, haettu 14.2.2022)

Tietojenkäsittelytieteen taustat löytyvät matematiikan ja tekniikan aloilta. Tietojenkäsittelytieteen katsotaan olevan osa viiden erillisen, mutta toisiinsa liittyvien tieteenalojen joukkoa. Muut viidestä ovat tietokonetekniikka (engl. computer engineering), tietojärjestelmätiede (engl. information systems), tietotekniikka (engl. information technology) ja ohjelmistotuotanto (engl. software engineering). (Encyclopedia Britannica, haettu 14.2.2022)

3.2.1 Tekoäly

Tietotekniikan sanakirjassa englanninkielinen sanapari artificial intelligence käännetään tekoälyksi ja keinoälyksi. Teos antaa käsitteelle seuraavan selitteen ”[t]ietojenkäsittelytieteestä alkunsa saanut kielitiedettä, kognitiivista psykologiaa ja eräitä muita tieteenaloja sivuava monitieteinen tutkimussuunta ja tekniikan ala, jonka tarkoituksena on tutkia älylliseksi luonnehdittavaa käyttäytymistä ja toimintoja, kehittää näille teoreettisia malleja ja rakentaa niihin perustuvia tietokoneohjelmia ja tietojärjestelmiä.” (Jaakohuhta, 2011)

Ennen kuin tekoälystä voidaan mielekkäästi keskustella, on pohdittava, mitä itse älykkyys on. Valitettavasti älykkyydelle ei ole olemassa yhtä yleisesti hyväksyttyä määritelmää. Olennaisina ominaisuuksina älykkyydelle pidetään oppimis- ja ongelmanratkaisukykyä sekä kykyä joustavasti sopeutua erilaisiin tilanteisiin. Älykkyyttä katsotaan myös olevan montaa eri lajia, kuten esimerkiksi kielellinen, sosiaalinen tai matemaattislooginen älykkyys. (Suomen Koodikoulu, 2018)

Älykkyyden voidaan myös katsoa sisältävän erilaisia toimintoja, joilla tietoa käsitellään. Oppiminen tarkoittaa tiedon vastaanottamista ja tarkastelua. Päättely on tiedon käyttöä. Ymmärrys on tiedon käyttämisen kautta saavutettua lisätietoa. Arviointi taas koskee tiedon todenperäisyyden tutkimista. Suhteiden havaitseminen tarkoittaa eri tietojen vertaamista yhteyksien löytämiseksi. Merkityksen tarkastelu on tiedon ja tilanteen yhteyden etsimistä. Tiedon ja luulon erottaminen toisistaan on sen määrittelemistä, voidaanko tietoa pitää johdonmukaisena todisteena jollekin asialle tai ilmiölle. (Mueller & Massaron, 2018)

Tekoäly perustuu ajatukseen siitä, että ihmisen älykäs toiminta on mallinnettavissa. Tietokone tarvitsee tarkat toimintaohjeet kyetäkseen simuloimaan erilaisia ilmiöitä. Mallintaminen on jonkin asian olennaisten osatekijöiden ja näiden suhteiden esittämistä mallin avulla. Esimerkiksi kartta on mallinnus geologisesta ympäristöstä. (Suomen Koodikoulu, 2018)

Älykkyyden mallintamisessa tietokoneella edellä mainitut älykkyyden sisältämät toiminnot voivat tuottaa seuraavan laisen komentosarjan: 1. Aseta tavoite, 2. Arvioi olemassa olevan tiedon

hyödyllisyyttä suhteessa tavoitteeseen, 3. Kerää lisää tietoa tavoitteen saavuttamisen tueksi, 4. Käsittele uusi tieto vastaamaan olemassa olevan tiedon muotoa, 5. Havainnoi suhteita ja yhteyksiä uuden ja olemassa olevan tiedon välillä, 6. Arvioi onko tavoite saavutettu, 7. Muokkaa tavoitetta uuden tiedon valossa, 8. Toista prosessin vaiheita 2-7, kunnes tavoite saavutetaan tai tavoitteen saavuttaminen todetaan mahdottomaksi. (Mueller & Massaron, 2018)

Termi tekoäly sai alkunsa vuoden 1956 kesällä, jolloin Yhdysvalloissa Dartmouth Collegessa järjestettiin kuusiviikkoinen tutkimusprojekti. Suurimpia rajoituksia tekoälyn kehitykselle asettivat tuolloin laitteiden laskentateho ja vasta kasvava käsityksemme ihmisälykkyydestä ja ihmisaivojen toiminnasta. (Mueller & Massaron, 2018)

Tekoäly on siis tietokoneohjelma tai tietojärjestelmä, joka tavoittelee ihmismäistä älykkyyttä. Perinteisesti tekoäly on jaettu tavoitteensa perusteella neljään eri kategoriaan: ihmismäinen käytös, ihmismäinen ajattelu, rationaalinen käytös ja rationaalinen ajattelu. (Mueller & Massaron, 2018)

Ihmismäiseen käytökseen ja ajatteluun pyrkivä tekoäly yrittää matkia ihmistä. Ihmismäisesti käyttäytyvä tekoäly läpäisee Alan Turingin (1912–1954) kehittämän kokeen, jolloin koneen kanssa kommunikoiva ihminen ei huomaa juttukaverinsa olevan tietokone. Ihmismäisesti ajatteleva tekoäly kykenee älykkyyttä vaativiin tehtäviin, kuten auton ajamiseen. Rationaaliseen ajatteluun ja käytökseen pyrkivä tekoäly ammentaa logiikasta ja tekoälyn kyvystä muodostaa raamit toiminnalleen sille syötetyn datan perusteella. (Mueller & Massaron, 2018)

Tekoälyä voidaan tarkastella myös sen kykyjen laajuuden näkökulmasta. Niin kutsuttu ’heikko’ tekoäly kykenee suoriutumaan vain juuri siitä tehtävästä, johon se on suunniteltu. ’Vahva’ tekoäly taas pystyy sopeutumaan ja toimimaan erilaisissa tilanteissa. (Mueller & Massaron, 2018)

Vielä toistaiseksi vahvaa tekoälyä ei ole kehitetty, vaan se on kunnianhimoisimpien tutkijoiden visio. Vain ihminen kykenee yleiseen, joustavaan ja monikäyttöiseen älykkyyteen. Sen sijaan heikko tekoäly on ohittanut inhimillisen pystymisen rajat, esimerkiksi matemaattisloogisissa tehtävissä. (Suomen Koodikoulu, 2018)

3.2.2 Koneoppiminen

Yksi tekoälyn osa-alueista on koneoppiminen (eng. machine learning). Siinä on kyse siitä, että kone oppii aineiston ja siihen liittyvien toistuvien tapahtumien pohjalta, ilman erillistä ihmisen antamaa ohjeistusta. Koneoppimisen tavoitteena on automatisoida koneen tiedon tulkintaa ja laajentaa sen havainnointikykyä. (Neittaanmäki & Tuominen, 2019)

Koneoppiminen voidaan jakaa kolmeen eri luokkaan oppimisen tyylin mukaan: ohjattu oppiminen, ohjaamaton oppiminen ja vahvistettu oppiminen. Ohjatussa oppimisessa (engl. supervised learning) konetta opetetaan tunnetun aineiston perusteella ja tavoitteena on, että kone voi uutta vastaavanlaista dataa kohdatessaan luokitella sitä kuten alkuperäistä aineistoa. (Neittaanmäki & Tuominen, 2019)

Ohjaamattomassa oppimisessa (engl. unsupervised learning) kone tunnistaa opetusaineistosta riippuvuuksia, suhteita ja samankaltaisuuksia. Tämä oppimistyyli jäljittelee tapaa, jolla ihminen oppii. Vahvistetussa oppimisessa (engl. reinforcement learning) kone oppii ympäristöltä saamansa positiivisen tai negatiivisen palautteen perusteella. Kone tekee valintoja jo kokeiltujen palkittujen ja tuntemattomien vaihtoehtojen välillä, pyrkien kasvattamaan positiivisen palautteen määrää. (Neittaanmäki & Tuominen, 2019)

Oppiakseen kone tarvitsee dataa, jota tarkastella. Käytetyn data-aineiston laajuus ja laatu vaikuttavat olennaisesti koneoppimisen tulokseen. Koneoppiminen perustuu algoritmeihin ja niitä on kehitetty useita jokaiselle oppimistyylille. Koneoppimisen kautta kone voi oppia esimerkiksi tunnistamaan, luokittelemaan ja ennustamaan asioita. (Neittaanmäki & Tuominen, 2019)

3.2.3 Luonnollisen kielen käsittely

Monien tekoälyyn perustuvien sovellusten on kyettävä kommunikoimaan ihmisen kanssa. Luonnollisen kielen käsittely (eng. natural language processing, NLP) on tekoälytutkimuksen osa-alue, joka keskittyy luonnollisen tekstin ja puheen koneelliseen analysointiin ja tuottamiseen. (Suomen Koodikoulu, 2018)

’Luonnollinen kieli’ on ihmisten kommunikoinnissaan käyttämiä äänteitä eli puhetta ja merkkejä eli tekstiä. Keinotekoisia kieliä ovat esimerkiksi ohjelmointikielet tai Star Trek scifimaailman klingon. Luonnollinen kieli on tullut olemaan ihmisten vuorovaikutuksen seurauksena, kun taas keinotekoisen kielen sanasto ja kielioppi ovat tietoisesti luotuja. (Lehtomäki & Kukkaniemi, 2020)

Käsitteen viimeinen sana ’käsittely’ tai sen sijasta usein käytetty, englannin kielestä lainattu, prosessointi, viittaa siihen, että tietokone ja tekoäly ei vielä ymmärrä kaikkia kielen nyansseja ja merkityksiä. Luonnollisen kielen käsittelyn keinoin teksteistä voidaan nykyään luoda referaatteja, käännöksiä ja hakea tietoa. Ymmärrystä seuraavia johtopäätöksiä ei kone ole vielä pystynyt tuottamaan. (Liddy, 2001)

