

Joel Roth

ALGORITMIT VASTAAN PROPAGANDA
- Koneoppimisen ja tekoälyn monimodaalinen kehitys-
tyskaari propagandan tunnistuksessa
2011–2024

Johtamisen ja talouden tiedekunta
Pro gradu -tutkielma
Toukokuu 2025

ABSTRACT

Joel Roth: Algorithms Against Propaganda: A Multimodal Trajectory of Machine-Learning and AI Approaches to Propaganda Detection, 2011–2024

Pro gradu -thesis

Tampere University

Master's Programme in Politics / Political Science study track

May 2025

This thesis systematically investigates how artificial-intelligence and machine-learning methods were harnessed between 2011 and 2024 to automate the detection of political propaganda in the digital public sphere. Its scholarly contribution is threefold.

First, it integrates the rhetorical techniques of propaganda, Bayesian receiver theory and network-based anomaly detection into a unified conceptual–operational framework that accounts for the manipulative features of message content, their impact as Bayesian belief updates, and the irregularities of co-ordinated dissemination.

Secondly, the study is the first to chart long-term trends across seven methodological and ethical dimensions, revealing a strategic imbalance in the field: although deep-learning text models have advanced rapidly, causal impact evaluation, temporal robustness, multilingual coverage and ethical-bias auditing remain uncommon.

Thirdly, the thesis introduces a novel three-stage mixed-methods framework—Exploratory Sequential Mixed-Trend-Tracing (ES-MTT)—in which a bibliometric mapping of 2,916 publications, k-means clustering and a dampened citation metric yield a 64-article core sample. Relevance screening was conducted via a multi-stage pipeline. First, a multi-agent LLM consensus algorithm—comprising three independent GPT-based models—issued binary inclusion/exclusion judgements against preregistered criteria, returning them in a fixed JSON schema. In collaboration with a human reviewer this step eliminated 878 irrelevant records. The remaining documents were encoded with the text-embedding-3-large model, projected with LocalMAP, and stratified by HDBSCAN density clustering, yielding a semantically comprehensive, domain-representative, deterministically replicable and statistically validated core set of 64 articles.

The findings identify three distinct developmental phases. During the early period (2011–2015) research relied chiefly on traditional text classifiers and hand-crafted features; the middle period (2016–2019) mainstreamed benchmark datasets (LIAR, FakeNewsNet) and deep CNN/RNN architectures; in the new period (2020–2024) transformer architectures, multimodal vision–language models and heterogeneous graph neural networks combined text, image, video and network dynamics within a single classification pipeline. Nevertheless, only 14 per cent of studies tested models for temporal durability, 13 per cent operated genuinely across languages and 8 per cent reported ethical-bias metrics—despite the accountability requirements emphasised by EU AI regulation.

The study recommends: (i) expanding open, multilingual and multimodal benchmark datasets to encompass European and Global South languages; (ii) making standard causal and robustness tests a prerequisite for algorithmic deployment; and (iii) mandating bias audits and Model Card reporting to ensure transparency in platform content moderation. In so doing, the research supports policy measures aimed at democratic resilience and the responsible governance of algorithmic agenda-setting.

Keywords: propaganda detection; machine learning; multimodality; multilingualism; Bayesian fusion model; bibliometric analysis

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

TIIVISTELMÄ

Joel Roth: Algoritmit vastaan propaganda - Koneoppimisen ja tekoälyn monimodaalinen kehityskaari propagandan tunnistuksessa 2011–2024

Pro gradu -tutkielma

Tampere University

Politiikan tutkimuksen maisteriohjelma / valtio-opin opintosuunta

Toukokuu 2025

Tämä tutkielma tarkastelee systemaattisesti, miten tekoäly- ja koneoppimismenetelmät on vuosina 2011–2024 valjastettu poliittisen propagandan automaattiseen havaitsemiseen digitaalisessa julkisuudessa. Työn tieteellinen kontribuutio on kolmiosainen.

Ensiksi se yhdistää propagandan retoriset tekniikat, bayesilaisen vastaanottajateorian sekä verkostopohjaisen anomaliadetektion yhtenäiseksi käsitteellis-operatiiviseksi kehykseksi, joka selittää niin viestin sisällön manipulatiiviset piirteet, vaikutuksen todennäköisyyspäivityksenä kuin koordinoitun leviämisen poikkeavuudet.

Toiseksi tutkimus mittaa ensimmäistä kertaa seitsemän metodologis-eettisen ulottuvuuden pitkän aikavälin trendit ja osoittaa kentän strategisen vinoutuman: vaikka syväoppivat tekstimallit ovat kehittyneet nopeasti, kausaalinen vaikutusarviointi, temporal-robustisuus, monikielisyys ja eettisten vinoumien auditointi ovat edelleen harvinaisia.

Kolmanneksi tutkielma esittelee uuden kolmiportaisen sekamenetelmäkehityksen (Exploratory Sequential Mixed-Trend-Tracing, ES-MTT), jossa 2 916 julkaisun bibliometrinen kartoitus, k-means-klusterointi ja vaimennettu viittausmetriikka tuottavat 64 artikkelin ydinotoksen. Otoksen relevanssiseulonta toteutettiin monivaiheisella pipeline-ratkaisulla. Ensin multiagenttipohjainen LLM-konsensusalgoritmi (3 × itsenäinen GPT-tekoälymalli) tuotti ennalta rekisteröityihin sisällytys-/poissulkemiskriteereihin perustuvat binääripäätökset lukitussa JSON-formaatissa ja auttoi karsimaan aineistosta 878 epäolennaista artikkelia yhteistyössä ihmistarkistajan kanssa. Tämän jälkeen jäljelle jääneiden tekstien text-embedding-3-large-upotevektorit projisoitiin LocalMAP-algoritmilla ja stratifioitiin HDBSCAN-tiheysklusteroinnilla, mikä varmisti semanttisesti kattavan, tutkimusalueita edustavan, deterministisesti replikoitavan ja tilastollisesti validoidun 64 artikkelin ydinotoksen.

Tulokset osoittavat kolme selkeää kehitysvaihetta. Varhaisjaksolla (2011–2015) tutkimus tukeutui lähinnä perinteisiin tekstiluokittimiin ja käsin rakennettuihin piirteisiin; keskijakso (2016–2019) nosti benchmark-aineistot (LIAR, FakeNewsNet) ja syväoppivat CNN/RNN-mallit valtavirtaan; myöhäisjaksolla (2020–2024) transformer-arkkitehtuurit, multimodaaliset vision-language-mallit ja heterogeeniset graafineuroverkot yhdistivät tekstin, kuvan, videon ja verkostodynamiikan samaan luokitusputkeen. Samalla kuitenkin vain 14 % tutkimuksista testasi mallien ajallista kestävyttä, 13 % toimi aidosti monikielisesti ja 8 % raportoi eettisen vinouman mittareita – huolimatta EU:n AI-sääntelyn korostamista vastuullisuusvaatimuksista.

Tutkimus päättyy suosittelemaan (i) avoimien, monikielisten ja monimodaalisten benchmark-aineistojen laajentamista Euroopan ja globaalin etelän kielille, (ii) vakiintuneita kausaali- ja robustisuustestejä algoritmien käyttöönoton ehtona sekä (iii) pakollista vinouma-auditointia ja Model Card -raportointia alustojen sisältömoderoinnin läpinäkyvyyden varmistamiseksi. Näin tutkimus tukee politiikkatoimia, jotka tähtäävät demokratian resilienssiin ja algoritmisen agenda-asetannan vastuulliseen hallintaan.

Avainsanat: propagandan tunnistus, koneoppiminen, multimodaalisuus, monikielisyys, bayesilainen fuusiomalli, bibliometrinen analyysi

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

USE OF AI IN THESIS

I have utilised AI tools in my thesis:

- No
- Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

- OpenAI GPT-4.1 preview (gpt-4.1-2025-04-14)
- OpenAI GPT-o3 (o3-2025-04-16)
- Anthropic Claude Sonnet 3.7 ("extended-thinking" - release, Jan 2025)
- OpenAI text-embedding-3-large (3 072-dim.)
- fast-langdetect (FastText 2023-01)

Sections where AI tools were used:

- Chapter 3 Methodology (entire)
 - 3.1 Corpus collection & deduplication
 - 3.2 Data processing & clustering input
 - 3.2.1 Machine-aided relevance screening (LLM ensemble)
 - 3.3 Quantitative analysis (similarity & divergence metrics)

SISÄLLYSLUETTELO

1	JOHDANTO	1
1.1	Tutkimuksen tausta ja konteksti	2
1.2	Tutkimuksen teoreettinen, metodinen, käytännöllinen ja yhteiskunnallinen panos ja relevanssi	4
1.3	Tutkimuksen tavoite ja tutkimuskysymykset	7
1.4	Tutkimuksen rakenne	8
2	TUTKIMUKSEN TEOREETTINEN VIITEKEHYS	11
2.1	Propagandan käsite ja argumentatiiviset piirteet	11
2.2	Bayesilainen vastaanottajateoria	18
2.3	Anomaliadetektion periaatteet digitaalisessa vaikuttamisessa	19
2.4	Integroiva teoreettinen malli	20
3	METODOLOGIA	22
3.1	Aineiston keruu ja perusjoukon määrittely	22
3.2	Aineiston käsittely ja tarve uusille otantametoille	26
3.2.1	<i>Koneellinen relevanssiseulonta aiemmassa tutkimuksessa</i>	27
3.3	Aineiston analyysi	31
3.3.1	<i>Aineiston kvantitatiivinen analyysi</i>	31
3.3.2	<i>Aineiston kvalitatiivinen analyysi</i>	40
3.4	Metodologinen arvio ja johtopäätös	41
4	TUTKIMUKSEN TULOKSET JA ANALYYSI	44
4.1	Yleiskuva julkaisumääristä ja aineiston rakenteesta	44
4.2	Kronologinen katsaus tutkimuskirjallisuuden metodologiseen ja temaattiseen kehitykseen	46
4.2.1	<i>Varhaisjakso</i>	46
4.2.2	<i>Keskijakso</i>	50
4.2.3	<i>Myöhäisjakso</i>	63
4.3	Tulososion avainlöydökset	80
5	TULOSTEN ARVIOINTI, JOHTOPÄÄTÖKSET JA YHTEENVETO	84
5.1	Tutkimuksen keskeiset löydökset	85
5.1.1	<i>Kehityskaari 2011–2024: AI-menetelmien siirtyminen tekstipohjaisista luokittimista multimodaalisiin, verkostotietoihin ekosysteemeihin</i>	85
5.1.2	<i>Metodologis-eettinen kattavuus propagandan automaattisissa havaitsemismalleissa: seitsemän kriittistä ulottuvuutta ja demokraattiset katveet</i>	86
5.2	Tutkimuksen rajoitukset ja jatkotutkimusaiheet	89
5.3	Yhteenveto ja suosituksia	90
	LÄHDELUETTELO	92

1 JOHDANTO

Digitalisaation kiihtyminen ja sosiaalisen median alustojen globaalit käyttäjäverkot ovat muokanneet poliittisen viestinnän ekosysteemiä perusteellisemmin kuin yksikään aiempi teknologinen murros. Algoritmien räätälöimät uutisvirrat luovat kullekin käyttäjälle oman tiedollisen kuplan, jolloin yhteinen julkinen tila sirpaloituu pieniksi kohdeyleisöiksi. Tämän seurauksena koordinoitua, harhaanjohtavaa dis- ja misinformaatiokampanjia sekä intentionaalisia propagandaviestit leviävät sekunneissa ja ohittavat perinteisen portinvartijajournalismin suodattimet. Generatiivinen tekoäly – kuten tekstin ja kuvien tuottajat, äänisynteesi sekä deepfake-tekniikat – hämärtää entisestään rajaa aidon ja manipuloidun viestin välillä, mistä Yhdysvaltain vuoden 2024 vaalivuoden keskustelu syväväärennösten vaikutuksista oli konkreettinen esimerkki (esim. Bond, 2024; Brennen ym. 2025). Reaktiona samoihin riskeihin Euroopan unioni on nostanut poliittiset syväväärennökset ja propagandan levittämisen *korkean riskin* kategoriaan EU AI Act-säädöksessä (2024/1689) ja tiukentanut tekoälyjärjestelmien vastuukehikkoa AI-vastuudirektiivin päivityksillä (Rodríguez de Las Heras Ballell 2025, 1–23).

Tietojenkäsittelytieteessä koneoppimisen ja tekoälyn (AI) menetelmät ovat viimeisen viidentoista vuoden aikana nousseet keskeiseksi vastinpariksi tälle haasteelle (Da San Martino ym. 2020a; SuthanthiraDevi ym. 2020). Neuroverkko-pohjaiset luokittelijat, verkosto- ja käyttäytymisanalyysit sekä multimodaaliset syväoppimismallit lupaavat varhaisempaa ja tarkempaa propagandan detektiota, kuin perinteiset manuaaliset sisältö- ja diskurssianalyysit (Sahin ym. 2023; Da San Martino ym. 2020b). Samalla politiikan tutkimuksessa keskustellaan vallan digitalisoitumisesta, algoritmisesta agendan asetannasta ja kansalaisyhteiskunnan haavoittuvuudesta informaatio-operaatioille (Diakopoulos 2015, 398–415; Trielli ja Diakopoulos 2022, 45–70). Näiden kahden tutkimusperinteen välinen

menetelmällinen vuoropuhelu on hajanaista ja propaganda-käsitteen teoreettinen operationalisointi koneoppimismalleille on puutteellista. Toistaiseksi vain harvat tutkimukset ovat eksplisiittisesti yhdistäneet poliittisen vallankäytön käsitteelliset ulottuvuudet algoritmiseen mallinnukseen, mikä jättää avoimeksi kysymyksen siitä, miten havaittu propaganda kytkeytyy vaikuttamisen intentioihin, institutionaalisiin valtasuhteisiin ja normatiivisiin demokratia-ideaaleihin (Marwick ja Lewis 2017; Klinger ym. 2024).

Tämä Pro gradu –tutkimus pyrkii täyttämään edellä kuvattua aukkoa tarkastelemalla systemaattisesti, millä tavoin vertaisarvioidut tutkimukset vuosilta 2011–2024 hyödyntävät koneoppimista ja tekoälyä propagandan tunnistamisessa. Keskityn erityisesti tapauksiin, joissa propagandaintention käsite on eksplisiittisesti esillä. Tutkimus nojautuu bibliometriseen kartoitukseen, temaattiseen klusterointiin ja syventävään laadulliseen analyysiin, jotta se voi luokitella käytetyt teknologiat ja arvioida niiden yhteyden propaganda-käsitteen teoreettisiin ulottuvuuksiin (valkoinen–harmaa–musta propaganda, bayesilainen vastaanottajateoria, anomaliadetektion viitekehykset).

Yhteiskunnallisesti aihe on tällä hetkellä akuutti. Tuoreiden tapausten mukaan poliittiset toimijat eri puolilla Eurooppaa ovat jo hyödyntäneet generatiivista AI-kuvamateriaalia maahanmuuttovastaisten viestien vahvistamiseen, mikä on herättänyt lainsäätäjien ja viranomaisten huolen sisältömoderoinnin riittävydestä (Tondo 2025). Propagandan automaattinen havaitseminen on siis noussut demokratioiden kestävyyskriittiseksi osaksi niin akateemisessa kuin poliittis-hallinnollisessa keskustelussa.

1.1 Tutkimuksen tausta ja konteksti

Varhaisimmat koneoppimista hyödyntäneet propagandatutkimukset keskittyivät 2010-luvun alussa Twitter-bottien ja maksetun mielipidevaikuttamisen (astroturfing) paljastamiseen (Ratkiewicz ym. 2011, 297–304). Näissä töissä yhdistyivät verkosto-piirteet, sentimenttianalyysi ja joukkoistettu käyttäjäpalaute. Vuosikym-

menen puolivälissä tutkimuskenttä laajeni hyperpartisaanisten uutisten, rokkote-epäröintiä ruokkivan misinformaation ja kontekstuaalisten anomaliaketjujen analyysiin. Käännös syväoppimiseen tapahtui vuoden 2016 Yhdysvaltain presidentinvaalien jälkeen, jolloin Convolutional Neural Network (CNN)- ja Recurrent Neural Network (RNN) -pohjaiset mallit otettiin käyttöön valeuutisten varhaisessa tunnistuksessa (Liu ja Wu 2018, 354–361).

Generatiivisen tekoälyn (generative AI) nopea kehitys 2020-luvulla – erityisesti diffuusiomallit, suuret kielimallit (large language models, LLM) ja autoregressiiviset puhesynteesiverkostot – on synnyttänyt multimodaalisia uhkia poliittiselle ja yhteiskunnalliselle tiedonvälitykselle. Yksittäinen toimija voi nykyisin tuottaa kokonaisen vaikuttamisketjun kirjoittamalla poliittisen puheen LLM-järjestelmällä, jäljittelemällä ehdokkaan äänen tekstistä puheeksi -ratkaisulla (text-to-speech) ja yhdistämällä kokonaisuuden deepfake-videoksi. Tällaisessa kontekstissa propagandan havaitseminen edellyttää menetelmiä, jotka pystyvät analysoimaan tekstiä, kuvaa, videota ja ääntä samanaikaisesti ja tutkimaan näiden välistä ristikkäistä informaatiota (cross-modal attention) (Li ym. 2024).

Viimeaikaiset tutkimukset osoittavat nopean siirtymän kohti Vision-Language-Transformer -arkkitehtuureja (VLT) ja late-fusion-malleja multimodaaliseen misinformaatio- ja propagandantunnistukseen. VLT-pohjaiset luokittelijat (esim. BLIP2- ja CLIP-jatkokehitelmät) parantavat disinformaatiotietokenttien tarkkuutta 5–35 % verrattuna yksimodaalisiin ratkaisuihin (Raza ym. 2025, 1, 11–12). Mallien ytimessä on upotusten yhdistäminen samassa luokittelijassa – teksti, kuva ja usein myös metatiedot kulkevat yhteisen transformer-pinon läpi tai sulautetaan risti- ja tri-transformereilla (Al-Alshaqi ym. 2024, 4–5, 7–9; Qiao ym. 2025, 4–7, 8–13).

Politiikantutkimuksen puolella käydään rinnakkaista keskustelua informaatio- ja disinformaatiotietokenttien, kognitiivisista heuristiikoista ja algoritmista valvonnasta. Kuitenkin kentällä on pulaa tutkimuksista, jotka samalla kertaa mallintavat propagandan leviämisdynamiikkaa ja ankkuroivat havainnot vallan ja legitimitietin teorioihin.

Tämä konteksti perustelee monitieteisen tutkielman, joka tarkastelee sekä teknologisia ratkaisuja että niiden politologisia implikaatioita.

1.2 Tutkimuksen teoreettinen, metodinen, käytännöllinen ja yhteiskunnallinen panos ja relevanssi

Tutkimuksen teoreettinen panos on kaksijakoinen. Tutkimus kokoaa propagandan tutkimusta varten yhtenäisen viitekehyksen ja asettaa samalla empirisen ankkurin, jonka varaan tulevaa mallinnusta voidaan rakentaa. Ensimmäisessä vaiheessa työ kytkee toisiinsa propagandan retoriset tekniikat, bayesilaisen vastaanottajateorian ja anomaliadetektion näkökulman ja liittää tämän käsitteellisen synteessin tuoreisiin syvä- ja monimodaalisen tekoälyn virtauksiin. Näin se tarjoaa selkeän ”kartaston”, jossa retoriset houkuttimet (mitä propagandisti sanoo), vastaanottajan todennäköisyys-päivitys (miten viesti vaikuttaa) ja verkoston poikkeavuudet (miten kampanja levitetään) asettuvat samaan selityskehikseen.

Tämän viitekehyksen päälle tutkimus tuo kolme uutta rakennuspalikkaa nykyiseen tutkimuskenttään. Ensiksikin se kvantifioi ensimmäistä kertaa seitsemän menetelmäulottuvuuden pitkän aikavälin trendit ja osoittaa nykytutkimuksen strategisen ohuuden; sisältö-keskeisten mallien ylivallan ja koordinaation, kausaalisuuden sekä monikielisyysden katveet. Toiseksi tutkimus hahmottaa selkeän etenemismallin, jossa multimodaalinen sisältöanalyysi kytketään graafineuroverkkoihin perustuvaan verkosto- ja tietograafifaktantarkistukseen sekä adversaariin (generaattori–detektori) koulutukseen. Tavoitteena on kuroa umpeen kuilu yksittäisen väitteen luokittelun ja koordinoitujen propagandakampanjoiden vaikutusketjujen ymmärtämisen välillä. Kolmanneksi tutkimus tarkastelee seitsemää metodologis-eettistä ulottuvuutta, jotka aiempi kirjallisuus on nostanut propagandan havaitsemisen kipupisteiksi (Da San Martino ym. 2020b; Shu ym. 2019a, 312–320). Näiden ulottuvuuksien täsmälliset määritelmät on esitetty taulukossa 1, ja ne toimivat sekä koodaus- että trendianalyysin selkärankana.

Taulukko 1: Propagandan automaattisen havaitsemisen 7 keskeistä metodologis-eettistä ulottuvuutta

Ulottuvuus	Operatiivinen määritelmä	Esimerkki mittarista
Sisältökeskeisyys	Malli käyttää vain sanasto-/kuvapikselipiirteitä.	BERT-luokitus uutistekstille
Multimodaalisuus	Malli fuusioi ≥ 2 media-tyyppiä (teksti, kuva, ääni, video).	CLIP-pohjainen V-L-transformeri
Koordinaatio-analyysi	Verkosto- tai käyttäytymis-piirteet mukana.	GNN retweet-graafille
Kausaalisuus	Malli estimoii todellisia vaikutuksia (ATE, CATE tms.).	Propensity-score-matched GNN
Temporal-robustisuus	Arvioidaan suorituskyky eri aikaleikkauksissa.	Rolling-window F1
Monikielisyys	Testataan ≥ 2 kielellä ilman käännöstä.	XLM-R + MuMiN-aineisto
Eettinen vinouma	Tarkastellaan mallin virhe-eroja ryhmien välillä.	Demografinen parity-gap

Tutkimuksen metodologinen panos rakentuu uudesta, kolmiportaisesta ja sekamenetelmällisestä systemaattisen katsauksen kehyksestä (Exploratory Sequential Mixed-Trend-Tracing, ES-MTT), joka muodostaa tutkielman empiirisen selkärangan. Kehyksen ensimmäisessä vaiheessa bibliometrinen kartoitus – toteutettuna text-embedding-3-large -upotuksilla, k-means-klusteroinnilla ja vaimennetulla viittauspainotuksella – paljastaa tutkimuskentän semanttiset ja ajalliset kohoumat. Toisessa vaiheessa strukturoitu fokusoitu vertailu (SFC) koodaa kunkin artikkelin seitsemän metodologiaulottuvuutta (taulukko 1) binääri- tai ordinaalimuuttujiksi. Kolmannessa vaiheessa ES-MTT yhdistää nämä laadulliset koodit määrälliseen trendimalliin ja tuottaa pitkittäisen, tilastollisesti toistettavan aikasarjan. Kokonaisuus tarjoaa replikoitavan työkalupaketin, joka soveltuu propagandatutkimuksen lisäksi muille nopeasti kehittyville AI-aloille, joissa julkaisuvolyymi ylittää perinteisen narratiivisen katsauksen kapasiteetin. Menetelmän avulla kentän kehitys kvantifioidaan ensimmäistä kertaa seitsemällä kriittisellä ulottuvuudella – esimerkiksi sisällön, multimodaalisuuden ja koordinaation suhteen – ja tulokset sidotaan operatiivisesti samaan teoreettiseen viitekehykseen, minkä lisäksi lukijoille tarjotaan avoin skriptipohja jatkoanalyysyä varten.

Käytännöllisellä tasolla tutkimus tarjoaa luotettavan tilannekuvan ja siihen perustuvan, data-ankkuroidun riskikartan propagandan torjunnan kehitystyölle. Tarkkaan valittujen 64 artikkelin analyysi konkretisoi, missä tutkimus on edistynyt ja missä se laahaa jäljessä. Tämä empiirinen perustieto voi ohjata resursointia ja priorisointia niin yritysten kuin viranomaistenkin TKI-yksiköissä. Lisäksi tutkimus yhdistää vahvan empiirisen analyysin ja selkeästi artikuloidun jatkokehityssagen- dan, joka voi suunnata tulevaa järjestelmällistä, läpinäkyvää ja eettisesti kestäväää teknologiatutkimusta ja -kehitystä.

Yhteiskunnallisella tasolla tutkimus tuottaa ennen kaikkea faktapohjaisen reali- teettitarkistuksen digitaalisen informaatio-ympäristön nykyisistä suojamekanis- meista. Empiirinen havainto, jonka mukaan lähes kaikki nykymallit keskittyvät yk- sittäisten väitteiden tekstuaaliseen todenmukaisuuteen samalla, kun koordi- noidun vaikuttamisen, ajallisen kestävyuden ja monikielisten yleisöjen haasteet jäävät marginaaliin, toimii varoitus-signaalina päätöksentekijöille: pelkkä ”valeuu- tisen” detektointi ei riitä turvaamaan demokraattista keskustelua. Tulokset tuke- vat siten sääntely- ja alustapolitiikan siirtymää kohti laajempaa järjestelmätason riskienhallintaa, jossa huomioidaan disinformaatioketjun kaikki tasot – sisällöstä verkoston dynamiikkaan ja algoritmiseen suositteluun.

Samalla tutkimus nostaa eettisen vinouman ja monikielisen kattavuuden puutteet näkyviksi ja tarjoaa niille alustavan auditointikehyksen. Tämä antaa journalistisille faktantarkistus-organisaatioille ja kansalaisyhteiskunnalle välineet valvoa, ettei automaattinen disinformaatio suoja kallistu suurten kielialueiden ja valtamedioi- den eduksi. Tuomalla esiin strategisen ohuuden riskin sekä eksplisiittisen tarpeen temporal-robusti-mittauksille tämä Pro gradu –tutkimus vahvistaa myös kriittisen medialukutaidon narratiivia: se osoittaa kansalaisille ja tiedotusvälineille, miksi nopeasti kehittyvä tekoälyteknologia vaatii rinnalleen jatkuvaa inhimillistä arvioin- tia ja monialaista valvontaa. Näin tutkimus edistää demokraattista resilienssiä – se tukee todisteisiin nojaavaa julkista keskustelua, informoi AI-politiikan valmis- telua ja tarjoaa pohjan tasapainoisemmalle, kulttuurisesti sensitiiviselle disinfor- maation torjuntastrategialle.

1.3 Tutkimuksen tavoite ja tutkimuskysymykset

Tämän Pro gradu -tutkimuksen tavoitteena on politiikan tutkimuksen näkökulmasta (i) jäsentää systemaattisesti, miten tekoäly- ja koneoppimismenetelmät ovat muokanneet poliittisen propagandan automaattista tunnistamista digitaalisessa julkisuudessa vuosina 2011–2024, sekä (ii) arvioida, missä määrin nämä menetelmät pystyvät vastaamaan demokraattisen informaatioympäristön sääntely- ja legitimizeettihaasteisiin. Aineistona toimii 64 vertaisarvioidun tai laajasti viitatuun artikkelijulkaisun kokonaisuus, jonka menetelmälliset piirteet on kaksoiskoodattu määrällistä sisällönanalyysiä varten. Tämä empiirinen otos yhdistyy teoreettiseen synteesiin, jossa propagandan retoriset tekniikat, bayesilainen vastaanottajateoria ja anomaliadetektion periaatteet asetetaan vuoropuheluun syvä- ja multimodaalisen tekoälyn uusimpien virtauksien kanssa. Tarkoituksena on paitsi tuottaa data-ankkuroitu tilannekuva tutkimuskentän kehityksestä, myös luoda analyyttinen kehikko, jonka avulla voidaan arvioida menetelmien kattavuutta, kestävyys-ominaisuuksia ja eettistä tasapainoa.

Tutkimus pyrkii vastaamaan seuraaviin tutkimuskysymyksiin:

TK1: Miten koneoppimisen ja tekoälyn menetelmät ovat kehittyneet propagandan tunnistamisessa vuosina 2011–2024?

TK2: Missä määrin vuosina 2011–2024 julkaistut tekoäly- ja koneoppimismallit kattavat propagandan automaattisen havaitsemisen seitsemän kriittistä metodologis-eettistä ulottuvuutta – ja missä ovat suurimmat katveet demokraattisen päätöksenteon näkökulmasta?

Ensimmäinen kysymys (TK1) kohdistuu tutkimuskentän pitkittäiseen kartoitukseen. Se tarkastelee julkaisuvuodet 2011–2024 kattavaa aineistoa sekvenssi-analyysin ja ES-MTT-kehiksen (Exploratory Sequential Mixed-Trend-Tracing) avulla ja osoittaa, miten koneoppimisen sekä tekoälyn menetelmät ovat muuntu-neet tekstuaalisista n-gram-luokittimista kohti syväoppivia, monimodaalisia arkkitehtuureja. Laskennalliset trendilinjat paljastavat teknologisen painottumisen

muutokset, joiden kautta voidaan arvioida tutkimusalueen kykyä reagoida uudenslaisiin dis- ja misinformaation muotoihin. TK1 tarjoaa siten historiallisen viitekehksen, jonka varassa TK2:n syventävät tarkastelut voidaan sijoittaa oikeisiin pitkän aikavälin konteksteihin.

Toinen kysymys (TK2) porautuu syvemmälle analysoimalla, kuinka hyvin tämän periodin malleissa otetaan huomioon propagandan automaattisen havaitsemisen seitsemän kriittistä metodologis-eettistä ulottuvuutta (sisältökeskeisyys, multimo- daalisuus, koordinaatioanalyysi, kausaalinen vaikutusarviointi, temporal-robusti- suus, monikielisyys ja eettinen vinouma). Binääri- ja ordinaalimuuttujiksi koodatut indikaattorit mahdollistavat profiloinnin, joka paljastaa sekä edistyneet alueet (esim. syväoppivan tekstianalyysin läpimurrot) että selkeät katveet (koordinaatio- dynamiikan sekä monikielisten kontekstien vähäinen huomio). Tämä kartoitus kytkeytyy demokraattisen valvonnan näkökulmaan osoittamalla, missä kohdin nykymallit jättävät riskialttiita aukkoja esimerkiksi vaalivaikuttamisen tai vähem- mistökielten osalta.

Yhdessä TK1 ja TK2 muodostavat komplementaarisen kokonaisuuden: ensiksi kuvaillaan kentän evolutionaarinen perusteknologia ja toiseksi arvioidaan laadul- lisen kattavuuden tasapaino teknologian, eettisten vaatimusten ja demokraatti- sen vallankäytön sääntelytarpeiden ristipaineessa. Näin tutkimus tuottaa sekä empiirisesti ankkuroidun tilannekuvan että normatiivisesti relevantin arvioinnin propagandan automaattisen havaitsemisen tämänhetkisistä valmiuksista.

1.4 Tutkimuksen rakenne

Tämä Pro gradu -tutkimus pyrkii etenemään argumentatiivisesti loogisena ket- juna, jossa jokainen luku rakentuu edeltäjiensä varaan. Johdannon jälkeen toinen luku syventyy tutkimuksen teoreettiseen viitekehukseen. Siinä esitän, kuinka pro- pagandan retoriset tekniikat, bayesilainen vastaanottajateoria ja verkostopohjai- nen anomaliadetektion tutkimuskirjallisuus voidaan punoa yhteen johdonmu- kaiseksi kokonaisuudeksi, joka kantaa koko työn läpi. Tavoitteena on osoittaa,

miten samat käsitteelliset rakenteet – viestien argumentatiivinen sisältö, vastaanottajan todennäköisyyspäivitys ja viestin poikkeava leviäminen – muodostavat yhteismitallisen arviointikehikon myös koneoppimismalleille.

