RAJAPINNOILLA-TEEMANUMERO

Aihemallinnuksen ja klusterianalyysin yhdistäminen aineiston esikäsittelyn ja mallinnuksen valintojen tutkimiseksi

Arho Toikka
Helsingin yliopisto
arho.toikka@helsinki.fi
https://orcid.org/0000-0003-1990-6008

Uusien tekstiaineistojen analyysimenetelmien käytön yhteiskuntatieteelliset käytännöt eivät ole vielä vakiintuneet. Yksi suosittu keskustelu- ja tekstiaineistojen mallinnustapa on aihemallinnus, jolla etsitään aineiston temaattista rakennetta sanojen yhteisesiintymisen avulla. Aihemallinnuksen tulokset vaihtelevat aineiston esikäsittelyn ja mallinnuksen parametrien myötä, ja työkalusta riippuen myös satunnaisesti. Tämä on yleensä tulkittu ongelmaksi, josta päästään eroon huolellisesti validoimalla ja valitsemalla yksi ”paras malli”. Sosiaalitieteilijän näkökulmasta mallinnuksen vaihtelut voivat kuitenkin olla myös erilaisia näkökulmia aineistoon tai vivahde-eroja, joita tulkitsemalla voidaan löytää aineiston ydin. Tässä artikkelissa käsitellään tutkimusprosessia, joka perustuu toistettuihin aihemallinnuksiin aineiston esivalmisteluja ja mallinnuksen parametreja vaihtelemalla. Kahden aiheen samankaltaisuus voidaan mitata ja lukuisista malleista tuotetut aiheet voidaan ryhmitellä klusterianalyysilla näiden samankaltaisuuksien avulla. Kun kaksi aihetta sijoittuu samaan ryhmään, voidaan niiden tulkita olennaisesti kuvaavan samaa aihetta, vaikka sanajakauma ei olekaan täysin sama. Nämä aiheiden ryhmät voidaan sitten nostaa analyysin keskiöön: jotkin aiheet löytyvät riippumatta alkuvalmisteluista, jotkin vain toisinaan ja jotkut aiheet löytyvät vain sattumalta ja jäävät yksin. Yhden mallin tulkinnan rinnalla ja sijasta voidaan tulkita näitä ryhmiä, ja samalla tehdä näkyväksi mallin reliabiliteettia ja tehtyjen valintojen vaikutusta tuloksiin.

Asiasanat: big data, tutkimusmenetelmät, triangulaatio (tutkimusmenetelmät), keskustelupalstat

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Pysyvä osoite: https://doi.org/10.23978/inf.107879

Johdanto

Digitaaliset aineistot, kuten sosiaalisen median aineistot tai uutissivustojen kommenttipalstat, ovat avanneet uusia mahdollisuuksia tuottaa yhteiskuntatieteellistä tietoa laskennallisilla menetelmillä. Koko tutkimusprosessi aineiston hankinnasta ja käsittelystä itse analyysiin ja tulosten raportointiin vaikuttaa (millä tahansa menetelmällä) siihen, minkälaista tietoa niillä tuotetaan. Laskennallisen menetelmien osalta yhteiskuntatieteilijät ovat aloittaneet näiden menetelmällisten valintojen merkityksen pohtimisen, ja tämä artikkeli jatkaa keskustelua. Artikkelissa esitetään aihemallinnuksen tutkimusprosessi, joka tekee analyysin läpinäkyvämmäksi ja ratkaisee eräitä menetelmään liittyviä ongelmia.

Aihemallinnus on suosittu keskustelu- ja tekstiaineistojen mallinnustapa, jossa useista dokumenteista koostuvasta tekstikorpuksesta etsitään teemoja sanojen esiintymistiheyksien avulla. Yhden aihemallin tuloksena on joukko aiheita. ”Aiheet” ovat sanojen todennäköisyysjakaumia. Malliperheen nimen – aihemallit – taustalla on intuitio siitä, että sanojen jakaumat muistuttavat ”aiheita”, siinä merkityksessä kuin dokumenttia lukeva ihminenkin tulkitsisi dokumenttien sisältöä (Blei, 2012b). Vaikkapa Science-lehteä selaava lukija voisi helposti huomata, että artikkeleissa esiintyy usein käsitteitä kuten {stars, astronomers, universe, galaxies, galaxy} ja tulkita tämän tähtitieteen aiheeksi – esimerkki on kaikki Science-lehden numerot aihemallintaneesta tutkimuksesta (Blei & Lafferty, 2007), jossa kuvattiin tieteen kehitystä.

Intuitiivisesti selkeät tulokset ovatkin saaneet yhteiskuntatieteilijät innokkaasti soveltamaan menetelmää. Yhden aihemallin tulkinnan keskiössä on pohdinta siitä, mikä on aiheen jakauman todennäköisimpiä sanoja yhdistävä teema. Aiheita validoidaan sisäisesti mallinnusprosessissa laskettavilla tilastollisilla tunnusluvuilla sekä ulkoisesti vertaamalla mallin tuloksia aiemmin tiedettyyn. Aihemallinnuksen tulokset vaihtelevat tekstimassan esikäsittelyn ja mallinnukseen liittyvien parametrien valinnan myötä, ja menetelmästä riippuen myös satunnaisesti – aihemallinnusmenetelmät perustuvat satunnaisesta alkuasetelmasta lähtevään paremman ratkaisun etsintään, joka lopetetaan, kun ratkaisua ei enää saada parannettua, ja kun ongelma on moniulotteinen, näin ei aina päädytä samaan lopputulokseen.

Vaihtelu on yleensä tulkittu ongelmaksi, josta päästään eroon huolellisesti validoimalla ja siten valitsemalla ”paras malli” (Wilkerson & Casas, 2017). Taustalla on siis ajatus siitä, että yksi malli tarjoaa parhaan, totuudenmukaisimman tiivistyksen aineistosta, ja voitaisiin määritellä jokin kriteeri, jolla tämä paremmuus mitataan. Kilpailevia mallin arvioinnin kriteerejä ja tapoja on kuitenkin monia, ja ne voivat nostaa parhaaksi eri mallit. Yhteiskunta-

tieteilijän näkökulmasta mallinnuksen vaihtelut voisivat kuitenkin olla myös erilaisia näkökulmia aineistoon tai vivahde-eroja, joita yhdessä tarkastelemalla voidaan löytää tulkinnallinen aineiston ydin.

Tässä artikkelissa esitellään tutkimusprosessia, joka perustuu toistettuihin aihemallinnuksiin vaihdellen aineiston esivalmisteluja ja mallinnuksen valintoja, kuten aiheiden määrää, ja toistettujen mallinnusten tulosten yhdistämiseen klusterianalyysilla[1]. Aiheiden samankaltaisuutta voidaan mitata – esimerkiksi laskemalla jokaiselle aiheparille niiden sanajakaumien kosinin samankaltaisuus, kuten tässä artikkelissa tehdään. Klusterianalyysissa näiden samankaltaisuuksien perusteella ryhmitellään aiheet. Riittävän samankaltaiset sanajakaumat laitetaan samaan ryhmään – vaikka sanajakauma jonkin verran vaihteleekin, laajemmassa tarkastelussa aiheet liittyvät kuitenkin samaan asiaan, ja voidaan tulkita saman aiheen kuvauksiksi.

Klusteroinnin tulokset paljastavat aineistosta jotain mitä ei yhdestä mallista voida nähdä: jokin teema nousee aiheeksi kaikissa malleissa, riippumatta siitä miten aineistoa valmistellaan tai mitä valintoja tehdään, joku toinen taas ainoastaan tietyillä esivalmisteluilla. Ensimmäiset voidaan määritellä aineiston ”ydinaiheiksi”, jälkimmäiset ”näkökulmiksi” aineistoon, kuten tämän artikkelin tutkimusprosessin kuvaus -luvussa tehdään. Jos teema esiintyy vain yhdessä toistetuista mallinnuksista, se ei todennäköisesti ole kovin mielenkiintoinen ja voidaan tulkita ”roska-aiheeksi”. Näin klusterointi antaa mahdollisuuden paitsi mallin arviointiin ja valintojen perusteluun, mutta toimii myös tuloksena itsessään: klusterien, niiden sanajakaumien ja suhteiden sosiaalitieteellinen tarkastelu avaa kolmannen tason yksittäisen aiheen ja alkuperäisten dokumenttien tarkastelun rinnalle.