Vaikka koko alaa kutsutaan luonnollisen kielen käsittelyksi, siinä voidaan havaita kaksi erillistä painopistettä – kielen käsittely ja kielen tuotto (eng. language generation). Ensimmäisessä pyritään

analysoinnin myötä luonnollisen kielen mallintamiseen ja jälkimmäisessä reagoimaan syötteeseen eli esimerkiksi vastaamaan chat-botille esitettyyn kysymykseen. (Liddy, 2001)

Luonnollisessa kielessä on ja sen käyttämiseen liittyy monia tasoja, joilla sitä voidaan analysoida ja mallintaa. Fonetiikan tasolla tutkitaan äänteitä. Morfologian tasolla keskitytään sanoihin ja osasiin, joista ne muodostuvat. Leksikaalisella eli sanastollisella tasolla analysoidaan sanojen merkityksiä. Syntaksin tasolla tutkitaan lauseiden rakenteita ja niiden keskinäisiä suhteita. Semantiikan tasolla pyritään määrittämään lauseen merkitys tarkastelemalla siinä käytettyjen sanojen merkityksiä. Diskurssin tasolla analysoidaan kokonaisen tekstin merkitystä, joka muodostuu siinä esiintyvien sanojen ja lauseiden suhteista. Pragmatiikan tasolla kyse on luonnollisen kielen ja kontekstin suhteesta, sekä siitä miten asioita voidaan ilmaista niin sanotusti rivien välistä. (Liddy, 2001)

Lähestymistapoja luonnollisen kielen käsittelyyn on kolme, symboliset, tilastolliset ja konnektionistiset. Symboliset lähestymistavat suorittavat kielellisten ilmiöiden syväanalyysia. Ihmisen tietokoneelle syöttämällä lisätiedolla on olennainen rooli symbolisissa lähestymistavoissa. Tekstin luokittelu, tiedon haku, sanaston kartuttaminen ja moniselitteisyyden selvittäminen ovat tehtäviä, joissa käytetään symbolisia lähestymistapoja. (Liddy, 2001)

Tilastolliset lähestymistavat käyttävät erilaisia matemaattisia tekniikoita ja usein suuria tekstiaineistoja. Lähestymistavassa pyritään muodostamaan yleistettäviä malleja kielellisistä ilmiöistä, jotka perustuvat tietokoneen tekstiaineistoista tekemiin havaintoihin, ilman ihmisen syöttämää lisätietoa kielestä tai maailmasta. Tilastollisia lähestymistapoja on tyypillisesti käytetty puheentunnistuksessa, sanaston kartuttamisessa, jäsentämisessä, sanaluokittelussa, kieliopin sääntöjen tunnistamisessa. (Liddy, 2001)

Konnektionistiset lähestymistavat kehittävät yleistettyjä malleja kielellisten ilmiöiden esimerkeistä. Konnektionistiset lähestymistavat yhdistävät tilastollisen oppimisen erilaisiin lauseenrakenteiden merkitysten esitysteorioihin. Konnektionistisia lähestymistapoja voidaan käyttää sanan merkityksen yksiselitteisyyden määrittelyyn, kielen luomiseen, syntaktiseen jäsennykseen ja käännöksiin. (Liddy, 2001)

3.3 Aihemallinnus

Aihemallinnus (eng. topic modeling) on ohjaamaton koneoppimismenetelmä, joka juontaa juurensa luonnollisen kielen käsittelystä ja on lähestymistavaltaan tilastollinen. Se ei ole yksittäinen menetelmä, vaan kyseessä on joukko algoritmeja, joilla voidaan löytää useista dokumenteista koostuvista tekstiaineistoista piileviä aiheita. (Budiarto et al., 2021. Stray, 2016)

Aihemallinnus on verrattain nuori menetelmä. Se kehitettiin vuosituhannen vaihteessa ja ensimmäiset laaja-alaiseen käyttöön tarkoitetut tieteelliset artikkelit julkaistiin 2010-luvun alussa (Nelimarkka, 2019). Aihemallinnuksessa käytettävä aineisto esikäsitellään tavalla, joka yksinkertaistaa sitä ja menetelmä tarkastelee aineistoa tilastollisten todennäköisyyksien kautta. Aihemallinnus havaitsee aineistosta yhdessä esiintyvien sanojen joukkoja eli aiheita sekä laskee aineiston jokaiselle sanalle ja dokumentille todennäköisyyden kuulua johonkin havaituista aiheista. (Ylä-Anttila, Eranti & Kukkonen, 2018)

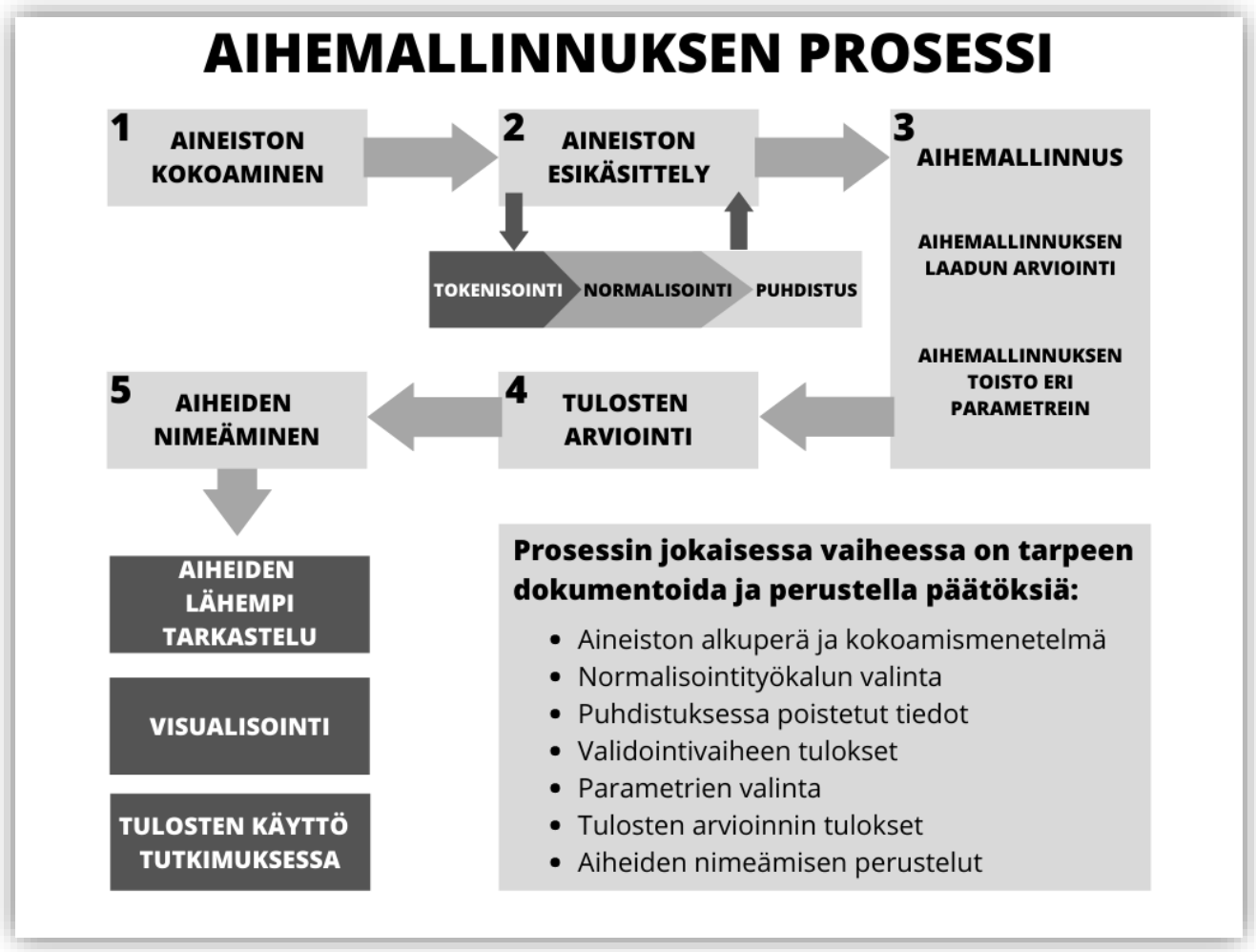
Aihemallinnus perustuu ajatukseen siitä, että aineistokokoelman eli korpuksen kukin tekstidokumentti sisältää aiheita (eng. topic). Näitä aiheita esiintyy dokumenteissa eri suhteissa ja dokumentissa esiintyvät sanat voidaan tilastollisen todennäköisyyden perusteella yhdistää tiettyyn aiheeseen. Aihemallinnus algoritmi käsittelee dokumentteja sanasäkkeinä (eng. bag-of-words) olettaen, ettei sanojen järjestys ole tärkeä, vaan vain niiden yhdessä esiintymisellä on väliä. (Blei, 2015, Makkonen & Loukasmäki, 2019)

Aihemallinnuksen hyöty syntyy siitä, että algoritmin aineistosta havaitsema piilevä aiherakenne muistuttaa lähiluvun keinoin löydettävää tekstin temaattista rakennetta. Aihemallinnus merkitsee jokaisen dokumentin tiettyyn aiheeseen kuuluvaksi ja siten automatisoi, käsin tehtynä aikaa sekä ihmisresursseja vievän, aineiston luokittelun ja teemoittelun. Näitä merkintöjä voidaan myöhemmin käyttää muun muassa tiedonhaussa, luokittelussa, järjestämisessä ja esimerkiksi korpustutkimuksessa. (Blei, 2015)