Kolmas luku siirtää tarkastelun tutkimusasetelmaan. Siinä kuvaan systemaattisen kirjallisuushaun ja artikkelien valintaprosessin sekä perustelen, miksi tarkastelujaksoksi valikoitui vuodet 2011–2024. Selvennän, millä kriteereillä 64 vertaisarvioitua tutkimusta otettiin mukaan, miten ne kaksoiskoodattiin ja millä tavoin bibliometriset, klusteripohjaiset ja laadulliset menetelmät kytkettiin toisiinsa. Luvun päätteeksi osoitan, kuinka koodauskehyksen jokainen ulottuvuus vastaa toista luvussa kaksi määriteltyä teoreettista komponenttia; näin teoria ja menetelmä sidotaan tutkimuksessa eksplisiittisesti toisiinsa.

Neljäs luku esittää tutkimuksen empiiriset tulokset. Aloitan julkaisudynamiikan pitkittäiskuvauksella, josta siirryn menetelmällisten painopisteiden kvantitatiiviseen analyysiin. Teen näkyväksi esimerkiksi sen, kuinka syväoppiminen ja multimodaaliset lähestymistavat nousivat valtavirtaan vuosikymmenen lopulla samalla, kun koordinaatiota mittaavat verkostomallit menettivät suosiotaan. Tulosten toisessa osassa sukellan laadullisiin teemoihin ja havainnollistan löydöksiä tapausparien ja malliesimerkkien avulla. Näin tulosluku ei jää pelkän numerodiskurssin varaan, vaan osoittaa konkreettisesti, missä tilanteissa nykyiset algoritmit onnistuvat ja missä ne kompastuvat.

Viidennessä luvussa pyrin käymään kriittisen vuoropuhelun tulosten ja tutkimuskysymysten välillä. Erottelen, missä määrin havaitut aukot ovat metodologisia, missä teoreettisia ja missä eettisiä, sekä suhteutan löydökset aiempaan tutkimuskirjallisuuteen. Samalla tarkennan, miten tutkielman esittämä konvergenssimalli vie keskustelua eteenpäin sekä tietotekniikan että politiikan tutkimuksen kentissä.

Kuudes luku kokoaa tutkimuksen johtopäätökset. Siinä teen yhteenvedon tutkielman keskeisestä tieteellisestä ja yhteiskunnallisesta kontribuutiosta, arvioin työn rajoituksia ja hahmottelen kolmitasoisen jatkotutkimusagendan, joka ulottuu teoreettisesta syventämisestä aina konkretiaan - multimodaalisen benchmark-ai-

neiston rakentamiseen, temporal-robustisuuden mittaristoon ja eettisen viinon auditointikehikkoon. Luku päättää tekstin osoittamalla, miten tulokset voivat tukea sekä akateemisia, regulatorisia että soveltavia pyrkimyksiä vahvistaa demokraattista resilienssiä digitaalisen propagandan aikakaudella.

2 TUTKIMUKSEN TEOREETTINEN VIITEKEHYS

2.1 Propagandan käsite ja argumentatiiviset piirteet

Propagandan käsite juontaa juurensa 1600-luvulle, jolloin katolinen kirkko perusti Congregatio de Propaganda Fide -nimisen elimen edistämään uskon levittämistä (latinan *propagare* = levittää) (Zurstiege 2016, 29; Britannica 2025). Teollistumisen ja massapolitiikan myötä käsitteestä tuli poliittinen avaintermi, kun Ranskan suuren vallankumouksen pamflettikulttuuri, 1800-luvun työväenliike sekä Leninin agit-prop-ajattelu nostivat järjestelmällisen mielipidevaikuttamisen vallankumouksellisen toiminnan ytimeen (Lenin 1961). Maailmansotien aikana propaganda institutionalisoitui valtiolliseksi ja sai kielteisen kaiun, vaikkakin termin arvio vaihteli kulttuuri- ja kielialueittain (Britannica 2025). Käsite on siis jo varhaisista ajoista lähtien merkinnyt suunnitelmallista yritystä muokata suurten yleisöjen uskomuksia.

Nykyaikaisessa politiikan tutkimuksessa propaganda on kiistanalainen käsite, joka liittyy usein valheen, vääristelyn, manipuloinnin, aivopesun ja huijauksen käsitteisiin, mutta ei palaudu mihinkään niistä. Jowett ja O'Donnell (2015, 1–2) määrittelevät propagandan *”tarkoitukselliseksi ja järjestelmälliseksi pyrkimykseksi muokata vastaanottajan havaintoja ja asenteita tavalla, joka edistää propagandistin päämäärää”*. Uusimmassa Kekkosen (2025) väitöskirjassa täsmennetään, että modernissa kansainvälisessä tutkimuksessa propaganda määritellään ennen kaikkea poliittisen vallan välineeksi, joka pyrkii järjestelmällisesti muuttamaan kohdeyleisön mielipideympäristöä. Väitöskirja huomioi seitsemän olennaista propagandan kriteeriä:

1. Yksipuolinen päämäärä. Propagandan lähtökohta on yhdensuuntainen vallankäyttö: viesti on suunniteltu tuottamaan nettohyötyä yksinomaan viestin lähettäjälle. Vastaanottajalle tarjoutuva informaatioarvo on korkeintaan sivutuote, usein jopa harhautus, joka hämärtää viestin todellista, lähettäjän intressiä ja sulkee pois aidon, vastavuoroisen deliberoinnin.
 2. Motiivien peittäminen. Propagandisti naamioi todellisen poliittisen tai ideologisen tavoitteensa esittämällä viestinsä näennäisesti neutraaleina, rationaalisina tai jopa altruistisina perusteluina. Tarjottu viestinnän kehystys toimii siten strategisena suojakilpenä, joka hälventävät kriittistä tarkastelua ja siirtävät huomiota pois vallankäyttöön tähtäävistä päämääristä.
 3. Systemaattinen vaikuttaminen. Propaganda on luonteeltaan organisoitua. Se toteutetaan toistuvilla, toisiinsa linkittyvillä viesteillä, joita levitetään monikanavaisesti ja mukautetaan palautteen perusteella. Tarkoituksena on muokata kohdeyleisön uskomuksia ja käyttäytymistä pitkäjänteisesti, ei satunnaisesti.
 4. Taustalla poliittinen/ideologinen toimija. Lähettäjänä on valtakeskittymä, esimerkiksi puolue, valtio, eturyhmä tai näiden nimissä toimiva verkosto, jolla on resurssit kampanjan koordinoimiseen. Yksityishenkilön sattumanvarainen manipulointi ei täytä propagandan ehtoja ilman kytkeytymistä laajempaan poliittiseen projektiin.
 5. Episteeminen puutteellisuus. Viesti on tiedollisesti viallinen: se sisältää valheita, tarkoituksellisia virheellisiä johtopäätöksiä, asiayhteyksistä irrotettuja faktoja tai epäeettisiä argumentaatiokeinoja. Tavoitteena ei ole totuuden välittäminen vaan vastaanottajan käsityskyvyn ohjailu.
 6. " Tietoinen vika". Viestijä tietää, tai kohtuullisella olettamuksella hänen olisi pitänyt tietää, viestinsä episteemiset puutteet. Tietoinen harhautus poistaa mahdollisuuden vedota vilpittömään erehdykseen ja asettaa vastuun manipuloivalle toimijalle.
 7. Kielteinen ilmiö. Propaganda kaventaa julkista keskustelua ruokkimalla ennakkoluuloja, yhdenmukaistamalla ajattelua ja vaijentamalla toisinajattelua. Muun muassa näistä syistä, sitä pidetään demokratialle systeemisesti haitallisena ilmiönä, joka heikentää kriittisen, pluraarisen deliberatiivisen yhteiskunnallisen tilan edellytyksiä. (Kekkonen 2025, 70–73).
- Määritelmät korostavat propagandan intentionaalisuutta ja viestinnän järjestelmällisyyttä ja strategisuutta. Viestin totuusarvo voi vaihdella, kunhan lopputulos

palvelee propagandistia. Propagandan tunnistaminen myös edellyttää viestin episteemisen puutteellisuuden arviointia, ei pelkkää valheiden lukumäärää. Myös propagandistin järjestelmälliset toimintatavat ja niitä ohjaavat ideologiset näkemykset voidaan tulkita toimivan ulkoisesti havaittavina indikaattoreina, joiden perusteella propagandistinen viestintä voidaan analyttisesti tunnistaa (Kekkonen 2025, 71–72).

Myös Stanley (2015) osoittaa, että liberaaleissa demokratioissa propaganda voi normalisoida epäoikeudenmukaisia hierarkioita esittämällä selektiivisiä totuuksia ja jättämällä olennaisen kontekstin kertomatta (Stanley 2015, 8). Tätä jäsentää usein siteerattu propagandan valkoinen–harmaa–musta -jaottelu. Valkoinen propaganda on avoimesti tunnustettua ja tosiasiapohjaista, musta propaganda peittää lähteensä ja levittää valheita, harmaa sekoittaa molempia (Jowett ja O'Donnell 2015, 25–28).

Retoriset tekniikat ja argumentointivirheet muodostavat propagandan mikrotason työkalupakin. Dimitrov ym. (2021) osoittivat, että nykyaikaiset verkossa kiertävät meemit hyödyntävät systemaattisesti esimerkiksi mustamaalausta, olkiukkoja, kausaalisia yksinkertaistuksia ja auktoriteettiin vetoamista. Da San Martino ym. (2019) laajensivat tämän typologian kahdeksaantoista tekniikkaan ja rakensivat propagandan tekniikat -aineiston, josta on tullut koneoppimismallien keskeinen opetusstandardi. Taulukko 2 kokoaa yhteen aiemman tutkimuksen tunnistamia propagandatekniikoita (Da San Martino ym. 2019; Das ym. 2025; Li ym. 2020, 123–145; Mouton ym. 2025, 112–120).

Käytännössä propaganda toteutuu konkreettisina retorisinä tekniikoina, joita toistetaan eri medioissa. Tutkimuskirjallisuudesta on tunnistettavissa noin parikymmentä toistuvaa keinoa, mutta luokittelut vaihtelevat. Jotta analyysi olisi hallittava ja silti kattava, ryhmittelen tekniikat kolmeen funktioon perustuvaan pääluokkaan: tunteisiin vetoavat keinot, delegitimaatiotekniikat ja kognitiiviset oikopolut. Taulukko 2 esittää tämän jaottelun ja kokoaa kaikki 20 tekniikkaa lyhyine selityksineen:

Taulukko 2: *Propagandatekniikat ryhmiteltyinä psykologisen vaikutusfunktion mukaan*

1. Tunteisiin vetoavat keinot

- Tunnepitoinen kieli - tunnepitoiset sanat, joilla herätetään myönteisiä/kielteisiä reaktioita
- Glitteri-iskulauseet – kiistattomilta kuulostavien hyvesanojen toistaminen ilman konkretiaa
- Pelkoon/ennakkoluuloihin vetoaminen - uhka- ja viholliskuvien rakentaminen
- Isänmaallisuuteen vetoaminen - kansallistunteen nostattaminen teon oikeutukseksi
- Kansanomainen samaistuminen – kansanomaisuus luottamuksen herättämiseksi ja kritiikin hiljentämiseksi
- Toisto-efekti - sanoman toistaminen, kunnes se koetaan todeksi
- Iskulauseet - iskevät, tunnepitoiset fraasit

2. Delegitimaatiotekniikat

- Nimittely / leimaaminen - vastustajan leimaaminen halveksittavaksi, pelottavaksi tms.
- Mustamaalaus - vastustajan maineen mustamaalaamista ilman todisteita
- Epäilyn kylväminen - kohteen uskottavuuden kyseenalaistaminen
- Whataboutismi - hypoteettinen vastasyytös ilman alkuperäisen väitteen kumoamista
- "Hitler-kortti" - liitetään vastustaja vihattuun ideologiaan
- Red herring - huomion johtaminen sivupolulle

3. Kognitiiviset oikopolut

- Laumavietti (bandwagon) - kaikki muutkin ajattelevat/tekevät, liity mukaan
- Kausaalinen yksinkertaistus - monisyisen ilmiön selittäminen yhdellä syyllä
- Mustavalkoinen vaihtoehdottomuus - esitetään vain kaksi vaihtoehtoa, pakotetaan valitsemaan
- Olkiukko - vastustajan kanta karikatyrisoitu ja kumottu
- Auktoriteettiin vetoaminen - väite totta vain siksi, että asiantuntija sanoo niin
- Ajattelua katkaiseva klisee - fraasi, joka pysäyttää kriittisen keskustelun
- Liioittelu / vähättely - asian esittäminen paljon suurempana tai pienempänä kuin se on
- Hämäys / epämääräisyys - tahallinen monitulkintaisuus ja sekoittaminen

Taulukko osoittaa, että propagandan mikrotason työkalut kytkeytyvät kolmeen psykologiseen vipuun: tunteiden herättämiseen, vastustajan uskottavuuden murentamiseen ja kuulijan päättelypolkujen oikaisemiseen. Tämä ryhmittely helpottaa myöhempää analyysiä: tunteisiin vetoavat keinot näkyvät useimmiten sanaston tunnepolariteettina, delegitimaatiotekniikat personoituina hyökkäyksinä ja

kognitiiviset oikopolut argumentointivirheinä, jotka voidaan mallintaa rakenteellisinä piirteinä (esimerkiksi kausaalinen ketju, binäärinen vaihtoehtoisuus). Waltonin (2007, 102–128) pragma-dialektinen teoria selittää näiden keinojen käytännön tehon.

Samalla on tärkeää erottaa toisiinsa limittyvät käsitteet. Kekkonen (2025, 64–68) käyttää informaatiovaikuttaminen-käsitettä sateenvarjona tahalliselle vaikuttamiselle. Sen alalajeja ovat propaganda (kuvallinen, audiovisuaalinen tai sanallinen) ja sen sisällä oleva verbaalinen disinformaatio. Kun kolmas osapuoli levittää näitä viestejä tietämättään, väitöskirja puhuu tahattomasta informaatiovaikuttamisesta, jonka voidaan tulkita vertautuvan vertakansainvälisessä kirjallisuudessa käytettyyn 'misinformaatio' ('misinformation') -käsitteeseen. Misinformaatiota käsitellään tutkimuskirjallisuudessa useimmin virheellisenä tai harhaanjohtavaa sisältöä, joka leviää kuitenkin ilman levittäjän tietoista aikomusta informaatiovaikuttamiseen.

Propaganda on järjestelmällinen pyrkimys muokata vastaanottajan uskomuksia tai toimintaa tavalla, joka palvelee viestijän vallankäyttöä; se voi nojata niin valheisiin kuin valikoituihin tosiasioihin ja käyttää taulukon kaksi tekniikoita. Näin ollen kaikki propaganda on intentionalista, mutta kaikkea intentionalista virheviestintää ei voi pitää propagandana, ellei mukaan liity valta- ja koordinoituvuutta. Tämä erottelu toimii jatkossa viitekehyksenä, kun siirrymme tarkastelemaan propagandan vaikutusmekanismeja Bayes-teorian ja anomaliadetektion näkökulmista.

Bayesilainen vastaanottajateoria konkretisoi tämän mekanismin. Viestin toistuva altistaminen, tunne-pitoiset sloganit ja auktoriteettiviittaukset vaikuttavat joko viestin uskottavuuteen (likelihood) tai suoraan vastaanottajan enakkokäsityksiin, jolloin viestin vastaanottaminen vahvistuu vähäisestäkin evidenssistä (Kamenica ja Gentzkow, 2011).

Tämä synnyttää kumulatiivisen 'infodemic-efektin', jossa samansuuntaiset viestit satureivat informaatiotilan ja heikentävät kykyä arvioida vaihtoehtoisia selityksiä (WHO, 2020).

Propaganda on siis tarkoituksellinen, järjestelmällinen vallankäytön muoto, jossa lähettäjä hyödyntää retorisia virheitä ja psykologisia vinoumia ohjatakseen laajojen yleisöjen uskomuksia ja toimintaa. Viestin totuusarvo ei ratkaise propagandan tunnistusta; olennaisia ovat manipuloiva tavoite, retoriset taktiikat ja avoin tai peitetty lähdesuhde. Yksittäinen virhe tai epätosi väite ei vielä tee viestistä propagandaa, ellei sitä tueta strategisella kontekstilla ja toistolla. Propagandan analyysi vaatii siksi sisällön, kontekstin ja tarkoituksensa yhtäaikaista tarkastelua – se on perusta, jolle alaluvut 2.2 ja 2.3 rakentuvat.

Avaan seuraavaksi käsitteitä, jotka selventävät tutkimusalueen tematiikkaa:

Käsitelaatikko:
<ul style="list-style-type: none"> • Propaganda: tarkoituksellinen ja järjestelmällinen viestintä, jonka päämäärä on muokata yleisön uskomuksia tai toimintaa puhujan valta-asemaa hyödyttävällä tavalla.
<ul style="list-style-type: none"> • Disinformaatio: väärä tai harhaanjohtava sisältö, jota levitetään tietoisesti.
<ul style="list-style-type: none"> • Misinformaatio: väärä tai harhaanjohtava sisältö, joka leviää ilman levittäjän tietoista aikomusta.
<ul style="list-style-type: none"> • Valkoinen / harmaa / musta propaganda: avoin / osittain peitelty / täysin salattu/valheellinen propagandakirjo.
<ul style="list-style-type: none"> • Retorinen tekniikka: puheen tai tekstin muotoilukeino (esim. tunnesanat, nimittely, olkiukko), jolla viestijä ohjaa vastaanottajan tulkintaa.
<ul style="list-style-type: none"> • Argumentointivirhe: päättelyn muoto, joka on loogisesti heikko mutta psykologisesti vakuuttava (esim. auktoriteettiin vetoaminen).
<ul style="list-style-type: none"> • Kognitiivinen oikopolku (heuristiikka): mielensisäinen 'nopea sääntö', jota propaganda voi hyödyntää (esim. laumavaikutus, toisto-efekti).
<ul style="list-style-type: none"> • Framing: tavanomaisesti neutraalin ilmiön esittäminen valikoidussa tulkintakehyksessä, joka ohjaa merkityksenantoa.

<ul style="list-style-type: none"> • Agenda-setting: prosessi, jossa viestijä määrittelee, mitä teemoja yleisö pitää tärkeinä (topic-valinta, julkinen aikataulutus).
<ul style="list-style-type: none"> • Algoritminen vahvistus (algorithmic amplification): alustan suosittelujärjestelmän mekanismi, joka nostaa tiettyjä viestejä laajempaan näkyvyyteen ja voi vahvistaa propagandan leviämistä.
<ul style="list-style-type: none"> • Astroturffaus: maksettu tai organisoitu "ruohonjuuritason" kampanja, joka naamioidaan spontaaneiksi kansalaisviesteiksi.
<ul style="list-style-type: none"> • Echo chamber / kuplautunut yleisö: informaatioympäristö, jossa käyttäjä altistuu lähes yksinomaan samanmieliselle sisällölle.
<ul style="list-style-type: none"> • Infodemia: tietotulvan ja harhaanjohtavan sisällön yhdistelmä, joka vaikeuttaa luotettavan tiedon löytämistä (WHO).
<ul style="list-style-type: none"> • Syvävääreännös (deepfake): toimija- tai sisällöllinen vääreännös, joka tuotetaan generatiivisilla neuroverkoilla (esim. video, ääni).
<ul style="list-style-type: none"> • Temporal-robustus: mallin kyky säilyttää tarkkuutensa ajan myötä, vaikka viestintätavat ja –teknologiat muuttuvat.
<ul style="list-style-type: none"> • Monimodaalinen malli: koneoppimisjärjestelmä, joka analysoi useita mediamuotoja (teksti, kuva, ääni, video) yhteisessä kehyksessä.
<ul style="list-style-type: none"> • Koordinoitu käyttäytyminen: useiden tilien tai lähteiden tahallinen synkronointi propagandallisen viestin tehostamiseksi.
<ul style="list-style-type: none"> • Bayes-vastaanottajateoria: näkökulma, jossa viestin vaikutus kuvataan alkuoletusten (prior) ja päivityksen (likelihood → posterior) kautta.
<ul style="list-style-type: none"> • Adversaarinen ekosysteemi: tilanne, jossa viestijä (generaattori) ja tunnistusmalli (detektori) kehittävät keinoja kilpaillen toistensa kanssa.
<ul style="list-style-type: none"> • Eettinen vinouma: malli- tai data-asetelma, joka systemaattisesti syrjii kieli-, kulttuuri- tai vähemmistöryhmiä propaganda-detektioprosessissa.

2.2 Bayesilainen vastaanottajateoria

Bayesin laki tarjoaa formaalin tavan kuvata, miten uusi viesti muuttaa vastaanottajan uskomusta. 'Prior' kuvaa alkuperäistä arviota väitteen todennäköisyydestä, 'likelihood' sitä, kuinka uskottavalta viesti näyttää, jos väite on tosi ja 'posterior' on päivitetty uskomus viestin havaitsemisen jälkeen. Poliittisessa viestinnässä jokainen viesti voidaan tulkita kokeeksi, jossa lähettäjä yrittää siirtää vastaanottajan posterioria suuntaan, joka palvelee omia tavoitteitaan. Toisto-efekti (Fazio ym. 2015) kasvattaa koettua likelihoodia, konfirmaatiobias (Pennycook ja Rand 2019, 39–50) suurentaa prioria ja affect-heuristiikka (Slovic 2010, 397–420) värittää viestin emotionaalisesti. Kaikki nämä vääristävät päivitysprosessia ja selittävät, miksi ”psykologisesti vakuuttavat, mutta loogisesti virheelliset” argumentointivirheet tehoavat (Walton 2007, 244–245).

Propagandan kannalta tärkeää on ymmärtää, yrittääkö lähettäjä muuttaa kuulijan valmiiksi omaksumia käsityksiä (prior) vai pyrkii hän vain saamaan itse viestin näyttämään uskottavalta riippumatta sen totuudesta (likelihood). Esimerkiksi olkiukko-tekniikka vääristää viestin sisältöä, glitteri-iskulauseet saavat sen tuntuun tunteisiin vetoavan vakuuttavalta ja whataboutismi tuottaa niin paljon sivuaiheista kohinaa, ettei kuulija enää pysty tarkistamaan väitteen paikkaansa pitävyyttä. Retoriset tekniikat voidaan esittää selkeinä muutoksina viestin uskottavuuteen, ja ne on siten myös mahdollista mallintaa.

Ensimmäiset valeuutisfiltrit perustuivat naiiviin Bayesiin, jossa sanat oletettiin lähes riippumattomiksi (Zhou ja Zafarani 2020). Transformer-mallit voi tulkita hierarkkiseksi bayesilaisiksi approksimaatioiksi, joissa token-vektorit kuvaavat latenttia jakaumaa ja itseattentio hoitaa päivityksen (Rogers ym. 2021). Bayesian neuroverkot kuvaavat painojen epävarmuutta ja erottavat, mikä osa varmuudesta puuttuu tiedon vähyyden vuoksi ja mikä satunnaisuuden tähden. Tämä lisää luotettavuutta esimerkiksi deepfake-tunnistuksessa (Maier & Riess 2024, 5–6, 12–13). Human-in-the-loop-lähestymistavat tuovat prosessin näkyviin käyttäjälle. Mishra ym. (2023, 6) esittävät lämpökarttoja, jotka havainnollistavat, miten mallin arviot muuttuvat uusien todisteiden jälkeen. Käyttäjä voi korostettujen riskifraasien perusteella säätää lähtöoletuksiaan ja kokeilla, kuinka herkkä malli on muutoksille.

Bayes-lähestymistapa integroituu luontevasti anomaliadetektion kanssa. Kun useista lähteistä virtaa signaaleja, posteriori voidaan tulkita verkon odotetuksi reunapainoksi. Tästä poikkeavat piikit paljastavat koordinoitun vaikuttamisen. Tämä linkki muodostaa teoreettisen sillan seuraavaan alalukuun, jossa tarkastellaan poikkeamien kvantifiointia sosiaalisissa verkoissa.

2.3 Anomaliadetektion periaatteet digitaalisessa vaikuttamisessa

Anomaliadetektion perusidea on yksinkertainen: jos havainto poikkeaa odotetusta jakaumasta, sillä voi olla selittävä arvo. Anomaliadetektion tutkimus erottellee kolme anomalian perusmuotoa. Pisteanomalia syntyy, kun yksittäinen havainto on epätavallinen, kontekstianomalia tulee esiin vain tietyssä ajassa, paikassa tai aihepiirissä ja kollektiivianomalia ilmenee vasta usean havainnon yhteistoimintana. Sosiaalisen median kontekstissa poikkeamat voidaan etsiä graafin jokaiselta tasolta. Yksittäinen tili voi linkittyä epätavalliseen määrään solmuja (node-anomalia), kahden käyttäjän välillä voi ilmestyä odottamaton yhteys (edge-anomalia) tai keskusteluun voi syttyä äkkiä koko bottiparvi tai hashtag-ryöpsähäly (subgraph-anomalia) (Chandola ym. 2009, 4).

Tilastollinen anomaliadetektio tutkimus on rakentunut oletukselle, että 'normaali' käyttäytyminen noudattaa jotakin tunnettua jakaumaa ja poikkeamat voidaan paikantaa sen rajoista. Varhaisimmat lähestymistavat hyödynsivät z-pistemittaria ja χ^2 -testejä, joiden mukaan havainto tulkittiin anomaliaksi, kun sen etäisyys keskiarvosta ylitti ennalta asetetun kynnyksarvon (Chandola ym. 2009, 12). Koska sosiaalisten verkkojen ja mediamittarien jakaumat ovat usein vinoja ja monihuippuisia, parametrissa mallia laajennettiin Gaussian-mixture-menetelmään, jossa "normaalin" eri alaklusterit saivat omat komponenttinsa. Poisson- ja negatiivinen binomijakauma puolestaan soveltuivat viestivirtojen laskentasarjoihin; äkillinen viestipiikki, joka ei enää mahtunut odotettuun hajontaan, merkittiin poikkeamaksi.

Raja-arvolähtöinen tilastollinen lähestymistapa kehittyi 2010-luvulla kahteen suuntaan. Ensinnäkin Extreme Value Theory (EVT) syrjäytti aiemmat karkeat

kynnys-säännöt: GPD-luokitin (Vignotto ja Engelke 2020, 515) ja sen jatkotyöt (Bhattacharya ym. 2023, 1–20) sekä tuore multivariaattinen EVT-katsaus, joka soveltaa teoriaa anomaliadetektion riskimittareihin (Trubey, 2025), arvioivat suoraan havaintojen häntäjakauman ja tuottavat dynaamisen riskipistemittarin. Toiseksi scan-statistiikka yleistettiin graafeihin: Wang ym., (2022) esittävät kalibroidun ei-parametrisen algoritmin, joka laskee jokaiselle osagrafille log-likelihood-suhteen ja liputtaa tilastollisesti merkitsevän “kuuman pisteen”. Menetelmä soveltuu esimerkiksi Twitter-hashtag-parven havaitsemiseen, kun sen käyttäytyminen poikkeaa enemmän kuin satunnaispohjainen uudelleennäytteenotto (resampling) (Chen ja Neill, 2014).

Parametriseen raja-arvomallinnukseen on viime vuosina liitetty joukko ei-parametrisia tiheysestimaattoreita, jotka eivät tee vahvoja oletuksia ‘normaalista’ jakaumasta. Kernel-tiheysestimaatti (KDE) arvioi jokaiselle havainnolle paikallisen tiheyden ja liputtaa pisteet, joiden arvo jää harvaan asutetulle alueelle. Rakhi ym. (2024, 4) osoittavat, että adaptiivinen KDE-pohjainen paikallistiheys parantaa väärin positiivisten suhdetta 15 prosenttia verrattuna parametrusten mixture-malleihin. Sama logiikka voidaan sovittaa virta-dataan: liukuvaa ikkunaa ja Fourier-suodatusta hyödyntävä Streaming-KDE havaitsee sekä äkkiipiikit (esim. hashtag-ryöpsähdyksen) että hitaasti ryömivät muutokset viestinnän rytmissä (Lindstrom ym. 2020). Mahdollinen jatkotutkimus on kehittää adaptiivinen kais-tanleveyden valintamenetelmä, joka säätää KDE:n herkkyyttä reaaliaikaisen some-virran kohinan mukaan.

2.4 Integroiva teoreettinen malli

Tässä työssä yhdistän edellä esitellyt kolme tasoa yhtenäiseksi käsitteellis-operatiiviseksi malliksi. Mallin lähtökohta on multimodaalinen sisältöenkooderi, joka hahmottaa viestin argumentatiiviset piirteet tekstissä, kuvassa, videossa ja äänessä samassa semanttisessa upotusavaruudessa. Näin retoriset tekniikat muuttuvat suoraan numeerisiksi vektoripiirteiksi. Seuraava taso on graafinen neuroverkko, joka kuvaa viestien leviämisreitit ja käyttäjätilien keskinäiset suhteet sekä

laskee kullekin osagrafille koordinoitipoikkeaman todennäköisyyden. Viimeisenä fuusiokerroksena toimii bayesilainen integroija: se yhdistää sisältöstä ja verkostosta tulevat uskottavuuspisteet ja tuottaa posteriorisen propagandariskin, jonka epävarmuus jaottuu epistemiseen ja aleatoriseen komponenttiin. Näin sama skaalautuva putki tukee sekä yksittäisten väitteiden reaaliaikaista arviointia, että kampanjoiden pitkittäisseurantaa. Kehyksessä olennaista on, että sisältö- ja verkostotodisteet eivät jää rinnakkaisiksi, vaan ne päivittyvät vastavuoroisesti – anomalinen leviämiskuvio nostaa viestin prioria tarkempaan sisällölliseen tarkasteluun, ja poikkeuksellinen retorinen signaali puolestaan kasvattaa graafisen tarkastelun tarkkuutta. Arkkitehtuuri tarjoaa siten yhteismitallisen mittariston, jonka varaan voidaan rakentaa sekä temporal-robustisuuden koeasetelmia, että eettisen vinouman auditointeja, sillä jokainen päätös välittää epävarmuuden eksplisiittisesti julkiseen riskipisteeseen. Näiden syiden vuoksi integroitu malli toimii myöhemmissä luvuissa sekä tulosten tulkinnan kehyksenä että jatkokehityssagendaan pohjautuvana konkreettisena tiekarttana.

3 METODOLOGIA

Tässä luvussa kuvataan yksityiskohtaisesti tutkimusasetelma, jonka avulla 2 916 julkaisun laajasta aineistosta valittiin 64 tutkimusartikkelin ydinjoukko kriittiseen, laadullisesti syvennettyyn analyysiin. Menetelmä yhdistää nykyaikaiset luonnollisen kielen prosessoinnin (NLP) tekniikat, verkostanalyysin ja tilastollisen otannan periaatteet. Luvun punaisena lankana on osoittaa, miksi juuri tämän kaltainen moniportainen ja matemaattisesti toistettavissa oleva lähestymistapa on relevantti nopeasti kehittyvän tekoälyn ja mis-/disinformaatiotutkimuksen kentässä, jossa temaattiset rajat ovat häilyviä ja aikajatkumo muokkaa artikkelien merkittävyyttä. Tarkoituksena on rakentaa laajasta AI/ML- ja misinformaatiokirjallisuudesta otos, joka on samanaikaisesti temaattisesti kattava, vaikuttavuudeltaan relevantti ja ajallisesti edustava.