Tutkimusotetta havainnollistetaan esimerkillä energiapoliittisesta kansalaiskeskustelusta Facebookin Uusi energiapolitiikka -ryhmässä. Aineisto kattaa keskustelut ryhmässä vuodesta 2014 vuoteen 2017, kaikkiaan yli 100 000 kommenttia ja lähes 7 000 keskustelua. Esimerkissä tarkastellaan kahden mallinnusvalinnan, aiheiden määrän ja sanaston perusmuotoistamisen vaikutusta aihemalleille. Artikkelin tutkimuskysymykset ovat siis a) miten sanaston esikäsittely perusmuotoistamalla vaikuttaa ymmärrykseen kansalaiskeskustelusta suomenkielisessä aineistossa? ja b) miten aihemallin aiheiden määrä valinta vaikuttaa ymmärrykseen kansalaiskeskustelusta? Englanninkielisessä sosiaalitieteellisessä kirjallisuudessa on päädytty siihen, että liiallinen käsittely saattaa heikentää mallinnuksen tuloksia (Schofield & Mimno, 2016),

1 Klusterianalyysiä kutsutaan usein suomeksi ryhmittelyanalyysiksi, mutta tällä käsitteellä on muitakin merkityksiä, joten yksiselitteisyyden vuoksi käytän klusterianalyysin käsitettä.

suomenkielisellä sitä on suositeltu (Nelimarkka, 2019), mutta suomenkielisellä aineistolla empiiristä vertailua ei ole aiemmin tehty.

Artikkeli etenee siten, että seuraavassa luvussa keskustellaan aihemallinnuksen yhteiskuntatieteellisistä tavoitteista ja haasteista, ja asetetaan tutkimusprosessi niiden kontekstiin. Yksityiskohtainen ja tekninen menetelmän kuvaus sekä ohjeita kiinnostuneille soveltajille on luvussa tutkimusprosessin kuvaus. Tämän jälkeen siirrytään esimerkin pariin. Johtopäätöksissä esitetään suosituksia aihemallintajille ja esitetystä tutkimusprosessista kiinnostuneille.

Aihemallinnuksen tutkimusstrategiat

Aihemallinnus on tilastollisten mallien perhe useista dokumenteista koostuvien tekstikorpuksien analysointiin. Mallit tarkastelevat sanojen esiintymistä dokumenttien sisällä: ne ovat nk. sanasäkki-malleja, eli ne eivät käsittele lainkaan lauseenjäseniä, -rakenteita tai muita kieliopillisia tekijöitä, ainoastaan sanojen esiintymistä dokumenteissa – ikään kuin teksti ei olisikaan teksti, vaan kuvainnollinen säkki, jossa tekstin sanat ovat, ja niitä voidaan sieltä satunnaisesti poimia. Näiden sanasäkkien avulla ne tuottavat kahdenlaisia todennäköisyysjakaumia: sanojen jakaumia ”aiheiden” sisällä ja ”aiheiden” jakaumia dokumenttien sisällä (Blei, 2012a). Menetelmiä ja variaatioita on useita: Latent Dirichlet Allocation (LDA), Correlated Topic Model (CTM) ja Structural Topic Model (STM) ovat ehkä laajimmin sovelletut aihemallinnuksen laskennalliset tavat. Vivahde-eroista huolimatta niiden tuottamat tulokset eli todennäköisyysjakaumat ovat samanlaisia. Tietojenkäsittelytieteen piirissä 2010-luvun alussa suosioon nousseet menetelmät herättivät nopeasti yhteiskuntatieteilijöiden ja humanistien kiinnostuksen: mitä nämä ”aiheet” oikeastaan ovat? Onko niille käyttöä lähiluentaan perustuvan laadullisen analyysin tukena tai jopa korvaajana (Purhonen & Toikka, 2016)? Minkälaisiin tutkimuskysymyksiin niillä voidaan vastata?

Pääkkönen ja Ylikoski (2020) analysoivat keskustelun lähestymistapoja ”aiherealismina”, jossa oletetaan aiheiden olevan sellaisinaan tulkittavissa yhteiskuntatieteellisesti relevanttien konseptien toteutumina, kuten kehyksinä, diskursseina tai narratiiveina, ja ”aiheinstrumentalismina”, jossa mallin pääasiallinen tavoite on organisoida hermeneuttista tulkintaa tai laadullista analyysia. Nämä lähestymistavat ovat ideaalityyppejä: esimerkiksi Ylä-Anttila ym. (2018) ehdottavat tutkimusotetta, jossa aihemallinnuksella löydetään kehysanalyysin kehyksiä, mutta vain suunnitelmalliseen aineiston esikäsittelyyn ja lähilukuun yhdistettynä – sanajakaumat siis edustivat kehyksiä, mutta

eivät aina, vaan ainoastaan osaavan tulkitsijan työn jälkeen. Tyypillinen aihemallinnustutkimus perustuukin siihen, että aiheita käytetään lähtökohtana, joille tulkitseva aihealueen tunteva analyytikko antaa merkitykset ja tulkinnalliset kehykset. Vaikka aihemallinnus on laskennallinen menetelmä, siinä on siis aina vahva rooli tulkitsijalle.

Ihmistulkintaan perustuvassa tutkimuksessa yleisesti haasteena on heikko läpinäkyvyys, useiden yhtä uskottavien tulkintojen mahdollisuus eli todistusaineiston heikko määrittävyys ja heikko skaalautuvuus (Pääkkönen & Ylikoski, 2020). Laskennalliset menetelmät ratkaisevat näistä viimeisen, mutta eivät aina kahta ensimmäistä: mallinnuksen yhteydessä ei aina raportoida kaikkia valintoja ja kaikkia tuloksia, ja niiden mahdollistamia erilaisia tulkintakehyksiä, vaan tiivistetään tulokset niin suppeiksi, että lukijan on mahdotonta saada kokonaiskuvaa siitä, mitä malli aineistosta kertoo.

Nimenomaan aihemallinnuksessa nämä haasteet muodostavat ainakin neljä mahdollista ongelmakohtaa. Nämä ovat 1) tapa raportoida vain osa aiheista, 2) tapa raportoida vain osa mallinnuksen valinnoista, 3) tapa käyttää epäsuorasti tutkimuksen mielenkiinnon kohteena olevia mittareita raportoinnin rajauksiin, ja 4) aihemallinnukseen liittyvä satunnaisuus.

Ensimmäinen aihemallinnuksen ongelmakohdista on se, että kaikki mallin tuottamat sanajakaumat eivät usein ole yhtä selkeitä tai tutkijalle mielenkiintoisia – ja valitettavan usein tapana on jättää tulkitsematta ja raportoimatta tällaiset aiheet kokonaan (Mimno ym., 2011), tai ainakin raportoida ne ”roska-aiheina” tai jargonina (Allen & Murdock, 2021). Kriteerejä sille, miksi yksi aihe on roskaa ja toinen mielenkiintoinen ei useinkaan anneta, vaan valinnat perustuvat laadulliseen tulkintaan, jota ei avata lukijalle.

Toinen ongelmakohta on se, että aineiston esikäsittely ja mallinnuksen valinnat vaikuttavat tutkimuksen tuloksiin. Denny & Spirling (2018) esittävät 7 erilaista esikäsittelyn valintaa välimerkkien poistosta harvinaisten käsitteiden poistoon. He käsittelevät valintoja yksinkertaisina kyllä/ei -valintoina, ja näin saadaan 128 erilaista tapaa tehdä aineiston esikäsittely. Käytännössä monet näistä valinnoista voidaan tehdä usealla tai mielivaltaisen monella tavalla numeerisia kriteerejä vaihdellen– esimerkiksi harvinaisten käsitteiden osalta voidaan poistaa kaikki sanat, jotka esiintyvät alle 0,5 % dokumenteista, alle 1 % dokumenteista, jne. Nämä 7 valintaa eivät myöskään kata kaikkia mahdollisia esikäsittelyn valintoja – esimerkiksi (Ylä-Anttila ym., 2021) poistivat lehtiaineistostaan kaiken muun paitsi poliittiset vaateet, ja tämäkin on yksi esikäsittelyn muoto. Kaikkien mahdollisten esikäsittelyjen yhdistelmien määrä on periaatteessa rajaton, ja realistisestikin vähintään tuhansissa. Näiden esikäsittelyjen lisäksi mallin parametreihin liittyy valintoja, joista tärkeimpänä

aiheiden määrä, joka tutkijan pitää asettaa. Tietoa siitä, mitä esivalmisteluja on kokeiltu ja millä lopulliseen valintaan päädytty ei useinkaan raportoida.

Nämä esikäsittelyn valinnat liittyvät kolmanteen ongelmakohtaan: mallinnuksen onnistumista mittaavaa mittatikkua ei usein ole olemassa – emme tiedä, mitkä korpuksen ”todelliset” aiheet ovat, vaan malli etsii latentteja, etukäteen tuntemattomia rakenteita, ja mittatikku onnistumiselle rakennetaan mallin ulkopuolella. Tilastollisilla tunnusluvuilla voidaan arvioida joko koko mallia esimerkiksi ennustusvoimaa arvioimalla (Wallach ym., 2009) tai yksittäisiä aiheita: (semanttinen) koherenssi (Mimno ym., 2011) mittaa aiheen sanojen esiintymistä samoissa dokumenteissa ja eksklusiivisuus sitä, ovatko todennäköiset sanat todennäköisiä nimenomaan tietyssä aiheessa (Roberts ym., 2016). Tuloksien tulkintaan voidaan rakentaa erilaisia käyttäjäkokeita, joissa voidaan laittaa joko aihemallinnusta tuntevia (Nelimarkka, 2019), sisältöasiantuntijoita (Chuang ym., 2013) tai maallikoita (Chang ym., 2009) pohtimaan aiheiden merkityksiä.