Aihemallinnus kehitettiin alun perin tietotekniikan sovelluksia ajatellen. Tilastollisen luonteensa vuoksi aihemallinnus on tapa havainnoida aineistoa ja voi kertoa siitä jotain sekä voi auttaa muodostamaan hypoteesin. Aihemallinnus on jo otettu käyttöön muun muassa lääketieteessä, politiikan tutkimuksessa, psykologiassa ja informaatiotutkimuksessa. Menetelmä on herättänyt kiinnostusta myös muilla aloilla, joissa tekstit ovat ensisijainen tutkimuksen kohde, kuten historia, sosiologia, oikeustiede ja kielitiede. (Blei, 2015)

3.3.1 Aihemallinnus käytännössä

Aihemallinnusalgoritmi laskee tilastollisen esiintymistodennäköisyyden teksteissä esiintyville sanoille ja luo esiintymistodennäköisyyksien perusteella listoja eli sanajoukkoja, joita kutsutaan menetelmän nimen mukaisesti aiheiksi. Aihemallinnus etsii aineistosta säännönmukaisuuksia, jotka voivat olla sellaisia piileviä rakenteita, joita lähiluku ei välttämättä paljasta. (Ylä-Anttila, Eranti & Kukkonen, 2018)



Kuva 2 Aihemallinnuksen prosessi

Aihemallinnuksen esittäminen prosessina tekee näkyväksi menetelmän käytön kokonaisuutena. Ensin kootaan aineisto ja sitten se esikäsitellään. Itse aihemallinnuksen yhteydessä on arvioitava mallinnuksen laatua ja mallinnusta toistetaan. Tämän jälkeen arvioidaan tuloksia ja voidaan nimetä löydetyt aiheet. Prosessin jokaisessa vaiheessa on pidettävä kirjaa eli dokumentoitava tehtyjä päätöksiä, sillä niillä on vaikutusta aihemallinnukseen.

Aineistoksi aihemallinnukseen kelpaa monenlainen data. Menetelmää on käytetty muun muassa geneettisen datan, kuvien ja sosiaalisen median sisältöjen tarkasteluun (Blei, 2012). Aihemallinnusta on käytetty myös esimerkiksi verkkosivujen, blogien, lehtiartikkelien, kirjojen ja tiedotteiden tutkimiseen. Aihemallinnuksessa käytetyn aineiston tai sen dokumenttien koolla ei ole ylärajaa. Erään aihemallinnuksen aineisto koostui reilusta 2500 kirjasta, toisen 77 miljoonasta twiitistä. (Asmussen & Møller, 2019)

Aineiston esikäsittely on tärkeä ja erittäin tarpeellinen vaihe aihemallinnuksen onnistumisen kannalta. Menetelmän sanasäkkiolettaman johdosta aineistossa esiintyvät sanat ovat ensimmäiseksi saneistettava eli tokenisoitava. Aineiston dokumentit puretaan listaksi, joka koostuu sen sisältämän tekstin sanoista. Kustakin tekstistä siis syntyy lista, jota aihemallinnusalgoritmi käsittelee sanasäkkinä. Seuraavaksi saneet (eng. token) on muutettava perusmuotoon. Tätä vaihetta kutsutaan normalisoinniksi. Esimerkiksi sanat ’koirat’, ’koiria’ ja ’koirille’ ovat johdettu perusmuodosta ’koira’. (Nelimarkka, 2019, Kukkaniemi & Lehtomäki 2020)

Sanojen normalisointi voidaan tehdä kahdella tavalla. Stemmaus (eng. stemming) tarkoittaa sanan katkaisemista kieliopillisten sääntöjen perusteella katkaistuun sanamuotoon. Lemmaus (eng. lemmatization) perustuu sanan perusmuodon etsimiseen sanakirjojen ja kielellisen analyysin kautta. Molempia toimenpiteitä varten on olemassa algoritmeja. *Natural Language Toolkit* (NLTK) on yleisesti käytetty stemmaustyökalu. Kielipankilta löytyy lemmaukseen ratkaisu suomen kielelle. (Nelimarkka, 2019. Kukkaniemi & Lehtomäki 2020)

Aineiston puhdistamisen pyrkimys on poistaa aineistosta sellainen data, joka saattaisi häiritä analyysiä. Tekstistä poistetaan välimerkit ja numerot. Teksti muutetaan pieneen kirjainkokoon. Yleisimmät sanat eli sulku- tai hukkasanat (eng. stopwords), jotka eivät välitä merkitystä, kuten konjunktiot ’ja’, ’tai’ ja ’myös’, sekä yleisimmät verbit, kuten ’olla’, poistetaan. (Nelimarkka, 2019. Makkonen & Loukasmäki, 2019)

Muitakin yleisimpiä ja harvinaisimpia sanoja voi olla aiheellista poistaa. Sanan yleisyys voi kertoa sen vähäisestä semanttisesta eli merkitystä välittävästä painoarvosta. Harvinaisuus myös voi olla viite sanan vähäisestä arvosta tekstin merkitykselle. Valitettavasti yleistä ohjeistusta ei ole olemassa, kuinka määritellä sanan yleisyys tai harvinaisuus. Nämä ovat päätöksiä, jotka aihemallinnuksen tekijän on dokumentoitava huolellisesti. (Nelimarkka, 2019. Asmussen & Møller, 2019. Särkiö, 2019)

Ennen aihemallinnusta valittava aiheiden määrä, jonka aihemallinnusohjelmisto tai -työkalu tuottaa. Aiheiden määrää kutsutaan usein parametriksi. Aihemallinnus tuottaa halutun määrän sanalistoja, joiden ’selkeyden’ pohjalta arvioidaan, onko valittu aihemäärä sopiva. Aihemallinnusta tulee toistaa eri parametrein, jotta löytyy se, joka tuottaa selkeimmät aiheet. Selkeydellä tarkoitetaan sitä, että sanalistan sanojen voidaan tulkita kuvastavan ja kuuluvan samaan aihealueeseen. (Nelimarkka, 2019. Makkonen & Loukasmäki, 2019)

Mikäli valittu parametri on liian pieni, aihemallinnus tuottaa sanalistoja, jonka sanoilla ei ole yhteistä nimittäjää. Liian suuri aiheiden määrä taas voi johtaa siihen, että aiheet pirstoutuvat ja useampi sanalista vaikuttaa liittyvän samaan asiaan. Aiheiden selkeyttä voidaan arvioida tilastollisilla

tunnusluvuilla, lasketuilla indikaattoreilla tai käyttäjäkokeilla. Kyseiset tunnusluvut ja indikaattorit toimivat myös aihemallinnuksen laadunarvioinnissa eli validoinnissa. (Toikka, 2021. Jacobi, Van Attenveldt & Welbers, 2015)

Aihemallinnustyökaluja löytyy niin korkeakoulujen kehittäminä tai yksittäisten henkilöiden luomina projekteina. Stanford Topic Modeling Toolbox (TMT) ja The Machine Learning for LanguagE Toolkit (MALLET) ovat yhdysvaltalaisissa yliopistoissa kehitettyjä ja suosittuja aihemallinnustyökaluja. GitHub-verkkopalvelusta löytyy useita avoimen lähdekoodin aihemallinnustyökaluja. (Barde & Bainwad, 2017)

Aihemallinnuksen tulos on algoritmin aineistosta havaitsema piilevä sanojen esiintymistodennäköisyyksiin perustuva rakenne. Tulokseksi syntyy valittu määrä sanalistoja eli aiheet sekä dokumenttien aihejakaumat ja aiheiden dokumenttijakaumat. Nämä ovat siis tilastollinen esitys aineiston sisältämästä datasta. (Nelimarkka, 2019)

Aihemallinnuksen tuloksena syntyviä sanalistoja tarkastellaan, arvioidaan ja ne pyritään nimeämään mielekkäästi. Aiheiden eli sanalistojen arviointi voi tapahtua esimerkiksi tarkastelemalla lähemmin sanalistan sanoihin yhteydessä olevien dokumenttien sisältöä. Sanalistat siis kertovat jotain aineiston teemoista. (Asmussen & Møller, 2019)

Dokumenttien aihejakauma eli se mitä aiheita kustakin aineiston dokumentista löytyy, kertoo aineiston dokumenttien sisällöllisestä moninaisuudesta. Aiheiden dokumenttijakauma eli mistä dokumenteista kukin aihe koostuu, kertoo itse aineiston moninaisuudesta. Mikäli aihemallinnettu aineisto sisältää esimerkiksi aika-, sijainti-, julkaisija- tai kirjoittajatietoja, voidaan näiden muuttujien avulla esimerkiksi visualisoida aihemallinnuksen tuloksia monin eri tavoin. (Puschmann & Scheffler, 2016)

3.3.2 Aihemallinnuksen anti

Aihemallinnus ei edellytä sitä, että sen käyttäjällä olisi minkäänlaista käsitystä aineistonsa sisällöstä. Menetelmää onkin yhteiskuntatieteissä usein käytetty käsitteiden ja teemojen tunnistamiseen aineistoista. Aihemallinnus ei selitä aineiston sisältöä, mutta voi antaa yleiskuvan aineiston sisältämistä teemoista. (Asmussen & Møller, 2019)

Aihemallinnusta voidaan käyttää aineiston koodaamisen apuna sisältöanalyysiä varten. Aihemallinnuksessa syntyvät sanalistat voivat toimia pohjana avainsana-analyysissä. Dokumenttien aihejakaumatiedot voivat auttaa kiinnostavien dokumenttiesimerkkien löytämistä aineistosta jatko tarkastelua varten. Aihemallinnus on tehokas ja edullinen tapa tarkastella ja saada

tietoa aineistosta, ennen perinteisiä ja vaativia sisältöanalyysin menetelmiä. (Jacobi, Van Attenveldt & Welbers, 2015)

Aihemallinnus on osoittautunut hyödylliseksi nimenomaan suurten aineistojen tarkastelussa ja tutkimuksessa. Asmussen & Møller (2019) listaavat artikkelissaan useita projekteja, joissa aihemallinnusta on käytetty niin twiittien, blogien, verkkosivujen, lehtiartikkelien, kirjojen, tiedotteiden ja puheiden tutkimiseen. Pienin listauksen aineisto sisälsi reilu 2500 tekstiä eli dokumenttia, kyseisessä tapauksessa kirjaa. Ja listauksen suurin aineisto koostui 77 miljoonasta twiitista.