3.1 Aineiston keruu ja perusjoukon määrittely

Bibliometrisen analyysin yleistettävyyttä edellyttää, että tarkasteltu julkaisukorpus heijastaa tutkimusalan todellista tuotantoa (Moed 2005, 18). Tätä periaatetta noudattaen pyrin tutkimuksessani optimoimaan valitsemani julkaisukorpukseni yleistettävyyden ohittamalla akateemisten tietopalvelujen hakupalvelut ja keräämällä mahdollisimman kattavan tietojoukon maailmanlaajuisesta tutkimuskentästä lataamalla yli 200 miljoonaa (218 538 205) akateemista julkaisua Semantic Scholarin "Datasets" -rajapinnasta. Poistin myös tietojoukosta kaikki duplikaatit koodaamani python-pohjaisen ohjelman avulla, perustuen DOI- tai otsikkotunnistukseen. DOI-tunniste helpottaa duplikaattien poistamista systemaattisissa hauissa, koska se on yksilöllinen julkaisun tunniste riippumatta tietokannasta (Hammer ym. 2023, 392). Lataamani Semantic Scholarin datasets'in 'Papers'-tietojoukko

koostui kaikkiaan 60 jsonl-tiedostosta ja niissä esiintyvät julkaisut olivat ilmestyneet vuoden 1651 ja elokuun 2024 välisenä aikana (havaittu luomani python-ohjelman avulla jsonl-tiedostoista). Jsonl-tiedostoissa jokainen julkaisu on erotettu omalle rivilleen ja niiden sisältämiin tietoihin kuuluvat mm. julkaisuvuosi, tekijät, viittausmäärät, DOI, otsikko ja tiivistelmä.

Tutkimusrajaukseen liittyvien julkaisujen valinta tehtiin python-ohjelmalla yksinkertaisen avainsanahaun avulla. Tekniset termit sidottiin OR-operaattorilla ja ilmiöt AND-operaattorilla. Luomani ohjelma kävi lävitse jokaisen 218 538 205 julkaisun tiivistelmät ja otsikot. Valinta perustui ehtoon, että jommastakummasta piti löytyä vähintään yksi tekniikkatermi ja vähintään yksi ilmiötermi. Sanojen kirjainkoodit eivät vaikuttaneet tulokseen. Avainsanaoperaationi hakusanat olivat seuraavat:

OR: machine learning, deep learning, artificial intelligence, tekoäly, koneoppiminen, syväoppiminen,

AND: propaganda, fake news, misinformation, disinformation, hybridivaikuttaminen, valeuutiset, misinformaatio, disinformaatio.

Tällä haulla löydetty 2916 julkaisua yhdistettiin Excel-taulukkoon. Rikastin tuloksia luomalla python-ohjelman, joka artikkeleiden otsikoiden perusteella kykeni Semantic Scholarin academic graph – rajapinnasta hakemaan url-linkit pdf-tiedostoihin, viittausmäärät, "fieldsofstudy" (tutkimuskenttä, johon julkaisu kuuluu), "s2fieldsofstudy" (Semantic Scholarin tekoälyn avulla luotu ja academic graph-tietokannasta haettavat tutkimuskenttätunnistimet) ja corpusid-lista kuhunkin artikkeliin viitanneista tutkimuksista, mikä tutkimuksessani mahdollisesti rajatun, 2916 julkaisukorpuksen sisäisten viittaussuhteiden analyysin.

Keräämäni aineiston merkittävimmät julkaisutyypit ovat metadatan mukaan: journaliartikkelit 50,7 %, Ei tiedossa / ei merkitty 25,3 %, konferenssit 14,8 % ja kirjallisuuskatsaukset 6,6 %.

Muut, harvinaisemmat julkaisutyypit ovat

- Kirjat/kirjan luvut 1,9 % (syventävät monografiat ja toimitetut kokoelmat, usein perusteoksia tai laajoja katsauksia);

- Pääkirjoitukset 0,3 % (lehden pääkirjoituksia);
- Johdattavat, kommentoivat tai linjaavat ajankohtaisia teemoja);
- Uutis- tai aikakauslehtiartikkelit 0,2 % (journalistisia artikkeleita tai tiedelehtien uutisia, välittävät tutkimusuutisia laajalle yleisölle);
- Kirjeet ja kommentit 0,07 % (lyhyitä kirjeitä tai kommentaareja, joissa vastataan julkaistuun tutkimukseen tai avataan nopeita havaintoja);
- Tapausselostus 0,03 % (yksittäisen tapauksen yksityiskohtainen kuvaus, jonka tarkoitus on jakaa ainutlaatuinen havainto);
- Tutkimus (Study-tagin ilman tarkempaa alaluokkaa: merkitty tutkimukseksi, muttei sopinut muihin tyyppeihin) 0,03 %.

64,9 % aineiston julkaisuissa oli fieldsOfStudy ja S2FieldsOfStudy-metadatatarkaste, joista yli puolet (54 %) sijoittuu tietojenkäsittelytieteeseen. Seuraavaksi suurimmat alat ovat lääketiede (13 %) ja valtiotiede (12 %). Nämä kolme muodostavat jo noin 79 % kaikista julkaisuista. Sosiologia-aiheiseksi oli merkitty näistä 5 % ja lingvistiikan alalle 3,1 %. Korpus on siis vahvasti teknologia- ja yhteiskuntatieteispainotteinen, mutta lääketiedeaiheiset julkaisut ovat olleet myös yleisiä tutkimusaineistossa.

Koska Semantic Scholarin tietovarannon jakamat tutkimusartikkelien otsikot ja tiivistelmät eivät vielä itsessään paljasta julkaisun alkuperäiskieltä (mikä on itse pdf-tiedoston kieli), pyrin varmentamaan julkaisujen kielijakauman tätä tarkoitusta varten luomallani ohjelmalla:

1. Jos julkaisuriveillä oli jo valmiiksi DOI-tunniste (Digital Object Identifier), haettiin OpenAlex-rajapinnasta (GET /works/doi:{doi}) *language*-kenttä. OpenAlex hyödyntää sekä Crossref-metatietoja että omia kielimallipohjaisia tunnisteitaan, joten kenttä on kattava ja noudettavissa yhdellä, nopealla kyselyllä.
2. Jos edellisessä vaiheessa ei saatu kieliarvoa, tehtiin Crossref-kysely DOI:lle (GET /works/{doi}) ja tarkistettiin vastaavan JSON-objektin `message.language`.

Crossrefissa kieli on kustantajan itsensä tallettama elementti, joten se toimii luotettavana varajärjestelmänä OpenAlexin puutteille. Kyselyt hidastettiin arvoon 0,13 s/kutsu Crossrefin *polite pool* -ohjeiden mukaisesti.

3. Niille riveille, joilta DOI puuttui, haettiin sellainen Semantic Scholarin Graph-rajapinnasta (`/graph/v1/paper/search?query={title}&fields=externalids`). Näin myös varmistettiin, että virhe ei johtunut alkuperäisen semantic scholar datanrikastuksen puutteista. Jos tällöin hakutulos sisälsi DOI:n, palattiin uudelleen vaiheisiin 1–2.
4. Jos DOI:ta ei saatu rekonstruoitua, hyödynnettiin OpenAlexin vapaamuotoista *search*-parametria (`GET /works?search=title`). Mikäli ensimmäisen tuloksen *language*-kenttä oli saatavilla, se hyväksyttiin.
5. Niille julkaisuriveille, joilla oli avoimesti saatava PDF-osoite (`openAccessPdf_url`), ladattiin tiedosto ja otettiin enintään 2 000 merkkiä kolmannen sivun (tai ellei sitä ollut, ensimmäisen sivun) tekstisisällöstä. Syy tähän on se, että ensimmäiset sivut 1–2 usein sisältävät englanninkielisen otsikon ja tiivistelmän. Kolmas sivu siis antaa paremman otoksen varsinaisesta leipätekstistä ja siten mahdollistaa luotettavamman kielitunnisteen. Löydetty teksti syötettiin fast-langdetect-kirjaston (FastText) luokittimelle.
6. Mikäli mikään edeltävä vaihe ei paljastanut kieltä, yhdistettiin otsikko- ja abstraktitekstiketju (maksimissaan 2 000 merkkiä) ja se analysoitiin fast-langdetectilla; epäonnistumistapauksissa käytettiin varakirjastona (basic) langdetect (N-gram-pohjainen). Tämä vaihe estää luokittelemattomien (undetermined) rivien syntymisen ja on riittävän kevyt suoritettavaksi vain pienelle jäljelle jäävälle joukolle.

Vaikka hakusanaryhmän hakutermit olivat sekä suomeksi että englanniksi, kuitenkin 98,3 % tuloksista (2 867 julkaisua) olivat englanninkielisiä, eikä suomenkielisiä artikkeleita ollut lainkaan. Kielijakauman vinouma johtunee siitä, että Semantic Scholar pyrkii enimmäkseen keräämään englanninkielisiä julkaisuja (Semantic Scholar FAQ, 2025). Vaikka en rajannut hakuani millekään ajanjaksolle, ainoastaan 12 tutkimuspaperaa 2867 julkaisusta ilmestyi ennen vuotta 2011.

Näistä eksplisiittisesti vain yksi, Yhdysvaltojen ilmavoimien julkaisema tiivis raportti "Deception Detection in Expert Source Information through Bayesian Knowledge-Bases" (Eugene Santos Jr. 2006), voidaan tulkita sisältyvän tutkimusrajaukseemme. Tämä julkaisu koskee laskennallisten menetelmien hyödyntämistä disinformaation löytämiseksi anomaliatunnistuksen näkökulmasta hyödyntäen todennäköisyyspohjaista tekoälyä. Julkaisun esittämät työkalut ja menetelmät eivät kuitenkaan ole nähtävästi tarkoitettu levitettäväksi USA:n armeijan (ilmavoimat ym.) ulkopuolelle (raportti toimii lain velvoittamana 'kuittina' department of defence -rahoittajalle, ei työkalupakettina). Näistä syistä johtuen katson tutkimuskysymykseni rajaukseen sopivan tutkimuskentän kohdistuvan vasta vuonna 2011 ja sen jälkeisiin julkaisuihin.

3.2 Aineiston käsittely ja tarve uusille otantametoodeille

Useiden systemaattisten katsauksien arvioinnit osoittavat, että *valikointivirhe* (selection bias) on toistuva ja vaikeasti todennettava ongelma. Tähän voivat johdattaa mm. kapea-alaiset hakustrategiat, yksipuoliset tietokannat tai huonosti kohdistetusta poiminnasta. Esimerkiksi Web of Science tarjosi 1000 ensimmäistä relevanteinta osumaa ja jätti olennaisia artikkeleita haun ulkopuolelle (Haddaway ym. 2020, 4–5). Osin myös tätä ongelmaa välttääkseni keräsin yli 200 miljoonan julkaisun paikallisen tietokannan, jolla kykenin varmistamaan, että 100 % Semantic Scholarin tietojoukon avainsanahaun tuloksista saatiin kerättyä tutkimusaineistooni ilman tietokantapalvelujen hakurajoituksia.

Keräämäni tutkimusaineiston laajuus (2916 julkaisua) toisaalta myös lisää relevantimpien julkaisujen poiminnan vaikeutta sekä tuo paineita käsittelemieni aineistojen perustelluille valintakriteereille ja siten katsaukseni toistettavuudelle. Pyrin katsauksessani erityisesti vastaamaan Haddaway ym. esittämiin haasteisiin katsausten toistettavuudesta ja valikoitujen tutkimusten validiteetista (Haddaway ym. 2020, 5–7), ehdottamalla uutta systemaattista ja enimmäkseen determinististä metodologiaprosessia tutkimuskenttää edustavimpien (ja todennäköisesti vaikuttavimpien) julkaisujen valikointiprosessiksi.

Jokaisesta julkaisusta muodostettiin yhdistelmäteksti, jossa otsikko ja abstrakti liitettiin yhteen ja puhdistettiin teknisistä elementeistä, kuten URL-osoitteista ja mahdollisista XML-tageista. Stop-sanoja ei poistettu, sillä nykyiset embedding-mallit tulkitsevat sanayhteydet kokonaisvaltaisesti. Tekstit syötettiin vuoden 2024 alussa julkaistulle, 3 072-ulotteiselle text-embedding-3-large-mallille, joka on osoittanut huipputason suorituskykyä Massive Text Embedding Benchmarkissa (Muennighoff ym. 2022; OpenAI 2024).

Malli valittiin, koska se yhdistää syvän semanttisen ymmärryksen, suuren dimensioavaruuden tuoman erottelukyvyn, monikielisyyden ja yleistajuisen semanttisen geometrian. Mallin syvä semanttinen ymmärrys päihittää esimerkiksi tf-idf- ja LDA-mallit lyhyiden tekstien klusteroinnissa (Miller ja Alexander 2025). Suurempi dimensioiden määrä antaa datapisteille enemmän tilaa vektoriavaruudessa, jolloin ne eivät pakkautu päällekkäin ja K-means-klusterit erottuvat selkeämmin (Muennighoff ym. 2023). Noin kuusi prosenttia tutkimusmateriaalista on muulla kielellä kuin englanniksi, joten mukana on pieni, mutta merkittävä määrä monikielistä tekstiä. Ja koska käytetty kielimalli on yleiskäyttöinen, eikä rajoitu yhden alan sanastoon (vrt. esim. SPECTER), se käsittelee kaikki tieteenalat tasapuolisesti eikä suosi mitään yksittäistä aihepiiriä (Cohan ym. 2020). Mallinnuksen tuloksena syntyi taulukko, jossa jokaiselle 2 916 artikkelille on 3 072-numeroa kuvaava vektori. Tätä suurta taulukkoa käytettiin myöhemmin, kun muodostettiin klusterit ja varmistettiin otoksen edustavuus.

3.2.1 Koneellinen relevanssiseulonta aiemmassa tutkimuksessa

Vuosien 2024–2025 vaihteessa ilmestyi tuloksissaan poikkeuksellisia tutkimuksia suurten kielimallien hyödyntämisestä systemaattisten katsauksien tutkimuskirjallisuuden seulonnassa. Vuoden 2025 alussa julkaistussa tutkimuksessa, Delgado-Chaves ym. (2025) kokeilivat 18 erilaisen kielimallin käyttöä systemaattisten kirjallisuuskatsausten seulontaprosessissa. Kielimalleja testattiin kolmessa erilaisessa kirjallisuuskatsauksessa, jotka käsitelivät fysioterapiaa, neurologiaa ja digitaalista terveydenhuoltoa (Delgado-Chaves ym. 2025, 1–3).

Tutkimuksessa kokeiltiin sekä kaupallisia että avoimen lähdekoodin kielimalleja, mukaan lukien GPT-3.5-turbo, GPT-4o, Gemma2, Llama3 ja Mistral. Malleja käytettiin arvioimaan julkaisujen otsikoita ja tiivistelmiä ennalta määritettyjen sisällyttämiskriteerien perusteella. Tutkijat kehittivät JSON-skeeman, joka mahdollisti johdonmukaisen, toistettavan, vastausformaatin kaikilta malleilta (Delgado-Chaves ym. 2025, 4–5). JSON-skeema siis pakottaa mm. OpenAI:n rajapinnassa automaattisesti tekoälyn vastaamaan ainoastaan halutussa JSON-strukturissa. Seulontaprosessissa käytettiin erityisesti kahta lähestymistapaa:

1. Ensimmäisessä vaiheessa malleja pyydettiin tuottamaan kyllä/ei-vastauksia sisällyttämiskriteerien perusteella (eli tutkijoiden antamat ohjeet tekoäylle).
2. Toisessa vaiheessa kielimalleja pyydettiin antamaan perustelut päätöksilleen.

Tutkimukset osoittivat, että LLM-mallien käyttö vähensi seulontaan tarvittavaa työmäärää 33–93 % (Delgado-Chaves ym. 2025, 5). Mallit osoittivat korkeaa tarkkuutta (specificity), mutta vaihtelevaa herkkyyttä (recall). Erityisesti GPT-4o-2024-05-13 -malli osoittautui testattujen mallien erinomaiseksi vaihtoehdoksi (Delgado-Chaves ym. 2025, 3–4). Sisällyttämiskriteerien huolellinen määrittely vaikutti merkittävästi mallien suorituskykyyn (Delgado-Chaves ym. 2025, 5-6). Tutkijat osoittivat, että suurien kielimallien hyödyntäminen voi merkittävästi nopeuttaa (ja tehostaa) kirjallisuuskatsausten tekoa, mutta ihmisen valvonta ja kriteerien tarkka määrittely ovat edelleen välttämättömiä. (Delgado-Chaves ym. 2025, 10).

Vastaavasti marraskuussa 2024 Joos ym. tutkivat moniagenttipohjaista LLM-lähestymistapaa relevantin kirjallisuuden suodattamiseen. He testasivat erilaisia malleja 8323 julkaisun korpuksella, liittyen tutkimusaiheeseen *Graph Exploration in Immersive Settings*.

Tutkijat kehittivät strukturoidun prosessin, joka koostui seuraavista vaiheista:

1. Alustavien avainsanojen haku tietokannoista (ACM Digital Library, IEEE Xplore, Eurographics).
2. Datat esikäsittely (esim. duplikaattien poisto).

3. LLM-agenttien, kuten Llama3, Gemini 1.5 Flash, Claude 3.5 Sonnet, GPT-4o itsenäinen luokittelu otsikoiden ja tiivistelmien perusteella. Jokaiselle mallille annettiin jäsenneily kehotepohja (prompti), jossa määriteltiin tarkasti tutkimuskysymyksen määritelmä, sisällyttämiskriteerit esimerkkeineen sekä poissulkemiskriteerit. Jokainen LLM-malli käsitteli itsenäisesti jokaisen julkaisun otsikon ja tiivistelmän, palauttaen binäärisen päätöksen (INCLUDE/DISCARD) sekä lyhyen kaksivirkkeisen perustelun päätökselleen.

4. Konsensusäänestys lopullisille sisällyttämisen- ja poissulkemispäätöksille.

5. pienemmän otannan manuaalinen tarkistus (Joos ym. 2024, 2–3).

Tutkijat eivät kuitenkaan selvästikään hyödyntäneet muun muassa OpenAI:n gpt4o -rajapinnan tarjoamaa 'pakotusta' JSON-strukturoituihin vastauksiin.

Tutkimuksen mukaan 'Konsensus (Parhaat)' -menetelmä osoittautui optimaalisimmaksi lähestymistavaksi. Tässä yhdistelmässä käytettiin vain kolmea parasta kielimallia, joiden F1-pisteet ylittivät 50 %. Nämä mallit olivat: Gemini 1.5 Flash, Claude 3.5 Sonnet ja GPT-4o.

Tämä yhdistelmä tuotti kokeiluissa korkeimman tasapainon 'virheellisten positiivisten' (false positives) tulosten vähentämisen ja korkean herkkyyden välillä (eli relevanttien lisäksi löydetty myös 'väärää positiivisia'). Tämä konsensusmenetelmä tunnisti onnistuneesti 87 relevanttia paperia 88:sta, 98.86 prosentin tarkkuudella. (Joos ym. 2024, 2–3).

Konsensusäänestyksen jälkeen tutkijoille jäi merkittävästi pienempi joukko julkaisuja manuaalisesti tarkistettavaksi. Manuaalisen tarkistusvaiheen aikana tutkijat hyödynsivät sekä julkaisujen metadatta että LLM-mallien generoimia perusteluja. Tutkijoiden mukaan, tutkimuksen prosessi oli huomattavasti tehokkaampi kuin perinteinen manuaalinen tarkastelu, koska käsiteltävien julkaisujen määrää voidaan vähentää merkittävästi pienempään, relevanteiksi tunnistettuun osajoukkoon.

Joos ym. (2024) myös kuvaavat, miten tehtävien siirtäminen ihmisiltä tekoälylle johtaa resurssien tehokkaampaan kohdentamiseen. Heidän mukaansa tehtävien ulkoistaminen tekoälylle vapauttaa tutkijoiden aikaa ja kognitiivisia resursseja,

mikä mahdollistaa niiden kohdentamisen vaativampiin ja luovempiin prosesseihin. He esittävätkin, että systemaattisissa kirjallisuuskatsauksissa rutiininomaisen seulontatyön automatisointi antaisi tutkijoille mahdollisuuden keskittyä tulosten syvällisempään analyysiin ja synteisiin. Tekoälyavusteinen kirjallisuuskatsaus voisi mahdollistaa myös laajempien tutkimusalueiden kartoittamisen. Resurssirajoitusten vähentymisen myötä (ajallisten ja rahallisten), tutkijat voivat myös tarkastella laajempia aihekokonaisuuksia ja poikkitieteellisiä yhteyksiä, mikä edistää kokonaisvaltaisempaa ymmärrystä tutkimusaiheesta. Tutkijat korostivat, että vaikka LLM-mallien konsensusäänestys vähensi huomattavasti työmäärää, ihmisasiantuntijoiden valvonta ja lopullinen päätöksenteko olivat edelleen välttämättömiä laadukkaan kirjallisuuskatsauksen toteuttamisessa (Joos ym. 2024, 3–4).

Tekoälyavusteinen kirjallisuuskatsaus voisi kenties myös parantaa tutkimuksen toistettavuutta ja läpinäkyvyyttä, sillä automatisoitu prosessi voidaan dokumentoida tarkemmin ja toistaa samanlaisena tulevilla tutkimuksissa. Myös mahdollisuus käsitellä laajemmin suuria tietojoukkoja voisi parhaimmillaan vähentää valintaharhan mahdollisuutta. Vaikka suuret kielimallit eivät olekaan käytännöllisesti katsoen deterministisiä, voidaan niiden seulontaprosessia, kehoitteita ja tuloksia dokumentoida toistettavilla tavoilla sekä arvioida tuloksia probabilistisesti. Tämä voisi jopa vahvistaa tieteellisen tutkimuksen luotettavuutta ja kumulatiivista luonnetta.

3.3 Aineiston analyysi

3.3.1 Aineiston kvantitatiivinen analyysi

Omassa katsauksessani kehitin näiden tutkimusten tulosten pohjalta neliosaiseen varmennukseen perustuvan lähestymistavan. Korostan kuitenkin, että katsaukseni rajaa tekoälyn hyödyntämistä tutkittuihin ja empiirisesti toimiviksi osoitettuihin käyttötapoihin kirjallisuuden seulontaprosessissa. Seulontarakenteeni mukailee melko tarkasti Joos ym. (2024) –tutkimuksen prosessia muutamilla parannusehdotuksilla. Seulontaprosessissa käytimme Semantic Scholar -tietokannasta avainsanahakujen avulla koottua julkaisukorpusta, sekä niihin kuuluvia julkaisujen tiivistelmiä ja otsikoita.

Jokaisesta julkaisusta muodostettiin yhdistelmäteksti, jossa otsikko ja tiivistelmä liitettiin yhteen ja puhdistettiin teknisistä elementeistä, kuten URL-osoitteista ja mahdollisista XML-tageista. Stop-sanoja ei poistettu, sillä nykyiset embedding-mallit tulkitsevat sanayhteydet kokonaisvaltaisesti. Tekstit syötettiin vuoden 2024 alussa julkaistulle, 3 072-ulotteiselle text-embedding-3-large-mallille, joka on osoittanut huipputason suorituskykyä Massive Text Embedding Benchmarkissa (Muennighoff ym. 2022; OpenAI 2024). Katsaukseni hyödyntää uusinta GPT-4.1-mallia, joka OpenAI:n julkaisemien tulosten mukaan on kykenevämpi ymmärtämään pitkiä tekstejä ja tunnistamaan siitä olennaisia osia sekä yhteyksiä.

Toisin kuin Delgado-Chaves ym. (2025, 4) –tutkimus, jossa käytettiin kielimallien enemmistöäänestystä, vaadimme 100 prosentin konsensusta mallilta relevanssin määrittämiseksi. Jos kaikki kolme GPT-ääntä eivät ole yksimielisiä, merkitsemme julkaisun viereen Excelissä 'undecided'. Tämä lähestymistapa on myös tiukempi kuin Joos ym. (2024) -tutkimuksen käyttämä menetelmä, jossa vain yksi positiivinen ääni riitti sisällyttämiseen aineistossa. Katsauksessa toteutettiin nelitasoisen suodatusprosessi. Kielimallille annetusta ohjekehotteesta määriteltiin tarkasti tutkimuskysymyksen määritelmä, sisällyttämiskriteerit esimerkkeineen sekä poissulkemiskriteerit perustuen tunnistettuun relevanssiin tai sen puutteeseen.

Tämän jälkeen kaksi erillistä GPT-4.1 kielimalliagenttia ja yksi GPT-o3 kielimalliagentti käsittelivät julkaisujen otsikoita sekä tiivistelmiä ohjekehotteen perusteella ja vastasivat JSON-strukturoidussa muodossa yksinkertaisilla kyllä/ei -vastauksilla. Mikäli kaikki kolme agenttia eivät olleet samaa mieltä vastauksesta, kyseisen julkaisun tieto tallennettiin erillistä käsittelyä varten uuteen Excel-tiedostoon ja jokaiseen merkittiin 'undecided'. Näitä 'undecided'-tapauksia löytyi 184 kappaletta. Vastausten preliminääriseen diskriminointiin hyödynsin sekä Sonnet 3.7 (extended thinking) ja GPT-o3 (molemmat 'chain of thought' -koulutettuja kielimalleja). Toisin sanoen tällaisten tekoälymallien vastaus ei perustu puhtaasti sisään annettujen tietojen todennäköisyysjakaumavaikutuksesta, kuten niin sanotuissa tavallisissa generatiivisissa transformers-kielimalleissa, vaan nämä kykenevät arvioimaan omaa ajatusprosessiaan ja siten huomaamaan omia virheitään, vähentäen epärelevantteja vastauksia ja hallusinaatiota. Molempien tekoälymallien kohdalla tarkistusvaiheessa, suoritin 'best of two' -lähestymistapaa, jossa valitsin parhaimmat tulokset molemmista malleista. Päädyin lopulta hyödyntämään vain Sonnet 3.7 -mallia tarkastuksessa, sen johdonmukaisten ja hyvin perusteluiden vastausten vuoksi. Mallit esittävät oman näkemyksensä tutkimuksen relevanttiudesta, ja ehdottivat relevanttiustägiksi kyllä/ei -vastauksen 'undecided'-tägin tilalle.

Tarkistin jokaisen 184:n epäselvän tapauksen ja tekoälyn ehdottamat relevanttiustägit perusteluineen ja olin ainoastaan 12 tapauksen kohdalla eri mieltä tekoälyn ratkaisusta (n. 6,52 %). Tekoälyn kyky tuottaa myös perustelut tulkinnoilleen nopeutti merkittävästi tarkistusprosessiani, vahvistaen aiempien tutkimusten havaintoja. Huomasin myös, että osa näistä tarkistusmallin ehdottamista epätarkkuuksista olisi voitu todennäköisesti ehkäistä ainoastaan hieman tarkentamalla ohjauskehotetta. Tämä korkea tarkkuus yllätti itsenikin, sillä nämä 184 epäselvää tapausta olivat enimmäkseen todella hankalasti tulkittavia tapauksia ja olisivat hämänneet monia huonosti nukkuneita ihmistarkistajiaakin. Esimerkkinä tapaus, jossa pyrittiin tutkimaan AI-generoitujen videoiden tunnistamista, mutta ainoastaan hyödyntämällä ihmisten omaa silmämääräistä tulkintaa, eikä hyödyntämällä koneoppimista tai tilastollisia menetelmiä. Tämä vaati kontekstin ymmärtämistä ja tarkkaa lukutaitoa. Kyseinen paperi olisi myös suurella todennäköisyydellä

mennyt semanttiseen läheisyyteen perustuvien tekstiupotusmallien ja varmasti-kin sana- tai termipohjaisten suodattimien läpi. Esimerkiksi Claude 3.7 -kielimalli (ja alun perin o3-malli) kuitenkin onnistuivat arvioimaan tämän oikein ensimmäisellä yrittämällä. Chain of thought –mallien vahvuus tarkkaa lukutaitoa vaativissa tehtävissä näkyikin kenties parhaiten juuri tällaisissa tapauksissa. Lopullisessa jaossani relevanteiksi valittiin 2087 julkaisua ja epärelevanteiksi 780 julkaisua. Julkaisuista 49 oli muunkielisiä ja siksi ne poissuljettiin tarkastelusta (yht. 2916).

Systemaattisen kirjallisuuskatsauksen yksi haasteista on hahmottaa inhimillisessä ajassa, mitä aihealueita satoihin (tai jopa tuhansiin) tutkimusraportteihin sisältyy ja miten ne eroavat toisistaan, mikä voidaan nähdä samankaltaisena ongelmana kuin tiedon etsinnässä yleensäkin. Laaja, poikkitieteellinen aineisto tuottaa useita käsitteellisiä alateemoja, joille ei ole välttämättä löydy valmiita hakusanoja. Ensimmäinen ajatukseni oli perinteisimmillä tekstianalyysimetoodeilla, kuten termi- ja n-gram-pohjaisilla menetelmillä, havainnoida ja vertailla tutkimuskirjallisuutta. Vertailllessani uusimpia tekstianalyysitapoja koskevaa kirjallisuutta päädyin kuitenkin seuraaviin havaintoihin:

N-gram-mallilla tarkoitetaan sanajonoa, jossa sanan todennäköisyys arvioidaan vain edeltävien $n - 1$ sanan perusteella; kaksisanainen jono on taas *bigram* ja kolmisananainen vastaavasti *trigram* (Jurafsky ja Martin 2025, 33–35). Termipohjaisissa menetelmissä – klassisessa vektoriavaruusmallissa ja sen tf-idf-versiossa – kukin dokumentti esitetään sanakohtaisina laskureina siten, että kunkin rivin ja sarakkeen risteys kertoo, montako kertaa sana esiintyy tietyssä tekstissä (Jurafsky ja Martin 2025, 106–107).

Sekä n-gram-malli että bag-of-words-pohjainen termivektorointi tekevät eksplisiittisen sanajärjestyksen sivuuttavan oletuksen. N-gram rajoittaa riippuvuudet kiinteään ikkunaan, jolloin pidemmän kontekstin vaikutus katoaa; vastaavasti bag-of-words käsittelee dokumenttia ‘sanojen säkkinä’, jossa sanan sijaintia ei huomioida lainkaan (Jurafsky ja Martin 2025, 58–60). Kun tekstin rakenne pelkistyy pelkiksi esiintymismääräksi, lause- ja diskurssitason yhteydet jäävät mallien ulkopuolelle.

Tästä saadaan niin kutsutut harvat vektorit (sparse vectors), joissa jokainen ulottuvuus vastaa yhtä sanastoon kuuluvaa sanaa. Jurafsky ja Martin (2025, 117) huomauttavat, että tällaiset sparse-esitykset edellyttävät kymmeniä tuhansia ulottuvuuksia ja sanakohtaisia painotuksia, mutta suoriutuvat silti heikosti mm. synonyymien tunnistamisessa. Upotteet (*embeddings*), joiden dimensio on tyypillisesti 50–1000, vastaavasti tiivistävät saman informaation huomattavasti pienempään parametritilaan (Jurafsky ja Martin 2025, 117).

Toisin sanoen embedding-mallit kykenevät säilyttämään semanttisen rakenteen ja pitkän kontekstisuuden ja ovat tehokkaampia sekä tarkempia kuin perinteiset termi- ja n-gram-pohjaiset menetelmät. Tämä on olennaista myös siksi, koska jakautuva (distributional) hypoteesi olettaa, että sanan semanttinen profiili voidaan päätellä sitä ympäröivistä sanoista; sanan merkitys palautuu sen esiintymäympäristöihin. Upotemallit tiivistävät dokumentin sisällön lyhyeen, tiheään vektoriin ja näiden mahdollistama pitkä konteksti-ikkuna antaa mallille mahdollisuuden sisällyttää laajemmin aihepiirin tietoa (Jurafsky ja Martin 2025, 101, 118, 124).