Mikään näistä ei kuitenkaan suoraan mittaa sitä, ovatko mallin aiheet ”aiheita” tutkijan tarkoittamassa mielessä, ja mittarit eivät välttämättä ole yhtä mieltä siitä, mikä malli olisi paras: Nelimarkka (2019) vertaa tilastollisia tunnuslukuja käyttäjäkokeessa ihmisten tekemiin tulkintoihin ja suosittelee niiden käyttöä ihmisten erilaisista tulkintakehikoista johtuvan heikon reliabiliteetin parantamiseksi, mutta toisaalta tiedetään, että näiden tunnuslukujen tulokset eivät välttämättä ole yhteydessä ihmisten tekemiin mallin tulkittavuuden tehtäviin (Chang ym., 2009) eivätkä ne välttämättä kaikki nosta samaa mallikandidaattia parhaaksi (Roberts ym., 2016). Toinen ja kolmas ongelmakohta yhdistyvät: kun aineiston käsittelyn tapoja on valtavasti ja kilpailevia mittatikkuja mallien arviointiin useita, parhaan tai oikeimman mallin löytämiseen perustuvat tutkimusstrategiat tulevat mahdottomiksi.

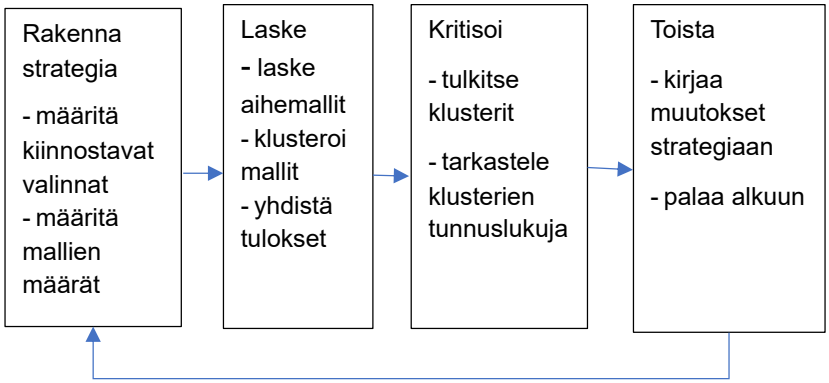
Aihemallinnuksessa neljäntenä ongelmakohtana on se, että tulokset eivät ole deterministisiä: mallinnus lähtee satunnaisesta alkuasetelmasta, siirtyy aina paremmin aineistoa kuvaavaan ratkaisuun, mutta ei aina löydä universaalisti parasta tulosta, edes mallinnuksessa sisäisesti käytettävillä tilastollisilla indikaattoreilla, vaan ”juuttuu” paikalliseen optimiin (Roberts ym., 2016). Tähän on kehitetty erilaisia ratkaisuja, kuten Roberts ym. (2016) kuvaama *spectral*-LDA -malli, joka poistaa tämän satunnaisuuden lisäämällä malliin erillisen alustusvaiheen. Malli vaatii kuitenkin ison aineiston, ja sen tuottamat tulokset eivät välttämättä ole suoraan vertailukelpoisia muiden aihemallien kanssa.

Tässä artikkelissa esitettävä tutkimusprosessi pyrkii ratkaisemaan näitä aihemallinnuksen ongelmakohtia. Roska-aiheille tulee kriteeri – jos toistetuista mallinnuksista vain yksi löytää aiheen eikä aihe siten klusteroidu, aihe

ei ole kovin mielenkiintoinen, ja siinä mielessä ”roskaa”. Esikäsittelyjen vaikutuksesta tulee eksplisiittinen osa tuloksia, kun tarkastellaan sitä, tuottaako erilainen esikäsittely siinä mielessä samat tulokset, että aiheet sijoittuvat kuitenkin samaan klusteriin. Useammasta kilpailevasta tilastollisesta tunnusluvusta ei tarvitse valita yhtä oikeaa, jonka perusteella malli valitaan, vaan voidaan raportoida usean mallin yhdistetyt tulokset, ja niihin liittyvät tunnusluvut, ja käyttää näitä tulkinnan tukena. Satunnaisuudestakin tulee positiivinen asia, kun klusterien avulla saadaan näkyväksi sitä, kuinka paljon aiheet satunnaisesti vaihtelevat. Kun yhden mallin sijaan tulkitaan aiheryhmiä useammasta mallista, saadaan avattua aihemallinnuksen mustaa laatikkoa.

Tutkimusprosessin kuvaus

Tutkimusprosessin tavoitteena on siis päästä syvemmälle aihemallinnuksen tuloksiin. Aihemalleissa, kuten latenttien muuttujien malleissa tai valvomattomassa oppimisessa usein, prosessi on toistettava rakenna-laske-kritisoi-toista -kehä. Seuraavassa kuvataan tutkimusprosessin vaiheet.



Kuva 1. Tutkimusprosessin piirteet.

1) Rakentaminen

Aihemallinnus alkaa sillä, että tekstikorpus muotoillaan sopiviksi sanasäkeiksi. Tyypillisesti aihemallinnuksessa tämä valmistelun vaihe päätyy yhteen spesifiin ratkaisuun, mutta nyt määritellään rajallinen joukko vaihtoehtoisia ratkaisuja. Tavoitteena on rajata niitä aineiston esikäsittelyn ja mallinnusprosessin valintoja, jotka juuri työn alla olevan analyysin tutkimus-

kysymysten osalta voisivat olla olennaisia, ja valmistella näiden pohjalta aineisto vaihtoehtoisin tavoin.

Kuten yllä todettiin, aihemallinnuksen esivalmistelussa on periaatteessa rajattomasti ja realistisestikin tuhansia erilaisia valintoja ja kilpailevia mallin hyvyyden mittareita useita, ja näin tavoite siitä, että koko esikäsittely-avaruutta tutkimalla voitaisiin löytää yksi ”paras” aineiston kuvaus on siis mahdoton, ainakin kunnes konsensus syntyy mallin hyvyyden mittareista. Aihemallinnuksen tutkimusstrategiassa on aina välttämätöntä keskittyä siihen, mitä tietynlaisella mallilla saadaan tehtyä aineistosta näkyväksi. Esikäsittelyn tavoitteet täytyykin sovittaa tähän: tutkijan pitää löytää valinnoista ne, jotka liittyvät juuri hänen tutkimuskysymyksiinsä, ja tarkastella niiden vaihtelun merkitystä tulosten sisällöllisessä tulkinnassa. Toisenlainen aineiston tiivistys ja esikäsittely vastaisi toisenlaisiin tutkimuskysymyksiin, eikä kyse ole siten siitä, mitkä valinnat olisivat yleisesti parhaat. Käytännössä tämä tarkoittaa sitä, että tutkija valitsee muutaman mallinnukseen liittyvän valinnan, jotka juuri tässä tutkimuksessa ovat kiinnostavia. Näitä valintoja voivat olla aiheiden määrä tai joku tietty esikäsittelyn toimenpide (tai periaatteessa voidaan olla vaihtelematta mitään, jos ollaan kiinnostuneita vain mallin ajojen välisen satunnaisuuden eli mallin sensitiivisyyden tarkastelusta).

Tavoitteena on suunnitella aihemallinnusstrategia eli päättää mitkä aihemallit seuraavaksi lasketaan. Jokaisesta vaihdeltavien toimenpiteiden yhdistelmästä kannattaa laskea useampia mallinnuskertoja, esimerkiksi 5–10, jotta menetelmään liittyvä satunnaisuus ei vaikuta klustereihin[2]. Määriä rajoittaa mielenkiinnon lisäksi käytettävissä olevat laskennalliset resurssit – nykykannettavalla operoivalle realistinen määrä voisi olla 30–200 erilaista mallia. Esimerkiksi jos halutaan verrata sitä, miten kahden erilaisen poistosanalistan käyttö ja kolmen erilaisen harvinaisten sanojen poistokriteerin (ei poisteta mitään, poistetaan sanat jotka esiintyvät alle 0,5 % dokumenteista tai alle 1 % dokumenteista) vaikuttaa, saadaan 2 poistosanalistaa * 3 harvinaisten sanojen kriteeriä eli 6 erilaista yhdistelmää, ja jos jokainen yhdistelmä ajetaan 10 kertaa, mallinnusstrategia koostuu 60 aihemallista.