Mikäli aihemallintaja tuntee aineistonsa sisällön läheisesti, on hyvin todennäköistä, että aihemallinnus ei paljasta aineistosta mitään uutta tai yllättävää. Aineiston mahdollisesti sisältämät aika-, sijainti-, julkaisija-, tai kirjoittajatiedot sen sijaan ovat sellaisia tekijöitä, joita hyödyntämällä voidaan tunnetustakin aineistosta tehdä mielenkiintoisia löytöjä. (Puschmann & Scheffler, 2016)

Aihemallinnuksen tuloksena syntyneitä aiheita eli sanalistoja esitetään usein visuaalisesti, koska se voi edesauttaa sanalistojen nimeämistä. Myös aineiston aihe- ja dokumenttijakaumista voidaan tuottaa visualisointeja, jotka paljastavat aineiston eri osien välisiä suhteita (de Waal & Mouton, 2013). Visualisoinnilla voidaan myös tehdä näkyväksi aineiston sisällön trendejä ja poikkeavuuksia (Günther & Quandt, 2015).

Eräs esimerkki (kuva 3) aihemallinnuksen tulosten visualisoinnista löytyy artikkelista *Gender Bias in the News: A Scalable Topic Modeling and Visualization Framework.* Tutkimuksessa aihemallinnettiin uutisartikkeleita kahden vuoden ajalta, 612 343 juttua. Jutuissa esiintyvien henkilöiden sukupuoli ja puheenvuorotyyppi pääteltiin olemassa olevalla Gender Gap Tracker -työkalulla. Aihemallinnus tunnisti aineistosta 15 aihetta, joihin kuuluvia uutisartikkeleita tarkasteltiin sukupuoli- ja puheenvuorotyyppitietoja vertaillen. (Rao & Taboada, 2021)

Kuva 3 Visualisointiesimerkki 1, Gender Bias in the News: A Scalable Topic Modeling and Visualization Framework -tutkimuksen kuvaaja (Rao & Taboada, 2021)

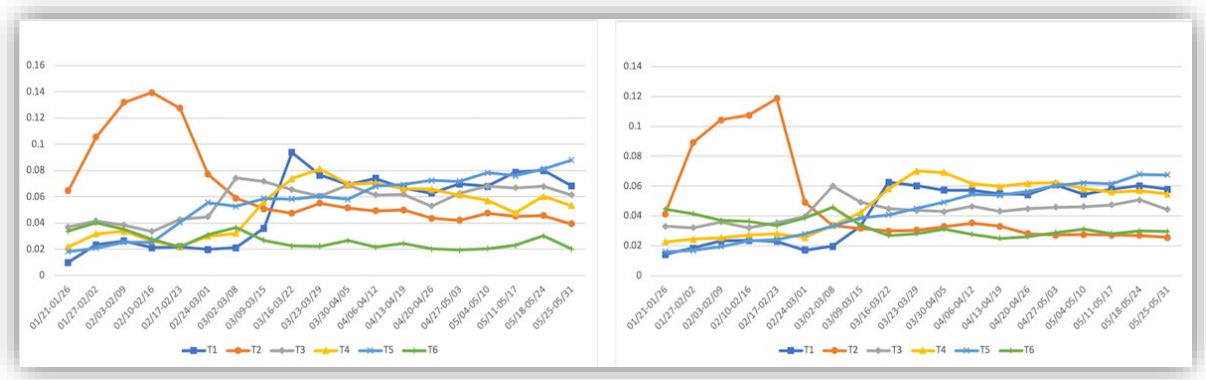
Tuloksista tunnistetuista aiheista yhdeksän visualisoitiin lämpökartta-kuvaajaksi, jossa on otettu huomioon myös alkuperäisaineiston dokumenttien julkaisuajankohta. Kuvaajassa punainen väri osoittaa, että jutussa esiintyy enemmän naishenkilöitä. Sininen taas merkitsee, että juttu sisältää enemmän mieshenkilöitä. Värin vahvuus kertoo, missä suhteessa jutuissa esiintyy kutakin sukupuolta. Vaalea sävy tarkoittaa, että jutussa esiintyy sekä miehiä että naisia, väri osoittaa kuitenkin kumpaa sukupuolta on enemmän. Tumma väri taas tarkoittaa, että jutuissa esiintyy pääosin vain yhden sukupuolen edustajia. (Rao & Taboada, 2021)

Tutkimuksen tarkoituksena oli selvittää median sukupuolittuneisuutta ja luoda visualisointimenetelmä, jonka kautta tulokset voitaisiin esittää kuvallisesti. Tutkimuksen tulokset kiteytyvät kuvaajassa, tehden tehokkaasti näkyväksi sen, kuinka tiettyjen aihealueiden uutisartikkeleissa esiintyy lähes yksinomaan tietyn sukupuolen edustajia. (Rao & Taboada, 2021)

Toinen esimerkki visualisoinnista (kuva 4) on *Tracking Covid-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis* -tutkimus. Tutkimuksessa aihemallinnettiin Pohjois-Amerikkalaisten Twitterissä käymää keskustelua koronapandemiaan liittyen. Tutkimuksen tarkoituksena oli jäljittää kansallisten pandemian hallintatoimien tehokkuutta. Aihemallinnuksessa twiiteistä tunnistettiin 20 aihetta, joista kuusi visualisoitiin twiittien määrä ja aika -akseleille. (Jang, Rempel, Roth, Carenini & Janjua, 2021)

Kuvaajassa ovat vierekkäin Yhdysvaltain (vasemmalla) ja Kanadan (oikealla) tulokset koskien kuutta aihemallinnuksen tunnistamaa aihetta. Kukin käyrä kuvaa kyseisestä aiheesta käydyn keskustelun vilkkautta. T1-aiheen twiitit koskevat sosiaalista ja fyysistä etäisyydenpitoa (social and physical distancing), T2-aiheen twiitit käsittelevät ilmailu- ja alueellisia matkustusrajoituksia ja

taudin leviämistä (air travel and regional border restrictions and outbreaks), T3-aiheen twiitit liittyvät käsienpesuun ja ennaltaehkäiseviin toimenpiteisiin (handwashing and preventive measures), T4-aiheen twiitit koskevat tarvetta pysytellä kotona ja pandemian vaikutuksia etulinjan työntekijöihin ja perheeseen (the need to stay home and impact of COVID-19 on essential workers and family), T5-aiheen twiitit liittyvät testien ja tautitapausten määrään (number of tests and cases) ja T6-aiheen twiitit käsittelevät maskeja ja kasvosuojaimia (masks and face coverings). (Jang et al., 2021)



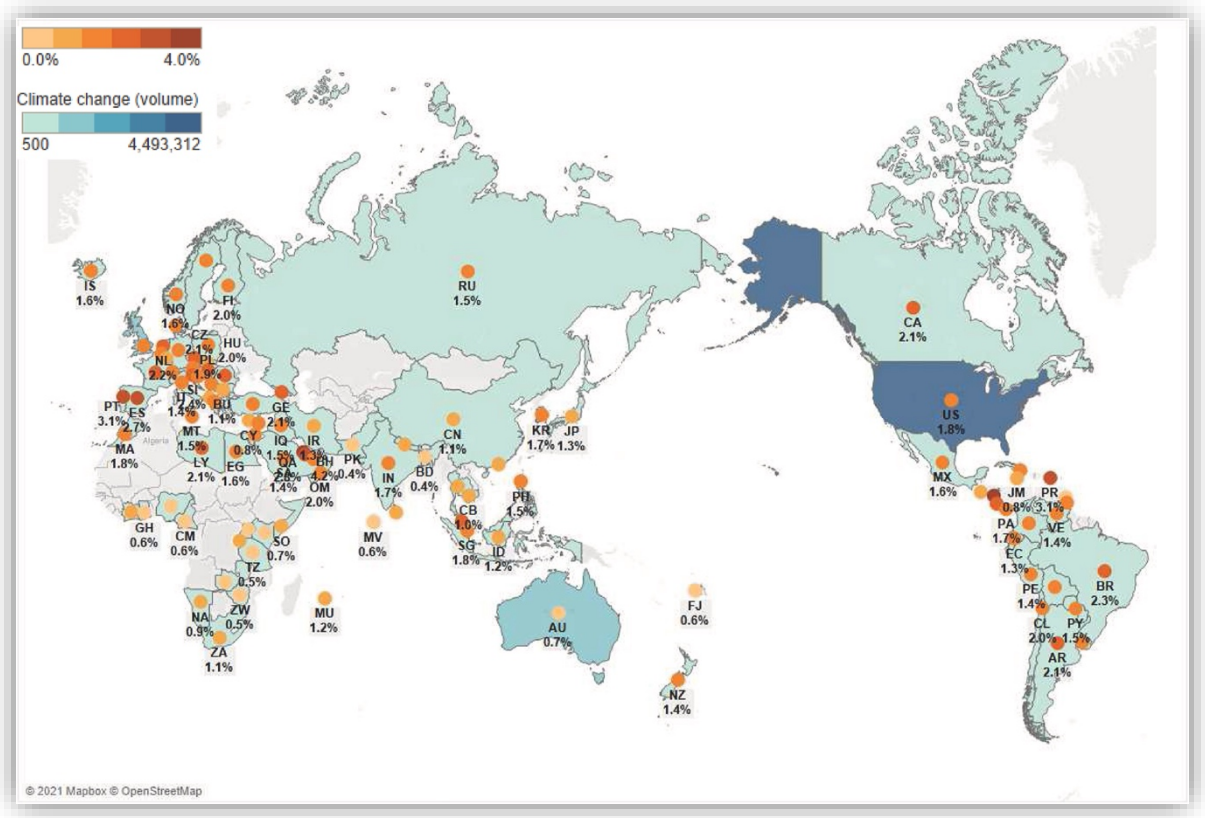
Kuva 4 Visualisointiesimerkki 2, Tracking Covid-19 Discourse on Twitter in North America, Yhdysvallat vasemmalla, Kanada oikealla (Jang, Rempel, Roth, Carenini & Janjua, 2021)