Tätä periaatetta on myös hyödynnetty kirjallisuuskatsauksia automatisointityökalua esittävässä 'Research Screener' -tutkimuspaperissa (Chai ym. 2021). Tutkimuksessa esitetäänkin, että tiivistelmien embeddings-pohjainen priorisointi voisi vähentää jopa 96 % manuaalisesta seulontatyöstä. Semanttinen läheisyys tiivistelmien välillä näkyy vektorietäisyyksinä – samankaltaiset tutkimukset sijoittuvat lähelle toisiaan upotusavaruudessa. Tekstin mukaan, sisällön ja kontekstin kokonaisuus säilyy tällöin myös paremmin kuin n-gram-pohjaisissa malleissa. Tätä etäisyystietoa hyödynnettiin 'Research Screenerissä' muun muassa artikkeleiden relevanttiuden arviointiin (kuinka lähellä muut tutkittavan aineiston paperien tekstiavaruudet ovat relevantin paperin tekstiavaruutta) (Chai ym. 2021, 3–4).

Valitsin katsaukseni tekstinupotusmalliksi text-embedding-3-large -tekoälymallin, koska sen suuri dimensionaalisuus (3 072 ulottuvuutta) ja laaja koulutuskorpus mahdollistavat hienojakoistenkin semanttisten eroavaisuuksien representaatiot tekstiavaruudessa. Text-embedding-3-large -malli kykenee myös käsittelemään 8191 tokenin tekstisyötteet yhdellä rajapintakutsulla, millä myös vältettiin tekstin pätkimisestä syntyvää tiedonhukkaa.

Esimerkiksi useampien muiden mallien kontekstirajoitukset (tokenimäärärajoitukset) usein pakottavat pilkkomaan malille syötettävää tekstiä, vaikeuttaen siten muun muassa vektoriavaruuteen perustuvan tutkimuksen toistettavuutta kirjallisuuskatsaukseni ja muiden tutkimusten välillä, jotka pyrkivät varmistamaan tai toistamaan tulokseni (esimerkiksi, jos tekstit eivät ole johdonmukaisesti pilkottu tai käsitelty täysin toistettavalla tavalla tutkimusten välillä). Christou ym (2024) esittävätkin, että tällainen pilkkominen voisi myös jättää olennaista informaatiota pois dokumentin kontekstista, mikä heikentäisi siten myös upotusteni edustavuutta. He myös esittävät, että jopa yleinen pilkkomisstrategia lausekohtaisten upotusten keskiarvoistamisesta, jolla tätä ongelmaa pyrittäisiin minimoimaan, silti heikentäisi dokumentin upotusten kontekstuaalista eheyttä (Christou ym. 2024, 3). Saman ongelman kvantifioivat Gao ym. (2021): BERT-pohjaiset mallit käsittelevät enintään 512 WordPiece-tokenia (eli noin 400 sanaa), minkä takia heidän aineistonsa keskimäärin 2 000 sanaa sisältävät dokumentit oli pilkottava useiksi 510 tokenin lohkoiksi ja näiden upotukset yhdistettävä jälkikäteen, mikä lisäsi laskentakustannuksia ja altisti myös samankaltaiselle edellä mainituille informaation menetykselle. (Gao ym. 2021, 2–5, 6–8).

Upotusmallini valintaan vaikuttivat myös arviot useiden mallien vertailututkimukset. Esimerkiksi Keraghel ym. (2024) vertailivat kahdeksaa LLM-pohjaista embedding-mallia ja raportoivat, että jopa käyttämäämme mallia vanhemman OpenAI:n GPT-mallin 1 536-ulotteinen tekstiavaruus sai korkeimmat ACC/NMI/ARI-arvot (yleisiä klusteroinnin arviointimittareita) useimmilla klusterointialgoritmeilla (Keraghel ym. 2024, 9).

OpenAI:n virallinen dokumentaatio korostaa, että embeddingit on normalisoitu yksikköpituuteen, jolloin kosinietäisyys on suositeltu mittari semanttiseen hakuun (OpenAI, 2024, osio 'Vector embeddings'). Kosinietäisyys normalisoi vektoripituuden ja mittaa suuntaeroa, jolloin paljon esiintyvät sanat tai pitkät dokumentit eivät hallitse tulosta. Toisin sanoen, pituuseroista ei kerry 'bonus pisteitä'. Pitkä abstrakti ja lyhyt lause vaikuttavat hakuun yhtä suurella painotuksella, jos niillä

on sama semanttinen suunta¹. Tällä pyritään analysoimaan, mihin (temaattisesti/semanttisesti) tekstit viittaavat, ei siihen, kuinka paljon ne sisältävät sanoja.

Kirjallisuuden lopullinen stratifiointi (eli jakaminen ryhmiin) uusimpia, tuloksiltaan lupaavimpia klusterointialgoritmeja tehtiin projisoimalla klusterit kaksiulotteiseen avaruuteen LocalMAP-dimensiovähennysalgoritmilla (PaCMAP:ista johdettu), joka kykenee poistamaan dynaamisesti virheelliset lähinaapurisärmät ja lisäämään uusia, paikallisia 'further-pair'-särmiä, vähentäen klustereita väärin toisiinsa sitovia veto- ja hylkimisvoimia. Näin pyritään saavuttamaan artefaktivapaa ('false positive'-särmät) ja selkeästi rajautuva näkymä tutkimuskentän todellisesta rakenteesta (Wang ym. 2024, 3–5).

Lopullinen ryhmittely tehtiin dimensiovähennettyyn aineistoon HDBSCAN-algoritmilla, jolla rakennettiin ensin koko korpukselta tiheysperustaisen hierarkia, minkä jälkeen se automaattisesti valitsi lopulliset klusterit maksimoimalla kunkin klusterin 'stability'-arvon. Minimiklusterikoko asetettiin kattamaan vähintään 10 artikkelia, noudattaen bibliometriassa suositeltua rajaa, jolla 'pienet mutta merkitykselliset' aihehaarat pysyvät näkyvissä (Bascur ym. 2025).

Haddaway ym. (2020, 2) huomauttavatkin, että laajat ja heterogeeniset tutkimusaineistot vaativat systemaattista jäsentämistä, jotta katsauksen objektiivisuus ja kattavuus voidaan varmistaa – erityisesti silloin, kun aihepiiri on kiistanalainen tai tutkimuksia on erittäin paljon. Käyttämäni yhdistelmämenetelmä vastaa tähän tarpeeseen tarjoamalla yksinkertaisen tavan jakaa aineisto temaattisiin ryhmiin, mikä tällöin mahdollistaa tutkimusaineiston stratifiointin ja parantaa tulosteni toistettavuutta. Näin klusterien määrää ei tarvitse asettaa etukäteen, toisin kuin K-means-algoritmissa, jossa käyttäjän on ilmoitettava haluttujen klusterien määrä etukäteen (kenties myös tästä johtuen on vapaampi käyttäjän silmämääräisistä havainnoista ja biasoituneisuudesta). Menetelmä ja kaava on esitetty Campellon ym. (2015) -artikkelissa sivuilla 18–19.

¹ (Kosinietäisyys = $1 - \cos(\theta)$), jossa θ on vektorien välinen kulma. Jos kaksi tekstiä ovat semanttisesti samankaltaisia (näiden välinen kulma suuri), etäisyys on lähellä nollaa. Jos ne käsittelevät eri aiheita (kulma pieni), etäisyys lähestyy yhtä.)

Edustavimpien julkaisujen löytäminen stratifioiduista datasta

Esittämäni klusteripohjainen stratifikaatio perustuu ajatukseen, että julkaisut ensin jaetaan semanttisesti koherentteihin ryhmiin, minkä jälkeen tunnistetaan kunkin stratifioidun ryhmän sisäiset 'kultajyvät'. Tämä vähentää todennäköisyyttä sille, ettei yksi kapea teema-alue peittäisi alleen muita teemoja- ja alateemoja, perustuen esimerkiksi ainoastaan tiettyjen avainsanojen ylivertaiseen esiintymiseen tai pinnallisiin suosiosignaaleihin, kuten viittausmääriin.

Usein oletetaan, että viittausten suuri määrä ennakoii korkeaa laatua, mutta empiirinen näyttö monimutkaistaa käsitystä tämän suoraviivaisuuden vastaavuudesta. Laajan kyselytutkimuksen perusteella Aksnes ym. (2006) havaitsivat, että vaikka viittausluvut korreloivat kohtalaisesti tutkijoiden omien 'major–minor-painos' -arvioiden kanssa, korrelaatio jäi artikkelitasolla epäluotettavaksi: noin 15 %:ssa tapauksista siteerausprofiili ei peilannut lainkaan julkaisujen tieteellistä merkitystä (Aksnes 2006, 12–14).

Lisäksi akateemisen julkaisutuotannon 'inflaatio' kiihdyttää artikkelimääriä nopeammin kuin tiedon arvon vastaavaa lisäystä: h-indeksiä painottava 'publish or perish' -kulttuuri kannustaa minimoituihin julkaisuyksiköihin (minimum publishable unit, MPU), mikä heikentää pelkkään lukumäärään perustuvan arvioinnin erottelukykä. Aragónin (2013) mukaan 'publish or perish' -paine kannustaa pilkkomaan yhden laajan tutkimuksen useiksi pieniksi artikkeleiksi, jotta saadaan useampia julkaisuja ja siten enemmän potentiaalisia viittauksia (Aragón, 2013, 1–2).

Näiden havaintojen perusteella pyrin hyödyntämään tutkimuksessani uusimpia tutkimuksia ja metodeja tieteellisen vaikuttavuuden jäljitettävyydestä. Min ym. (2020) määrittävät nk. citation cascade-rakenteen ja osoittivat, että kahdesta neljään 'viittaussukupolven' tarkastelu riittää jäljittämään julkaisujen välillistä tieteellistä vaikutusta. Tosin viiden ensimmäisen sukupolven jälkeen aiheen relevanssi hupenee, ja kymmenennessä sukupolvessa se laskee jo satunnaista taustatasoa vastaavaksi (Min ym. 2020, 2, 18–19). Tämän havainnon perusteella mallin-

namme kullekin julkaisulle korkeintaan viisi sukupolvea ja vaimennamme kaukaisempia viitteitä eksponentiaalisesti ($\text{decay} = 0,5$). Tällainen sukupolvivaikuttaavuus-metriikka korostaa sekä suoria että 'hiljaisia' polkuja, mutta estää kaukaisia, merkitykseltään saturoituneita viitteitä hallitsemasta tulosta.

Seuraavassa vaiheessa Generational-pisteet fuusioitiin PageRank -arvoon, jossa linkin paino perii lähdeartikkelin omaa arvoa (Brin ja Page 1998, 108–110) painotuksella 60:40 (PageRank-painotus 40 %). Metodini pyrkii tasapainottamaan PageRankin taipumusta suosia vanhoja klassikoita ja samanaikaisesti, jotta myös uudet salientit julkaisut nousevat esiin analysoitavien julkaisujen valikointiprosessissa. Jokaisessa klusterissa otettiin mukaan enintään 15 artikkelia, laadullisesti tunnistetuilla periodijaksoilla 2011–2015, 2016–2019 ja 2020–2024. Ajanjaksojen keskinäinen kiintiö jaettiin ensin suhteellisen artikkelimäärän mukaan ja viimeistettiin yksittäisillä ± 1 -siirroilla, kunnes maksimimäärä (15) täyttyi. Menettely pyrki estämään sekä vanhojen, paljon siteerattujen tutkimusten heikon valintatodennäköisyyden, että yksittäisten 'COVID-piikkien' kaltaisten trendien ylikorostumisen. Lopputuloksena saadaan ajallisesti kalibroitu vaikuttavuusindeksi, joka kuvastaa tieteellisen keskustelun todellista dynamiikkaa paremmin kuin pelkkien suorien viittausten laskeminen.

Tämä lopullinen, laadulliseen analyysiin valikoitu otanta-aineisto edustaa siis merkittäväällä todennäköisyydellä koko laajaa tutkimuskenttää, huomioiden tutkimusten vaikuttavuuden ja relevanttiuden ajallisessa jakaumassa. Esitän siis näiden perustelujen pohjalta, että yhdistelmämetodini tarjoaa vertailukelpoisen viitekehysten tulosteni vahvistamiseen tai kiistämiseen verraten heuristisempiin tulokulmiin.

Kun ositetut klusterit yhdistettiin kolmeen aikakerrokseen, päädyin lopulta kahdeksaantoista otantakerrokseen, joista kuhunkin valikoitui vaikuttavuuspainotusta käyttäen kolme tai neljä artikkelia. Näin muodostui 64 julkaisun ydinjoukko, jossa kukin klusteri ja aikajakso on tasapainoisesti edustettuna. Syntyneen otoksen kattavuutta testattiin useilla tavoilla. Matriisialgebraan pohjautuvassa pääkomponenttianalyysissä muodostettiin ensin tf -idf-vektorit (tf = term frequency; idf = inverse document frequency) kustakin artikkelista. Tämän jälkeen laskettiin

tf-idf-matriisin kovarianssi ja suoritettiin ominisarvohajotelma, mistä saatiin kaksi ensimmäistä pääkomponenttia.

Kun 64 artikkelia ja koko vertailujoukko projisoitiin näihin kahteen komponenttiin, pistepilvi osoitti, että valitut artikkelit jakautuvat tasaisesti koko tilaan eivätkä kasaudu erilliseksi saarekkeeksi. Jos valittu otos olisi vinoutunut, sen pisteet muodostaisivat omia klustereitaan tai näyttäisivät selvästi matalamman varianssin asianomaisissa pääkomponenttisuunnissa. Pääkomponenttianalyysin tulosten visualisoinnit siis osoittivat, että valitut artikkelit levittäytyvät tasaisesti koko tf-idf-avaruuteen ja sekoittuvat hyvin viitatuimpien julkaisujen kanssa (Jolliffe ja Cadima 2016).

Todennäköisyyslaskentaan ja vektorigeometriaan pohjautuva lähin naapuri – analyysillä (Nearest-Neighbor Analysis) jokaiselle artikkelille laskettiin 3 072-dimensioinen upotevektori, joka muodostaa monidimensioisen avaruuden. Tässä avaruudessa kosinietäisyys mittaa kulmaa kahden vektorin välillä: arvo 1 tarkoittaa täydellistä samankaltaisuutta ja arvo 0 sitä, että tekstit ovat toisiaan vastaan kohtisuorassa eli sisällöllisesti kaukana. Menettelyssä otettiin eniten viitatus joukon jokainen artikkeli ja etsittiin laskettujen vektorien joukosta lähin piste 64 artikkelin otoksesta. Jos otos kattaa semanttisen tilan hyvin, jokaiselle vertailuartikkelille löytyy läheinen naapuri pienen kulman päässä. Laskettu mediaanisamankaltaisuus oli 0,71. Näin ollen otos muodostaa edelleen varsin tiheän peiteverkon semanttisessa avaruudessa. Jos koko aineistossa olisi teemoja, joita otokseen ei sisälly, näiden artikkelien kosinietäisyydet lähimpiin otosartikkeleihin olisivat selvästi pienempiä ja jakauman hännät venyisivät kohti nollaa. Näin korkea minimi- ja mediaanisamankaltaisuus osoittaa tilastollisesti, että 64 artikkelin otos muodostaa tiheän peiteverkon semanttisessa avaruudessa ja kuvaa kattavasti koko relevantiksi rajatun aineiston aihepiirit.

Lopuksi sanastollinen yhdenmukaisuus mitattiin Jensen–Shannon-divergenssillä (Lin 1991), joka mittaa kahden diskreetin todennäköisyysjakauman välistä etäisyyttä Shannonin informaatioteorian pohjalta. Tässä tutkimuksessa jokaiselle sanalle laskettiin suhteellinen frekvenssi sekä 64 artikkelin otoksessa, että eniten

viitattujen artikkelien joukossa stop-sanat poistettuna. Jensen–Shannon-divergenssi laskettiin näiden kahden sanaprobabiliteettivektorin välillä. Saatu arvo 0,082 sijoittuu selvästi lähelle nollaa (täydellinen yhdenmukaisuus), kun teoreettinen maksimi on $\ln 2 \approx 0,693$. Näin sanastollinen jakauma otoksessa muistuttaa erittäin läheisesti vertailujoukkoa, joten kieli ja aiheet vastaavan hyvin koko aineistoa.

3.3.2 Aineiston kvalitatiivinen analyysi

Kvantitatiivisen esikartoituksen tuottamat 64 artikkelia analysoitiin syventävästi soveltaen strukturoitua fokusoitua vertailua (SFC), jossa kukin artikkeli käsitellään itsenäisenä tapauksena, mutta sama ennalta määritetty kysymysjoukko kohdistetaan niihin kaikkiin (George ja Bennett 2005, 67–72, 73–86). Strukturoitu fokusoitu vertailu on politiikan tutkimuksen klassinen laadullinen menetelmä, joka mahdollistaa useiden kohteiden vertailun latistamatta kontekstia. Kullakin artikkelilla käytetty koodausmatriisi sisälsi neljä pääulottuvuutta:

1. propagandan teoreettinen operatiivisuus,
2. käytetty ML-arkkitehtuuri ja datamodaliteetit,
3. raportoitu suorituskyky ja robustius-testit sekä
4. mainitut eettiset tai normatiiviset pohdinnat.

Lisäksi jokaiselle artikkelille annettiin erillinen binaarinen tai ordinaalinen arvo kaikille seitsemälle metodologia-akselille (Taulukko 1): sisältökeskeisyys, multimodaalisuus, koordinaatio, kausaalisuus, temporal-robustus, monikielisyys ja eettinen vinouma. Nämä akselit muodostavat myöhemmin ES-MTT-trendimallinnuksen selittäjämuuttujat; esimerkiksi 'multimodaalisuus' koodattiin arvoilla 0 = ei multimodaalia, 1 = teksti + kuva/ääni ja 2 = täysi V-L-fuusio. Yhtenäinen kysymyspatteristo varmisti, että artikkelit olivat keskenään vertailukelpoisia terminologisista eroista huolimatta.

Koodauksen sisäinen luotettavuus varmistettiin kaksoiskoodauksella vapaaehtoiseksi suostuneen opiskelijan kanssa. Ensin kumpikin koodari prosessoi samat kymmenen pilottiartikkeliä; Cohenin κ laskettiin jokaiselle pääulottuvuudelle (keskiarvo 0,82 \rightarrow ‘erinomainen’; (Miles ym. 2014, 315). Sen jälkeen epäselvät koodiselitteet täsmennettiin. Loput 54 artikkelia tutkielman tekijä analysoi ja tarkisti itse, koska näin laajan aineiston pareittainen ristiin tarkistus olisi liian työläs tehtävä vapaaehtoisena tehtävänä. Menettely noudattaa politiikan tutkimuksen suositusta iteratiivisesta spesifikaatiosta, jossa koodausohje päivittyy pilotin havaintojen perusteella mutta lukitaan ennen varsinaista analyysia (Della Porta ja Keating 2008, 308).

Lopulta laadullinen koodaus integroitiin kvantitatiiviseen trendiosioon exploratory sequential mixed-trend-tracing (ES-MTT) -menetelmän mukaisesti: kunkin artikkelin strukturoidun fokusoidun vertailukoodit (SFC-koodit) muunnettiin binäärisiksi tai ordinaalisiksi tunnusluvuiksi, jotka syötettiin dynaamiseen aihemalliin (ks. luku 3.3.1). Tämän käänteisen kvantifioinnin kautta voitiin testata tilastollisesti, missä määrin tietyt laadulliset piirteet (esim. bayes-integrointi tai GNN-koordinointimittaus) selittävät menetelmien ajallista esiintymistä. Menettely ilmentää George ja Bennettin (2005, 19) kehotusta yhdistää systemaattinen vertailu ja teoriaohjattu hypoteesien testaaminen saman tutkimusasetelman sisällä.

3.4 Metodologinen arvio ja johtopäätös

Esitetty kaksivaiheinen (kvantitatiivinen + kvalitatiivinen) prosessi tuottaa läpinäkyvän ja toistettavan rungon 64 artikkelin ydinjoukon muodostamiseksi. Laaja semanttinen ja ajallinen osittaminen varmistaa, että otos peittää kentän koko variaation eikä painotu vain näkyvimpiin tai tuoreimpiin teemoihin. Kun tähän lisätään geometrisesti vaimennettu verkostovaikutus, esiin nousevat tutkimukset, joiden vaikuttavuus resonoi laajimmin koko viittausverkossa. Kvantitatiivinen esivaihe sidotaan laadulliseen strukturoituun fokusoituun vertailuun (SFC), jota syvennetään “etsivän sekventiaalisen sekoitetun trendijäljityksen” (ES-MTT) periaatteella; näin tutkimukselle muodostuu aidosti yhdistetty sekametodinen (mixed methods) otantakehys.

Kvantitatiivinen validiteetti perustuu kolmoistarkistukseen. Pääkomponenttianalyysi (PCA) osoittaa, että otos sijoittuu tasaisesti koko tf-idf-vektoritilaan, lähimmän naapurin analyysi vahvistaa otoksen semanttisen peittoverkon korkealla kosinisimilariteetilla (mediaanikosinietäisyys 0,92) ja Jensen–Shannon-divergenssi (0,078 puolestaan osoittaa otoksen sanaston vastaavan koko aineistoa. Siluetti-arvo 0,41 tukee kuuden klusterin ratkaisua, ja geometrisesti vaimennettu viittausmittari korreloi altmetriikkaan ($\rho = 0,62$, $p < 0,001$), mikä vahvistaa konstruktiiviteetin. Laadullisen koodauksen reliabiliteetti varmistettiin kaksoiskoodauksella; pilottiartikkelien Cohenin κ oli 0,82, minkä jälkeen epäselviä koodiselitteitä tarkennettiin ja loput artikkelit koodasi tutkija ohjeeseen sitoutuen.

Esitetty moniportainen prosessi tuottaa luotettavan ja toistettavan rungon 64 artikkelin ydinjoukon muodostamiseksi alkujaan lähes kolmesta tuhannesta julkaisusta. Menetelmän vahvuus on sen kaksoisrakente. Ensin laaja semanttinen ja ajallinen stratifikaatio takaa, että otos peittää kentän koko variaation eikä painotu vain näkyvimpiin tai tuoreimpiin teemoihin. Tämän jälkeen geometrisesti vaimennettu verkostovaikutus nostaa esiin julkaisut, joiden vaikutus kantautuu laajimmin tutkimusverkkoon. Tuloksena syntyy otos, joka on samanaikaisesti aiheiltaan monipuolinen ja tieteelliseltä painoarvoltaan merkittävä.

Tutkimuksen keskeiset rajoitteet liittyvät datalähteeseen ja algoritmeihin. Semantic Scholar -kattavuus voi painottua englanninkielisiin julkaisuihin, ja k-means olettaa pallomaisia klustereita. Huomioitavaa on kuitenkin, että hierarkkinen Ward-klusterointi antoi kuitenkin lähes identtisen ryhmärakenteen (Adjusted Rand 0,88). Verkostovaimentimen parametri α vaihteli herkkyydestä (0,4–0,6) vaikuttaen otokseen vain neljän artikkelin verran, mikä osoittaa parametrin stabiiliuden. Laadullisen koodauksen yksintekijäosuus sisältää riskin intensiivisestä tulkintaharhasta, mitä pyrittiin vähentämään parina tehtävällä pilottiartikkelien yhteiskoodauksella ja koodien täsmentämisellä.

Yhteenvetona voisi todeta, että tutkimuksen kvantitatiivinen otantakehys tuottaa tilastollisesti edustavan ja parametriseltaan vakaan semanttisen kartan, kun taas strukturoitu fokusoitu vertailu lisää teoreettista syvyyttä ja tarkentaa propagandan

operatiivisia ulottuvuuksia. ES-MTT-integraatio mahdollistaa hypoteesien koettelun sekä trendi- että tapaustasolla. Näin muodostunut 64 artikkelin otos tarjoaa riittävän laajan mutta hallittavan aineistopohjan tuloslukuja varten, ja tutkimusasetelma on dokumentoitu siten, että se voidaan toistaa tai laajentaa uusiin julkaisuvuosiin.

4 TUTKIMUKSEN TULOKSET JA ANALYYSI

Tässä luvussa esitetään systemaattisen kirjallisuuskatsauksen tulokset sekä niihin perustuva analyttinen tulkinta. Luvun tavoitteena on vastata tutkimuskysymyksiin TK1 (menetelmien kehityskaari 2011–2024) ja TK2 (menetelmien kattavuus, heikkoudet ja vahvuudet). Esitystapa noudattaa kolmea analyttistä aikajännettä (varhaisjakso, keskijakso eli konsolidoitumisvaihe ja myöhäisjakso) ja peilaa jokaisen vaiheen löydökset johdannossa ja teoriassa määriteltyihin kolmeen tarkastelutasoon (retoriset tekniikat, bayes-vastaanottaja, anomaliaverkosto). Luku päättyy poikkileikkaavaan metatriadiin (4.3), jossa kytken tulokset seitsemään menetelmäakseliin (sisältö, multimodaalisuus, koordinaatio, kausaalisuus, temporal-robustisuus, monikielisyys ja eettinen vinouma).

4.1 Yleiskuva julkaisumääristä ja aineiston rakenteesta

Katsaukseen sisältyi 64 vertaisarvioitua ja laajasti viitattua artikkelia vuosilta 2011–2024. Tutkimusaineisto jäsentyy selkeästi kolmeen vaiheeseen – varhaisjakso (2011–2015), keskijakso (2016–2019) ja myöhäisjakso (2020–2024) – joiden väliset erot näkyvät sekä määrällisesti että laadullisesti.

Tarkastelujaksolla julkaisujen määrä kasvoi seuraavasti: varhaisjaksolla (2011–2015) julkaistiin 14 artikkelia, keskijaksolla (2016–2019) 33 artikkelia ja myöhäisjaksolla (2020–2024) 17 artikkelia. Kasvu hidastui vuosien 2020–2024 taitteessa, mikä heijastaa kahta ilmiötä: koronapandemian aikaista konferenssiaikataulujen viivettä, sekä uusimman jakson tutkimusten siirtymistä konferenssijulkaisuista hitaamman julkaisuaikataulun monialatiedelehtiin tai arXiv-preprinteiksi.

Varhaisjaksolla 86 % julkaisuista ilmestyi konferenssijulkaisuissa. Myöhäisjaksolla luku putosi 47 %:iin. Artikkelin ensimmäisen kirjoittajan tausta sijoittui 66 %:ssa tietojenkäsittelytieteeseen, 22 %:ssa yhteiskunta-tieteisiin ja 12 %:ssa informaatiotieteisiin. Tieteenalajakauman perusteella kenttä on siis yhä selvästi tietojenkäsittelytiedevetoinen, vaikka tutkimusongelmat ovat vahvasti yhteiskunnallisia.

Tutkimusten laadullinen tarkastelu osoitti myös selviä painopiste-eroja eri jaksosten välillä, kuten taulukosta 3 voidaan havaita:

Taulukko 3.

Mittari	2011–2015	2016–2019	2020–2024
Syväoppimista hyödyntävät	1/14 (7 %)	20/33 (61 %)	17/17 (100 %)
Monimodaaliset (teksti + kuva/-video)	1/14 (7 %)	5/33 (15 %)	5/17 (29 %)
Tietopohjaiset (KG, fact-check)	0/14 (0 %)	4/33 (12 %)	6/17 (35 %)
Binääri-luokitus pätehtävänä	14/14 (100 %)	27/33 (82 %)	12/17 (71 %)
Mediaani data-setin koko	2 900	38 400	112 000
Mediaani sitaatit/art. ²	185	78	21

Vaikka absoluuttinen luku nousi eniten keskijaksolla, myöhäisjakso korostui laadullisesti. Lähes jokaisessa siihen kuuluvassa artikkelissa hyödynnettiin syväoppimista ja yhä useammin monimodaalisia sekä tietopohjaisia ratkaisuja. Cochran–Armitage-trendianalyysi kolmelle aikajaksolle vahvistaa taulukon kuvaamat muutokset lineaarisesti tilastollisiksi: syväoppimista hyödyntävien tutkimusten osuus kasvaa voimakkaasti ($Z = 5.17$, $p < 0.000001$), samoin tietopohjaisten mallien osuus ($Z = 2.72$, $p = 0.006$). Binääriseen luokitteluun keskittyvien töiden suhteellinen määrä puolestaan laskee merkitsevästi ($Z = -2.12$, $p = 0.034$). Monimodaalisuuden osuus kasvaa taulukossa, mutta Cochran–Armitage-testin tulos ($Z = 1,65$; $p = 0,099$) kertoo, että nousu ei tilastollisesti ole merkitsevä

tavanomaisella 5 prosentin riskitasolla – suunta on myönteinen, mutta vielä epävarma.

4.2 Kronologinen katsaus tutkimuskirjallisuuden metodologiseen ja temaattiseen kehitykseen

4.2.1 Varhaisjakso

Vuosien 2011–2015 tutkimuskirjallisuudessa dis- ja misinformaation tunnistaminen oli vielä hajanaista, ja ilmiölle annettiin erilaisia kontekstisidonnaisia nimiä. Poliittisia, ruohonjuuritasolta lavastettuja kampanjoita kutsuttiin usein astroturffingiksi (Ratkiewicz ym. 2011, 297–299), nopeasti leviäviä vahvistamattomia väitteitä taas huhuiksi (Zhao ym. 2014), kun taas jihadistien verkkojulkaisuihin viitattiin propagandana (Kaati ym. 2015; Ashcroft ym. 2015, 161–162). Myös terveystiedon kohdalla puhuttiin “misinformaatiosta”, mikä näkyi esimerkiksi rokotekestelussa (Dunn ym. 2015, e144). Yhtenäistä jaettua dis-/misinformaatiokäsitteistöä ei siis ollut, mikä heijastui myös tutkimusasetelmiin ja datankeruuseen.

Mikäli tutkimus painottui tiedon leviämisen dynamiikkaan, hyödynnettiin verkko- ja aikaterminologiaa, kuten 'poikkeava tiedon tunnistaminen' (anomaly detection), ajallinen mallinnus (temporal patterns) (Zhao ym. 2014) ja vastaavasti 'veracity of information' (tiedon todenmukaisuus) kun tutkimusongelmana oli tiedon tai viestinnän totuudenmukaisuuden tunnistaminen, erityisesti kriisitilanteissa, kuten maanjäristyksissä jne.) (Bodnar ym. 2014, 206–207). Poliittisen (ammattimaisen) keinotekoisien vaikuttamisen tunnistamiseen sosiaalisissa verkostoissa käytettiin jopa harvinaista² termiä 'astroturffing' (Ratkiewicz ym. 2011, 297–299) ja ääriyhmien viestintään esimerkiksi jihadistien viestintä sosiaalisessa mediassa (Kaati ym. 2015; Ashcroft ym. 2015, 161–162). Terveysaiheisessa analyysissä painotettiin usein mielipiteiden ja tunnetilojen (sentiment) tutkimiseen, erityisesti terveysaiheisen misinformaation näkökulmasta (Dunn ym. 2015, e144).³

² Mainittu 12 julkaisussa 2011–2024 välisenä aikana. 1kpl 2011, 1kpl 2013, 1kpl 2016.

³ Tutkimus kuitenkin avainsanahaun perusteella ainoa terveysaiheinen julkaisu ennen vuotta 2017

Termi misinformaatio esiintyi useissa julkaisussa, mutta termi oli kuitenkin hyvin harvinainen itse otsikossa ja tiivistelmässä, viitaten sen verrattaiseen vakiintumattomuuteen. Tähän viittaa myös esimerkiksi termin implisiittinen merkitys Ratkiewiczin ym. (2011) tutkimuksessa tarkoituksellisesti levitettävänä vääränä tietona, kun taas Dunn ym. (2015) tutkimuksessa, termillä on motiivineutraalimpi käyttötapa (Ratkiewicz ym. 2011, 297–299; Dunn ym. 2015, e144). Burst-analyysin⁴ perusteella, tälle ajanjaksolle nouseviksi ilmiötermeiksi muodotui ”anomalous”, ”grassroots”, ”analysts”, ”hpv”, ”campaign”, ”microblogging” (vanha termi esim. tweeteille), ”astroturf”, ”message”, ”radical”, ”exposure”. Campaign ja grassroots liittyvät erityisesti Ratkiewiczin ym. (2011) tutkimukseen, jossa pyritään erottamaan ”grassroots” eli luonnollisesti tapahtuva viestinä keinotekoisesta ja epätavallisesta (esimerkiksi astroturffauskampanja).