2) Mallinnus

Mallinnusstrategian toteutus koostuu strategian mallien suorittamisesta, niiden tulosten klusteroinnista, ja klusteroinnin tulkinnasta. Aihemallien tuloksissa jokainen aihe on todennäköisyysjakauma sanaston sanoista – millä

2 Poikkeuksena tästä Roberts ym. (2016) esittämä *spectral*-LDA -malli, joka tuottaa ajosta riippumatta samat tulokset, jolloin tietysti yksi mallinnus esikäsittelyjen yhdistelmää kohden riittää.

todennäköisyydellä tästä aiheesta kirjoitettaessa käytettäisiin kyseistä sanaa. Tässä artikkelissa ei kuvata itse aihemallinnuksen toteuttamista tarkemmin – Nelimarkka (2019) on hyvä suomenkielien johdatus aiheeseen – vaan keskitytään aiheiden käsittelyyn mallinnuksen jälkeen. Jokaiselle kahden aiheen parille (kunhan sanasto pysyy samana tai sanastot ovat jälkikäteen yhdistettävissä) voidaan laskea erilaisia samanlaisuuden tai erilaisuuden indikaattoreita, ja kun samankaltaisuudet lasketaan kaikille mahdollisille pareille, aiheita voidaan sitten klusteroida erilaisin metodein.

Tässä artikkelissa käytetään aiheiden samankaltaisuutta kuvaamaan kosinin samankaltaisuutta, joka tiivistää aiheparin samankaltaisuudeksi numeerisesti asteikolle [0, 1]. Vaihtoehtoisesti samankaltaisuutta voitaisiin laskea myös esimerkiksi siitä, kuinka monta samaa sanaa on kahden aiheen 10 yleisimmän sanan joukossa, ja algoritmeja samanlaisuuden (tai erilaisuuden) mittaamiseen on muitakin – Mantyla ym. (2018) vertaavat kolmea erilaista metriikkaa, Chuang ym. (2013) neljää, joista käyttöön vakiintuneista metriikoista kosini ennusti parhaiten ihmisten arvioita aiheiden samankaltaisuudesta. Kosinin samankaltaisuuden arvot eivät myöskään riipu tekstien pituudesta eikä sitä käyttäessä tarvitse rajautua yleisimpiin sanoihin, vaan voidaan tarkastella koko todennäköisyysjakaumaa. Wilkerson ja Casas (2017) käyttävät myös samaa samankaltaisuuden mittaria.

Muutamissa aiemmissa tutkimuksissa on ehdotettu useampien aihemallinnusajojen aiheiden klusterointia. Klusterianalyysiin on runsaasti erilaisia työkaluja, joiden valinta liittyy analyysin tavoitteisiin. Aihemalleja aiemmin klusteroinneista Chuang ym. (2015) keskittyvät mallien stabiliteettiin ja käyttävät hierarkkista agglomerativiista klusterointia, jossa ryhmien määrää ei tarvitse tietää etukäteen, mutta jokaisesta mallista kuhunkin klusteriin voi kuulua enintään yksi aihe. Wilkerson ja Casas (2017) käyttävät Spectral Clustering -menetelmää ja Mantyla ym. (2018) käyttävät K-mediaanien klusterointia – näissä molemmissa täytyy määritellä klusterien määrä etukäteen ja jokainen aihe sijoitetaan johonkin klusteriin.

Tämän artikkelin tavoitteisiin klusterointimenetelmäksi sopii HDBSCAN -menetelmä (Hiearchical Density-Based Spatial Clustering of Applications with Noise, Campello ym., 2013). Tekstianalytiikkaan tätä menetelmää ovat soveltaneet ainakin Crisan ym. (2019). Menetelmä palvelee kolmea klusteroinnin tavoitetta: ensiksi, tavoitteena on tunnistaa aiheita, jotka syntyvät vain sattumalta tai tietystä alkuasetelmasta. On siis tärkeää, että menetelmä mahdollistaa sen, että jokin aihe jää kokonaan ryhmittelemättä. Toinen tavoite on se, että yhdestä mallinnusajosta samaan aiheklusteriin voi sijoittua useampi aihe: yhtä teemaa koskeva sanasto voi sijoittua yhdessä aihemallissa kokonaan saman aiheen alle, mutta toisessa aihemallissa jakautuakin kahteen

aiheeseen. Kaksi näkökulmaa teemaan päästävät tutkijan syvemmälle aineiston tulkintaan, ja siksi tällaisten aiheiden tulisi sijoittua samaan klusteriin. Kolmas tavoite on se, että aiheryhmien määrää ei tiedetä etukäteen. HDBSCAN-klusteroinnissa täytyy määritellä yksi parametri, klusterin minimikoko, ei esimerkiksi klusterien määrää, joten menetelmä ei rajaa sitä, minkälaisia ryhmiä se voi aiheista tuottaa.

Klusteroinnin tuloksena on siis joukko klustereita eli samanlaisten aiheiden ryhmiä, joille lasketaan klusterin sanojen todennäköisyysjakauma siihen kuuluvien aiheiden todennäköisyysjakaumien summana. Tämän avulla klusteria voidaan tulkita samoin kuin aiheitakin tyypillisesti on tapana, eli nostaa 10 tai 50 todennäköisintä sanaa tarkastelun kohteeksi, ja tulkita mitä klusteri ”tarkoittaa”. Lisäksi voidaan laskea sitä, kuinka moni mallinnusajoista tuotti aiheen, joka sijoittui kyseiseen klusteriin ja sitä, kuinka monta aihetta samasta mallista sijoittui aihe kyseiseen klusteriin.

3) Kritiikki

Kuten yllä todettiin, aihemallien yhteiskuntatieteelliselle merkitykselle ei ole yksikäsitteistä tulkintaa ja siten mallinnuksen onnistumisen arviointi ei ole myöskään yksikäsitteistä. Klusterit avaavat uuden perspektiivin aihemallien sisällölliseen tulkintaan ja mallin validointiin mahdollistamalla aiheiden luokittelun. Aihemallinnuksessa on aiemmin pyritty luokittelemaan yksittäisiä aiheita eri tavoin: esimerkiksi Chuang ym. (2013) käyttävät asiantuntijatiedon perusteella koostettuja referenssikategorioita kuvaamaan sitä, mitä aineistosta olisi ”pitänyt” löytyä, ja kehittävät näistä typologian.

Klusterointi tuo tällaiseen luokitteluun kuitenkin uuden tason: aiheiden tulkinnassa voidaan käyttää apuna sitä, kuinka usein ja minkälaisista malleista tiettyyn klusteriin sijoittuu aihe tai aiheita. Näin syntyy typologia aiheiden kuvaamiseen. Jos tiettyyn klusteriin sijoittuu jokaisesta mallinnuskerrasta (esikäsittelystä riippumatta) yksi ja vain yksi aihe, klusteri selvästi kuvaa jotain aineiston hyvin keskeistä rakennetta. Nämä aiheet määrittelen ”ydinaiheiksi”. Aivan ideaalitapauksessa kaikki aiheet olisivat ydinaiheita, eli riippumatta esivalmistelusta ja sattumasta samat aiheet löytyisivät aina, ja voitaisiin todeta mallin hyvin luotettavasti kuvaavaan aineistoa. Näin ei kuitenkaan tyypillisesti tapahdu, vaan toisinaan syntyy klustereita, joihin yhdistyy yksittäisistä malleista useampia aiheita. Kutsun näitä ”yhdistelmiksi”. Joskus klusteriin päätyy monista malleista aihe, mutta ei kaikista – näitä määrittelen ”näkökulmiksi”. Yhdistelmät ja näkökulmat eivät välttämättä ole mallin validiteetin kannalta ongelmallisia: klustereiden sisällöllisessä tulkinnassa voidaan huomata, että ne kuvaavat tulkitsijan kannalta mielenkiintoisesti aineiston moni-

mutkaisia rakenteita. Yksittäinen aihe voi myös jäädä kokonaan klusterien ulkopuolelle –”roska-aiheet” eli aiheet, jotka löytyivät vain yhdessä mallinnusajossa, ovat todennäköisesti löytyneet vain sattumalta, eikä niiden tulkinnalle kannata antaa juuri painoarvoa. Aihemallinnuksessa on aiemminkin tarkasteltu epäonnistuneita aiheita, mutta aiheet on yleensä tuomittu roskaksi subjektiivisessa tulkinnassa. Kun roska-aiheet määritellään sellaisiksi, jotka löytyvät mallinnuksia toistettaessa vain kerran, saadaan läpinäkyvä määritelmä niille. Roska-aiheiden tarkastelu on kuitenkin hyvä työkalu mallin validointiin – jos toistetut mallinnukset tuottavat runsaasti kaikista muista poikkeavia aiheita, malli ei syystä tai toisesta luotettavasti löydä aineiston rakennetta. Klusterityyppejä määrittävissä kriteereissä kannattaa olla jonkin verran joustoa – yksittäiset malliajot eivät välttämättä löydä ydinaihettakaan, joten klusterien tyypittelyssä ”ydinaiheen” rajaksi kannattaa asettaa esimerkiksi se, että vähintään 90 % mallinnuksista aiheen löytää.