Kuvaajasta voimme nähdä, että Twitterissä käyty keskustelu sekä Yhdysvalloissa että Kanadassa oli tarkasteluajanjaksolla trendeiltään samansuuntaista (Jang et al., 2021). On kuitenkin huomattava, että vasemmanpuoleisen kuvaajan pystyakselin arvot nousevat korkeammaksi kuin oikeanpuoleisen kuvaajan. Tästä aiheutuu harhaanjohtava vaikutelma kuvaajien samankaltaisuudesta, vaikka todellisuudessa pandemiasta twiittaaminen Kanadassa ei ollut yhtä vilkasta kuin Yhdysvalloissa. Kahden kuvaajan esittäminen rinnakkain on tehokas esitystapa. Tätä visualisointia voidaan pitää myös varoittavana esimerkkinä kuvaajien mahdollisesta harhaanjohtavuudesta.

Kolmas esimerkki visualisoinnista (kuva 5) löytyy tutkimuksesta *Twitter conversations reveal issue salience of aviation in the broader context of climate change*. Tutkimuksessa tarkasteltiin loppuvuodesta 2019 julkaistuja ilmastonmuutosaiheisia twiitteja ja erityisesti ilmailuun ja lentoliikenteeseen liittyvien aiheiden esiintymistä kyseisissä twiiteissa. (Becken, Stantic, Chen & Connoly, 2021)

Twiittien sisältämien sijaintitietojen avulla tutkijat loivat karttavisualisoinnin. Kartan maat ovat väritetty sinisiksi ilmastonmuutokseen liittyvien twiittien määrän mukaan. Mitä tummempi sininen, sitä enemmän twiittejä. Oranssin ja punaisen sävyiset ympyrät kartan maiden päällä symboloivat

ilmailuaiheisten twiittien osuutta ilmastonmuutosaiheisista twiiteistä. Mitä tummempi ympyrä, sitä suurempi osuus maan twiiteista koski ilmailualaa. (Becken et al., 2021)



Kuva 5 Visualisointiesimerkki 3 Twitter conversations reveal issue salience of aviation in the broader context of climate change (Becken, Stantic, Chen & Connoly, 2021)

Kuvaaja tekee näkyväksi sen, kuinka osa ilmastonmuutokseen liittyvistä twiiteistä käsittelee nimenomaan ilmailualaa. Kuvaajasta voimme myös havaita, että esimerkiksi Yhdysvalloissa ja Australiassa twiitattiin ilmastonmuutoksesta enemmän kuin muualla maailmassa. Sen sijaan esimerkiksi Brasiliassa twiiteistä suurempi osuus koski ilmailua, kuin Yhdysvalloissa. Tutkimuksen aineisto kerättiin englanninkielisiä avainsanoja käyttäen, mikä aiheutti sen, että englanninkielisistä maista aineistoa kertyi enemmän. (Becken et al., 2021)

3.3.3 Aihemallinnuksen sudenkuopat

Aihemallinnus on jatkuvasti kehittyvä ala, jonka käytännöt edelleen muuttuvat ja mukautuvat. Siksi on tärkeää, että aihemallinnukseen ryhtyvä on perehtynyt menetelmään, sen kehitykseen ja siitä käytyyn keskusteluun. Menetelmää käyttäessään aihemallintajan on kyettävä varmistamaan työnsä pätevyys ja uskottavuus. Aihemallinnuksen uskottavuutta, eli reliabiliteettia voidaan arvioida tilastollisilla tunnusluvuilla, laskennallisilla mittareilla tai käyttäjäkokeilla. Pätevyydellä eli

validiteetilla tarkoitetaan sitä, että tutkittava ilmiö on kyetty kuvailemaan järkevästi ja että tutkimuksessa mitataan oikeaa asiaa. (Nelimarkka, 2019. Toikka, 2021)

Nelimarkka antaa kolme suositusta artikkelissaan *Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa – kriittisiä havaintoja* (2019). 1) Aihemallinnuksen parametrien (aiheiden määrä) valinnassa tulee käyttää laskennallisia mittareita. 2) Sanalistojen tulkinnassa ja jatkokäytössä tulee käyttää joko aiheesta olemassa olevaa kirjallisuutta, yhdistää tulkintaan muita analyysimenetelmiä tai kuvata aihemallinnuksen tuloksista tehtyjä havaintoja systemaattisesti. 3) Tulosten raportoinnin ja dokumentoinnin tulee olla selkeää ja läpinäkyvää. Nelimarkka peräänkuuluttaa refleksiivistä otetta, jossa aihemallintaja tuo esiin omaa taustaansa ja pohtii mahdollisia syitä tulkinnoilleen. (Nelimarkka, 2019)

4 Aihemallinnuksen mahdollisuudet journalismissa

Tässä opinnäytetyön osiossa siirrytään aihemallinnus-menetelmän kuvailusta pohtimaan aihemallinnuksen tarjoamia mahdollisuuksia journalismille. Luvussa pyritään vastaamaan opinnäytetyön tutkimuskysymykseen: miten aihemallinnusta voisi käyttää journalismin työkaluna? Samalla pyritään esimerkkien kautta ideoimaan ja pohtimaan aihemallinnuksen journalistista käyttöä.

Luvussa pyritään myös vastaamaan tutkimuskysymyksen alakysymyksiin: missä vaiheessa jutuntekoa aihemallinnusta voidaan käyttää, minkälaisia aineistoja journalisti voisi aihemallintaa ja minkälaisiin kysymyksiin journalisti voisi aihemallinnuksen keinoin etsiä vastauksia?

4.1 Aineistoon tutustuminen

Aihemallinnusta on tituleerattu uudeksi mikroskoopiksi ja etäluennan mahdollistavaksi työkaluksi (Nelimarkka 2019, Blei, 2012). Menetelmä osoittaakin hyödyllisyytensä, kun kyseessä on valtava aineisto, josta ei ehkä vielä tiedetä juurikaan. Aihemallinnuksella voidaan ottaa selvää, mistä kaikesta aineistossa on kyse. Menetelmällä voidaankin luoda eräänlainen kartta aineistosta ennen lähilukuun ryhtymistä.

Menetelmä jäsentää aineistoa tavalla, joka voi helpottaa halutun tai kiinnostavan tiedon ja aineiston osan äärelle löytämistä. Menetelmän avulla journalisti voi tarkastella aineistoaan kokonaisuuden, aihemallinnuksen tunnistamien aiheiden tai yksittäisten dokumenttien tasolla. Menetelmän avulla aineistoa voidaan järjestellä ja jäsentää.

Erityisesti tutkivassa journalismissa aihemallinnuksen käyttö voi olla hyödyllistä, etenkin silloin kun aineisto on digitaalisessa muodossa ja niin suuri, että lähilukien tutustumiseen kuluisi liikaa arvokasta aikaa tai tarvittaisiin koko joukko ihmisiä. Tietty, aihemallinnus ei korvaa aineistoon tutustumista lähiluvun keinoin, mutta se voi edesauttaa olennaisen äärelle pääsyä.

Aineistoon tutustuminen tapahtuu yleensä journalistisen jutuntekoprosessin alkuvaiheessa. Tällöin journalisti on jo kehittänyt juttuidean ja ryhtyy sitten etsimään tietoa juttuaan varten. Menetelmällä voi saada nopeasti alustavan käsityksen valtavan aineiston sisällöstä ja löytää sen mitä on etsimässä.

Aihemallinnusta voi myös käyttää juttuidean etsimiseen aineistosta, joka on herättänyt journalistin mielenkiinnon. Aihe juttuun voi löytyä pinoista taloudellisista tiedoksiantoja, oikeudenkäyntipöytäkirjoja, lainsäädäntökäsittelyasiakirjoja, virkamiesten kalentereita tai kokousmuistiinpanoja tai sääntelyviranomaisten sähköpostikeskusteluista. Valtavat aineistot

kuitenkin aiheuttavat haasteen suuruudessaan, koska journalistilla ei välttämättä ole mahdollisuuksia käyttää paljoa aikaa koko aineiston läpi käymiseen (Stray, 2016). Aihemallinnus voi siis nopeuttaa aineistoon tutustumista.

Aihemallinnus tunnistaa aineistosta piileviä rakenteita ja tämän ominaisuuden takia sen käyttö myös jo tunnettujen aineistojen kohdalla voi olla hyödyllistä. Menetelmän avulla voidaankin tunnistaa trendejä, yhteyksiä, samankaltaisuuksia tai eroavaisuuksia aineistoista.

Aihemallinnuksen tuloksina syntyvät aihe- ja dokumenttijakaumat kertovat aineiston sisällöllisestä moninaisuudesta. Nämä tulokset voivat osoittaa, että aineiston tietyt osat ovat jollain tavalla yhteydessä toisiinsa, vaikka yksittäisten dokumenttien teksteinä ne käsittelisivät eri asioita. Tulokset voivat myös osoittaa, että jokin sanastollinen elementti tai esimerkiksi fraasi esiintyy tietyissä yhteyksissä.