Datasetit kerättiin pääosin alustojen (Twitter, Weibo) avoimien API-rajapintojen kautta, ja aineisto rajattiin tyypillisesti yhteen ilmiöön tai yksittäiseen kriisitapahintaan (esim. Sandy-myrsky, Ebola-paniikki). Mediaanikoko oli pieni ($\approx 2\,900$ riviä) ja luokkajakauma usein epätasapainoinen, mikä rajoitti mallien yleistettävyyttä jo rakenteellisesti. Lisäksi tutkimusten toistettavuus kärsi, koska raakadata (tweet-ID:t, annotaatiot) julkaistiin avoimesti vain harvoissa töissä, kuten Ratkiewiczin ym. (2011) ja Dunnin ym. (2015) tutkimuksissa.

Menetelmällisesti alan lähtöpiste oli vahvasti tietopohjainen. Valvotut koneoppimismallit perustuivat inhimilliseen annotointiin, jota toteuttivat sekä asiantuntijat että joukkoistetut vapaaehtoiset. Asiantuntijat tarkistivat esimerkiksi jihadististen Twitter-tilien sisällön ennen kuin tilit poistettiin (Ashcroft ym. 2015, 161–162), ja rokotekeskustelussa he luokittelivat twiittejä sentimentin mukaan (Dunn ym. 2015, e144). Tutkimukset tunnistivat kuitenkin ihmistyön manuaalisen pullonkaulan; manuaalinen labelointi hidasti skaalautumista nopeasti kasvaviin somevirtoihin. Truthy-järjestelmässä automaattinen klusterointi yhdistettiin käyttäjäh-

⁴ Burst-analyysi (engl. *burst detection*, *burst analysis*) etsii ”purskeita”; yliaikaisia jaksoja, jolloin jonkin termin, aiheen tai ilmiön esiintymistiheys nousee selvästi taustatasoa korkeammaksi, eli sanat alkavat esiintyä odottamatonta vauhtia. Algoritmi on yleistynyt bibliometriikassa, uutisvirtojen seurannassa ja sosiaalisen median trendien tunnistamisessa.

teisön jälkiarviointiin, jolloin massojen viisaus toimi validointikerroksena (Ratkiewicz ym. 2011, 297–299). Lähteiden luotettavuutta hierarkisoitiin selkeästi. Viranomais- ja uutislähteet muodostivat niin sanotun “kultastandardin”, joihin sosiaalisen median dataa verrattiin (Bodnar ym. 2014, 206–207). Tämä hierarkkinen tietokäsitys – jossa tietyt lähteet nähtiin lähtökohtaisesti luotettavampina kuin toiset – edusti perinteistä episteemistä lähestymistapaa misinformaation torjunnassa.

Tutkimuksissa data jaettiin lähes aina sattumanvaraisesti kahteen osaan – harjoitteluun ja kokeiluun (i.i.d.-jako). Sen sijaan ei testattu, miten malli pärjäisi myöhemmin kerätylle aineistolle (ajallinen hold-out) tai kokonaan toiselta some-alustalta tuleville viesteille (cross-platform-siirtotesti). Siksi ei syntynyt tietoa, pysyvätkö tulokset yhtä luotettavina ajan kuluessa (temporal-robustisuus) tai eri platform-ympäristöissä (domain-transfer).

Tekstipohjaisessa analytiikassa hyödynnettiin tuolloin pääasiassa perinteisiä NLP-piirteitä: sentimenttiä, stilometrisia tunnuslukuja ja n-grammeja. Koneoppimis-yhdistelmät rajoittuivat lähinnä SVM-, Naive Bayes- ja AdaBoost-malleihin; syväverkkoja ei vielä sovellettu. AdaBoost-luokitin erotti poliittisesti motivoitunut astroturffausviestit tavallisista twiiteistä 96 prosentin tarkkuudella (Ratkiewicz ym. 2011, 297–299), ja jihadistiviestien luokitus nousi jopa yli 99 prosenttiin, kun stilometriset ja ajalliset piirteet yhdistettiin (Ashcroft ym. 2015, 161–162). Korkeat luvut osin selittyivät kuitenkin datariippuvuudella: mallit toimivat hyvin siinä ympäristössä, josta opetusdata oli kerätty, mutta siirrettävyydestä muihin konteksteihin oli vähän näyttöä. Arviointimetrikka rajoittui käytännössä tarkkuuteen (accuracy) ja alustavaan F1-lukuun; AUC-, makro-F1- tai robustiusmittauksia ei käytetty, mikä vaikeuttaa tulosten vertailua myöhempisiin syväoppimis-tutkimuksiin.

Koska sisältöanalyysi ei yksin riittänyt, tutkijat alkoivat tarkastella myös verkostoja käyttäjäpiirteitä. Replikoitavuuden kannalta olennaista raakadataa kuitenkin tarjottiin harvoin avoimena – poikkeuksina Ratkiewicz ym. (2011) sekä Dunn ym. (2015), jotka julkaisivat sekä tweet-ID:t että annotaatiot. Tämä osaltaan rajasi menetelmävertailujen mahdollisuuksia. Retweet-topologiat, viestien leviämisenopeus ja tilien aikataulutettu synkronointi toimivat vahvoina signaaleina koordinoitua toiminnasta (Ratkiewicz ym. 2011, 297–299; Zhao ym. 2014). Profiili- ja

historiatietoihin perustuvat luotettavuusmallit laskivat käyttäjälle pisteitä sen mukaan, kuinka johdonmukaisesti tämä oli aikaisemmin jakanut paikkansapitävää informaatiota (Bodnar ym. 2014, 206–207). Sama ajatus sovellettiin jihadistien “media-mujahideen” -verkostoon: tili luokiteltiin propagandan moninkertaistajaksi, jos sen verkosto- ja julkaisuaktiivisuus poikkesi selvästi tavanomaisesta (Kaati ym. 2015).

Huomionarvoista on, että kaikki tämä tapahtui lähes yksinomaan tekstuaalisen datan varassa. Kuvat, videot ja URL-linkkien sisältämä konteksti kirjattiin kyllä metatietona, mutta automaattista visuaalianalyysiä ei vielä implementoitu – tutkimusten oma reflektio tunnisti multimodaalisuuden avoimeksi kysymykseksi ja tutkimusaukoksi, mutta niitä ei vielä systemaattisesti analysoitu. Esimerkiksi Ashcroft ym. (2015) mainitsivat jihadistien YouTube-videot keskeisenä propagandakanavana (Ashcroft ym. 2015, 161–162), mutta varsinainen kuvien tai videoiden automaattinen sisältöanalyysi jäi myöhempien vuosien haasteeksi. Tietopohjaiset menetelmät nojasivat siis vahvasti inhimilliseen panokseen – joko asiantuntijoiden tai joukkoistettujen tarkistusten muodossa.

Piirteisiin pohjautuvat analyysimenetelmät lähestyivät tunnistusmetodologiaa identifioimalla ja laskemalla erilaisia piirrekokonaisuuksia-tai avaruuksia. Nämä piirteet voidaan karkeasti jakaa sisältö-, verkosto- ja käyttäjäperäisiin. Usein tutkimuksissa konstruointiin malleja, jotka hyödynsivät yhtä tai useampaa piirretyyppiä arvioidakseen sisällön tai toiminnan autenttisuutta. Vuosina 2011–2015 lupaavimmat menetelmät usein yhdistivät eri piirteitä – kuten sisältöanalyysin tuloksia ja sosiaalisen verkoston ominaisuuksia parantaakseen tunnistuksen tarkkuutta.

Yhteenvetona varhaisjakson metodologia nojaa kahteen tukipilariin. Ensimmäinen on human-in-the-loop: ilman ihmisen tekemää annotointia koneoppiminen ei kyennyt tuottamaan käyttökelpoisia malleja, ja asiantuntija- tai joukkoistustaso ratkaisi lopullisen totuusarvion. Toinen on käsin suunniteltu piirre-painotteisuus: suorituskyky syntyi lingvistisistä, verkostollisista ja käyttäjäkohtaisista muuttujista, ei automaattisesta piirreoppimisesta. Vaikka tulokset vaikuttivat lupaavilta –

jopa hämmästyttävän tarkoilta – ne osoittautuivat herkästi kontekstidatasta riipuvaisiksi, ja visuaalisen sekä monimodaalisen disinformaation tunnistaminen jäi vielä kartoittamattomaksi alueeksi. Tämä perintö määritti reunaehdot myöhemmille, syväoppimisen ja multimodaalisuuden aikakausille. Erityisesti Ratkiewicz ym. (2011) saavutti aiemmin kuvatulla polkuvaimennetulla viittausvaikutuskavalla laskettuna noin 1107 vaikuttavuuspistettä 2916 julkaisun joukossa. Myös useimmat tällä ajanjaksolla ilmestyneet julkaisut saavuttivat huomattavasti keskimääräistä suuremmat vaikuttavuuspisteet.

4.2.2 Keskijakso

Maturiteettivaiheen tutkimuksissa (2016–2019) tunnistettiin tarve standardoiduille ja avoimille aineistoille. Samaan aikaan julkaisut siirtyivät ensimmäistä kertaa satojen tai jopa kymmenien tuhansien esimerkkien datasetteihin – LIAR (~12 k väitettä) ja FakeNewsNet (~23 k artikkelia + 1,6 M twiittiä) nostivat mittakaavan kymmenkertaiseksi verrattuna varhaisjaksoon, mahdollistaen koneoppimismallien tehokkaamman ja vertailukelpoisemman kehittämisen.

Koneoppimismallien kehittyminen loi sekä uusia haasteita ja mahdollisuuksia. Tämän ajanjakson loppupuolella ensimmäistä kertaa aineistossani ilmeni termi "deepfake", viitaten tekoälyllä manipuloituun mediaan, erityisesti kasvojen vaihtoihin (4kpl /2019). Tällainen koneoppimis pohjainen tekniikka yleistyi noin vuosina 2018–2019 (Sabir ym. 2019,1; Mirsky ja Lee, 2021,1). Tämä osaltaan loi painetta tietojoukkojen luomiseen tunnistusmallien koulutusta varten sekä tunnistusmallien kehittämiseen.

Muuttuva maailmanpolitiikka ja erityisesti Yhdysvaltojen presidentinvaalit merkitsivät erittäin suurta temaattista muutosta terminologiassa, tutkimusaineistoissa, tutkimuspainotuksissa ja metodologiassa. Tämän voimme nähdä mm. tutkimusaineistoommekin heijastuneessa ilmiössä, jossa termi vale uutiset (fake news) räjähti populaariin tietoisuuteen. Termin suosio näkyy noin 12-kertaisena nousuna google trends⁵ -tilastoissa vuosien 2016–2017 välillä ja tutkimusaineistossani

⁵ <https://trends.google.com/trends/explore/TIMESERIES/1746437400?hl=en-GB&tz=-180&date=all&hl=en-GB&q=fake+news&sni=3>

tämä termi ilmestyi ensimmäisen kerran vuonna 2017 ja kasvoi räjähdysmäisesti seuraavina vuosina (10 kpl/2017; 53 kpl/2018; 122 kpl/2019).

Tutkimuksissa esitetyt eksplisiittiset ja implisiittiset määritelmät valeuutisista vaihtelivat usein. Esimekiksi valeuutisia kuvattiin joko tarkoituksella harhaanjohtavana tai vääränä tietona, joka tarkoituksellisesti esitetään oikeellisena. Tällöin määritelmällisesti silkkä väärä tieto olisi 'virheellisiä uutisia' (false news). Toisaalta termiä hyvin usein käytettiin kuvaamaan 'epätotta tietoa' jonka tutkijat joko manuaalisesti arvioivat tai vertaavat faktantarkastustietojoukkohin, usein ilman että tutkimuksessa pyrittäisiin suoraan havainnoimaan tarkoituksellisuutta (Roy ym, 2018,1). Voidaan toisaalta päätellä, että tätä voidaankin olettaa tietyssä kontekstisidonnaisuudessa, esimerkiksi Syyrian sodasta levitettävä väärä tieto oli sellaista, että alkuperäisen levittäjä ei olisi tahattomasti voinut erehtyä - tai itse informaatio oli manipuloitua. Termi myös liittyi vahvasti vuoden 2016 Yhdysvaltain presidentinvaaleja ympäröineeseen julkiseen keskusteluun ja huoleen vaali-vaikuttamisesta (Wang 2017,1).

Tällä ajanjaksolla suosiota saanut disinformaation käsite (tahallinen harhaanjohtaminen) erotettiin misinformaatiosta (joka voi olla tahatonta), vaikkakin käytännössä tutkimuskentällä terminologinen ero saattoi jäädä häilyväksi. Thorne ja Vlachos (2018) korostavat myös automatisoitujen mallien vaikeutta todentaa näiden välistä eroa: "Disinformaatio edellyttää lisäksi pahantahtoista tarkoitusta johtaa yleisö harhaan..." ...Automaattinen faktantarkistus... ei pysty erottamaan sitä misinformaatiosta (kirjoittajan vapaa suomennus). (Thorne ja Vlachos 2018, 3). Disinformaatio-termi myös osin ohitti valeuutistermin politisoituneisuutta: Thorne ja Vlachos (2018) huomauttavatkin, että valeuutis-termin merkitys on hämärtynyt ja sitä käytetään poliittisen vastapuolen nimittelyyn tai sitä yhdistetään usein virheellisesti termiin 'vihapuhe' (Thorne ja Vlachos 2018, 3). Misinformaatio terminä vakiintui kuitenkin yhä enemmän termiksi väärälle tai harhaanjohtavalle tiedolle, jonka tarkoitusperä on tuntematon ja/tai todistamaton. Esimerkiksi Terveysalan kriiseissä, kuten Zika-epidemiassa, käytettiin usein termejä 'misinformaatio' ja 'huhut' (kokonaiset 17 ja 128 kertaa artikkelissa), sikäli kun väärän tiedon levittämisen tarkoituksellisuus ei ollut tutkimuksen kohteena (Ghenai ja Mejova 2018, 1).

Tälle temaattiselle muutokselle johdonmukaisesti yhä useammat tutkimukset kehystettiin automaattisen todentamisen ja faktantarkistuksen metodologian kehittämiseksi ja arvioinniksi ja terminologia siirtyi enemmässä määrin termeihin faktantarkistus /fact verification (Thorne & Vlachos, 2018, 1), automatisoitu faktantarkistus /automated fact-checking (Graves, 2018, 1), väitteiden arviointiin/ claim, käytettyjen todisteiden arviointiin/ evidence ja analyysin tulosten nimeämiseen sekä jäsentelyyn (verdict) (Aly ym., 2021, 1). Myös mielipiteen tunnistus /stance detection, sekä väitteiden ja todisteiden välisten suhteiden analyysi kuuluivat tähän trendiin (Rakholia ja Bhargava 2017, 1).

Tutkimuskenttä siirtyikin yhä enemmän intentionaalisuuden arvioinnista implisiitisiin oletuksiin, eivät ottaneet kantaa tai voineet todistaa tutkimuskehityksen valossa viestinnän intentiota. Keskijakson havaintometodologiat yhä useimmin pyrkivät arvioimaan tekstin ”totuusarvoa” tai lähteen esittäjän yleistä luotettavuutta. Tosin voidaan loogisesti päätellä, että vaikka lähteen luotettavuudelle voidaan arvioida todennäköisyys, kuten varhaisessakin tutkimuksessa tutkittiin, tämä ei silti todista epäluotettavuuden syytä. Kuten teoriaosuudessa sivusimme, argumentaatiovirheet ovat inhimillisiä, eivätkä välttämättä osoita tahtoa tarkoitukselliseen harhaanjohtamiseen tai valehteluun.

Tietopohjaiset menetelmät ja tietojoukot

Tällä ajanjaksolla tutkijat alkoivat kerätä ja käyttää erityisesti valeuutisten tunnistamiseen tarkoitettuja tietojoukkoja, jotka usein koottiin faktantarkistussivustoilta (kuten PolitiFact, Snopes) tai merkittiin manuaalisesti erilaisista lähteistä. Esi-merkkejä ovat hienojakoisiin poliittisiin väitteisiin keskittyvä LIAR (Wang 2017), otsikon ja leipätekstin vastaavuutta tarkasteleva FNC (käytössä mm. Rakholia ja Bhargava 2017) sekä ISOT (käytössä mm. Jiang ym. 2021) -aineisto. LIAR-tietojoukon julkistanut tutkimuspaperi on myös tämän aineiston eniten vaikuttavuuspisteitä kerännyt julkaisu (2680.5), mikä on noin kaksi kertaa enemmän, kuin toiseksi eniten pisteitä saanut. LIAR-tietojoukon kaltaiset avoimet, benchmarkattavat, ja siten helposti vertailtavat tietojoukot nousivat merkittäväksi lähteeksi useimpiin havaintomalleja kehittäviin tutkimuksiin.

Nämä aineistot sisälsivät pääasiassa tekstimuotoista uutissisältöä ja totuusarvo-leimoja, mikä ohjasi tutkimusta kohti tekstipohjaisia koneoppimis- ja syväoppimisluokittelijoita. Toisaalta myös tunnistettiin, että pelkkä uutistekstin sisältö ei useinkaan riitä, mikä johti pyrkimyksiin sisällyttää sosiaalista kontekstia. FakeNewsNet (Shu ym. 2019b) oli keskeinen tietovaranto, joka yhdisti uutissisällön (PolitiFact, GossipCop) sosiaaliseen kontekstiin Twitteristä (käyttäjäprofiilit, twiitit, seuraajat jne.). Tämä mahdollisti sosiaalisen verkoston ja käyttäjäpiirteiden laajemman hyödyntämisen (Vosoughi ym. 2018; Della Vedova ym. 2018). Myös ensimmäiset multimodaalisuuteen keskittyvät datajoukot (esimerkiksi deepfake-aineistot kuten FaceForensics++) (Sabir ym. 2019) alkoivat ilmestyä tämän vaiheen loppupuolella. Datan keruussa faktantarkistussivustojen hyödyntäminen yleistyi, ja luotettiin asiantuntijoiden tekemään todentamiseen (Wang 2017; Shu ym. 2019b).

Tällaiset **asiantuntijaperusteiset menetelmät** nojasivat siis usein faktantarkistusorganisaatioiden (kuten PolitiFact) tuottamiin arvioihin, joita käytettiin opetusdatana koneoppimismalleille. Tutkimuskirjallisuudessani suurimman vaikutuspistemäärän saaneen (2680.5) tutkimuksen kokoama LIAR-datasetti perustui juuri tähän PolitiFactista kerättyyn dataan (Wang 2017, 1). Vaikka asiantuntijapohjainen-annotointi on tarkkaa, sen hitaus ja vaatimus suurelle määrälle manuaalista työtä tunnistettiin edelleen haasteeksi (Graves 2018, 3). On huomioitava myös, että LIAR on rajattu angloamerikkalaiseen poliittiseen kontekstiin, mikä rajoittaa mallien yleistettävyyttä muihin kielialueisiin. Tämän datan pohjalta koulutetut mallit voivat siis toimia heikosti poliittisessa keskustelussa, jossa kieli, kuvakulttuuri ja argumentaatiotavat ovat monikielisiä tai LIAR – koulutusdataan nähden ‘liian eriävää’. Nämä uudet tutkimukset edustivat muuttuvaa tutkimuskenttää; asiantuntija-arvio toimii nyt vain enemmässä määrin ainoastaan referenssipisteenä, jonka jälkeen prosessia pyritään skaalaamaan automaattisesti miljooniin dokumentteihin, esimerkiksi someviesteihin. Tänä ajanjaksona faktantarkistustyön helpottamiseksi syntyi sekä asiantuntija-alustoja (PolitiFact, Snopes) että myös suoraan joukkoistettuja työkaluja (Fiskkit, TextThresher). Samalla ajanjaksolla toistettiin myös varhaisen tutkimusten havainto siitä, että käsin tehtävä faktantarkistus ei skaalaudu hyvin sosiaalisen median valtavaan julkaisutahtiin. (Zhou ym. 2019, 2).

Syväoppiminen kehitys eri modaliteettiipiirteiden analyysissa

Maturiteettivaiheessa, tekstianalyysin syväoppimismetodien skaala laajentui merkittävästi tutkimusalueella syväoppimisen kehityksen myötä. Vaikka stilometriset piirteet (kielen kompleksisuus, havaittavat tunnetilat) tunnistettiin edelleen hyödyllisiksi (Ahmed ym. 2018, 2; Horne & Adali 2017), pääpaino siirtyi piirteiden automaattiseen oppimiseen esimerkiksi CNN- ja RNN-verkoilla. Royn ym. (2018) esittämä automaattiseen oppimiseen perustuva 'deep-ensemble'-lähestymistapa, jossa CNN- ja Bi-LSTM-esikäsitteily yhdistettiin MLP-luokittimeen, nosti kuu-siluokkaisessa LIAR-datasetin kokeessa havainnointitarkkuuden 44.87 prosenttiin, joka kolminkertaisti satunnaisluokittelun tason ja ohitti kirjoittajien mukaan siihen mennessä aiemmin parhaimmat tulokset saaneen CNN-mallin (Roy ym. 2018, 1). Royn ym. (2018) esittämän ensemble-mallin aiempiin artikkeleihin verrattuna alhaiselta vaikuttavat tulokset voidaan kenties selittää sillä, että binäärisen luokittelun yksinkertaisuus mahdollistaa paperilla tarkemmat tulokset, kuin LIAR-datasetin 6-luokkainen jaottelu. Esimerkiksi 'barely true' ja 'half true' sekoituvat helposti keskenään (Roy ym. 2018, 5). Tämä käy järkeen myös siksi, koska binäärinen luokittelu ei ota kantaa totuusarvon nuansseihin tai ylipäättään sen luonteeseen (esimerkiksi satiiri, eli totuudellinen mutta harhaanjohtava, tai valheellinen, joka sisältää totuudellisia piirteitä). Toisin sanoen kaikki moniluokkainen (enemmän kuin kaksi) luokittelu on vaikeammin laskettavissa. Ajao ym. (2018) tuovat esille myös konvoluutioverkkoja ja pitkän muistin toistoverkkojen (aiemmin mainitut CNN ja LSTM) hyödyt twiittien binäärisessä luokittelussa (82 prosentin tarkkuus pelkällä LSTM – neuroverkolla 5800 twiitin PHEME-tietojoukolla).

Myös teoriapohjaisiin ja manuaalisesti annotoituihin tekstipiirteisiin pohjautuviin tunnistusmenetelmiin kehitettiin samanaikaisesti uusia lähestymistapoja. Nämä eivät kuitenkaan saaneet samaa suosiota tällä ajanjaksolla kuin automaattisesti laskettujen tekstipiirteiden analyysit. Zhou ym. (2019) kehittivät motiivipohjaisen piirrejoukon, jossa kahden keskeisen huijauspsykologisen selitysmallin hypoteesit muutettiin mitattaviksi kieli-indikaattoreiksi. Ensimmäisenä viitekehystenä toimi nk. 'Reality Monitoring' -teoria (Johnson ja Raye, 1981), jonka mukaan

todellisiin muistoihin sisältyy runsaasti aistikokemuksia ja kontekstuaalisia yksityiskohtia, kun taas keksityt kertomukset ovat abstraktimpia.

Teorian ennustetta testatakseen Zhou ym. (2019) laskivat LIWC-sanaston avulla näköön (see), kuuloon (hear) ja tuntoaistiin (feel) viittaavien sanojen suhteelliset frekvenssit sekä kognitioprosesseja kuvaavien ilmausten, kuten insight ja causation esiintymisen (Zhou ym. 2019, 9). Toisena perustana oli nk. 'Four-Factor Theory' (Zuckerman ym. 1981), joka selittää huijaukseen liittyvää vireystilan kohoamista, kognitiivista kuormaa, emootioiden hallintaa ja käyttäytymiskontrollia.

Tähän teoriaan pohjautuen tutkijat keräsivät tekstuaalisia piirteitä, jotka heijastivat epävarmuutta ja ristiriitaa (esim. should, perhaps, always), emotionaalista kuormitusta (negatiivisten tunne-sanojen osuus) sekä itseilmaisun kontrollia (erottavien konnektorien ja possessiivien käyttö). Nämä koottiin Specificity-luokkaan osana laajempaa Disinformation-related Attributes -sarjaa (Zhou ym. 2019, 21–22). Kaikki uutisartikkelien tekstit muunnettiin näin neljän kielitason — leksikonin, syntaksin, semantiikan ja diskurssin — teorialähtöisiksi mittareiksi. Yhteensä yli kolmesataa käsin määriteltyä ominaisuutta syötettiin tukivektorikoneeseen, satunnaismetsään (random forest) ja XGBoost-luokittimeen. Malli saavutti PolitiFact- ja BuzzFeed-datan tarkkuudeksi 0,892 – 0,879 ja kykeni siten tunnistamaan valeuutisen jo pelkän otsikon ja ensimmäisen kappaleen perusteella ($F1 \approx 0,80$), siis ennen verkkoleviämisen alkamista (Zhou ym., 2019, 19). Teorialähtöinen lähestymistapa tarjoaa kaksi merkittävää etua: ensinnäkin se lisää tulkittavuutta, sillä jokainen indikaattori on sidottu eksplisiittiseen psykologiseen oletukseen, ja toiseksi se parantaa siirrettävyyttä, koska piirteet eivät perustu yksittäisen korpuksen n-gram-tilastoihin, vaan yleistettäviin kognitiivislingvistisiin mekanismeihin. Monissa artikkeleissa, joissa keskityttiin ihmiseen liittyviin tekijöihin, hyödynnettiin kognitiivisia tai affektiivisiä termejä (Janze ja Risius 2017, 2).

Joukkoistuksen kehitys ja tunnistetut haasteet

Uusimmissa joukkoistamismetodeissa usein pyrittiin esittämään keinoja skaalata aiemmin hitaaksi ja tutkimusten pullonkaulaksi tunnistettua tiedon (someviestien

yms.) annotointia. Ghenai ja Mejova (2018, 1, 5) yhdistivät asiantuntijatiedon Zika-huhuista ja joukkoistetun annotoinnin Twitter-datasta, saavuttaen korkean F1-tuloksen (94,5 %) huhujen ja oikaisujen luokittelussa. Tacchini ym. (2017) puolestaan osoittivat, että Facebookin tykkäyskuviot voivat toimia myös implisiittisenä joukkoistussignaalinä: heidän mallinsa erotti valeuutissivujen postaukset aidoista lähes täydellisellä AUC-arvolla ($> 0,99$) pelkkien tykkääjien perusteella (Tacchini ym. 2017, 1; Della Vedova ym. 2018, 3 viittaavat Tacchinin ym. (2017) $AUC > 0.99$ tulokseen). Tämä korostaa käyttäjäyhteisöjen merkitystä, mutta sen yleistettävyyden voi olla rajallista. Vaikka joukkoistaminen tarjoaa nopean keinon tuottaa suuria määriä merkittävää dataa, käyttäjäraportit ovat väistämättä ”kohinalisia” – osa lukijoista merkitsee aitoja uutisia virheellisesti valeuutisiksi ja perustelut voivat jäädä pinnallisiksi (Wang ym. 2019, 2). WeFEND-kehityksen kehittäjät ratkaisivat ongelman RL-valitsimella (reinforced selector), joka poisti noin 25 % heikoista näytteistä, ennen kuin lopullinen BERT-pohjainen tunnistin saavutti parhaat tuloksensa (Wang ym. 2019, 3, 5–7). Sama kohinallisuusongelma näkyy LIAR-vertailuissa: vaikka datasetti on ”järjestyksessään suuruusluokkaa aiempia laajempi”, sen riippuvuus Yhdysvaltain poliittisesta kontekstista synnyttää siirto-ongelman monikielisiin ympäristöihin (Wang 2017, 2–3).

Graves (2018) puolestaan muistuttaa, että automaattinen faktantarkistus on edelleen altis virheille uskottavistakin lähteistä ja vaatii siksi ihmisvalvontaa - etenkin, kun päätökset voivat rajoittaa poliittista ilmaisua (Graves 2018, 1–2). Tällä ajankaksolla esitetyt joukkoistamismetodit pyrkivät siis vastaamaan skaalaushaasteen, jonka Ashcroft ym. (2015) pitivät keskeisenä ongelmana, mutta samalla siirsivät virhe- ja vinoumariskin eksperteiltä joukkoistetulle signaaleille. Ratkaisuksi ehdotettiin automaattisia selektoreita ja monitasoisia varmistusprosesseja (esimerkiksi ihmisvalvonta).

Kuva- ja videoväärennosten (deepfake) uudet haasteet havaintometodien multimodaalisuudelle

Tekoälypohjaiset kuva- ja videogeneraatiot nousivat 2016–2020 keskeiseksi haasteeksi dis- ja misinformaation torjunnassa. GAN-generaattorien yleistyessä sekä täysin uudet kasvokuvat että syväväärennös videot muuttuivat ihmissilmien lähes erottamattomiksi — ilmiö, jonka Kai Nakamura ym. (2020, 1) luonnehtivat

“seeing is no longer believing”. Teknologiset läpimurrot, erityisesti tekoälyssä, olivat merkittävä uusien termien lähde. Generatiivisten verkkojen (GAN) ja syväoppimismallien kehittyminen mediasynteesissä johti suoraan termin ‘deepfake’ syntyyn ja käyttöönottoon noin vuosina 2018–2019 (Mirsky ja Lee, 2021, 1; Sabir ym. 2019, 1).

Jo ennen StyleGAN:n varsinaista julkaisua GAN-generaattorit alkoivat tuottaa lähes virheettömiä kasvokuvia. Käännekohtana pidetään Progressive Growing of GANs -työtä (PGGAN), jonka arXiv-versio (12/2017) osoitti ensimmäisenä, että 1024×1024 pikselin resoluutio on saavutettavissa ilman selviä artefakteja (Karras ym. 2018, 5). Samaan aikaan Redditissä levinneet niin kutsutut deepfake-videot (12/2017) nostivat syväväärengökset suuren yleisön ja tutkijoiden tietoisuuteen, mikä siirsi painopistettä puhtaasti tekstuaalisista tarkistusmenetelmistä multimodaalisiin ratkaisuihin.

Empiirisen tutkimuksen kiihdytin oli dataset-ekosysteemin nopea laajeneminen vuosina 2018–2019. ‘FaceForensics’ ja sen laajennettu versio ‘FaceForensics++’ kokosivat yhteensä yli 1,8 miljoonaa manipuloitua kasvokuvaa valvotun oppimisen tarpeisiin (Rössler ym. 2019, 2), ja Facebookin julkaisema Deepfake Detection Challenge Preview laajensi mittakaavan 5 000 videoon kattaen useita generointitekniikoita (Dolhansky ym. 2019, 1).

Detektio tutkimus erkani kahteen päähaaraan.

Ensimmäisessä hyödynnettiin artefaktisignaaleja: Li ja Lyu (2018, 2) paljastivat syväväärengökset havaitsemalla epärealistisia silmänräpäysrytmejä, kun taas Afchar ym. (2018, 1) esittelivät MesoNet-arkkitehtuurin, joka tunnistaa manipuloinnin mesotasoisista tekstuuripiirteistä.

Toisessa haarassa kehitettiin multimodaalisia fuusiomalleja; esimerkiksi Event Adversarial Neural Network (EANN) yhdisti otsikon, leipätekstin ja upotetun kuvan samaan latentiin tilaan, jolloin vale uutisen tunnistus perustui samanaikaisesti kieli- ja kuvapiirteisiin (Wang ym. 2018, 3).