Klusteritypologian tulkinta ei vielä yksin riitä, vaan klusterien tulkinnassa täytyy kurkistaa niiden sisällekin. Laadullisesti tämä on samanlaista tulkitsijavirtuoosin työtä (Pääkkönen & Ylikoski, 2020) kuin yksittäisenkin mallin tarkastelu, ja samat laskennalliset tai kokeelliset työkalut validointiin ovat käytettävissä klusterin kuin yhden aiheenkin osalta (yleiskuvauksena näistä esim. Nelimarkka (2019)). Lisäksi voidaan hyödyntää klusteria koskevia tietoja: kun käytössä on klusterin todennäköisimpien sanojen jakauma sekä kunkin siihen kuuluvan aiheen sanojen jakauma, näitä voidaan vertailla ja tarkastella – tutkija voi siis käydä läpi esimerkiksi seuraavia kysymyksiä: ovatko klusterin yleisimmät sanat systemaattisesti myös sen jäsenten yleisimpiä, vai löytyykö yksityiskohdissa eroja? Yksittäisten aiheiden tunnusluvut voidaan laskea keskimäärin klusterille ja vertailla – ovatko jotkut klusteriin sijoitetut aiheet koherentimpia tai eksklusiivisempia kuin muut? Klusteriin sijoittamattomia aiheita kannattaa tarkastella – miksi näitä aiheita ei saatu ryhmittelyyn mukaan? Myös yksittäisten mallinnusten tilastollisia tunnuslukuja voidaan hyödyntää – klusteroituvatko parempien tunnuslukuarvojen malleista tuotetut aiheet paremmin tai huonommin kuin muut? Parantaako tunnusluvun käyttö näin analyysin reliabiliteettiä?

4) Toisto

Kuten kaikessa latenttien muuttujien mallinnuksessa, prosessi voi hyvin toistua, kunnes päästään tuloksiin, jotka tutkijan arvion mukaan vastaavat tutkimuskysymyksiin ja kiinnostavat tutkijayhteisöä. Mallinnusstrategiaa voi siis hioa klusteroinnista opitun perusteella.

Esimerkki tutkimusprosessista

Tutkimusotteen havainnollistamiseksi esitetään esimerkki sosiaalisen median aineiston analysoinnista. Analyysin aineisto on kerätty Facebookin keskusteluryhmästä Uusi energiapolitiikka. Ryhmä perustettiin vuonna 2014 ja aineisto kattaa kaikki ryhmän keskustelut kevääseen 2017 asti. Keskustelussa kirjoitettiin yli 100000 viestiä 7000 keskusteluketjussa. Keskustelun temaattinen rakenne on mielenkiintoinen sosiaalisen median poliittisten keskustelujen merkityksen ymmärtämiseksi, mutta nyt sitä käytetään esimerkkinä menetelmän toiminnasta. Olemme kiinnostuineita siitä, miten aihemäärän valinta vaikuttaa mallista saatavaan ymmärrykseen sekä sitä, onko sanojen perusmuotoistamisella merkitystä, ja vaikuttaako se aiheiden määrään.

Aineisto kerättiin sittemmin väärinkäytösten takia suljetulla Facebook Graph API-rajapinnalla. Sosiaalisen median tutkimuskäytössä eettiset kysymykset ovat aina sidoksissa tutkimukseen kontekstiin ja kokonaisuuteen (Laaksonen, 2021) ja läpäisevät tutkimusprosessin aineistonkeruusta analyyseihin ja niiden julkaisuun (Franzke ym., 2020). Nyt käytetyssä aineistossa ei ole mitään henkilötietoja (vaikka se ei olekaan anonyymiä, sillä keskustelut ovat edelleen julkisessa internetissä) ja se on kerätty julkisesta ryhmästä, jonka säännöissä erikseen huomautetaan keskustelun julkisuudesta. Ryhmän aihe ei ole erityisen sensitiivinen – keskustelun aiheena on politiikkatyökalut ja niiden toiminta. Analyysin raportoinnissa ei käytetä lainauksia tai muita ilmaisutapoja, josta olisi helppo yhdistää yksittäiseen kommenttiin. Mahdolliset haitat keskusteluun osallistuville ovat näin minimaaliset.

1) Rakentaminen

Aineiston esivalmistelussa osa toimenpiteistä pidetään samoina, eli niiden merkityksestä ei tällä kertaa olla kiinnostuneita: aineistosta poistettiin numerot, välimerkit, alle kolmen merkin mittaiset sanat (ennen perusmuotoistamista), yleiset ns. poistosanat, ja kaikki sanat, jotka esiintyivät alle 1 % dokumenteista sekä manuaalisen tarkastelun perusteella joukko englanninkielisiä sanoja ja keskustelijoiden erisnimiä. Aineistoksi tuli 6721 keskusteluketjua ja niissä 720956 sanaa. Sanastosta tehtiin kolme versiota: ilman perusmuotoistamista (3643 erilaista sanaa), stemmauksella eli typistämisellä sanat niiden vartaloon R:n SnowballC-paketilla (Bouchet-Valat, 2020, 2115 erilaista stemmaa) ja lemmauksella eli perusmuotoistamisella Voikko[3] -ohjelmiston Python-työkaluilla (1542 lemmaa).

3 https://voikko.puimula.org/

Lemmaus ei ole yksikäsitteistä, vaan samaan taivutusmuotoon voidaan päätyä useammasta perusmuodosta – Nelimarkan (2019) esimerkkitaulukossa lemmatisoiva työkalu on tulkinnut "vaalit" sanan "vaalia" muodoksi, vaikka ilmeisesti kyse on ollut vaaleista. Lemmatisoitu sanasto pitää siis aina manuaalisesti tarkastaa ja tutkijan valita todennäköisin mahdollisista lemmoista. Nyt jokaiselle sanalle valittiin yksi, kontekstiin sopivin, lemma koko aineistoon. Samaa käsitettä käytetään tietysti eri merkityksissä aineistossa, ja täysin virheetön lemmatisointi edellyttäisi koko aineiston läpikäymistä manuaalisesti.

Tämän jälkeen toistettiin aihemallinnuksia näillä kolmella erilaisella esikäsittelyllä vaihdellen aihemääriä 7, 10, 15, 20 ja 30 aiheen välillä. Kukin sanaston ja aihemäärän yhdistelmä suoritettiin 10 kertaa, eli näin saatiin 150 mallia. Mallinnuksessa käytettiin Tieteen tietotekniikan keskuksen suurteholaskennan palveluita.

Esimerkissä on käytetty esikäsittelyyn R:n quanteda-pakettia ja aihemallinnukseen R:n stm-paketin structural topic model -mallia. Klusterointiin käytettiin dbscan-pakettia.

Klusteroitavien aiheiden sanastojen tulee olla samat. Nyt sanastot piti yhdistää mallinnuksen jälkeen laskemalla todennäköisyyksien summia lemmojen mukaan: käsittelemättömästä mallista siis laskettiin termien {sähkö, sähkön, sähköä..} todennäköisyydet yhteen lemman {sähkö} todennäköisyydeksi. Koska esikäsittelyn jälkeen jokaista käsitettä vastaa aina sama lemma, tämä laskutoimitus on yksiselitteinen. Stemmojen osalta näin ei täysin ole, vaan samaan stemmaan voidaan päätyä useammasta eri sanan taivutusmuodosta – *käsitän* ja *käsite* ovat molemmat stemmattuna *käsit*, mutta lemma on eri. Yksinkertaisuuden vuoksi nyt laskettiin todennäköisyys ensimmäisen havaitun stemma-lemma parin mukaan.

2) Mallinnus

Mallinnuksesta saadaan 2460 aihetta, joista jokaista kuvaa 1542 lemman todennäköisyysjakauma. Nämä aiheet klusterointiin erikseen aiheiden määrän ryhmissä, eli esimerkiksi 7 aiheen malleja oli 10 toistoa kutakin 3 käsittelyä kohti – kaikkinensa 210 aihetta. Taulukossa 1 on kaikki näin syntyneet klusterit aihemäärien 7–20 osalta. Liitetaulukoissa (verkkosivuilla) on vastaava 30 aiheen rakenne sekä kaikki yksittäiset aiheet klustereittain.

3) Kritiikki

Esimerkkianalyysissa oltiin kiinnostuneita kahdesta tutkimuskysymyksestä: perusmuotoistamisen ja aiheiden määrän vaikutuksesta aineistosta saatavaan käsitykseen.