Aihemallinnuksella voidaan myös havainnoida ja vertailla aineiston määrällisiä ominaisuuksia, kuten dokumenttien tekstien sanamääriä ja sanastollista laajuutta. Sen avulla voidaan havaita, että jotain asiaa käsitellään aineiston teksteissä määrällisesti laajemmin tai monisanaisemmin kuin toista.

Aineiston aika-metatietojen kautta voidaan havaita ja tehdä näkyväksi, että sanasto ja ilmaisuntavat vaihtelevat ajan kuluessa. Yhdistämällä julkaisija- ja kirjoittajatiedot aihemallinnukseen, voidaan havaita ja tehdä näkyväksi, miten kullakin julkaisijalla tai kirjoittajalla on oma tyylinsä ja ominainen sanasto.

4.2 Millaisia aineistoja on jo aihemallinnettu?

Aihemallinnusta voidaan käyttää monenlaisten aineistojen tarkasteluun. Viranomaiset, poliittiset toimijat, erilaiset organisaatiot ja yritykset tuottavat nykypäivänä paljon sellaista dataa, jota journalisti voisi tarkastella aihemallinnuksen keinoin. Kokousmuistiot, päätösasiakirjat, puheenvuorot, oikeudenkäyntiasiakirjat, tiedotteet ja katsaukset ovat esimerkkejä mahdollisista aihemallinnettavista aineistoista.

Esimerkiksi Makkonen & Loukasmäki (2019) aihemallinsivat eduskunnan täysistuntopuheenvuoroja tarkastellen, miten hallitus- ja oppositiopuolueiden edustajien puheenvuorot eroavat toisistaan. Nelimarkka (2019) käytti aihemallinnusta suomalaisten puolueiden yleisohjelmien tarkasteluun.

Visualisointiesimerkissä 2, Rao ja Taboada aihemallinsivat uutisartikkeleita ja tarkastelivat juttujen sisältämien henkilöiden sukupuolijakaumaa (Rao & Taboada, 2021). Ylä-Anttila ja kumppanit

(2018) käsittelivät aihemallinnuksella New York Times- ja The Hindu-lehtien ilmastonmuutosaiheisia artikkeleja tarkastellen, miten kussakin julkaisussa aiheesta kirjoitetaan. (Ylä-Anttila, Eranti & Kukkonen, 2018)

Piikkilä (2020) aihemallinsi maisterintutkielmaansa varten suomalaisen lehdistön juttuja tekoälyä koskien. Hän havaitsi tutkimuksessaan, että tekoälystä puhuttaessa ääneen pääsivät useimmiten yritysmaailmaa edustavat, yhteiskunnallisesti korkeassa asemassa toimivat miehet, joilla on odotuksia hyötyä alan kehityksestä. (Piikkilä, 2020)

Marjanen ja kumppanit aihemallinsivat sanoma- ja aikakauslehtiä vuosilta 1854–1917, tarkastellen diskurssin muutoksia uskontoon liittyvässä keskustelussa. Tutkimuksessaan he visualisoivat aihemallinnuksen löytämiä aiheita ja niiden muutoksia aikajanalla. (Marjanen, Zosa, Hengchen, Pivovarova, & Tolonen, 2021)

Internetin erilaisilta keskustelufoorumeilta kerättyä aineistoa voi myös käyttää aihemallinnuksen aineistona. Ylisiurua keräsi aineistonsa Suomi24-verkkopalstalta ja tarkasteli erityisesti Terveys-palstalle kirjoitettuja viestejä. Aihemallinnuksella havaittiin palstan neljä suosituinta aihetta ja Ylisiurua toteaa, että menetelmällä voidaan löytää suosituimmat keskusteluaiheet sosiaalisen median aineistosta. (Ylisiurua, 2017)

Myös Särkiö aihemallinsi keskustelufoorumeilta kerättyä dataa tunnistaakseen trendejä ja ajankohtaisia puheenaiheita. Aineisto kerättiin vauva.fi ja Suomi24-keskustelufoorumeilta. Maisterintutkielmassaan hän kehitti aihemallinnukseen perustuvan trendintunnistusmenetelmän, jolla voi tunnistaa keskustelufoorumidatasta merkittäviä tapahtumia. (Särkiö, 2019)

Sosiaalisesta mediasta kerätyt aineistot ovat myös käyttökelpoisia aihemallinnuksen kohteita. Toikka (2021) aihemallinsi Facebookin Uusi energiapolitiikka -keskusteluryhmän sisällön. Visualisointiesimerkkinä (3) käytetyssä tutkimuksessa aihemallinnettiin twiittejä tarkastellen ilmailu- ja lentoliikenneaiheisten twiittien osuutta ilmastonmuutosaiheissa Twitter-keskustelussa (Becken et al., 2021).

Aihemallinnuksen aineistoksi käy mikä tahansa tekstiaineisto. Esimerkiksi Stranius (2019) käytti aihemallinnuksessaan suomenkielisten hittibiisien lyriikoita. Hän yhdisti aihemallinnukseensa Suomen Virallisen Listan tilaston tietoja ja saattoi sen myötä tarkastella aihemallinnuksen havaitsemien aiheiden suosiota ja suosion jakautumista vuodenaikoina ja eri vuosina.

Székely ja vom Brocke (2017) käyttivät aihemallinnuksensa aineistona yrityksien kestävän kehityksen raportteja. Työssään he keskittyivät tarkastelemaan yritysten toimenpiteiden kehitystä koskien kestävää kehitystä, raporttien käsittelemiä aiheita ja eri alojen eroja raportoinnissa.

4.3 Mitä voisi aihemallintaa?

Straniuksen (2019) esimerkin mukaisesti journalisti voisi aihemallintaa esimerkiksi kirjaston lainaustilastojen suosituimpia teoksia ja havainnoida, minkälaisia teemoja suomalaisten lainaamista ja lukemista kirjoista nousee ja ovatko aiheet muuttuneet vuosien kuluessa. Journalisti voisi jutussaan kysyä, onko lainatuimmilla teoksilla jotain yhteistä ja pohtia, mitä tämä kertoo suomalaisten lukumieltymyksistä.

Käsittelyyn voisi myös ottaa vaikkapa Finlandia-palkinnon saaneet ja ehdolla olleet teokset ja tarkastella niistä nousevia aiheita. Aihemallinnuksesta voisi saada selville, onko palkituilla teoksilla jotain yhteistä teemaa tai onko se muuttunut vuosien varrella? Entä palkintoa vaille jääneillä teoksilla? Yhdistämällä analyysiin tietoja ehdokkaat valitsevasta lautakunnasta ja tuomarista, voitaisiin tarkastella esimerkiksi, onko raadin jäsenten iällä, äidinkielellä tai sukupuolella ollut vaikutusta valintoihin.

Facebookissa on useita suosittuja ryhmiä, joiden sisältöä aihemallintamalla journalisti voi esimerkiksi paikallistaa mielenkiintoisia keskustelunaloituksia ja kommenttiketjuja, joista etsiä aihetta juttuun. Twitterissä tiettyä aihetunnistetta käyttävä keskustelu voisi aihemallinnettuna valottaa kyseistä puheenaihetta. Twitterillä on erityinen palvelu, jossa esimerkiksi tutkijat ja journalistit voivat tehdä datapyyntöjä tai kerätä dataa reaaliajassa (Toivanen, Huhtamäki, Valaskivi & Tikka, 2020).

Mielenkiintoista voisi olla myös aihemallintaa paljon luettujen ja runsaasti keskustelua kirvoittaneiden uutisartikkelien kommentteja. Tällainen aihemallinnus voisi valottaa artikkelista ja sen aiheesta käydyn keskustelun piirteitä. Mikäli kommentoijilta on vaadittu palveluun kirjautumista, voidaan heistä löytyvää metatietoa käyttää hyväksi, esimerkiksi tarkastelemalla kommentoinnin maantieteellistä tai ajallista aktiivisuutta tai selvittämällä miten eri ikäisten ihmisten kommentit eroavat toisistaan.

Marjasen ja kumppanien esimerkin mukaan lehdistön omat arkistot voivat olla aarreaitta. Esimerkiksi lukijoiden mielipidekirjoitusten tai vain tietyn palstan juttujen aihemallintaminen ja aikajanalla tarkastelu voisi olla mielekästä. Millaisia aiheita löytyy, jos aihemallinnettaisiin lehden suosituimmat jutut? Mitä aiheita tuottaa vähiten lukijoita kiinnostaneiden juttujen aihemallinnus?

Puolueiden ja yksittäisten poliitikkojen ulostulojen aihemallintaminen voisi avata poliittista liikehdintää. Aihemallinnuksella voisi tarkastella, miten puolueen tai poliitikon viestintä eroaa hallituskaudella oppositiossa oloon. Miten tietyn puolueen kuntapoliitikkojen kannanotot eroavat eduskunnassa toimivien poliitikkojen kannanotoista? Yksittäisen puolueen tai poliitikon viestintää voisi tarkastella myös suhteessa kannatuslukuihin tai uran kehitykseen.

Yksi mahdollisuus voisi olla kansalaisaloite.fi-verkkopalveluun julkaistujen kansalaisaloitteiden aihemallintaminen. Kaikkien aloitteiden aihemallintamisella voitaisiin tarkastella, mitä kansalaisaloitteet ovat käsitelleet. Yhdistämällä analyysiin kansalaisaloitteiden kannatusluvut, voisi saada selville miten suositut ja vähemmän suositut aloitteet eroavat toisistaan. Analyysi saattaisi paljastaa, mitä yhteistä on eduskunnan käsittelyyn päätymättömillä aloitteilla ja miten ne ehkä eroavat niistä aloitteista, jotka saivat vaaditun määrän kannattajia.