Varhaiset tulokset olivat lupaavia mutta datariippuvaisia. FaceForensics-pohjaiset konvoluutioverkot ylittivät 90 prosentin tarkkuuden, kun testi- ja koulutusdata oli tuotettu samalla generaattorilla, mutta generalisoituivat huonosti uusiin väärennöstekniikoihin – rajoite, joka nousi tutkimusagendalle enemmissä määrin vasta 2020-luvulla.

Verkostoalgoritmit, varhainen tunnistus ja käyttäjäpiirteet

Informaation leviämisdynamiikan analysoinnin tutkimus myös eteni tällä ajanjaksolla, vaikkakin tämä tutkimuksen haara ei myöskään saavuttanut suurinta suosiota tutkimuskentällä. Tutkimusten keskittyessä tiedon leviämisen dynamiikkaan, käytettiin usein verkko- ja aikaterminologiaa, kuten 'leviämisspolut' (propagation paths) (Liu ja Wu, 2018, 1). Liu ja Wu (2018) kehittivät tällä ajanjaksolla verkostopohjaisen mallin, joka kykenee tunnistamaan valeuutisten leviämisen hyvin varhaisessa leviämisvaiheessa sosiaalisessa mediassa. Mainittavaa on myös termin 'valeuutiset' käyttö, mikä terminä on puuttunut tutkimuskentän varhaisimmista tutkimuksista. Tutkimus osoitti, että vain viidessä minuutissa julkaisun jälkeen malli saavuttaa 85 % tarkkuuden Twitterissä (monikielinen) ja jopa 92 % tarkkuuden Sina Weibossa (Kiinalainen microbloggausalue) valeuutisten tunnistuksessa. Mallin tehokkuus perustui erityisesti käyttäjäprofiileista johdettuihin piirteisiin ja verkostoanalyysiin, joiden avulla valeuutiset voitiin tunnistaa ennen laajan leviämisen tapahtumista (Liu ja Wu, 2018, 1).

Tämä toistaa tutkimuskentällä aiemmin mainittuja johtopäätöksiä, varhaisen tunnistuksen merkityksestä ja sen suhteesta epäorgaanisen viestinnän havaitsemisen tarkkuuteen. Tutkimuksessa esiteltiin uusi aikajaksoluokitukseen perustuva lähestymistapa, jossa yhdistetään takaisinkytkettyä ja konvoluutioneuroverkkoa (recurrent and convolutional networks). Kukin uutisen leviämisspolku mallinnettiin ajalliseksi monimuuttujaisarjaksi, jonka jokainen askel on uutista eteenpäin leviävän käyttäjän piirteitä kuvaava numeerinen vektori (Liu ja Wu 2018, 2–3).

Esitetty malli on neliosainen:

1. Leviämispolun (propagation path) muodostus ja muunnos,
2. takaisinkytketty neuroverkko (RNN) oppii käyttäjäpiirteiden globaaleja muutostrendejä polun varrella,
3. konvoluutioneuroverkko (CNN) puolestaan sieppaa paikallisia vaihteluita käyttäjäjonossa
4. lopuksi nämä neljä osaa yhdistetään leviämispolun yhdeksi klassifikaatioksi (Liu ja Wu, 2018, 3–4).

Liu ja Wu (2018) tutkivat valeutisten leviämistä siis yksilöiden käyttäytymisen ja verkostorakenteen kautta. He perustelevat käyttäjäpiirteisiin keskittymistä sillä, että varhaisessa vaiheessa valeutista jaettaessa käyttäjät yleensä vain retwiittaavat lisäämättä omia kommenttejaan, jolloin tekstipohjaiset tai rakenteelliset vihjeet ovat pinnallisia. Lisäksi kaikki twiitit eivät edes sisällä käyttäjän omaa tekstiä tai tekstiä lainkaan, esimerkiksi kuvia. Lisäksi käyttäjäkommentteja voidaan manipuloida helposti, kun taas käyttäjätilin ominaisuuksia on vaikeampi manipuloida. Näin ollen käyttäjien profiili- ja käyttäytymispiirteet tarjoavat luotettavamman perustan varhaiselle tunnistukselle kuin kielelliset vihjeet. (Liu ja Wu 2018, 1–2). Kirjoittajien mukaan ehdotettu malli on yleistettävämpi ja kestävämpi varhaisessa valeutisten tunnistuksessa, koska se tukeutuu ainoastaan yleisiin käyttäjäominaisuuksiin. Malli ei siis nojaa lainkaan viestin tekstisisältöön, vaan ainoastaan käyttäjäprofiilien ominaisuuksiin ja niiden jakamisjärjestykseen (Liu ja Wu 2018, 2–3).

Myös Graph Neural Network (GNN) -mallit yleistyivät tällä ajankaksolla; Monti ym. (2019, 1, 5) osoittivat GCN-mallin saavuttavan ~93 prosentin AUC-arvon yhdistämällä sisällön, käyttäjät ja verkostorakenteen, kyeten jo varhaiseen valeutisten tunnistamiseen. Verkostopiirteiden (laajemman, nopeamman ja syvemmän leviämisen) todettiin myös olevan vahvoja indikaattoreita (valheet leviävät eri tavoin, yleensä paremmin, kuin totuudellinen tieto) (Zhao ym. 2018, 2). Tutkimukset, jotka käsittelivät tiedon leviämistä ja sen dynamiikkaa, käyttivätkin usein relevantteja, usein totuusarvon määrittäviä termejä kuten 'toden ja väärän uutisen leviäminen' (spread) (Zhao ym. 2018,1; Vosoughi ym. 2018). Esimerkiksi

Vosoughi ym. (2018) suorittivatkin laajan analyysin perustuen leviämisdynamiikan analyysiin (sosiaalisen verkoston piirteisiin) käyttäen faktantarkistettuja Twitter-huhuaaineistoja. Tutkimus myös osoitti, että mittarit kuten kerrosten lukumäärä, kerrosten suhde, ominaispolunpituus ja heterogeenisyys erottavat keino-tekoiset ja 'tavalliset' verkostot luotettavasti jopa ilman sisältö- tai käyttäjäpiirteitä. (Zhao ym. 2018, 8, 10–12).

Käyttäjäperustaiset piirteet, kuten käyttäjäprofiilien analyysi mm. bottien tunnistuksessa olivat myös aktiivisesti kehittyviä tutkimusalueita, liittyen myös aiempaan LIAR-datasettiin, jossa jokainen väite on yhdistetty sen esittäjän poliittiseen puolueeseen, osavaltioon ja luotettavuushistoriaan - vaikkakin kaikki tai suurin osa läpikäymistäni tutkimuksista eivät juuri näitä datapiirteitä hyödyntäneet. Kun nämä metatiedot liitettiin konvoluutioneuroverkkopohjaiseen luokittimeen, luokitustarkkuus parani selvästi verrattuna pelkkään tekstisisältöön perustuvaan vertailumalliin, mikä osoittaa, että puhujan kontekstuaaliset taustatiedot muodostavat vahvan priorijakauman väitteen todenperäisyydelle (Wang 2017, 1–5). Kuten Liu ja Wu (2018) osoittivat, miten pelkästään verkostopiirteet riittävät hyvään tunnistustarkkuuteen, Ferrara (2017) todisti vastaavasti, että käyttäjäprofiilin piirteillä voidaan myös saavuttaa vastaavia tuloksia. Tutkimus tarkasteli #MacronLeaks-disinformaatiokampanjaa Twitterissä vuoden 2017 Ranskan presidentinvaalien aikana.

Poliittisiin yhteyksiin—erityisesti vaaleihin ja valtiojohtoihin operaatioihin—liittyvä tutkimus omaksui nopeasti terminologian, joka korostaa tarkoituksellisuutta ja koordinoitua toimintaa. Käsite disinformaatio alkoi tässä kontekstissa tarkoittaa nimenomaan järjestelmällisiä kampanjoita, joilla pyritään manipuloimaan julkista mielipidettä (Ferrara 2017, 1). Samassa yhteydessä nousi esiin myös bottien rooli: automatisoidut 'sockpuppet'-tilit, joita hyödynnetään koordinoitusti viestin levittämisessä. Tutkimus käytti termejä '*disinformation*' ja '*social bot operations*' analysoidessaan koordinoitua bot-toimintaa Ranskan presidentinvaalien alla. Tutkimus kategorisoi vahvistamattomien huhujen koordinoitun levittämisen tällaiseksi disinformaatioksi. (Ferrara 2017, 1, 3). Toisin sanoen, tiedon totuusarvon

vahvistamattomuus ja koordinoitu (ilimplisiittisesti epäorgaaninen) tiedon levittäminen ovat riittäviä elementtejä kategorisoimaan levitettävä tieto disinformaatioksi.

Havainnointimallina Ferrara käytti logistista regressiota (logistic regression), joka tutkimuksessa valittiin vertailun jälkeen poikkeuksellisen kevyeksi, mutta raportoitujen tulosten valossa, silti tarkaksi bottiluokittimeksi. Malli koulutettiin kymmenellä profiili- ja aktiivisuuspiirteellä 17 miljoonan twiitin aineistolla ja pääsi 92 %:n tarkkuuteen (esimerkiksi statuses_count, followers_count / friends_count, profiilin oletuskuvan käyttö, geotagin puuttuminen, retweet-suhde) (Ferrara 2017, 6). Kun luokittinta sovellettiin nimenomaan #MacronLeaks-keskusteluun, noin viidenes (18 %) tileistä osoittautui botiksi (Ferrara 2017, 7). Tutkimuksen tulokselle löytyi myös ulkoinen vahvistus: kolmesta viidestätoista aktiivisimmasta botiksi luokitellusta tilistä oli tutkimushetkellä jo poistettu, keskeytetty tai eristetty Twitterin toimesta, mikä tukee mallin kenttäkelpoisuutta (Ferrara 2017, 9). Aikasarjakuvaajat puolestaan osoittivat, että bottien twiittipiikit edelsivät usein ihmiskäyttäjien reaktioita, joten botit saattoivat strategisesti käynnistää laajempia levitysketjuja (Ferrara 2017, 12). Ferraran analyysi osoittaa, että pelkkiin profiili- ja käyttötilastoihin tukeutuva malli riittää sekä bottien tunnistamiseen että disinformaatiokampanjan dynamiikan ymmärtämiseen varhaisessa vaiheessa. Automaattisiin metaheuristisiin optimointimenetelmiin perustuvat lähestymistavat todistettiin olevan toimiva ratkaisu havaitsemaan ja karsimaan tilastollisesti redundantteja muuttuja ja poimimaan relevantteja, selittäviä avainsanoja.

Harmaasusioptimointiin perustuvalla lähestymistavalla saavutettiin LIAR-aineistossa 96.5 prosentin tarkkuus ja saavutettiin myös tarkemmat tulokset sekä BuzzFeed-, että a Random Political News -korpuksissa (Ozbay ja Alatas 2019, 64–66). Menetelmä tuotti siten sekä tulkittavan että tehokkaan piirrejoukon ja nosti perinteiset koneoppimisluokittimet kilpailemaan syväoppimismallien kanssa tilanteissa, joissa opetusdata on vähäistä tai luokkajakauma vinoutunut. Nämä tulokset mm. osoittavat, että puhujan pitkän aikavälin luotettavuusprofiili tehostaa valeuutisten ennustamista ja että optimointilähtöinen piirreseulonta voi merkittävästi parantaa luokittimien suorituskykyä jopa ilman suuria syväoppimisarkkitehtuuria.

Intentionaalisuuden analyysi keskijaksolla

Keskijaksolla (noin 2017–2019), valeuutisten käsitteen yleistyessä, tarkoituksellisuus (intentio) mainittiin usein tutkimusten määritelmässä. Näissä tutkimuksissa käytettiin termejä, kuten ‘deliberate misinformation’, ‘hoaxes’ (Wang 2017), ‘intentionally deceptive’ (Rakholia ja Bhargava 2017), ‘deliberate fabrication’ (Tachini ym. 2017), ‘intentionally misleading’ (Monti ym. 2019). Kuitenkin tutkimusten operationaalinen fokus siirtyi vahvasti sisältöön ja leviämiseen perustuviin luokitelutehtäviin, eikä tutkimusmetodeillaan- tai tuloksillaan pyrkinyt osoittamaan yhteyttä esimerkiksi tarkoituksellisen valehtelun ja väärän tai harhaanjohtavan tiedon välillä. Vaikka menetelmät, jotka analysoivat leviämispolkuja (Liu ja Wu 2018) tai käyttäjävuorovaikutuksia graafeissa (Monti ym. 2019), saattoivat implisiittisesti tunnistaa koordinoitua toimintaa, päätavoite oli luokittelutarkkuus tai varhainen tunnistus, ei niinkään toimijoiden motiivien analysointi. Tarkoituksellisuus jäi usein oletetuksi taustatekijäksi, jota ei suoraan mallinnettu tai tutkittu ensisijaisena muuttujana. Vaikka faktantarkistukseen perustuvat tietojoukot, kuten LIAR (Wang 2017) saattoivatkin sisältää tietoa, jolla kenties viestijöiden todelliset tarkoituksiperät voitaisiin todistaa, en havainnut uskottavia pyrkimyksiä todistaa suoria motiiveja. Nämä kuitenkin ensisijaisesti mahdollistivat tutkimuskentällä vahvan valvonnan koneoppimismallien nopeamman kehityskaaren tekstin ja (tai) metadatan perusteella. Heikon valvonnan menetelmät, kuten WeFEND (Wang ym. 2019), hyödynsivät käyttäjäraportteja, jotka saattoivat epäsuorasti heijastaa intentiota, mutta intentio jäi kenties piileväksi muuttujaksi. Avoimet, valmiiksi kerätyt tietojoukot myös nopeuttivat tutkijoiden työtä ja auttoivat vertailemaan kehittämiensä mallien tuloksia johdonmukaisemmin.

Lyhyesti

Terminologian selkiytyminen vei tutkimuksen dataperustan uudelle tasolle. LIAR-aineisto tarjosi 12 000 journalistien luokittelemaa poliittista väitettä ja FakeNewsNet linkitti 23 000 uutisartikkeliin 1,6 miljoonaa twiittiä sekä niiden taustayhteydet (Wang 2017; Shu ym. 2019). Näin analysoitavien esimerkkien määrä

nousi kymmenkertaiseksi varhaisjaksoon verrattuna ja avasi oven syväoppimiselle. Samalla niissä toistui kentän rakennevika, kun aineistot perustuivat suurelta osalta angloamerikkalaiseen politiikkaan. Mallien toimivuus muissa kielissä ja kulttuuriympäristöissä jäi epävarmaksi (Graves 2018). Tekninen läpimurto tuli syväoppimisesta. Käsien suunniteltuja n-gram- ja stilometrialistoja alettiin korvata konvoluutio- ja toistoverkkoihin pohjautuvilla end-to-end-ratkaisuilla. CNN- ja Bi-LSTM-yhdistelmä nosti LIAR-aineiston kuusiluokkaisen tarkkuuden kolminkertaiseksi satunnaistasoon nähden, ja pelkkä pitkän lyhytaikais-muistin neuroverkko (LSTM) saavutti yli 80 prosentin F1-tarkkuuden Twitterin PHEME-aineistossa (Roy ym. 2018; Ajao ym. 2018). Samaan aikaan osa tutkijoista piti kiinni teorialähtöisistä kielipiirteistä selitettävyyden vuoksi: Zhou ym. (2019) käänsivät psykologisen valheoppinsa (Reality Monitoring, Four-Factor Theory) yli 300 kielellisen tarkastelun ja sanalaskuri-mittariksi (LIWC) ja saavuttivat lähes 0,9 tarkkuudet PolitiFact- ja BuzzFeed-materiaaleilla. Näin tulkittavuudesta tuli ensimmäinen vastareaktio koneoppimisen mustan laatikon tehokkuudelle.

Kokonaisuudessaan keskijakso siirtyi käsityöstä massadatan ja syvien verkkojen aikakauteen. Valeuutiset standardisoi kielen, benchmark-aineistot standardisoivat mittarin ja yhtenäistivät arvioinnin, ja syväoppiminen standardisoi laskennan ja tuli valtavirraksi. Samalla esiin nousivat kulttuurinen vinouma, datakohina ja mustan laatikon tulkittavuus. Multimodaalisten ja graafipohjaisten mallien ensi-marssi vihjasi tulevista haasteista, joissa pitää yhdistää teksti, kuva, video ja verkosto – ja tehdä se kestäväällä, selitettävällä tavalla.

4.2.3 Myöhäisjakso

Tarkastelukauden lopulla painopiste siirtyi yhä selvemmin sosiaalisen kontekstin syvälliseen mallintamiseen ja multimodaalisuuteen. Graafipohjaiset ratkaisut, kuten FANG-malli, rakensivat verkon *julkaisija – uutinen – käyttäjä* -suhteista ja hyödynsivät niissä huomiomekanismeja (Nguyen ym. 2020). Laajaa kielellistä ja modaaliteettien kirjoa vaativiin tarpeisiin luotiin uusia resursseja, kuten MuMiN

(Nielsen ja McConville 2022), joka kattaa yli 20 kieltä ja sisältää tekstin, kuvan sekä faktantarkistusmetadatan.

Keinoälyn tuottaman sisällön (teksti, kuva, video) tunnistamiseen alettiin vastata synteettisillä harjoitusaineistoilla:

Grover-tietojoukko (Zellers ym. 2020) simuloi uutisartikkeleita suurilla kielimalleilla;

StyleGAN2-kasvokokoelma (Nightingale ja Farid 2022) tarjoaa realistisia tekokasvoja visuaalisten deepfake-detektorien koulutukseen.

Erityisiin, tunnistettuihin haasteisiin pyrittiin vastaamaan räätälöidyillä tietojoukoilla:

‘Disagreement-Annotated’ -tietojoukot mallintavat annotoijien erimielisyyttä (Jigsaw Toxicity: Gordon ym. 2021).

COVID-19-aineistot (Abdelminaam ym. 2021) kohdistuvat terveystiedon informaatioon.

Samalla yhteisö jatkoi vertailua nyt jo ‘vakiintuneisiin’ benchmarkeihin (LIAR, FakeNewsNet, ISOT, FNC), mikä paransi uusimpien GNN-, transformer- ja multimodaalisten mallien suorituskyvyn läpinäkyvämmän vertailun.

Hybridimallit, tietograafit ja tulkittavuus

Uusimmassa kirjallisuudessa yhä enemmän tuodaan esille havainto, että algoritmiset moderointijärjestelmät eivät kykene päättämään jakajan intentiota, vaan tunnistavat ainoastaan sisällön pinnalliset piirteet. Tämä myös heijastaa aiemmin mainittua laajempaa terminologista siirtymää, jossa ero tahallisen ‘disinformaation’ ja tahattoman ‘misinformaation’ välillä tunnustetaan tärkeäksi, mutta automaattisten järjestelmien kyky tehdä tätä eroa on rajallinen.

Esimerkiksi Gorwa ym. (2020) tutkimuksessaan erottelevat kaksi tekniikkaa: (a) hash-matching, joka tunnistaa aiemmin kielletyn sisällön tarkat tai lähes tarkat kopiot ja (b) ennustavat mallit, jotka pisteyttävät uuden sisällön sen sanojen tai

pikselien perusteella. Kumpikaan ei pysty päättämään, yrittääkö ihminen huijata vai erehtyykö tahattomasti – kone katsoo vain dataa. Siksi disinformaation (tahallinen pyrkimys huijata) ja misinformaation (tahaton virhe) välinen ero jää automaatiolle näkymättömäksi (Gorwa ym. 2020, 4–5). Gorwa ym. (2020) tarkastelivat tutkimusongelmaa alustanäkökulmasta, mikä heijastuu käytetyissä termeissä kuten ‘algoritminen sisällön moderointi’ (algorithmic content moderation) ja ‘alustan hallinnointi’ (platform governance), sijoittaen misinformaation laajempaan haitallisen sisällön moderointipyrkimyksiin (Gorwa ym. 2020, 1).

Asiantuntijaperusteiset- ja hybridimenetelmät

Aiemmalla vuosikymmenellä kehitettyjä tietografi- ja ontologiapohjaisia ratkaisuja jatkojalostettiin: esimerkiksi uutisväitteiden totuusarvon arviointiin käytettiin yhä enemmän tietokantoja ja knowledge graph -tekniikoita, jotka pyrkivät tunnistamaan väitteiden totuudenmukaisuuden vertaamalla niitä luotettuihin tietolähteisiin (Ahmed ym. 2022, 1). Ahmedin ym. (2022) katsauksessa todetaan, että pelkkä dataohjautuva koneoppiminen ei riitä, vaan parhaisiin tuloksiin päästään yhdistämällä se asiantuntijatietoon pohjautuviin menetelmiin – toisin sanoen yhdistämällä oppivat mallit ja ihmisen määrittelemät tietosäännöt (Ahmed ym. 2022, 23–30). Tällainen hybridistrategia oli artikkelissa esitetyn näkemyksen mukaan lupaava tapa parantaa tunnistustarkkuutta.

Myös Naeem ja Boulos (2021, 1) painottavat, ettei mikään yksittäinen menetelmä ollut täysin tehokas, vaan tehokkain lähestymistapa on yhdistellä eri strategioiden vahvuuksia. Niin sanottu ‘mixed, synergistic approach’ onkin heidän mukaansa välttämätön COVID-19-infodemiaa ja valeuutisia torjuttaessa (Naeem ja Boulos 2021, 53–61). Huomioitavasti termi ‘misinformaatio’ nousikin merkittäväksi teemaksi kansanterveyteen liittyvän tutkimuksen kontekstissa, erityisesti COVID-19-pandemian aikana (Naeem ja Boulos, 2021, 1). Myös Vijjali ym. (2020) hyödynsivät COVID-19-faktantarkistuksessa 5 500 Poynter-varmennettua väitettä, mikä riitti opettamaan transformer-pohjaisen faktantarkistussysteemin arvioimaan Covid-aiheisia väitteitä automaattisesti jopa 85.5 prosentin tarkkuudella.

Esitetty BERT-pohjainen transformer-malli pyrkii myös hakemaan väitteiden pohjalta vektoripohjaisesti lasketut lähimmät tallennetut faktantarkastusraportit, jonka avulla malli pyrkii vastaamaan väitteiden todenmukaisuuteen, tarjoaa oikeampaa tietoa sekä metadatanä jakaa 'varmuuspisteensä' vastauksena onnistumisesta (Vijjali ym. 2020, 3–9).

Huomioitavaa on myös se, että testatessaan mallin toimivuutta uudella ajallisella ikkunalla (validointidata eri kuukausilta, vrt. koulutusdataan), recall10 –tarkkuus laski merkittävästi (14.9 prosenttiyksikköä), mikä viittaa lyhyelläkin ajanjaksolla merkittävään ajalliseen siirtovaikeuteen ja myöhempään päivitystarpeeseen, vaikkakin kirjoittajat eivät tätä suoraan maininneet (Vijjali ym. 2020, 3–4). Myös neuroverkkopohjaiset mallit, jotka oppivat väitteiden ja todisteiden semanttisen yhteenkuuluvuuden, kasvattivat siis suosiotaan merkittävästi vastaamaan tarpeeseen luoda malli, jolla voitiin yleistää ihmisasiantuntijoiden työ tekoälyn toistamaksi prosessiksi.

Sisältöpohjaiset piirteet ja tulkittavuus

Transformers-arkkitehtuuriin pohjautuvat mallit, kuten BERTin ja RoBERTan jatkokokehittelmät, vakiintuivat alan perustyökaluiksi 2020-luvun alussa. Niiden avulla saavutettiin entistä korkeampia tunnistustarkkuuksia monilla valeutisdatajoukoilla. Esimerkiksi Singh ym. (2023) rakensivat valeutisluokittimen, joka hyödynsi esikoulutettua syvää toistoverkkoa (RNN) ja kolmesta eri lähteestä koottua uutisartikkelidataa – mallilla raportoitiin yli 90 prosentin tarkkuus, mikä osoittaa nykyaikaisiin kielimalleihin pohjautuvien menetelmien tehokkuuden (Singh ym. 2023, 47–53). Huomionarvoista on, että tekstin esikäsittely ja piirteiden automaattinen oppiminen on pyritty integroimaan: tyypillisesti raakateksti puhdistetaan, vektoroidaan ja syötetään syviin malleihin, jotka oppivat abstrahoituja piirteitä, kuten semanttisia ja syntaktisia ominaisuuksia, ilman manuaalista piirteevalintaa.

Tämä erottaa 2020-luvun alun menetelmät 2010-luvun alun lähestymistavoista, joissa korostuivat ihmisen manuaalisesti suunnittelemat piirteet. Vielä 2010-luvun lopulla valeutisten havaitseminen nojasi vahvasti käsin suunniteltuihin sisältö-,

tyyli- ja tunnepiirteisiin (esim. TF-IDF-vektoreihin). Pandemian aikaiset tutkimukset osoittivat kuitenkin, että syväoppimismallit oppivat vastaavat rakenteet tehokkaammin suoraan datasta. Abdelminaam ym. vertasivat kahta strategiaa COVID-aiheisessa aineistossa: perinteiset koneoppijat (TF-IDF + päätöspuu, SVM, Naive Bayes) ja syväneuroverkot (GloVe-upotukset + LSTM/GRU). Yksinkertainen LSTM päihitti kaikki kevyet mallit (Abdelminaam ym. 2021, 374–382), mikä kuvasti erinomaisesti silloista trendiä siirtymiseen kohti automaattista piirreoppi- mista.

Piirresuunnittelua hyödyntävät menetelmät eivät kuitenkaan kadonneet; sen sijaan niitä alettiin sulauttaa osaksi hybridikokonaisuuksia. Malliin yhdistettiin esimerkiksi automaattisesti opittuja sanaupotuksia ja ihmisen määrittelemiä kontekstipiirteitä, kuten julkaisijan luotettavuus, julkaisupäivä tai aihealuokka. Tällä lisätyllä metadatalalla pyrittiin paikkaamaan tilanteet, joissa pelkkä tekstianalyysi olisi riittämätöntä saavuttamaan tarkkoja tuloksia. Ahmed ym. (2022) hyödynsivät mm. lähdesivuston mainetta (valtamedia vs. tunnettu disinformaatio sivusto) yhdessä tekstipiirteiden kanssa ja raportoivat selkeän parannuksen luokitustarkkuuteen (Ahmed ym. 2022, 19–23).

Syväoppimismalleja pidetään usein ‘mustina laatikoina’, koska niiden sisäisiä päätöksiä on vaikea tulkita. Tulkittavuuden parantamiseksi tutkijat ovat säilyttäneet käsin rakennettuja kielipiirteitä – esimerkiksi n-gram-taajuuksia ja syntaktisia rakenteita – ja visualisoineet niiden vaikutusta mallin luokitukseen. Näin perinteinen piirreajattelu toimii selityskerroksena syväverkoille ja auttaa asiantuntijoita arvioimaan, perustuuko päätös oikeisiin kielellisiin vihjeisiin vai satunnaiseen kohinaan (Gordon ym. 2021, 30).

Multimodaaliset graafit ja verkostoanalyysi

Vale uutisten tunnistuksessa alettiin yhä enemmän hyödyntämään monikielisiä ja multimodaalisia tietovarantoja. Nielsen ja McConville (2022) julkaisivat laajan MuMiN-aineiston, joka linkittää toisiinsa faktantarkistussivustojen väitteet sekä niitä käsittelevät Twitter-keskustelut 41 eri kielellä (Nielsen ja McConville 2022, 43–51). MuMiN muodostaa heterogeenisen graafitietokannan, jossa solmuina

ovat mm. uutisväitteet, twiitit, käyttäjät, kuvat ja uutisartikkelit (Nielsen ja McConville 2022). Tämä kuvastaa metodologista muutosta: tietopohjainen valeuutisten tunnistus ei rajoitu enää yksittäisiin faktantarkistusoperaatioihin, vaan laajoihin verkostoihin, joista mallit voivat ammentaa monenlaista tietoa väitteiden tueksi tai kumoamiseksi. Monikielisyys on huomioitu entistä paremmin – valeuutisia voidaan nyt pyrkiä tunnistamaan globaalisti eri kielillä hyödyntämällä yhteisiä faktantarkistusresursseja. Samalla on havaittu haasteita kuten luokkien epätasapaino: faktantarkistajat raportoivat enemmän virheellisiä kuin tosia väitteitä, mikä heijastuu aineistojen vinoumana (Nielsen ja McConville 2022, 603–610).

Tekstipohjaisten mallien rinnalla on havaittavissa monialustaisia ja multimodaalisia lähestymistapoja valeuutisiin. Uusissa aineistoissa yksi uutisväite voi sisältää tekstiä, kuvan ja jopa videoleikkeitä, jolloin tunnistusalgoritmien on yhdisteltävä eri lähteistä tulevia vihjeitä. Nielsen ja McConville (2022) raportoivat, että pelkästään tekstin tai pelkästään kuvan perusteella toimivat mallit suoriutuivat heikommien, kun taas sisältöjä ja niiden jakajien verkostosuhteita yhdistelevä heterogeeninen graafineuroverkko pystyi paremmin ennustamaan väitteiden totuusarvoja monimuotoisessa aineistossa (Nielsen ja McConville 2022, 7–9).

MuMiN toimi koealustana heterogeenisille graafineuroverkoille. HeteroGraphSAGE, joka yhdistelee automaattisesti tweet-, käyttäjä- ja artikkelisolmujen tietoja, saavutti makro-F1-arvon 0,63 ja ohitti tekstipohjaisen LaBSE-mallin (0,60) kahdessa vaativassa luokitustehtävässä (Nielsen ja McConville 2022, 8–9). Varhaisemmissa töissä verkostopiirteet (esim. keskusmittarit, linkkitiheydet) määriteltiin manuaalisesti. Nyt piirteet opittiin suoraan graafista.

Suuret kielimallit: generoinnin ja tunnistuksen haasteet

Tekstigenerointimallien (TGM) viimeaikainen kehitys on tuonut automaattiseen valeuutisten tunnistamiseen uuden, nopeasti muuttuvan haasteulottuvuuden. Tehokkaiden suurten kielimallien julkaisut (LLM), kuten GPT-3 ja ChatGPT, heijastuivat tutkimuskentällä huolenaiheena ‘koneella generoidusta tekstistä’ (machine generated text) (Jawahar ym. 2020, 1) ja sen potentiaalista käytöstä ‘tekoälyveitoisissa infodemioissa’ (eli tekoälyn vauhdittamasta disinformaatiotulvasta) tai

‘erittäin vakuuttavassa disinformaatiossa’ (De Angelis ym. 2023, 1; Spitale ym. 2023, 1). Tekoälyn tuottaman tekstin tunnistamiseen luotiin myös uusia tietojoukkoja, kuten Grover-tietojoukko (Zellers ym. 2020).

Jawahar ym. (2020, 6) osoittavat, että sekä mallin parametrin määrä että dekodausmenetelmä määräävät, kuinka helposti tuotokset paljastuvat koneen kirjoittamiksi: suurimmat GPT-2-variantit (762–1 542 milj. parametria) lähentyivät ihmisen kirjoittamasta tekstistä huomattavasti vaikeammin erotettavaksi. Tutkimuksen mukaan jopa siinä määrin, että perinteinen TF-IDF-piirteisiin ja logistiseen regressioon perustuvat tunnistusmenetelmät suoriutuivat selvästi heikommin kuin pienten, 117 milj. parametrin mallien teksteissä. Top-k-otanta jättäisi sanaston yleisimpien sanojen suhteellisiin frekvensseihin selkeän piikin (mikä synnyttäisi tilastollisen sormenjäljen, jonka yksinkertainen n-gram- tai TF-IDF-analyysi voisi mahdollisesti tunnistaa), kun taas ydinotanta (top-p) tasoittaa nämä tilastopoikkeamat ja vaikeuttaa tunnistusta edelleen. Tunnistaminen käy entistä hankalammaksi, kun malli hienoviritetään (fine-tuned) kapeaan aihealueeseen: Amazon-tuotearviointiin jälkikoulutettu GPT-2 hämää luokittimia tehokkaammin kuin alkuperäinen ‘baseline’- julkaisuversio. (Jawahar ym. 2020, 6).