Ensimmäiseen kysymykseen vastaus on teknisempi: perusmuotoistamisen tapoja vaihtelemalla haluttiin varmistaa, että tuloksissa ei ole kyse vain esikäsittelyssä syntyneestä artefaktista. Nyt perusmuotoistaminen ei juurikaan vaikuttanut aihemallinnuksen tuloksiin: pääpiirteissään lemman {ydinvoima} tai stemman {ydinvoim} todennäköisyys osuu samaan kuin raakakäsitteiden {ydinvoima, ydinvoiman, ydinvoimalla..} todennäköisyyksien summa. Klusteroinnin kautta tämä käy ilmi tarkastelemalla taulukon 1 tietoja siitä, miten erityyppisellä sanastolla tuotetuista malleista sijoittui aiheita erilaisiin klustereihin: jos perusmuotoistaminen muuttaisi mallinnuksen tuloksia radikaalisti, ääritapauksessa syntyisi vain näkökulma-aiheita, joissa lemmatun sanaston mallien aiheet menisivät yhteen ryhmään eli tuottaisivat yhden näkökulman, stemmatut toiseen, ja perusmuotoistamattomat kolmanteen. Näin ei kuitenkaan nyt käy: eri sanastoilla tuotetut mallit jakautuvat erilaisiin klustereihin, ja ydinaiheita – aiheita, jotka löytyvä (miltei) jokaisella ajolla on useita. Muutamassa ajossa tiiviimmällä lemma-sanastolla syntyy ”eksentrinen” malli, jossa keskimääräistä isompi osa aiheista poikkesi kaikista muista malleista, mikä näkyy hieman isommassa roska-aiheiden osuudessa, mikä voisi vihjata siihen, että sanaston tiivistäminen tekee tuloksista vaihtelevampia, mutta vaikutus on ainakin tässä analyysissä hyvin pieni.

Toiseen tutkimuskysymykseen vastaamiseksi tarkastellaan klustereita aihemäärän mukaan. Seitsemän aiheen malleista syntyy kahdeksan klusteria. Kaksi ydinaihetta, polttoaineita ja ydinvoimaa koskevat sanastot, löytyivät riippumatta perusmuotoistamisesta ja sattumasta. Näiden lisäksi syntyi pieniä näkökulmaklustereita: nämä yhdistelivät välillisesti toisiinsa liittyviä teemoja, kuten sähkön ja lämmön tuotantoa ja sähköautoilua. Tällaisia tulkinnallisesti erilaisia aiheita yhdisteleviä näkökulmia syntyy, kun mallin aihemäärä on liian pieni: kun aineistossa on enemmän teemoja kuin malleissa aiheita, malli joutuu yhdistelemään lähekkäin olevia teemoja.

10 aiheen malleista syntyy 4 ydinaiheklusteria: sähköautoja, päästöjä ja polttoaineita, lämpöä ja kaukolämpöä sekä ydinvoimaa koskevat aiheet löytyivät riippumatta malliajosta ja sanastosta. Sähkön hintaa ja uusiutuvia energianlähteitä koskeva aihe löytyi myös aina, mutta viidestä mallista sijoittui klusteriin kaksi aihetta, jakaen sanaston uusiutuvien hintaa koskevaan aiheeseen ja hintaa, tukia ja veroja koskevaan aiheeseen. Malleissa syntyi 2 näkökulmaklusteria: 16 mallia (30:stä) löysi tuulivoimaa, vesivoimaa, Pohjoismaita ja

Saksaa yhdistelevän aiheen, 8 mallia taas aiheen, joka keskittyi vain maiden nimiin, ja näissä uusiutuvia koskeva sanasto oli erikseen sähköaiheissa.

15 aiheella ydinaiheet pysyvät muuten samoina, mutta päästöjen ja polttoaineiden aiheklusterista tulee yhdistelmä: 10 malliajoa tuotti kaksi aihetta tähän klusteriin. Näissä aiheissa oli päästöjä koskevaa sanastoa ja hiileen, öljyyn, kaasuun, puuhun ja turpeeseen liittyvää sanastoa. Jos malli tuotti kaksi aihetta klusteriin, sanasto jakautui kahteen teemaan: tyypillisesti kotimaista polttoainekeskustelua kuvaavaan aiheeseen, kuten {Suomi, turve, puu, metsä}, ja kansainvälistä keskustelua kuvaavaan aiheeseen, jossa korostuivat {hiili, öljy, Kiina, Saksa, ilmastonmuutos}.

20 aiheella klusterit sirpaloituvat: samat ydinaiheet löytyivät kuin pienemmilläkin aihemäärillä, mutta nyt syntyi 10 klusteria, joihin sijoittui aihe 6–19 mallinnuksesta (30:stä). Yksittäiset aiheet ovat tulkinnaltaan spesifimpiä kuin pienemmällä aihemäärällä. Esimerkiksi sähköä koskevia klustereita syntyy kulutuksen, tehon ja verkon sanaston kanssa, samoin kuin hintaa koskevan sanaston kanssa. 30 aiheen mallissa (liitetiedostona osoitteessa `https://github.com/arhot/aihemallien_klusterointi/`) sama trendi jatkuu – mielekkäästi tulkittavia klustereita syntyy, mutta yhä suurempi osa niistä löytyy n. 20–70 % mallinnusajoista.

Aiheiden määrän lisääntyessä potentiaalisten aineistoon otettavien näkökulmien määrä siis kasvaa: aiheet pilkkoutuvat aina hienojakoisemmiksi tulkinoiksi. Analyysin tuloksena voidaan todeta, että 7 aihetta oli liian vähän, 30 ehkä liian paljon: 7 aihetta yhdistää tulkinnallisesti poikkeavia teemoja, 30 aihetta hajauttaa mahdolliset tulkinnat liian pitkälle. Tältä väliltä löytyy kuitenkin useita uskottavia tapoja kuvata keskustelua ja on epätodennäköistä, että mikään mallin hyvyyden tilastollinen mittari ratkaisisi sitä, mikä aihemäärä on universaalisti paras aineiston kuvaus. Aihemäärän ratkaisu pitäisikin kiinnittää tutkimuskysymykseen: onko tarkoitus antaa lukijalle mahdollisimman tarkka ymmärrys korpuksen rakenteesta ja keskustella erilaisista näkökulmista, vai ehkä käyttää kvantitatiivista tietoa aiheiden osuuksien muutoksista dokumenteissa jatkoanalyyseissa.

Riippumatta aiheiden määrästä klusterointi synnytti yhden suuren klusterin, jossa on kokoelma ”keskustelusana-aiheita”: erilaisia aineiston substanssiin liittymättömiä, yleisiä keskustelun rakennetta tukevia sanoja – {paljon, vähän, ongelma, esimerkki, tehdä}. Klusterointi auttoi näin myös erottamaan aineistoon spesifisti liittyviä aiheita yleisistä puhetapoihin liittyvistä aiheista.

Olisiko samat johtopäätökset voitu saavuttaa tarkastelemalla analyysin tilastollisia tunnuslukuja? Taulukossa 1 on klusterityypeittäin laskettu klusterin aiheiden keskimääräinen eksklusiivisuus ja koherenssi. Erot ovat hyvin pieniä, mutta roska-aiheet ovat keskimäärin hieman koherentimpia, mutta

vähemmän eksklusiivisia kuin muut. Tämä kuvaa mittarien ominaisuuksia: hyvin yleisten sanojen aihe on koherentti, mutta ei tietenkään kovin eksklusiivinen, eikä erotu massasta klusterin jäseneksi. Näkökulma-klusterit eli vain harvoin aineistosta löytyvät teemat ovat puolestaan eksklusiivisempia. Tämäkin on luonnollista: aihe, joka eristää aineistosta hyvin pienen teeman ja sitä koskevaa sanastoa on eksklusiivinen, mutta ei välttämättä kovinkaan toistettavissa. Aihetason mittareita ei siis voi käyttää mallin valinnassa, sillä ne saattavat ohjata väärään suuntaan: koherenssin käyttö voi valita malleja, joissa hyvin yleistä sanastoa liitetään temaattiseen sanastoon, ja eksklusiivisuuden käyttö puolestaan sattumalta poikkeavia malleja.

4) Toisto

Tämän analyysin perusteella voisi lähteä uudelle mallinnuskierrokselle, jossa rajattaisiin aihemääriä tarkemmin: nyt huomattiin, että 7 aihetta oli liian vähän ja 30 aihetta liikaa, ja voitaisiin aloittaa alusta esimerkiksi niin, että siirryttäisiin 10, 12, 14, 16, 18, ja 20 aiheen malleihin, jotta saadaan yhä tarkempaa kuvaa aineistosta. Toisaalta tieto ja esitys aiheiden määrän rajoista klusteroinnin puitteissa on mielenkiintoinen tulos jo itsessään, ja nyt prosessia ei ole tarpeen lähteä toistamaan.

Taulukko 1. Klusteroinnin tulokset. M= Mallien määrä, joista ainakin 1 aihe sjoittui klusteriin. T=malllien määrä, joista sijoittui useampi aihe klusteriin. L=lemmatut mallit, joista ainakin yksi klusterissa. S=stemmatut mallit, joista ainakin 1 klusterissa. R=raakakäsitteiden mallit, joista ainakin 1 klusterissa. E=klusterin aiheiden keskimääräinen eksklusiivisuus. K=klusterin sanojen keskimääräinen koherenssi.