Kansalaisaloite.fi -verkkopalvelu on osa oikeusministeriön demokratia.fi verkkopalvelua. Kyseisestä palvelusta löytyy muun muassa kuntalaisaloite.fi, otakantaa.fi ja lausuntopalvelu.fi. Myös näiltä sivustoilta löytyviä aineistoja voisi olla mielenkiintoista aihemallintaa. Aineistoksi sivustoilta voisi kerätä aloitteita, lausuntoja ja sivuille kirjoitettuja kommentteja.

Myös yritykset tuottavat aineistoa, jota aihemallintaa. Székely ja vom Brocke (2017) aihemallinsivat yritysten kestävän kehityksen raportteja – mitä aiheita suomalaisten yritysten vastaavista julkaisuista nousisi aihemallinnuksessa?

5 Pohdinta

Aihemallinnuksella on käyttökohteensa mielestäni myös journalismissa. Menetelmänä aihemallinnus on vaativa, mutta huolellisesti tehtynä, se voi paljastaa aineistosta trendejä tai osoittaa journalistin mielenkiintoisen dokumentin ja sen sisältävän tarinan äärelle. Aihemallinnus on ennen kaikkea työkalu, jolla journalisti voi tehdä havaintoja aineiston ominaisuuksista ja jäsentää sitä.

Aihemallinnuksen tulosten tulkinta on sen sijaan mutkikkaampi asia. Tilastollisena menetelmänä aihemallinnuksen tulokset eli sanalistat ja aihe- ja dokumenttijakaumat ovat kvantitatiivista tietoa. Tuon tiedon tulkitseminen ja siitä yleistysten tai johtopäätösten teko vaatii tulosten yhdistämistä olemassa olevaan tutkimustietoon ja tämä on kyseiseen aiheeseen perehtyneen tutkijan työtä.

5.1 Miten aihemallinnusta voisi hyödyntää journalismissa?

Journalismissa pyritään toteamaan ja selittämään ilmiöitä. Journalismin pyrkimys on palastella ja esittää tietoa, jotta yleisö voisi tehdä siitä itse tulkintoja. Aihemallinnus on ennen kaikkea havainnointityökalu ja siten sopiva journalismin käyttöön.

Aihemallinnus voi nopeuttaa suureen aineistoon tutustumista. Menetelmä voi paljastaa aineistosta trendin tai poikkeavuuden, johon journalisti voi tarttua. Se voi auttaa löytämään neulan heinäsuovasta, niin sanotusti. Aihemallinnuksen tuottamien havaintojen perusteella journalisti voi löytää uusia kysymyksiä, joita esittää asiantuntijalle haastattelussa.

Erilaisia aihemallinnettavia aineistoja on valtavasti, mutta journalistisesti mielenkiintoisten kysymysten esittäminen aineistolle tai aihemallinnuksen tulosten pohjalta on tärkeintä. Aihemallinnuksesta ja sen tuloksista yksinään ei todennäköisesti saa laadittua juttua, vaan journalistin on luotava mielekäs kokonaisuus sen ympärille. Esimerkiksi yhteiskuntatieteilijät, kielen ja politiikan tutkijat sekä muut asiantuntijat voivat haastateltavina antaa juttuun tulkintoja ja pohdintoja, joilla kehystää aihemallinnuksen tuottamia havaintoja.

Aineistojen sisältämät metatiedot luovat myös hyödynnettävän ulottuvuuden aihemallinnuksen tuloksiin. Visualisoinnit ovat voimallinen tapa esittää tietoa, ja aihemallinnuksesta on olemassa useita esimerkkejä, joissa tuloksien visualisoinnissa tiivistyy mielenkiintoinen havainto. Datajournalismissa visualisoinnit ovat yleisessä käytössä ja myös siksi aihemallinnuksella voisi olla käyttöä alalla.

Aihemallinnus on lupaava menetelmä, mutta samaan aikaan hyvin vaativa. Aihemallinnuksen tekijän on kyettävä arvioimaan mallinnuksensa laatua, eikä sen käyttöön kannata ryhtyä

perehtymättä asiaan kunnolla. Huolellisesti toteutettuna aihemallinnus kuitenkin voi paljastaa aineistosta, jotain mitä lähiluvussa ihminen ei havaitse tai koe merkittäväksi.

5.2 Opinnäytetyön tavoitteet ja tulos

Asetin opinnäytetyöni tavoitteeksi vastata seuraaviin kysymyksiin: Miten aihemallinnusta voisi käyttää journalismin työkaluna? Mitä aihemallinnus on ja miten sitä käytetään? Missä vaiheessa jutuntekoa aihemallinnusta voidaan käyttää? Minkälaisia aineistoja journalisti voisi aihemallintaa? Minkälaisiin kysymyksiin journalisti voisi aihemallinnuksen keinoin etsiä vastauksia?

Työn sisältö etenee mielestäni mielekkäästi journalismista ja datasta, kohti aihemallinnusta, palaten lopulta jälleen journalismin pariin. Opinnäytetyön tietoperustassa tein kuvailevan katsauksen kirjallisuuteen ja esittelen olennaiset käsitteet. Journalismia, dataa ja datajournalismia käsittelevä tietoperustan osuus toimii pohjustuksena luvussa 4 tekemälleni ideoinnille aihemallinnuksen käytöstä journalismissa.

Aihemallinnuksen esittelyä edeltävissä kappaleissa tietojenkäsittelystä, tekoälystä ja luonnollisen kielenkäsittelystä pyrin johdattamaan lukijan loogisesti yleisempien käsitteiden kautta kohti erikoissanastoa. Tietojenkäsittelyn maailma on monimutkainen ja käsitteet kytkeytyvät toisiinsa mitä monituisimmilla tavoilla.

Vaikka opinnäytetyöni aihe, aihemallinnus, on kovin tekninen, koen onnistuneeni esittelemään menetelmän selkeästi ja ymmärrettävästi. Esittelen menetelmää teorian ja käytännön kautta sekä havainnollistan esimerkein. Uskon tämän opinnäytetyön helpottavan aihemallinnuksesta kiinnostuneen perehtymistä korkeammilla koulutusasteilla tehtyihin opinnäytteisiin ja tutkimuksiin.

Pohdin aihemallinnuksen mahdollisuuksia journalismissa ja ideoin mahdollisia käyttökohteita. Aihemallinnus pääsee oikeuksiinsa suurten digitaalisten aineistojen käsittelyssä ja menetelmä voi tuottaa monenlaisia havaintoja. Mainitsen luvussa 4 esimerkkejä, siitä minkälaisia aineistoja aihemallinnuksella on käsitelty, antaakseni lukijalle käsityksen siitä, mitä on jo tehty. Lopuksi pohdin aihemallinnuksen käyttöä journalismissa esimerkkejä hyödyntäen ja ideoin mahdollisia käyttöaiheita ja potentiaalisia aineistoja.

Koen vastanneeni alussa esitettyihin tutkimuskysymyksiin, mutta tunnustan opinnäytetyöprojektini edetessä usein tajunneeni, kuinka vähän oikeasti aiheestani ymmärrän. Aihemallinnus on suosittu sekä monella tutkimuksen alalla käytetty menetelmä ja sen käyttö vaatii erityistä perehtyneisyyttä. En voi hyvällä omalla tunnolla suositella aihemallinnukseen ryhtymistä vain tämän opinnäytetyön pohjalta.

5.3 Opinnäytetyöprosessi ja oppiminen

Opinnäytetyöni aiheenvalintaa ohjasi ennen kaikkea kiinnostus datajournalismiin ja sen saloihin. Kiinnostuin aihemallinnuksesta menetelmänä nimenomaan sen takia, että sillä voidaan käsitellä valtavia tekstiaineistoja. Opintojen aikana datajournalismiin tutustuttiin lähinnä numerodatan kautta, mikä on vain yksi monista lähestymiskulmista dataa kohtaan.

Aihemallinnuksen valikoiduttua aiheekseni, ryhdyin tutustua tietojenkäsittelyn maailmaan. Tietojenkäsittely ja tietotekniikka ovat nykyään osa jokapäiväistä elämäämme kodinkoneiden ja viestintälaitteiden muodossa ja minusta tuntuu, ettemme ymmärrä kyseisten laitteiden toimintaa tarpeeksi. Katsaukseni tekniikan maailmaan osoitti, että alan kieli ja ilmaisutapa voisi olla monin tavoin saavutettavampaa. Toivon selittäneeni opinnäytetyöni aihetta ymmärrettävästi ja selkeästi, ja siten, että esimerkiksi journalismin opiskelija voisi ymmärtää, mistä on kyse.

Ryhdyin innokkaasti tiedonhankintaan ja putosin kuvainnolliseen kaninkoloon. Vietin noin puolet opinnäytetyöni tekemiseen kuluneesta ajasta siellä. Kulunut aika ei kuitenkaan mennyt hukkaan, koska harras perehtyminen lähteisiin auttoi kiteyttämään aiheeni ja löytämään sopivimmat esimerkit pohdintojeni tueksi.

Kirjoittamisvaiheen käynnistäminen vaati järjestelmällisyyttä ja uudenlaista itsekuria. Sain luotua itselleni työskentelyä tukevan rutiinin, jonka ylläpitoa helpottivat opinnäytetyöohjaajan kanssa sovitut välideadlinet.

Lähdin työstämään opinnäytetyötä otsikolla Aihemallinnus journalismin työkaluna: miksi, miten ja mitä odottaa? Otsikko palveli myös tutkimuskysymyksen roolissa pitkälle prosessissa. En voi suositella kenellekään tutkimuskysymysten sivuuttamista opinnäytetyöprosessin loppupuolelle. Kysymykset ovat työtä ja sen muodostumista ohjaavia apuvälineitä ja niiden tärkeyttä ei pidä vähätellä. Lopullinen otsikko muodostui vasta loppumetreillä.