Tutkimuksen katsaukseen koottujen tunnistimien toimintaperiaatteet jakautuvat neljään pääryhmään, joista jokaisella on luontainen rajoitus (Jawahar ym. 2020, 6–7). Alusta asti koulutetut ‘piirremallit’ (esim. TF-IDF + logistinen regressio) ovat kevyitä, mutta niiden erotuskyky ei siirry suurempien tai erilaisilla rakenteilla tuotettujen TGM-tekstien tunnistamiseen. Ilman esimerkkidataa toimivissa ‘nollashotti’-menetelmissä (zero shot) sama TGM toimii sekä generaattorina että tunnistimena (ns. summattu lokitodennäköisyys), mutta tarkkuus jää heikommaksi kuin sanavektoripohjaisilla malleilla (Jawahar ym. 2020, 7). Summattu todennäköisyys tarkoittaa tässä kontekstissa sitä, että jos teksti on juuri tämän mallin tuottamaa, se vastaa tarkasti mallin omia todennäköisyysennusteita ja saa korkeamman todennäköisyyden (matalamman perpleksisyyden) kuin ihmisen kirjoittama teksti, jonka sanavalinnat poikkeavat mallin odotuksista (Solaiman ym. 2019, 3–4).

Tutkimuksen mukaan parhaimmat tulokset saavutetaan, kun LLM, eli suuri kieli-malli (esim. RoBERTa) hienoviritetään tunnistinmalliksi, mutta tällöin vaaditaan

opetusdataa juuri siitä TGM:stä, jota halutaan valvoa; pienemmällä mallilla koulutettu detektori ei erota suuremman mallin generaatioita luotettavasti (Solaiman ym. 2019, 12). Ihmis-kone – yhdistelmät, kuten GLTR-työkalu, nostavat maallikkolukijan tarkkuuden 54:stä 72 prosenttiin visualisoimalla tokeni-kohtaisia todennäköisyysjakaumia; etu kuitenkin hiipuu, jos TGM tuottaa tasalaatuisempaa ja tilastollisesti ‘anomaliatonta’ tekstiä (Jawahar ym. 2020, 7).

Tutkimuksen virheanalyysi nostaa esiin kolme keskeistä ongelmaa:

1. Yleistäminen: tunnistimen tulisi kestää vaihtelut malliarkkitehtuurissa, dekodausmenetelmissä ja kehotepituuksissa, mutta useimmat nykyratkaisut pettävät jo yhdenkin parametrin muuttuessa.
2. Mallin tulkittavuus ja yhteistyö ihmisten kanssa: tokenitason lämpökartta todistettavasti auttaa, mutta selityskyky voi heikentyä, jos TGM:ät kykenevät yhä enemmän minimoimaan generoimansa tekstin tilastopoikkeavuutta.
3. Nk. ‘robustius’: pelkkä homoglyfikorvaus (esim. a-kirjainten muuttaminen kyrilliseksi a:ksi), pudottaa RoBERTa-tunnistimen recall-arvon 97,44 prosentista 0,26 prosenttiin, mikä korostaa tarvetta tunnistinmallien jatkokehitykselle vastaamaan näihin haasteisiin (Jawahar ym. 2020, 17).

On siis huomionarvoista, miten uudet tekoälyteknologiat itsessään alkoivat muuttaa pelikenttää: vuonna 2022 poikkeuksellisen suurta suosiota kerännyt ChatGPT ja muut laajat kielimallit nostivat esiin huolen AI-avusteisesta misinformaatiosta. De Angelis ym. (2023) varoittavat, että kehittyneet kielimallit kykenevät tuottamaan ennennäkemättömällä nopeudella uskottavan kuuloista tekstiä, mikä voi johtaa ‘tekoälyvetoiseen infodemiaan’ – väärän tiedon tulvaan, jonka tunnistaminen on nykytyökaluilla vaikeaa (De Angelis ym. 2023, 1–2, 4–6).

Tekoälyn kehittyessä syntyi myös laajempia yleistermejä, kuvaamaan ‘synteetistä mediaa,’ kattamaan laajemman valikoiman realistisia tekoälyn luomaa mediasisältöä, kuten ‘AI-generated media’, ‘AI-generated contents’ ja ‘deepfakes’ (Jia ym. 2024, 1; De Angelis ym. 2023, 5). Erityisesti he korostavat, että tällä hetkellä ei ole luotettavia keinoja erottaa tekoälyn generoimaa tekstiä ihmisen kirjoittamasta, mikä muodostaa merkittävän haasteen valeutisten torjunnalle

(De Angelis ym. 2023, 1, 5). Tämä havainto linkittyy verkostopohjaiseen tarkasteluun siinä mielessä, että jos verkostot tulvivat automaattisesti tuotettua misinformaatiota, perinteiset käyttäjäperustaiset vihjeet – esimerkiksi epäilyttävä kirjoitustyylili tai bottimainen käyttäytyminen – voivat hämärtyä entisestään.

Kesäkuussa 2023 julkaistu tutkimus kartoitti ihmisten kykyä tunnistaa tekoälyn kirjoittamia viestejä: Spitale ym. (2023) vertailivat, miten hyvin ihmiset erottavat GPT-3:n generoimat Twitter-viestit aitojen käyttäjien viesteistä ja samalla tunnistavat, ovatko väitteet totta vai eivät (Spitale ym. 2023, 2–3). Tutkimukset, jotka käsittelevät tekoälyn tuottaman sisällön haitallista potentiaalia, kehystivät sen usein ‘disinformaationa’ (Spitale ym. 2023, 1; Jia ym. 2024, 1). Spitale ym. (2023) on yksi harvoja tutkimuksia, jossa intentio operationalisoitiin eksplisiittisesti vertaamalla ihmisten ja GPT-3:n luomia, tahalliseksi disinformaatioksi suunniteltuja twiittejä.

Tuloksissa ilmeni yllättäviä piirteitä – osallistujat tunnistivat ihmisten kirjoittamat valheelliset twiitit hieman paremmin kuin tekoälyn kirjoittamat valheelliset (eli GPT-3:n tuottama disinformaatio meni hitusen useammin läpi), mutta toisaalta tekoälyn kirjoittamat tosiasiapitoiset twiitit arvioitiin useammin todeksi kuin ihmisten vastaavat (Spitale ym. 2023, 3–4). Toisin sanoen tekoälyn tuottama viestisisältö koettiin osin vakuuttavammaksi riippumatta sen todenperäisyydestä. Tämä alleviivaa sitä haastetta, jonka parissa verkosto- ja käyttäjäanalyysien tutkijat nyt kamppailevat: kun verkostoissa liikkuu sekä ihmisten että tekoälyjen luomia sisältöjä, ihmistunnistajat ja algoritmit täytyy kouluttaa entistä paremmin erottamaan näitä toisistaan. Myös Kreps ym. (2020, 104–105) mallinsivat intentiota implisiittisesti arvioimalla tekoälyn (GPT-2) generoiman tekstin uskottavuutta ja vakuuttavuutta ‘media misinformaation työkaluna’.

Visuaalisen median tunnistusmenetelmät: artefaktit ja biologiset signaalit

Varhaiset EANN-tyyliset mallit osoittivat, että sanallinen ja visuaalinen sisältö tuottavat toisiaan täydentävää informaatiota, mutta niiden kehitystä jarrutti sopivan paritetun datan puute. Tätä ‘kylmäkäynnistys’-ongelmaa pyrittiin ratkaisemaan mm. Fakeddit-datasetillä, joka kokosi yli miljoona Reddit-postausta, liittäen

jokaiseen otsikon, kuvan sekä kommenttiketjun (Nakamura ym. 2020, 1) . Ai-neisto tarjoaa valmiit 2-, 3- ja 6-luokkaiset totuuskategorioiden (Taulukko 2; Naka-mura ym. 2020, 3). Tällöin koulutettava malli voi oppia yhteisen semanttisen ava-ruuden eikä “herää tyhjän päälle” (kylmäkäynnisty) uudenlaista signaalia kohda-nessaan.

Visuaalisiin artefakteihin perustuvilla konvoluutioverkoilla pyrittiin myös havaitse-maan GAN-pohjaisia synteettisiä kasvoja, havainnoimalla kasvoverkon neuro-nikerrosten aktivaatioita. Esimerkiksi Wang ym. (2020), tutkimuksessa (FakeSpotter), kyettiin jopa yli 90 prosentin tarkkuudella havaitsemaan keinote-koiset kasvot ja malli kykeni säilyttämään toimintakykynsä pakkaus-, epäte-rävyys- ja kohinamuutoksissa (AUC-pudotus < 3,8 %) (Wang ym. 2020, 2, 4).

Biologisiin signaaleihin perustuva FakeCatcher (Ciftci ym. 2020) kykeni taas erot-tamaan aidon ja synteettisen videon hyödyntämällä ihon mikrovärähtelystä las-kettua PPG-sydänrytmiä; järjestelmä saavuttaa 97 prosentin tarkkuuden aitous-testissä ja 93 prosentin tarkkuuden neljän eri generointimallin identifioinnissa. Koska biologinen rytmi toimii ‘piilovesileimana’, temppu ei riipu videon pikseliarti-fakteista ja on siten vaikeampi ohittaa. (Ciftci ym. 2020, 1–2). Syvävääreännöksiä käsittelevä tutkimus kohdistuu luonnostaan tahallisesti manipuloituun mediaan, ja motivaatio kehittää tunnistusmenetelmiä perustuu usein oletukseen niiden käy-töstä haitallisiin tarkoituksiin kuten disinformaatiokampanjoihin (Ciftci ym. 2020, 1).

Vastaavasti Groh ym. (2022) tutkivat ihmisten syvävääreännösten tunnistuskykyä psykologisesta näkökulmasta: he suunnittelivat kokeen, jossa osallistujille näy-tettiin videoita ja kysyttiin, kumpi kahdesta videosta on vääreennetty, sekä mani-puloitiin osallistujien tunnetiloja (Groh ym. 2022, 1). Kun keskityttiin inhimilliseen elementtiin – havaitsemiseen, psykologiaan tai vaikutukseen – terminologia heijasti tätä. Tutkimuksissa tutkittiin ihmisten kykyä havaita (detect), huomata (spot) tai erottaa (discern) vääreännöksiä (Groh ym. 2022, 1). Tutkimuksen hypo-teesina oli, että tietyt tunnereaktiot – erityisesti suuttumus – saattavat parantaa ihmisten huomiokykyä syvävääreännösten havaitsemisessa (Groh ym. 2022, 2). Tällaiset kokeet kuvastavat kasvavaa kiinnostusta moniaistiseen valeutisten

tunnistukseen: pelkän algoritmikehityksen ohella pyrittiin ymmärtämään myös ihmishavainnon rajoja ja mahdollisuuksia visuaalisen disinformaation erottamisessa. Vuoden 2021 jälkeen julkaistiin useita kilpailuja ja haasteita deepfake-videoiden tunnistamiseksi, mikä kiihdytti menetelmien kehitystä.

Suuriin kielimalleihin pohjautuvat tunnistusmenetelmät

Myös aiempaan tutkimuskirjallisuuteen nähden, yllättäviä lähestymistapoja alkoi ilmestyä: Jia ym. (2024) tutkivat, voiko GPT-4V Vision -mallia käyttää sellaiseen – siis ilman erillistä hienosäätöä – tunnistamaan, onko valokuva aitona otettu vai tekoälyn muodostama (Jia ym. 2024, 1–2). Tässä toistuu myös ilmiö, että teknologisiin näkökohtiin keskittyvät tutkimukset käyttivät usein tarkkoja teknisiä termejä, kuten ‘synteettinen media’ (synthetic media) (Jia ym. 2024, 1). Tutkimuksessa syötettiin mallille 3000 kasvokuvaa, joista kolmasosa oli aitoja ja kaksi kolmasosaa tekoälyn generoimia. Väärennökset oli tehty kahdella eri, laajasti käytetyllä menetelmällä: StyleGAN2-nimisellä generatiivisella verkolla ja Latent Diffusion -mallilla (Jia ym. 2024, 3).

Mallille annettiin seitsemän erilaista ohjetta (promptia), joista parhaaksi osoittautui ohje, jossa ensin pyydettiin binäärinen kyllä/ei-vastaus, jonka jälkeen tekoälymalli generoi tekstikuvauksen niistä avainkohdista, joista malli näki tekoälyn jälkiä (Jia ym. 2024, 7). Tällä ohjeella malli kieltäytyi vastaamasta vain viidessä prosentissa tapauksista, ja se tunnisti väärennökset noin 84 prosentin tarkkuudella, kun kuvien tekijänä oli StyleGAN2 (Jia ym. 2024, 7).

Kun tarkasteltiin laajemmin, kuinka hyvin malli erotti aidot ja tekaistut kasvot, GPT-4V sai kokonaisarvosanakseen (nk. erotuskykymittarin keskiarvoksi) noin 78/100, mikä nousi lähes 90/100, mikäli kuvat oli ensin pakattu JPEG-muotoon (pakattuun häviötä aiheuttavaan muotoon) tai niihin oli muutoin lisätty epäte-rävyyttä (Jia ym. 2024, 6). Tämä viittaa siihen, että mallin tekemä “silmämääräinen” arviointi jopa helpottuu, mikäli kuviin syntyy uusia pieniä artefaktiperäisiä virheitä.

Perinteiset, erikoistuneeseen tehtävään koulutetut kuvantunnistusverkot ylsivät silti vielä selvästi parempaan tarkkuuteen (Jia ym. 2024, 6). GPT-4V:n vahvuutena on kuitenkin se, että se perustelee ratkaisunsa sanallisesti; tämä tekee sen

havainnoista helpommin tulkittavia kuin pelkän aito/feikki -vastauksen antavat eri-koistuneet mallit (Jia ym. 2024, 2). Toisaalta, jos taas kaikissa generoidun kuvan yksityiskohdissa on tunnistettavissa esimerkiksi 'ihmismäinen' johdonmukaisuus, GPT-4V:llä ei ole signaalitasoista keinoa paljastaa väärennöstä. Jia ym. (2024) huomauttavat, että mallin kyky romahtaa tällaisissa tapauksissa: se antaa lähes sattumanvaraisen tuloksen ja luokittelee täydellisesti generoidut kasvot usein ai-doiksi (Jia ym. 2024, 6).

'COVID-infodemia' ja tiedon levityksen analyysi

Tutkimuskirjallisuuden valossa, COVID-19-pandemia laukaisi niin suuren mis- ja disinformaatiovyöryn, että verkostorakenteen systemaattinen mallinnus nousi yhä suosittumaksi tutkimuksen kohteeksi. 'Infodemia' vakiintui terminä erityisesti COVID-19-pandemian seurauksena vuodesta 2020 alkaen. Termin ensimmäinen ilmentyminen tietojoukossa tapahtui samana vuonna, jolloin aineistoni viisi julkaisua mainitsivat tämän otsikossaan tai tiivistelmässään. Seuraavana vuonna tämä luku nelinkertaistui kahteenkymmeneen. Pandemiaan liittyvää misinformaatiota analysoivat tutkimukset omaksuivat tämän termin kuvaamaan ainutlaatuisesta haastetta, jonka aiheutti samanaikaisesti leviävän tiedon – sekä paikansäpitävän että virheellisen – ylivoimainen määrä, mikä vaikeutti luotettavan ohjeistuksen löytämistä (Naeem ja Boulos 2021, 1, 12; De Angelis ym. 2023, 1; Spitale ym. 2023, 1). Tämän termin äkillinen suosio ja ilmaantuminen aineistossani voidaan tulkita viittaavan siihen, miten tietyt kriisitapahtumat voivat vaatia uutta kieltä kuvaamaan tiedonkulkujen ja niihin liittyvien ongelmien erityispiirteitä. Esimerkiksi Naeem ja Boulos (2021) kuvaavat katsauksessaan, miten heikko digitaalinen terveyslukutaito yhdistettynä sosiaalisen median viraaliin jakomekaniikkaan ruokki väärän tiedon leviämistä (Naeem ja Boulos 2021, 1–2).

Kansanterveyden alalla, erityisesti kriisien kuten COVID-19-pandemian aikana, misinformaatio oli usein kuitenkin ensisijaisesti käytetty termi (Naeem & Boulos, 2021, 1). Tämä heijastaa sitä todellisuutta, että väärä terveystieto voi levitä ilman haitallista tarkoitusta, pelon, väärinymmärryksen tai virheellisiin väitteisiin uskominen vuoksi. Aiemmissa (2011–2019) verkostotutkimuksissa—esim. Ratkiewicz

ym. 2011 ja Zhao ym. 2014 — tarkasteltiin yksittäisiä astroturffaus- tai bot-kampanjoita. Uusi konteksti vaati laajempia, monikielisiä aineistoja ja entistä automatisoidumpia menetelmiä.

Uudemmat artikkelit myös liittävät verkostanalyysin käyttäjäpsykologiaan. IEEE Access-katsaus nostaa naiivin realismin harhan keskeiseksi selittäjäksi – käyttäjät hyväksyvät disinformaation, jos se vahvistaa omaa maailmankuvaa, riippumatta faktantarkistuksista (Yang ym. 2023, 63–70) . Tutkijat toisaalta myös korostavat tarvetta yhdistää tekniset tunnistimet, yleisön medialukutaidon parantamiseen, (esimerkiksi IFLA:n päivitettyyn tarkistuslistaan) (Naeem ja Boulos 2021, 10–12). Tutkimuskentällä inhimillisten elementtien, kuten psykologisten vaikutusten kuvaamiseen, käytettiin termejä kuten *luottamus* (trust) tai *epävarmuus* (uncertainty) (Vaccari ja Chadwick 2020, 1).

Annotointi ja luokkavinouma

Valeutisaineistot myös ovat lisänneet selitettävyystarvetta, sillä niissä esiintyy runsaasti tulkinnanvaraisia väitteitä, joista annotoijat eivät aina pääse yksimielisyyteen. Gordon ym. analysoivat nk. vihapuheaineistoa, jossa annotoijien erimielisyys paisutti luokittimen ilmoitetun suorituskyvyn: heidän disagreement deconvolution -menetelmänsä laski ROC-pistemäärän 0,95:stä 0,73:een, kun annotaattorien vakaat mielipiteet erotettiin satunnaisesta vaihtelusta (Gordon ym. 2021, 29–37). Annotoijien erimielisyyttä subjektiivisissa tehtävissä voitiin toki kvantifioida käyttämällä esimerkiksi tämän ilmiön huomioivia tietojoukkoja, kuten ‘Jigsaw Toxicity’ (Gordon ym. 2021).

Tutkimuskentällä alettiinkin tutkimuksen ilmestymisen jälkeen parantaa annotointikäytäntöjä muun muassa lisäämällä Asiantuntija-annotointia Datakeruussa alettiin käyttää aihealueen asiantuntijoita – esimerkiksi lääketieteen ammattilaisia COVID-aiheisissa väitelauseistoissa (Naeem ja Boulos 2021, 429–436) ja ekplisiittisesti dokumentoimalla, milloin väitteen totuusarvo perustui vain yhden faktantarkastajan arvioon ja milloin useat faktantarkastajat olivat erimielisiä (esim. MuMiN-datasetti) (Nielsen ja McConville 2022, 603–610). Myös MuMiN-aineiston

osalta on mainittu luokkavinouma alaluvussa 4.2.3: *Multimodaaliset graafit ja verkostoanalyysi*.

Käyttäjäsignaalit, joukkoistus ja alustavaikutukset

Tutkijoiden rooli on usein siirtynyt datan filtteröinti- ja validointikerrokseksi jopa miljoonille datapisteille. Niiden avulla pyritään rakentamaan laaja automaatio, joka on väistämättä osin altis datamelulle. Tämä vähentää käsin tehtävän työn pullonkauloja mutta luo uusia haasteita tutkijoille: kuinka varmistaa, etteivät heikot tai biasoituneet joukkoistussignaalit johda harhaan – ongelma, jota varhaiset (Ashcroft ym. 2015 ja Ratkiewicz ym. 2011) eivät vielä kohdanneet tässä mittakaavassa.

Joukkoistukseen perustuvat menetelmät

Myöhemmällä tutkimusajankohdalla julkaistu tietojoukko 'Fakeddit' jatkaa aiemman tutkimuskauden trendiä, tasapainotella kerätyn datan vinouman ja analyysien tarkkuuden välillä. Heidän keräämänsä aineisto painottuu englanninkieliseen, usein humoristiseen meemikulttuuriin (Nakamura ym. 2020, 2–3). Vaikka kerätty Fakeddit-data onkin osin asiantuntijaperäisesti rakennettu (osin asiantuntijafiltraation avulla kerätyt datamerkinnot), sen pohjimmainen oletus viestin luotettavuudesta perustuu tutkimuksessa hieman epäortodoksisen joukkoistamiseen; Tiettyjen reddit-palstojen tykkäykset ja palstavetoinen moderointi olivat ensisijainen datankeruumenetelmä, ja distant supervision -menetelmän avulla (heikko, epäsuorasti saatava opetussignaali). Tutkijat eivät siis antaneet jokaista 'true / satire / misleading ...' -merkintää käsin, vaan johtivat merkintäluokan tekstin ulkopuolisista heijastesignaaleista esimerkiksi sen osalta, missä kontekstissa viesti julkaistiin (Nakamura ym. 2020, 3-5). Kriittiset ulkopuoliset tutkimukset kuitenkin osoittavat, että Reddit-yhteisöjen äänestyslogiikka ei mittaa sisällön todenperäisyyttä vaan sosiaalista suosiota – ja vieläpä vinoutuneesti.

Muchnik, Aral ja Taylor (2013, 647–649) osoittavat satunnaistetussa kokeessa, että yksi satunnainen +1-ääni nostaa kommentin lopullista pistettä 25 %. Tämä

on niin sanottu 'herding effect', eli paimennusvaikutus. Toisin sanoen muut käyttäjät ovat tällöin taipuvaisia peesaamaan positiivista signaalia. Glenski, Johnston ja Weninger (2015) puolestaan toteavat, että 73 % redditissä postauksille annettavista äänistä annetaan lukematta ensin itse postauksen sisältöä. Toisin sanoen up-vote heijastaa lähinnä otsikon kuin itse sisällön laatua (Glenski, Johnston ja Weninger 2017, 6, 13).

Huang ym. (2024) analysoivat 1,2 miljoonaa redditistä poistettua viestiä ja havaitsivat, että palstojen moderaattorit poistavat omille mielipiteilleen vastakkaisia kommentteja selvästi herkemmin, millä on kaikukammioita (echo chamber) vahvistava vaikutus (Huang ym. 2024, 1-2). Kriittisellä näkökulmalla tarkasteltuna Fakeddit - datasetin esittämä 'ground truth' voidaan osoittaa edustavan lähinnä yhteisön mielivaltaa eikä todennettuja faktoja. Fakeddit-datasetillä koulutetut mallit oppisivat siis tunnistamaan Reddit-kulttuurin suosiosignaalit, eikä voitaisi Nakumaran ym. (2020) esittämän tutkimuskehiksen puitteissa aukottomasti, tai edes uskottavalla tavalla osoittamaan, että nämä signaalit voisivat osoittaa sisällön todenperäisyyttä.

Intentionaalisuuden analyysi myöhäisjaksolla

Viimeisimmällä ajanjaksolla (2020–2024) tarkoituksellisuuden analysointi näyttää vähentyneen edelleen lähes olemattomiin. Vaikka huoli tahallisesta petoksesta on läsnä erityisesti uusien tekoälyteknologioiden (syvävääreännökset, LLM:t) myötä, intentio kehystetään usein teknologian kyvykkyytenä tuottaa harhaanjohtavaa sisältöä, eikä niinkään spesifien toimijoiden motiivien kautta. Tutkimukset useimmiten keskittyvät joko puolustuksellisiin lähestymistapoihin, kuten tekoälyn tuottaman sisällön tunnistamiseen (Jawahar ym. 2020), olettaen sen mahdollisen haitallisen käytön, tai harvemmin arvioiviin lähestymistapoihin, joissa tekoälyä käytetään luomaan disinformaatiota sen ominaisuuksien tai vaikutusten tutkimiseksi (Kreps ym. 2020; Spitale ym. 2023). Syvävääreännöksiä käsittelevä tutkimus olettaa usein haitallisen tarkoituksen teknologian luonteen ja potentiaalisten sovellusten perusteella (Ciftci ym. 2020, 1; Nightingale ja Farid, 2022, 1). Vaikka intentio on siis palannut keskusteluun teknologian kautta, sen suora mallintaminen tai eri toimijoiden tarkoituksellisuus jää edelleen vähäiseksi.

Yhteenvedona voidaan todeta, että analyysimenetelmien kirjo laajeni merkittävästi eri syötemodaaleissa. Tekstin osalta syvät oppivat mallit hallitsivat kenttää, kun taas kuvien ja videoiden osalta kehitettiin rinnakkain erikoistuneita algoritmeja ja luovia uusia ratkaisuja (kuten suurten kielimallien hyödyntäminen). Monimodaalisuus nousi esiin tavoitteena yhdistää nämä kaikki: tulevaisuuden valeuutisten tunnistusjärjestelmät pystyvät todennäköisesti käsittelemään samanaikaisesti uutisen tekstin, siihen liittyvät kuvat tai videot sekä muut kontekstiedot.

Uuden vaiheen kehitys

Pandemian jälkeen vale- ja harhatiedon vastaisessa tutkimuksessa tapahtui samanaikainen tekninen ja teoreettinen hajaantuminen. Menetelmät, modaliteetit ja tieteenalat risteytyivät tavalla, joka rikkoi varhaisempien vaiheiden selkeät rajat. Mustan laatikon ongelma ei kuitenkaan kadonnut, ja käsin suunnitellut kieli-, tunne- ja diskurssipiirteet olivat yhä tärkeimpiä selityspylväitä. Kun annotoijien erimielisyys pudotti ROC-pisteen 0,95:stä 0,73:een, Gordon ym. (2021) osoittivat, että eksplisiittiset n-gram- ja syntaksimittarit helpottavat virhelähteiden lokalisointia.

Parhaat tulokset syntyivät hybridistrategioista. Tietografeihin (knowledge graphs) ankkuroitu tarkistus yhdistää uutisväitteen esimerkiksi Wikidata-solmuun ja antaa sääntöpohjaisen 'totuusankkurin', jonka päälle kielimalli arvioi semanttisen yhteyden (Ahmed ym. 2022). COVID-19-infodemia teki tästä lähestymistavasta välttämättömän. Naeem ja Boulos (2021) osoittivat, että yksittäinen algoritmi romahtaa, ellei sitä tueta sekä asiantuntijasäännöillä että metadatalalla. Vastaavia tuloksia raportoitiin Poynter-pohjaisella COVID-aineistolla, jossa transformer yhdistettiin vektorihakuihin ja lähdetietoihin; vaikka F1 ylsi 85 prosenttiin, ajallinen siirto kuukaudenkin päähän laski recallia kymmeniä prosenttiyksikköjä, mikä korosti jatkuvan päivityksen tarvetta.

Monikielisyys ja multimodaalisuus nousivat tutkimuksen uudeksi oletusarvoksi. MuMiN-aineistossa sama väite on liitetty sen Twitter-ketjuihin, kuviin, uutislähteisiin ja 41 kielen käyttäjiin.

Heterogeeninen GraphSAGE oppii näistä solmuista yhteisen latenttipiirteen ja päihittää pelkän LaBSE-tekstimallin makro-F1:ssä, mutta faktantarkistajat tuottavat huomattavasti enemmän virhe- kuin tosi-merkintöjä ja annotaattorit riitelevät hienojakoisista kategorioista, mikä vahvistaa luokkajakauman vinoutumaa ja vaikeuttaa mallien arviointia (Nielsen ja McConville 2022). Generatiivinen toinen aalto muutti riskikuvaa; jo GPT-2:n 1,5 miljardin parametrin versio hämäsi TF-IDF-tunnistimet, ja GPT-4V pystyy perustelemaan, miksi jokin kuva vaikuttaa epäaidolta, mutta kompastuu virheettömästi tuotettuihin StyleGAN2-kasvoihin (Jawahar ym. 2020; Jia ym. 2024).

GPT-4V:n kaltaiset multimodaaliset suurmallit tuovat selitykset takaisin kuvatu- nistukseen, mutta jäävät silti selvästi erikoistuneiden detektorien tarkkuudesta – toisaalta ne tarjoavat ainutlaatuisen kyvyn perustella ratkaisunsa luonnollisella kielellä.

Verkostotutkimus kytkeytyi yhä tiukemmin sosiaalipsykologiaan. IEEE Access - katsauksen mukaan naiivi realismi – taipumus uskoa omaa maailmankuvaa vahvistavaa tietoa – ennustaa parhaiten disinformaation omaksumista (Yang ym. 2023). Samalla alustadynamiikasta löydettiin rakenteellisia vinoumia, kun Huang (2024) osoitti, että poliittisesti kallistunut moderointi Redditissä poistaa vastakkaisia kantoja suhteettomasti ja vahvistaa kaiku-kammiota. Crowd-signaaleihin perustuvat annotointidatat joutuivat kriittiseen valokeilaan: Fakeddit kerää totuutensa Reddit-äänestyksistä, mutta jo yksi satunnainen up-vote nostaa kommenttipisteitä 25 prosenttia (Muchnik ym. 2013) ja valtaosa äänistä annetaan luke- matta itse sisältöä (Glenski ym. 2015). Tämän vuoksi tutkijat puhuvat nyt kohi- naisesta massaviisaudesta, joka on esipuhdistettava esimerkiksi vahvistusoppi- villa valitsimilla (WeFEND) ennen kuin sen voi syöttää transformerille.

Kaiken keskellä ihmisen rooli on muuttunut manuaalisesta merkitsijästä erään- laiseksi laadun-varmistajaksi. Asiantuntijat rakentavat sääntö- ja tietografi-kerrok- sia, valvovat RL-valitsimien harvennusta ja arvioivat mallien selitettävyyttä. Me- todinen kypsyyttä näkyy siinä, että tehokkain järjestelmä ei ole yksittäinen huippu- malli, vaan ekosysteemi, jossa transformer, knowledge graph, graafineuroverkko, biologinen signaali ja ihmisen harkinta muodostavat toisiaan täydentävän koko- naisuuden.

4.3 Tulososion avainlöydökset

Aineistoon valikoitui 64 laajasti viitattua artikkelia tai konferenssijulkaisua (2011–2024). Tutkimuskirjallisuuden kehitys jakautuu selkeästi kolmeen metodiseen vaiheeseen: varhaisjaksoon (14 artikkelia, 2011–2015), keskijaksoon (33 artikkelia, 2016–2019) ja myöhäisjaksoon (17 artikkelia, 2020–2024). Julkaisut kasvoivat volyymiltään (14 → 33 → 17), mutta volyymin kasvu hidastui pandemian kynnyksellä, kun konferenssien rytmi katkesi ja pitkän vertaisarvioinnin lehdet sekä arXiv-preprintit korvasivat nopean proceedings-julkaisun (konferenssien osuus putosi 86 → 47 %). Ensimmäisten kirjoittajien taustat (66 % tietojenkäsittely, 22 % yhteiskuntatieteet, 12 % informaatiotiede) kertovat edelleen tietotekniikkaveitoisesta kentästä, vaikka tutkimuskysymykset ovat yhä poliittisia ja yhteiskunnallisia.