K	Tyyppi	Klusterin yleisimmät sanat	M	T	L	S	R	E	K
7	Roska-aihe	suomi, energia, paljo, tehdä, lisätä, tuottaa, hinta, sähkö, mennä, alkaa	21	14	12	10	13	9,1	-106
	Ydinaihe	päästö, suomi, hiili, öljy, puu, käyttö, energia, fossiilinen, polttoaine, turve	29	0	10	10	9	9,5	-115
	Ydinaihe	suomi, tehdä, ydinvoima, venäjä, laitos, mennä, fortum, uutinen, voimala, jälki	30	0	10	10	10	9,4	-117
	Näkökulma	tuki, suomi, raha, valtio, tehdä, maksaa, yritys, haluta, mieli, investointi	6	0	2	4	0	9,3	-108
	Näkökulma	sähkö, energia, lämpö, tuottaa, vesi, kaukolämpö, käyttää, aurinko, verkko, tarvita	14	0	6	4	4	9,2	-110
	Näkökulma	lämpö, sähkö, sähköauto, tehdä, auto, vesi, paljo, tuottaa, hyötysuhde, tarvita	6	0	1	2	3	8,7	-106
	Näkökulma	sähköauto, auto, hinta, tehdä, paljo, liikenne, tesla, akku, öljy, ajaa	14	0	5	4	5	9,4	-108

	Yhdistelmä	sähkö, hinta, tuulivoima, suomi, tuotanto, ruotsi, saksa, tuki, paljo, tuuli	30	6	11	13	12	9,5	-96
	Yhdistelmä	paljo, energia, vähä, tosi, mieli, esimerkki, tehdä, ihminen, ongelma, varma	29	11	13	13	14	8,8	-106
10	Roska-aihe	suomi, energia, paljo, tuulivoima, sähkö, tuottaa, tuki, lisätä, tuotanto, tehdä	26	24	24	14	12	9,6	-106
	Ydinaihe	sähköauto, auto, liikenne, tesla, tehdä, akku, aku, paljo, hinta, ajaa	27	0	8	9	10	9,5	-110
	Ydinaihe	päästö, hiili, suomi, puu, öljy, turve, polttoaine, fossiilinen, käyttö, polttaa	29	1	9	10	11	9,6	-118
	Ydinaihe	lämpö, energia, sähkö, vesi, kaukolämpö, helsinki, tuottaa, käyttää, lämpöpumppu, teho	28	0	8	10	10	9,4	-115
	Ydinaihe	tehdä, venäjä, fortum, suomi, laitos, fennovoima, hanke, uutinen, ydinvoima, voimala	30	2	12	10	10	9,6	-115
	Näkökulma	suomi, ruotsi, saksa, maa, eurooppa, norja, uusiutuva, paljo, lisätä, tanska	8	0	2	3	3	9,7	-106
	Näkökulma	suomi, sähkö, tuulivoima, ruotsi, saksa, tuotanto, paljo, vesivoima, norja, tuuli	16	0	5	6	5	9,7	-98
	Yhdistelmä	paljo, tosi, vähä, tehdä, mieli, ydinvoima, energia, varma, ongelma, ihminen	30	46	22	27	27	9,3	-105
	Yhdistelmä	sähkö, hinta, maksaa, tuki, tuotanto, markkina, euro, kustannus, tuulivoima,	28	5	10	11	12	9,7	-102
15	Roska-aihe	energia, suomi, sähkö, tuottaa, paljo, tuotanto, hinta, uusiutuva, tehdä, tuulivoima	28	62	31	29	30	9,7	-106
	Ydinaihe	sähköauto, auto, liikenne, tesla, akku, aku, autto, tehdä, diesel, ajaa	29	0	10	9	10	9,7	-109
	Ydinaihe	lämpö, kaukolämpö, helsinki, energia, vesi, sähkö, lämpöpumppu, lämmitys, talo, helen	30	0	10	10	10	9,7	-105
	Ydinaihe	venäjä, fortum, fennovoima, tehdä, laitos, voimala, ydinvoimala, hanke, ydinvoima, rakentaa	30	2	11	11	10	9,8	-113
	Näkökulma	energia, uusiutuva, tuotanto, kasvu, saksa, kasvaa, käyttö, kulutus, lisätä, tavoite	7	0	1	3	3	9,7	-114
	Näkökulma	tuki, valtio, raha, yritys, vero, suomi, investointi, maksaa, tehdä, miljardi	20	0	6	5	9	9,7	-107
	Näkökulma	hinta, sähkö, maksaa, markkina, euro, tuki, halpa, kustannus, tuotanto, investointi	24	0	7	8	9	9,8	-106
	Näkökulma	sähkö, aurinkosähkö, aurinko, tuotanto, verkko, tuottaa, aurinkoenergia, kulutus, varastointi, paneeli	15	0	4	5	6	9,6	-100
	Näkökulma	suomi, ruotsi, maa, uutinen, alue, eurooppa, alkaa, lisätä, norja, saksa	23	1	9	7	8	9,7	-110
	Näkökulma	tuulivoima, sähkö, suomi, ruotsi, vesivoima, tuotanto, tuuli, paljo, saksa, rakentaa	21	0	6	6	9	9,7	-99
	Yhdistelmä	päästö, hiili, puu, turve, suomi, kivihiili, fossiilinen, polttaa, polttoaine, metsä	30	10	14	14	12	9,7	-113
	Yhdistelmä	paljo, ydinvoima, vähä, tehdä, tosi, mieli, ongelma, varma, esimerkki, ihminen	30	88	41	43	34	9,6	-105
20	Roska-aihe	suomi, sähkö, energia, tuottaa, paljo, tuotanto, tehdä, hinta, saksa, lisätä	30	139	62	62	45	9,7	-105
	Ydinaihe	sähköauto, auto, liikenne, tesla, akku, aku, autto, diesel, ajaa, tehdä	30	2	11	10	11	9,8	-108
	Ydinaihe	fortum, venäjä, fennovoima, hanke, tehdä, laitos, voimala, rosatom, uutinen, ydinvoimala	27	0	10	8	9	9,8	-100

	Näkökulma	öljy, tuotanto, kaasu, polttoaine, käyttää, tuottaa, raaka, fossiilinen, biokaasu, maakaasu	6	0	0	3	3	9,8	-108
	Näkökulma	tuulivoima, tuuli, rakentaa, vesivoima, voimala, tuulivoimala, tuottaa, paljo, suomi, tanska	18	0	6	3	9	9,7	-101
	Näkökulma	suomi, maa, alue, uutinen, alkaa, lisätä, paljo, viikko, eurooppa, tieto	7	0	2	3	2	9,8	-118
	Näkökulma	suomi, ruotsi, norja, sähkö, maa, eurooppa, venäjä, paljo, lisätä, alue	17	0	7	4	6	9,8	-104
	Näkökulma	aurinko, aurinkosähkö, tuottaa, aurinkoenergia, tuotanto, tuuli, varastointi, sähkö, energia, paneeli	11	0	2	6	3	9,7	-98
	Näkökulma	sähkö, kulutus, verkko, teho, tuotanto, tunti, silta, käyttö, sähköverkko, päivä	8	0	2	1	5	9,8	-99
	Näkökulma	hinta, sähkö, maksaa, markkina, euro, halpa, kustannus, tuotanto, laskea, investointi	19	0	4	6	9	9,8	-106
	Näkökulma	tuki, valtio, raha, maksaa, vero, investointi, hinta, euro, teollisuus, yritys	25	0	8	9	8	9,8	-111
	Näkökulma	valtio, raha, yritys, yhtiö, hallitus, kunta, tehdä, toiminta, päätös, haluta	9	0	3	2	4	9,7	-98
	Näkökulma	öljy, maailma, kasvu, kasvaa, kiina, talous, kulutus, tuotanto, hinta, maa	6	0	3	2	1	9,8	-104
	Näkökulma	ydinvoima, saksa, ydinvoimala, uusiutuva, rakentaa, tehdä, laitos, ongelma, power, reaktori	22	2	7	6	11	9,8	-108
	Näkökulma	energia, uusiutuva, tulevaisuus, tarvita, suomi, tehdä, tuottaa, alkaa, tuotanto, energiapolitiikka	18	1	5	7	7	9,8	-102
	Yhdistelmä	lämpö, kaukolämpö, helsinki, vesi, energia, lämpöpumppu, helen, sähkö, lämmitys, talo	29	3	9	11	12	9,8	-106
	Yhdistelmä	päästö, hiili, puu, turve, fossiilinen, kivihiili, vähentää, käyttö, suomi, polttoaine	30	17	14	15	18	9,8	-114
	Yhdistelmä	paljo, vähä, tehdä, tosi, varma, mieli, esimerkki, ongelma, ilma, mennä	30	94	45	42	37	9,72	-108

Johtopäätökset

Tässä artikkelissa on käsitelty aihemallinnuksen tutkimusprosessia, jossa yhden aihemallin valinnan ja raportoinnin sijasta tehdään vaihtoehtoisia mallinnuksia, yhdistetään näiden tuottamat aiheet klusterianalyysillä, ja käytetään tätä ryhmittelyä niin raportoinnissa kuin validoinnin ja sensitiivisyysanalyysien työkaluna. Esitetty tutkimusprosessi ammentaa aiemmista aihemallinnuksen ja klusteroinnin yhdistelmistä (Chuang ym., 2015; Mantyla ym., 2018; Wilkerson & Casas, 2017), mutta laajentaa sitä, mihin tällaista yhdistelmää voidaan käyttää. Aiemmin fokus on ollut mallin stabiiliuden arvioinnissa, mutta nyt esitetyllä tavalla voidaan klusterit kytkeä myös aineiston kuvailevaan analyysiin, suoraan sisällöllisiin tutkimuskysymyksiin, tai käyttää spesifien valintojen validoinnin työkaluna.