Sen lisäksi, että olen oppinut opinnäytetyöni aiheesta valtavasti, olen oppinut itsestäni, työskentelytavoistani ja siitä mikä minua motivoi. Opin prosessin aikana sen, että tekeminen voittaa tuskailun ja että kunnianhimoisimmatkin haaveet saavutetaan askel kerrallaan etenemällä.

Lähteet

Asmussen, C.B., Møller, C. 2019. Smart literature review: a practical topic modelling approach to exploratory literature review. J Big Data 6, 93. https://doi.org/10.1186/s40537-019-0255-7.

Barde, B. V. & Bainwad, A. M. 2017. An overview of topic modeling methods and tools. International Conference on Intelligent Computing and Control Systems (ICICCS), 2017, pp. 745-750, doi: 10.1109/ICCONS.2017.8250563.

Becken, S., Stantic, B., Chen, J. & Connoly, R., 2021. Twitter conversations reveal issue salience of aviation in the broader context of climate change. Journal of Air Transport Management, volume 98, 2022, 102157, ISSN 0969-6997, https://doi.org/10.1016/j.jairtraman.2021.102157.

Blei, D. 2015. 'Probabilistic Topic Models' (2012) Communications of the ACM, 55(4), pp. 77–84. doi: 10.1145/2133806.2133826.

Budiarto, A., Rahutomo, R., Putra, H., Cenggoro, T., Kacamarga, M. & Pardamean, B. 2021. Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering. Procedia Computer Science. volume 179, 40-46, https://doi.org/10.1016/j.procs.2020.12.007.

Cambridge Dictionary, hakutermi data. https://dictionary.cambridge.org/dictionary/english/data haettu 26.1.2022.

Coddington, M. 2015 Clarifying Journalism's Quantitative Turn, Digital Journalism, 3:3, 331-348, DOI: 10.1080/21670811.2014.976400.

de Waal, A. & Mouton, F. 2013. Topic modelling in the information warfare domain. International Conference on Adaptive Science and Technology, pp. 1-7, doi: 10.1109/ICASTech.2013.6707492.

Deuze, M. 2005. What is journalism? Professional Identity and Ideology of Journalists Reconsidered. Journalism. 6. 442-464. 10.1177/1464884905056815.

Encyclopedia Britannica, hakutermi computer science. https://www.britannica.com/science/computer-science. haettu 14.2.2022.

Günther, E. & Quandt, T. 2016 Word Counts and Topic Models. Digital Journalism, 4:1, 75-88, DOI: 10.1080/21670811.2015.1093270.

Jaakkola, M. 2013. Hyvä journalismi: Käytännön opas kirjoittajalle. [Helsinki]: Kansanvalistusseura.

Jaakohuhta, H. 2011. Tietotekniikan sanakirja: suomi – englanti – suomi. [Vaajakoski]: Readme.fi.

Jacobi, C., van Atteveldt, W. & Welbers, K. 2016. Quantitative analysis of large amounts of journalistic texts using topic modelling, Digital Journalism, 4:1, 89-106, DOI: 10.1080/21670811.2015.1093271.

Jang, H., Rempel, E., Roth, D., Carenini, G. & Janjua N. 2021. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. J Med Internet Res 2021;23(2):e25431. URL: https://www.jmir.org/2021/2/e25431. DOI: 10.2196/25431.

Kherwa, P. & Bansal, P. 2019. Topic Modeling: A Comprehensive Review. EAI Endorsed Transactions on Scalable Information Systems, 7(24), EAI, SIS. doi: 10.4108/eai.13-7-2018.159623.

Kielitoimisto, hakutermi data. https://www.kielitoimistonsanakirja.fi/#/data?searchMode=all .

Lehtomäki, E. & Kukkaniemi, R. 2020. Luonnollisen suomen kielen ymmärtäminen koneellisesti. [Jyväskylä]: Jyväskylän yliopisto.

Lewis, N. P., McAdams, M. & Stalph, F. 2020. Data Journalism. Journalism & Mass Communication Educator, 75(1), pp. 16-21. doi:10.1177/1077695820904971.

Liddy, E.D. 2001. Natural Language Processing. In Encyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc.

Makkonen, K. ja Loukasmäki, P. (2019) Eduskunnan täysistunnon puheenaiheet 1999–2014: Miten käsitellä LDA-aihemalleja? Politiikka, 61(2), ss. 127–159. https://journal.fi/politiikka/article/view/77163.

Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L. & Tolonen, M. 2021. Topic Modelling Discourse Dynamics in Historical Newspapers. in Reinsone, S., Skadiņa, I., Baklāne, A. & Daugavietis, J. (eds), Digital Humanities in the Nordic Countries 2020: Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020). CEUR Workshop Proceedings, no. 2865, CEUR-WS.org, Aachen, pp. 63-77, Digital Humanities in the Nordic Countries, [Online event] 21/10/2020. https://dblp.org/rec/conf/dhn/MarjanenZHPT20.

Merriam-Webster, hakutermi data. https://www.merriam-webster.com/dictionary/data haettu 26.1.2022.

Mueller, JP. & Massaron, L. 2018. Artificial Intelligence for Dummies, John Wiley & Sons, Incorporated, Newark.

Neittaanmäki, P. & Tuominen, H. 2019. Tekoälyn perusteita ja sovelluksia. [Jyväskylä]: Jyväskylän yliopisto. http://urn.fi/URN:ISBN:978-951-39-7796-2.

Nelimarkka, M. 2019. Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. Politiikka, 61(1), ss. 6–33. Saatavissa: https://journal.fi/politiikka/article/view/79629.

Nieminen, A. 2015 Uuden ajan journalismia: Datajournalismin määritelmä, merkitys ja tila Suomessa. [Tampere]: Tampereen yliopisto.

Piikkilä, M. 2020. Tekoälyä koskeva julkinen keskustelu Suomessa vuosina 1994 – 2019. [Helsinki]: Helsingin yliopisto http://urn.fi/URN:NBN:fi:hulib-202004201854.

Puschmann, C. & Scheffler, T. 2016. Topic Modeling for Media and Communication Research: A Short Primer. HIIG Discussion Paper Series No. 2016-05, Available at SSRN: https://ssrn.com/abstract=2836478 or http://dx.doi.org/10.2139/ssrn.2836478.

Rao, P. & Taboada, M. 2021. Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework. Frontiers in Artificial Intelligence, vol. 4, https://www.frontiersin.org/article/10.3389/frai.2021.664737. DOI=10.3389/frai.2021.664737.

Rogers, S. 2010. The Guardian. Florence Nightingale, datajournalist: information has always been beautiful. https://www.theguardian.com/news/datablog/2010/aug/13/florence-nightingale-graphics.

Rydenfelt, H., Haapanen, L. ja Lehtiniemi, T. 2021. Dataa näkyvissä: Läpinäkyvyys algoritmien ja datan journalistisessa hyödyntämisessä", Media & viestintä, 44(2), ss. 1–22. doi: 10.23983/mv.109857.

Salmela, S. 2021. Automaatiota uutistoimistossa: tapaustutkimus STT:n toimittajien näkemyksistä tietojenkäsittelyjournalismista. [Jyväskylä]: Jyväskylän yliopisto.

Salminen, A. 2011. Mikä kirjallisuuskatsaus? Vaasan yliopisto, opetusjulkaisu.

Stranius, J. 2019 Mitä kansa kuuntelee? : Suomenkielisten hittibiisien aiheet vuosina 1996–2018. [Helsinki]: Helsingin yliopisto.

Stray, J. 2016. What do Journalists do with Documents? Field Notes for Natural Language Processing Researchers. https://api.semanticscholar.org/CorpusID:30829303.

Suomen Koodikoulu, 2019. Johdatus tekoälyyn.

Székely, N. & vom Brocke, J. 2017. What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique. PLoS ONE 12(4): e0174807. https://doi.org/10.1371/journal.pone.0174807.

Särkiö, I. 2019. Topic modelling of Finnish Internet discussion forums as a tool for trend identification and marketing applications. [Helsinki]: Aalto yliopisto. http://urn.fi/URN:NBN:fi:aalto-201903172292.

Toikka, A. 2021. Aihemallinnuksen ja klusterianalyysin yhdistäminen aineiston esikäsittelyn ja mallinnuksen valintojen tutkimiseksi. Informaatiotutkimus, 40(3), ss. 142–162. doi: 10.23978/inf.107879.

Uskali, T. & Kuutti, H. 2016 Datajournalismin työkäytännöt. Tampere: Vastapaino.

Veglis, A. and Pomportsis, A. 2014. Journalists in the Age of ICTs: Work Demands and Educational Needs. Journalism & Mass Communication Educator, 69(1), pp. 61–75. doi: 10.1177/1077695813513766.

Ylisiurua, M. 2017. Aihemallinnuksen mahdollisuudet sosiaalisen median aineistojen jäsentämisessä: terveyskeskustelu Suomi24-verkkopalstalla. Kulutustutkimus.Nyt: Kulutustutkimuksen seuran julkaisu. Vuosikerta. 11, Nro 2, Sivut 44-67. http://www.kulutustutkimus.net/wp-content/uploads/2017/11/Aihemallinnuksen-mahdollisuudet-sosiaalisen-median-aineistojen-j%C3%A4sent%C3%A4misess%C3%A4-terveyskeskustelu-Suomi24-verkkopalstalla.pdf.

Ylä-Anttila, T., Eranti, V. & Kukkonen, A. 2018. Aihemallinnuksesta kehitysmallinnukseen. Politiikka: Valtiotieteellisen yhdistyksen julkaisu, vol. 60, no. 2, pp. 148-156.