Joskus aihemallintaja haluaa vain kuvata aineiston temaattisen rakenteen yleiselle, ja tällaisessa eksploratiivisessa tutkimusasetelmassa klusterointi

toimii tukena aineiston tiivistykselle: näkökulma-aiheiden käsittely vaihtoehtoisina, mutta valideina, tulkintoina aineistolle näyttää lukijalle, minkälaisten vaihtoehtojen välillä aihemalli ”valitsee”. Nyt esimerkiksi ydinvoimakeskustelu oli aina oma teemansa, mutta esimerkiksi päästöjä koskevan sanaston voi jakaa suomalaiseen ja kansainväliseen keskusteluun – tai käsitellä nämä yhdessä. Tämä on tulos itsessään, ei ongelma mallinnuksessa. Nyt toinen tutkimuskysymys esimerkissä, aiheiden määrän vaikutus, oli tämän tyyppinen: klusterointi mahdollisti systemaattisen tarkastelun siitä, mitkä teemat ovat aina esille, mitkä pulpahtavat vain välillä esille.

Prosessin voi kytkeä tiukemminkin sisällöllisiin tutkimuskysymyksiin mallinnusstrategialla: nyt olisi esimerkiksi voitu laadullisella tutkimuksella erotella aineistosta poliittiset vaateet, kuten Ylä-Anttila ym. (2018) tekevät, ja sitten verrata poliittisten vaateiden aiheklustereita kaiken puheen aiheklustereihin, ja tehdä näkyväksi sitä, miten vaadepuhe poikkeaa muusta puheesta.

Tutkimusprosessi toimii myös validoinnin työkaluna. Aihemallinnuksen parametrivalinnat vaikuttavat tuloksiin, ja klusterointi tekee tästä läpinäkyvää, ja aiheiden pysyvyyttä erilaisten mallien välillä voidaan arvioida. Nyt ensimmäinen tutkimuskysymys perusmuotoistamisesta liittyi tällaiseen validointiin. Aiempaa empiiristä tutkimusta perusmuotoistamisen vaikutuksesta suomenkielisellä aineistolla ei ollut, joten asia nostettiin omaksi tutkimuskysymyksekseen, mutta tutkija voisi myös vain varmistaa, että tulokset pysyvät tulkinnallisesti samoina siitä riippumatta – tällä aineistolla näin oli, ja tulos on tältä osaltaan luotettava.

Tällaisten validointien osalta tutkimusprosessi voisi toimia tukena koko aihemallinnuksen kentälle, työkaluna merkityksellisten esivalmistelutyökalujen etsinnässä. Käytännössä nykyisin päätöksen poistosanojen tai numeroiden tai välimerkkien joudutaan tekemään vakiintuneisiin käytäntöihin nojaten, mutta näitä käytänteitä ei varsinaisesti ole empiirisesti perusteltu. Tutkimusprosessia itseäänkin voisi edelleen kehittää – tässä artikkelissa käsiteltiin vain yhtä aiheiden samankaltaisuuden mittaria, kosinin samankaltaisuutta. Tutkimusprosessi toimisi samalla tavalla toisella samankaltaisuuden mittarillakin, ja jatkotutkimuksissa olisi hyvä verrata näitä.

Tämän esityksen tavoitteena on tietysti innostaa aihemallinnuksen soveltajia tällaisen tutkimusprosessin käyttöön (tai sen edelleen kehittämiseen!). Kiinnostuneille soveltajille esimerkin tuottavat R-koodit ja analyysin liitetaulukko ovat saatavilla osoitteessa `https://github.com/arhot/aihemallien_klusterointi/`. Vastaavat työkalut ovat saatavilla myös ainakin Python-kielelle ja aihemallinnusta tekevien ohjelmistojen, kuten MALLETin, tuloksia pystyy jälkikäsittelemään jossain toisessa ympäristössä. Aihemallinnus ja siihen liittyvät oheistoimenpiteet (kuten perusmuotoistaminen)

asettavat helposti korkeita vaatimuksia käytettävissä olevien tietokoneiden laskutehon suhteen. Nyt esitetty analyysi suoritettiin Tieteen tietotekniikan palvelimilla, mutta tämän analyysin voisi vielä toteuttaa myös tutkijan omalla koneella – yksittäisen malliajon kesto tavallisella Windows-kannettavalla oli n. 15 minuuttia, eli koko analyysi kestäisi 38 tuntia – pitkään, mutta analyysit saisi kuitenkin viikonlopun aikana ajettua. Tieteen tietotekniikan keskus tarjoaa myös tarkempia lauseenjäseniin perustuvia lemmatisointimenetelmiä, joilla voitaisiin parantaa lemmojen tunnistusta taivutuista sanoista. Nyt tavoitteena oli kuitenkin esittää tutkimusprosessi muodossa, joka olisi saavutettavissa ilman näiden palveluiden käyttöönottoa, ja aloittelevallekin aihemallintajalle kohtalaisen saavutettavana, vaikka asiaan liittyy monenlaisia teknisiä yksityiskohtia.

Lähteet

Allen, C., & Murdock, J. (2021). LDA Topic Modeling: Contexts for the History & Philosophy of Science. Teoksessa G. Ramsey & A. De Block (toim.), *The Dynamics of Science: Computational Frontiers in History and Philosophy of Science*. Pittsburgh University Press.

Blei, D. M. (2012a). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M. (2012b). Topic Modeling and Digital Humanities. *Journal of Digital Humanities, 21*(1), 8–11.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17–35. https://doi.org/10.1214/07-aoas114

Bouchet-Valat, M. (2020). *SnowballC: Snowball Stemmers Based on the C "libstemmer" UTF-8 Library. R package version 0.7.0*. https://CRAN.R-project.org/package=SnowballC

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Teoksessa J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (toim.), *Advances in Knowledge Discovery and Data Mining. PAKDD 2013* (s. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296. Noudettu osoitteesta http://machinelearning.wustl.edu/mlpapers/paper_files/NIPS2009_0125.pdf

Chuang, J., Gupta, S., Manning, C. D., & Heer, J. (2013). Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. *Proceedings of the 30th International Conference on Machine Learning*, 9.

Chuang, J., Roberts, M. E., Stewart, B. M., Weiss, R., Tingley, D., Grimmer, J., & Heer, J. (2015). TopicCheck: Interactive Alignment for Assessing Topic Model Stability. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–184. https://doi.org/10.3115/v1/N15-1018

Crisan, A., Munzner, T., & Gardy, J. L. (2019). Adjutant: An R-based tool to support topic discovery for systematic and literature reviews. *Bioinformatics, 35*(6), 1070–1072. https://doi.org/10.1093/bioinformatics/bty722

Denny, M. J., & Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis, 26*(2), 168–189. https://doi.org/10.1017/PAN.2017.44

franzke, a. s., Bechmann, A., Zimmer, M., Ess, C. M., & Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Laaksonen, S.-M. (2021). Sosiaalinen media tutkimusaineistona. Teoksessa T. Kallinen & T. Kinnunen (toim.), *Laadullisen tutkimuksen verkkokäsikirja*. Yhteiskuntatieteellinen tietoarkisto.

Mantyla, M. V., Claes, M., & Farooq, U. (2018). Measuring LDA topic stability from clusters of replicated runs. *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–4. https://doi.org/10.1145/3239235.3267435

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 262–272.

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

Pääkkönen, J., & Ylikoski, P. (2020). Humanistic interpretation and machine learning. *Synthese*. https://doi.org/10.1007/s11229-020-02806-w

Purhonen, S., & Toikka, A. (2016). "Big datan" haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät. Esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935-2015. *Sosiologia, 53*(1), 6–27.

Roberts, M., Stewart, B., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. Teoksessa R. M. Alvarez (toim.), *Computational social science: Discovery and prediction* (s. 51-97). Cambridge University Press. https://doi.org/10.1017/CBO9781316257340.004

Schofield, A., & Mimno, D. (2016). Comparing apples to apple: The effects of stemmers on topic models. *Transactions of the Association for Computational Linguistics, 4*, 287–300. https://doi.org/10.1162/tacl_a_00099

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105–1112.

Ylä-Anttila, T., Eranti, V., & Kukkonen, A. (2018). Aihemallinnuksesta kehysmallinnukseen. *Politiikka, 60*(2), 148–156.