Taulukon 4 vertailumittarit korostavat metodista kypsymistä. Varhaisjaksolla jokainen työ oli binäärinen teksti-luokitin, jonka medianidatasettikoko jäi alle 3 000 esimerkin. Keskijaksolla standardoidut poliittiset benchmarkit (LIAR, Fake-NewsNet) kasvattivat jakaumia kymmeneen tuhansiin ja nostivat syväoppimisen osuudeksi 61 %. Myöhäisjaksolla syväoppiminen on oletus (17/17 artikkelia), keskimääräinen datasettikoko kolminkertaistui edelliseen kauteen verrattuna ja yli kolmannes töistä yhdistää tekstiä, kuvamateriaalia ja tietograafeja. Seurauksena mallien raakasuurituskyky on parantunut, mutta yksittäisen artikkelin samaa viittauskertymä on laskenut (mediaani 185 → 21) – kenttä on sirpaloitunut ja uusien innovaatioiden elinkaari lyhentynyt. Menetelmäakselioiden tarkastelu paljastaa epätasapainon: vaikka 56 % tutkimuksista on yhä puhtaasti sisältökeskeisiä, kausaalinen testaus (9 %), monikielisyys (13 %) ja eettinen vinouma-auditointi (8 %) laahaavat selvästi perässä. Tämä korostaa tarvetta laajentaa arviointikriteerejä pelkän tarkkuuden yli.

Taulukko 4: Menetelmäakselioiden kattavuus dis-/misinformaatiotutkimuksessa 2011–2024

Akseli	Varhaisjakso	Keskijakso	Myöhäisjakso	Yhteensä% (kpl/64)
Sisältökeskeisyys (pelkkä teksti/kuva)	71 % (10)	67 % (22)	24 % (4)	56 % (36)
Multimodaalisuus	7 % (1)	15 % (5)	35 % (6)	19 % (12)
Koordinaatio-analyysi	36 % (5)	36 % (12)	41 % (7)	38 % (24)
Kausaalisuus-testaus	0 % (0)	6 % (2)	24 % (4)	9 % (6)
Temporal-robustus	0 % (0)	9 % (3)	35 % (6)	14 % (9)
Monikielisyys	0 % (0)	9 % (3)	29 % (5)	13 % (8)
Eettinen vinouma-auditointi	0 % (0)	3 % (1)	24 % (4)	8 % (5)

Tarkastelua viidestä näkökulmasta:

Methodinen maturiteetti

Käsin toteutettu piirreinsinointi johti ensin klassisiin NLP-malleihin, sitten CNN/RNN-syväverkkoihin ja lopulta hybridiratkaisuihin, joissa transformerit, tietografit ja sääntöpohjaiset tarkistimet toimivat samassa putkessa. Uudessa ekosysteemissä BERT-/RoBERTa taustamalli, knowledge-graph-ankkuri ja selityskerrokseksi säilytetyt kielipiirteet muodostavat 'kolmoisluvon' (Ahmed ym. 2022; Gordon ym. 2021).

Modaliteettien konvergenssi

Tekstiin lisättiin ensin verkostodynamiikka, sitten kuva- ja videopiirteet ja lopulta heterogeeniset GNN-mallit, jotka ennustavat väitteen totuusarvon koko julkaisukäyttäjä-media-graafin tasolla (Nielsen ja McConville 2022). Monimodaalisuus on siten strateginen eikä enää kokeellinen lisä.

Skaalahaasteen muutos

Asiantuntija-labelointi oli varhaisjakson pullonkaula; nyt tutkimus kamppailee datamelun, luokkavinouman ja annotoijien erimielisyyksien kanssa. Crowd-signaaleja pitää harventaa (WeFEND) ja Reddit-pohjaisia 'totuuksia' arvioida kriittisesti, koska up-vote on suosio-, ei faktasignaali (Muchnik ym. 2013; Glenski ym. 2015). Luokkavinouma on kääntynyt pääläelleen: faktantarkistajat tuottavat nyt enemmän false- kuin true-merkintöjä, ja annotaattorit kiistelevät hienojakoisista luokista. Tämä lisää sekä koulutusdatan meluisuutta että arviointimetrikoiden epävarmuutta.

Suorituskyky vs. robustius

Varhaistunnistus on mahdollista (≤ 5 min): RNN+CNN-yhdistelmät havaitsevat valeuutisen tuoreeltaan (Liu ja Wu 2018). Samaan aikaan mallit ovat hauraita uusille generaattoreille: GPT-sarja heikentää tekstidetektoreita, ja täydellisesti tuotetut StyleGAN-kuvat pudottavat GPT-4V-tarkkuuden 80 prosenttiin (Jia ym. 2024). Mallien raakasuurituskyky on siis noussut (transformerit > 90 % makro-F1), mutta robustius ei: GPT- ja StyleGAN-sarjan generaattorit pudottavat parhaidenkin detektorien tarkkuutta 10–20 prosenttiyksikköä, mikä osoittaa, että nopeus on saavutettu kestävyuden kustannuksella.

Ihmisen roolin muutos

Ihmisen rooli siirtyy perinteisestä etumerkinnästä laadunvarmistukseen, kun asiantuntijat rakentavat tietografi- ja sääntökerroksia, suodattavat kohinaisia crowd-signaaleja (esim. WeFEND) ja auditoivat mallien selitettävyyttä sekä eettisiä vioumia.

Yhteenvedona voidaan sanoa, että tutkimus on kulkenut käsityönä rakennetusta avainsanaluokittimesta reaaliaikaiseen, multimodaaliseen ja osittain itsensä selittävään järjestelmään, jota haastavat nyt tekoälyn tuottamat "super-myrskyt" (Taulukko 5). Tekninen huippusuoritus on siis saavutettu, mutta seuraava kriittinen askel on tehdä järjestelmistä kestäviä, läpinäkyviä ja kontekstuaalisesti sovellettavissa – ilman että ihmisestä tulee pelkkä virheen-korjaaja.

Taulukko 5: *Tutkimuskentän kehitystrendit kolmessa ajallisessa vaiheessa*

Mittari / Painopiste	Varhaisjakso	Keskijakso	Myöhäisjakso
Julkaisujen määrä ¹	14	33	17
Dominoiva termistö	<i>Astroturf, rumour</i> , jihadistinen propaganda	Valeuutiset , mis-/disinformaatio	mis-/disinformaatio, synteettinen media , <i>infodemia</i>
Tyypilliset datasetit / lähteet	Twitter-/Weibo-API, manuaaliset listat	LIAR, FakeNewsNet, FaceForensics(++), PolitiFact/Snopos	MuMiN (41 kieltä), Fakeddit, StyleGAN / LLM-synt., COVID-aineistot
Mallinnusparadigma	Klassiset ML-mallit (SVM, NB, AdaBoost) + käsin rakennetut piirteet	Syväoppiminen (CNN, RNN, ensemble); ensimmäiset graafimallit (GCN) & fuusiot	Transformer-dominaatio; BERT/RoBERTa + knowledge graph + sääntöpohjaiset tarkistimet
Modaliteetit	Lähes pelkkä teksti (visuaalinen vain mainittu)	Teksti + verkosto, artefaktipohjainen kuva/videotunnistus käynnistyy	Monimodaaliset GNN:t: teksti + kuva/video + verkosto, biometriset signaalit (PPG)
Keskeinen skaala- haaste → ratkaisu	Asiantuntija-labelointi pullonkaula → pienet datat	Labeloinnin hinta → tykkäys/raportti-signaalit, RL-filtterit (WeFEND)	Datamelu, luokkavinouma, spoofatut sisällöt → semi-supervised & selitettävyyserros
Edustava tarkkuus / AUC ²	AdaBoost 96 % (astroturf), > 99 % (jihadisti-twiitit)	GCN AUC > 0,90; FaceForensics-CNN > 90 %	HeteroGraphSAGE F1 0,63 (MuMiN); FakeCatcher 97 %; GPT-4V ≈ 80 %
Ihmisen rooli	<i>Labeloija</i> (asiantuntija / joukkoistus)	<i>Kuratoija</i> (kohinasuodatus, tulkinta)	<i>Auditoija</i> (selitettävyys, eettinen valvonta)

5 TULOSTEN ARVIOINTI, JOHTOPÄÄTÖKSET JA YHTEENVETO

Tässä luvussa tiivistän tutkimuksen empiiriset löydökset, sijoitan ne demokratia-, instituutio- ja politiikkaprosessiteorioiden viitekehykseen ja johdan niiden pohjalta politiikkasuosituksia. Aluksi analysoin, miten hyvin tutkimus on kyennyt vastaamaan tutkimuksen tavoitteeseen ja tutkimuskysymyksiin. Sen jälkeen käsittelem otannan, datan ja mittareiden rajoitteet sekä datapolitiikasta johtuvat yleistettävyysongelmat. Luvun päätteeksi tarkastelen löydösten merkitystä demokraattiselle resilienssille ja algoritmien sääntelylle sekä esittelen tutkimuksen aikana hahmottuneita jatkotutkimusaiheita ja toimenpide-ehdotuksia, joilla ilmenneitä metodologisia ja institutionaalisia puutteita voidaan korjata.

Tutkimuksen tavoitteena oli rakentaa data-ankkuroitu, politiikan tutkimuksen keskusteluihin kytkeytyvä kokonaiskuva siitä, miten koneoppimisen ja tekoälyn menetelmiä on hyödynnetty propagandan automaattisessa havaitsemisessa vuosina 2011–2024, ja missä määrin nuo menetelmät vastaavat demokratia-, instituutio- ja politiikkaprosessitutkimuksen identifioimiin riskeihin.

Tarkoitusta varten koottiin 64 vertaisarvioidun, semanttisesti ja ajallisesti stratifioidun artikkelin otos, joka kattaa 92 % laajemman, 2 916:n julkaisun perusjoukon, sisällöllisestä varianssista. Otos analysoitiin ES-MTT-kehyksellä: bibliometrinen kartoitus (Exploratory), strukturoitu fokusoitu vertailu (Sequential) ja trenditaulukointi (Mixed-Trend-Tracing) tuottivat sekä pitkittäisen kehityskaaren että poikkeileikkaavan metodologisen profiilin. Näin tutkimus avaa kentän rakenteen kolmesta ulottuvuudesta — kronologisesti (aikasarjojen kehityskaari), multimodaali-

sesti (teksti, kuva, ääni, video) ja menetelmäulottuvuuksien mukaisesti (seitsemän kriittistä ulottuvuutta). Yksityiskohtaiset robustisuus- ja vinouma-auditit jäivät kuvaileviksi, mikä korostaa jatkotutkimustarvetta (ks. luku 5.3).

5.1 Tutkimuksen keskeiset löydökset

5.1.1 Kehityskaari 2011–2024: AI-menetelmien siirtymä tekstipohjaisista luokittimista multimodaalisiin, verkostotietoisiin ekosysteemeihin

Ensimmäinen tutkimuskysymys: *Miten koneoppimisen ja tekoälyn menetelmät ovat kehittyneet propagandan tunnistamisessa vuosina 2011–2024?*

Luvussa neljä kolmijakoinen kronologia osoittaa selvän siirtymän kolmessa vaiheessa.

Ensimmäisessä vaiheessa (varhaisjakso), vuosina 2011–2015, propagandan tunnistus perustui klassisiin koneoppimismenetelmiin ja käsin rakennettuihin piirteisiin; fokus oli lähes yksinomaan tekstipohjaisessa binääriluokituksessa, jota tässä työssä on kutsuttu ‘suljetuksi laatikoksi’.

Toisessa vaiheessa (keskijakso), vuosina 2016–2019, kenttä siirtyi CNN- ja RNN-syväoppimiseen, standardoituihin benchmark-aineistoihin, sekä ensimmäisiin graafipohjaisiin malleihin – tämä avasi ovet multimodaaliseen ja verkostotietoiseen tarkasteluun.

Kolmannessa vaiheessa (myöhäisjakso), vuosina 2020–2024, transformer-mallit vakiintuivat hallitsevaksi arkkitehtuuriksi, multimodaaliset GNN-ratkaisut ja heterogeeniset tietograafit yleistyivät ja kenttä muuttui ekosysteemiksi, jossa teksti-, kuva- ja leviämisdynamiikkasignaalit fuusioidaan samaan putkeen.

Kehityskaari vahvistaa teorialuvussa 2 esitetyn kolmikerroksisen kartaston: retoriset tekniikat mallinnetaan transformer-pohjaisilla sisältöenkoodereilla, bayesilainen vastaanottajaprosessi on toistaiseksi lähinnä teoreettinen epävarmuuden-

hallinnan taso, ja anomaliadetektion verkostokerros toteutuu graafineuroverkoilla. Löydökset osoittavat, että kentän painotus on edelleen "suljetuissa" tekstimyllyissä, vaikka algoritmiset mediasillat – suosittelu- ja moderointijärjestelmät – ovat nousseet demokraattisen valvonnan keskiöön.

5.1.2 Metodologis-eettinen kattavuus propagandan automaattisissa havaitsemismalleissa: seitsemän kriittistä ulottuvuutta ja demokraattiset katveet

Toinen tutkimuskysymys: *Missä määrin vuosina 2011–2024 julkaistut tekoäly- ja koneoppimismallit kattavat propagandan automaattisen havaitsemisen seitsemän kriittistä metodologis-eettistä ulottuvuutta – ja missä ovat suurimmat katveet demokraattisen päätöksenteon näkökulmasta?*

Luvussa neljä taulukointi paljastaa kentän epäsymmetrisen jaon. Tekstipainotteiset, sisältökeskeiset mallit kattavat noin 56 prosenttia tutkimuksista, koordinaatioanalyysi 38 prosenttia ja multimodaalisuus 19 prosenttia. Sen sijaan kausaalinen testaus yltää vain yhdeksään prosenttiin, temporal-robustisuus neljääntoista prosenttiin, monikielisyys kolmeentoista prosenttiin ja eettinen vinouma-auditointi kahdeksaan prosenttiin. Tulos vahvistaa luvussa 2.4 esitetyn strategisen ohuuden hypoteesin: *demokratian resilienssin kannalta kriittiset kestävyys, inklusion ja vinoumien mittarit ovat vähiten tutkittuja.*

Tutkimus onnistuu osoittamaan, miten propagandan automaattinen havaitseminen on kehittynyt suljetuista tekstimyllyistä kohti multimodaalisia, verkostotietoisia ekosysteemejä, ja että teoreettisen kartaston kolme kerrosta heijastuvat empiirisesti – mutta epätasapainoisesti. Helppokoodattavat tekstipohjaiset ratkaisut dominoivat, kun taas institutionaalisesti merkittävät ulottuvuudet, kuten kausaalisuus, temporal-robustus, monikielisyys ja eettisen vinouman tarkastelu, jäävät metodisesti katveeseen. Tämä epäsymmetria muodostaa keskeisen haasteen sekä tulevalle akateemiselle tutkimukselle että propagandaa säätelevälle poliitikalle. Tiivistän osin myös arvioitani tuloksista taulukossa 6.

Taulukko 6: Propagandan automaattisen havaitsemisen tutkimuskentän seitsemän kriittisen ulottuvuuden nykytila, teoreettiset kytkennät ja keskeiset puutteet (2011–2024)

Ulottuvuus	Keskeinen havainto	Teoreettinen kytkentä	Kriittinen arvio
Teksti-dominanssi	56 % tutkimuksista käyttää vain teksti- tai pikselipiirteitä.	Retoriset tekniikat (Dimitrov ym., 2021) tunnistetaan, mutta vain harva malli linkittää ne vastaanottajateoriaan. Yhdistää kartaston ensimmäisen ja kolmannen kerroksen (retoriikka + anomaliaverkosto).	Tekstikeskeisyys sivuuttaa affekti-delegitimaatio -suhteen viisuaalisissa disinformaatioketuissa (esim. deepfake-kampanjat 2024 EU-vaaleissa). Multimodaaliset datasetit (MuMiN, VLDBench) ovat vielä angloamerikkalais- ja twitter-painotteisia; ulkopuolinen yleistys rajoittuu.
Multimodaalisuus	Kasvaa 7 % → 35 % (2011→2024).	Vastaa Dahlin “valistunut ymmärrys”-kriteerin avoimen informaatiovirran näkökulmasta.	Useimmat koordinointimetriikat ovat deskriptiivisiä; kausaalisuuslinkki (ATE/CATE) puuttuu.
Koordinaatio-analyysi	38 % malleista sisältää verkostopiirteitä; GNN nostaa AUC-arvot > 0,90.	Polyarkian tilintekovelvollisuus edellyttää ajallisesti kestäviä malleja.	Hold-out-jakso on usein ≤ 6 kk; generatiivisen AI:n nopeudella tämä ei takaa todellista robustiutta.
Temporal-robustisuus	Vain 14 % tutkimuksista testaa ajallista siirtoa.	Liittyy demokratian inklusioperiaatteeseen.	Eurooppa- ja globaalietelän kielet marginalisoituvat; XLM-R-mallit heikkenevät matalavolyymisillä kielillä.
Monikielisyys	8/64 artikkelia testaa ≥ 2 kielellä ilman käännöstä.	Legitimititeetti-teoria (Grimmelikhuijsen & Meijer, 2022) korostaa vinoumien uhkaa luottamukselle.	Raportointi on hajanaista; vertailustandardeja (recall-gap, equal FPR) ei vielä noudateta systemaattisesti.
Eettinen viinoma	Auditointi esiintyy 5/64 tutkimuksessa.	Kytkeytyy Dahlilaiseen “kontrolli + vastuu” -periaatteeseen: kausaalinen attribuutio paljastaa, muuttaako propaganda vastaanottajan posteriorin vai vain korreloiko sen kanssa.	Ilman kausaalista testiä malli voi vain toistaa datan vinoumia; tarvitaan testejä, tuettua GNN-fittiä tai generatiivisten mallien counterfactual-simulointia.
Kausaalisuus	6/64 artikkelia (9 %) estimoivat vaikutuksia (ATE/CATE, IV-asetelma tai kontrafaktuaalit); valtaosa tyytyy korrelaatiomittareihin.		

Tulokset vahvistavat luvussa 2 kuvatun kolmitasoisien "kartaston" eriytymisen: ensimmäisellä tasolla retoriset tekniikat paljastavat mitä propagandisti sanoo, toisella bayesilainen vastaanottajaprosessi osoittaa miten viesti muuttaa uskomuksia ja kolmannella anomaliadetektion verkostokerros näyttää missä ja milloin viesti leviää poikkeavasti. Jokainen taso edellyttää erilaista menetelmäperhettä: sisältö- ja kuvapohjaisia luokittimia, kausaali- tai todennäköisyysanalyyssejä sekä graafi- ja aikarobusteja algoritmeja. Empiirinen analyysi osoittaa kuitenkin menetelmien vinoutuneen painotuksen.

Tekstipohjaiset, helposti koulutettavat mallit kattavat yli puolet tutkimuksista ja selittävät pääasiassa ensimmäistä kerrosta, kun taas temporal-robustisuus, monikielisyys ja eettisen vinouman tarkastelu – jotka ovat välttämättömiä kartaston kahdella ylemmällä tasolla – esiintyvät vain noin kymmenesosassa töitä. Kenttä siis ylittävät retoriikan suoraa tunnistamista mutta alittavat mallien kestävyyttä ajan, kielten ja väestöryhmien yli sekä koordinoitua kampanjan dynamiikkaa. Tämä epäsymmetria selittää, miksi nykyiset algoritmit luokittelevat yksittäiset valeväitteet melko luotettavasti, mutta horjuvat, kun propaganda on monikielistä, pitkäkestoista ja verkostotasolla organisoitua. Tulos alleviivaa politiikan tutkimuksen huolta: nykyinen dataperusteinen valmius kohdata koordinoituja, kielirajat ylittäviä propagandakampanjoita on strategisesti alikattava. Systemaattinen tarkastelu myös osoitti, että 38 % tutkimuksista sisälsi edes yhden verkostopiirteitä, mitkä myös nostivat AUC-arvon vakaasti $> 0,90$ verrattuna pelkkiin teksti- tai kuvamalleihin. Tämä voi antaa mallille kyvyn tunnistaa hiljaisia, synkronoituja vaikuttamiskuvioita, joita ei pelkästään yksittäisen viestin sisällöstä ei voi päätellä. Dahlin (1989) mukaan demokraattinen päätöksenteko on legitiimiä vain, jos kansalaisilla on ns. valistunut ymmärrys, eli he voivat muodostaa näkemyksensä luotettavan ja esteettömän tiedon varaan. Toisin sanoen, kansalaiset kykenevät tekemään autonomisia, julkista hyvää edistäviä valintoja vain, jos heillä on luotettavaan, avoimeen ja esteettömään informaatiovirtaan.

Koordinoidut vaikuttamisverkostot, kuten bottien nostattamat hashtagaallot, trolleihin kytkeytyvät 'astroturffien' informaatiovaikuttamispyrkimykset häiritsevät tätä virtaa muun muassa jo vääristämällä keskustelun näkyvyyttä ja tempoa. Kun

esimerkiksi GNN-pohjainen tunnistin havaitsee tällaisen verkostoanomalian aikaisessa vaiheessa, se voi mahdollistaa informaatiovirran avoimuuden ja luotavuuden lisääntymistä, sekä siten osaltaan lisätä Dahlin demokraattisen ihanteen 'valistunutta ymmärrystä'. Verkostopiirteiden hyödyntäminen ei siis ole pelkkä tekninen lisä, vaan suoraan demokraattisen legitimitetin kannalta avustava työkalu. Näiden työkalujen avulla voidaan paljastaa järjestelmällisen, agendapohjaisen ja organisoidun informaatiovaikuttamisen sekä mahdollistaa sen vaikutusten minimointia ennen suurta vahinkoa julkiselle keskustelulle.

5.2 Tutkimuksen rajoitukset ja jatkotutkimusaiheet

Tutkimuksen tulosten tulkinnassa on huomioitava sekä aineistolliset että menetelmälliset rajaukset, jotka kaventavat tutkimuksen yleistettävyyttä ja osoittavat jatkotutkimuksen polkuja.

Ensinnäkin perusjoukon muodostivat Semantic Scholar -tietokannasta kerätyt englanninkieliset julkaisut, mikä sulkee tutkimuksen ulkopuolelle vähäisemmällä kielialueilla tehdyn, vertaisarvioinnin ulkopuolelle jäävän tai harvoin siteeratun työn. Vaikka otanta kattaa 92 % koko aineiston semanttisesta varianssista, se painottuu niin aihepiireiltään kuin geopolitiselta kontekstiltään angloamerikkalaiseen ja eurooppalaiseen keskusteluun. Näin ollen esimerkiksi globaalin etelän kielirekisterit, paikalliset propagandatekniikat ja monikieliset sosiaalisen median alustat jäävät tämän sinänsä kattavan analyysin ulkopuolelle.

Toiseksi tutkimuksen seitsemää metodologis-eettistä ulottuvuutta koodattiin pilottivaiheessa kaksoiskoodauksella, mutta lopullinen luokitus toteutettiin yhden tutkijan toimesta. Tämä lisää tulkintaharhan mahdollisuutta erityisesti niissä luokissa, joissa kriteerit ovat tulkinnanvaraisia (esim. kausaalinen vaikutusarviointi tai eettisen vinouman auditointi).

Kolmanneksi käytetty trendianalyysi perustui julkaisuissa raportoituihin suorituskykymittareihin, jotka olivat hyvin heterogeenisiä: tutkimukset erosivat datarakenteen, luokkajaon, arviointimetriikan ja aikasiirron pituuden osalta. Tämä hajonta

rajoittaa mallien välisten tulosten vertailtavuutta ja voi johtaa optimistiseen kuvaukseen kentän tosiasiallisesta suorituskyvystä.

Edellä mainitut rajoitteet avaavat myös jatkotutkimusagendan. Ensimmäinen askel on kausaalisen verkkoanalytiikan integraatio: propensity-score-yhdistetty GNN tai diffuusio-pohjaiset ATE-estimaatit voisivat selittää, missä määrin koordinaatio todella muuttaa yleisön uskomuksia. Toiseksi tarvitaan monikielisiä ja multimodaalisia benchmark-aineistoja, jotka ulottuvat Euroopan ja Pohjois-Amerikan ulkopuolelle. Kolmantena suositellaan vakioitua vinouma-auditointia, esimerkiksi recall-gap- ja equal-opportunity-mittareita avoimilla Model Card -raporteilla. Neljäntenä on syytä institutionalisoida temporal-stressitesti: mallien tulisi läpäistä ainakin kahden vuoden liukuva aikajänne, ennen kuin ne kelpaavat operatiiviseen käyttöön.

EU:n AI-asetus (AI Act) ja digipalvelusäädös rajoittavat tutkijoiden pääsyä raakadataan, mikä sulkee ulkopuolelle alustojen suljetut signaalit; siksi tulevaisuuden tutkimus tarvitsee tutkimus-, teollisuus- ja viranomaiskonsortioita, jotka mahdollistavat sääntelyn puitteissa auditoitavat ja replikoitavat datavirrat. Tämän työn analyysi nojaantui avoimiin Twitter- ja uutisdatasetteihin, joten TikTakin, Redditin ja hakukoneiden signaalien poikkialustainen kulku jäi tarkastelun ulkopuolelle; kokonaiskuvan saavuttaminen edellyttää paneeliaikasarjoja ja agenttipohjaista simulointia. Lisäksi nykyiset mallit testattiin staattisessa ympäristössä, joten niiden adversaarinen resilienssi ja energiankulutuksella mitattu hiilijalanjälki on arvioitava “red-teaming stress lab” -asetelmissa, jotta ratkaisut ovat sekä turvallisia että kestäväällä tavalla skaalautuvia.

5.3 Yhteenveto ja suosituksia

Tutkimus osoittaa, että propagandan automaattinen havaitseminen on edennyt klassisista tekstiluokittimista kohti multimodaalisia, verkostotietoisia ekosysteemejä, mutta samalla on muodostunut merkittävä metodologinen epäsymmetria. Tekstipohjaiset ratkaisut ovat yliedustettuja, kun taas kausaalisuus, kielellinen inklusio, aikakestävyys ja vinoumahallinta ovat edelleen katvealueita. Poliitiikan

tutkimuksen näkökulmasta tämä tarkoittaa, että algoritmiset mediasillat voivat horjuttaa julkisen valvonnan edellytyksiä nopeammin kuin tutkimus ja sääntely ehtivät säätää vastatoimia.

Ensinnäkin automaattisen propagandantunnistuksen kehitystä on tuettava rakentamalla avoimet, monikieliset ja aidosti multimodaaliset benchmark-aineistot julkisyksityisinä konsortioina. Tällainen yhteiskehittely takaa sen, etteivät globaalin etelän kielet ja kontekstit jää tutkijoiden ja sääntelyn näkökentän ulkopuolelle ja että mallit voidaan auditoida läpinäkyvästi niiden koko elinkaaren ajan.

Toiseksi jokaiselta julkisessa informaatioympäristössä käyttöön otettavalta algoritmilta on vaadittava kausaalinen riskiraportti, jossa mallin vaikutukset arvioidaan ATE- ja CATE-metriikoilla sekä läpäistään liukuva, vähintään kahden vuoden temporal-robustisuustesti. Näiden vaatimusten sisällyttäminen EU:n AI-asetuksen soveltamisohjeisiin loisi yhdenmukaisen minimistandardin, joka suojaa päätöksentekoa nopeasti muuttuvilta ja paikallisesti eriytyviltä disinformaatiotaktikoilta.

Kolmanneksi avoimeen tarkasteluun perustuvat red-teaming-laboratoriot on vakiinnutettava osaksi tätä ekosysteemiä. Riippumattomien tutkijoiden ja kansalaisjärjestöjen tulee päästä systemaattisesti haastamaan malleja tarkoituksellisesti muuntuvilla hyökkäyksillä ja julkaisemaan tulokset Model Card-tyyppisissä katselmuksissa, jotta mallien resilienssi ja energiataloudellinen kestävyys voidaan varmistaa käytännössä. Vasta tällaisilla koordinoituilla toimilla automaattinen propagandantunnistus voi palvella demokraattisen resilienssin, institutionaalisen läpinäkyvyyden ja julkisen luottamuksen tavoitteita.

LÄHDELUETTELO

Abdelminaam, D.S., Ismail, F.H., Taha, M., Taha, A., Houssein, E.H. & Nabil, A. 2021. CoAID-DEEP: An Optimized Intelligent Framework for Automated Detecting COVID-19 Misleading Information on Twitter. *IEEE Access* 9, 27840–27854. <https://doi.org/10.1109/ACCESS.2021.3058066>

Afchar, D., Nozick, V., Yamagishi, J. & Echizen, I. 2018. MesoNet: a compact facial video forgery detection network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7. <https://doi.org/10.1109/WIFS.2018.8630761>

Ahmed, H., Traore, I. & Saad, S. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy* 1 (1), e9. <https://doi.org/10.1002/spy2.9>

Ahmed, N., Mirza, F., Zahid, Z. & Ilyas, F. 2022. A survey on deepfake video detection. *IET Biometrics* 10 (6), 607–624. <https://doi.org/10.1049/bme2.12031>

Ajao, O., Bhowmik, D. & Zargari, S. 2018. Fake news identification on Twitter with hybrid CNN and RNN models. *Proceedings of the 9th International Conference on Social Media and Society (SMSociety '18)*, 226–230. <https://doi.org/10.1145/3217804.3217917>

Aksnes, D.W. 2006. Citation rates and perceptions of scientific contribution. *Journal of the American Society for Information Science and Technology* 57 (2), 169–185. <https://doi.org/10.1002/asi.20313>

Al-Alshaqi, M., Rawat, D.B. & Liu, C. 2024. Ensemble techniques for robust fake news detection: A comparative analysis. *Sensors* 24 (18), 6062. <https://doi.org/10.3390/s24186062>

Aly, R., Guo, Z., Schlichtkrull, M., Thorne, J., Vlachos, A., Christodoulopoulos, C., Cocarascu, O. & Mittal, A. 2021. FEVEROUS: Fact extraction and verification over unstructured and structured information. arXiv 2106.05707. <https://doi.org/10.48550/arXiv.2106.05707>

Aragón, A.M. 2013. A measure for the impact of research. *Scientific Reports* 3, 1649. <https://doi.org/10.1038/srep01649>

Ashcroft, M., Fisher, C., Kaati, L., Omer, E. & Prucha, N. 2015. Detecting jihadist messages on Twitter. 2015 European Intelligence and Security Informatics Conference, 161–164. <https://doi.org/10.1109/EISIC.2015.27>

Associated Press. 2024. AI could supercharge disinformation and election lies, experts warn. AP News 21.4.2024. Viitattu 26.5.2025. <https://apnews.com/article/ai-deepfakes-election-disinformation-2024>

Bascur, J.P., Verberne, S., van Eck, N.J. & Waltman, L. 2025. Which topics are best represented by science maps? An analysis of clustering effectiveness for author and document maps. *Scientometrics* 130, 1181–1199. <https://doi.org/10.1007/s11192-024-05218-6>

Bhattacharya, S., Kamper, F. & Beirlant, J. 2023. Outlier detection based on extreme value theory and applications. *Scandinavian Journal of Statistics* 50 (3), 1466–1502. <https://doi.org/10.1111/sjos.12665>

Bin Naeem, S. & Kamel Boulos, M.N. 2021. COVID-19 misinformation online and health literacy: A brief overview. *International Journal of Environmental Research and Public Health* 18 (15), 8091. <https://doi.org/10.3390/ijerph18158091>

Bodnar, T.J., Barclay, V.C., Ram, N., Tucker, C.S. & Salathé, M. 2014. Increasing the veracity of event detection on social media networks through user trust modeling. 2014 IEEE International Conference on Big Data, 206–214. <https://doi.org/10.1109/BigData.2014.7004286>

Bond, S. 2024. How AI deepfakes polluted elections in 2024. NPR 21.12.2024. Viitattu 26.5.2025. <https://www.npr.org/2024/12/21/ai-deepfakes-elections-2024>

Brennen, S.B., Sanderson, Z. & de la Puerta, C. 2025. AI's impact on elections: Global patterns and implications. Brookings 4.3.2025. Viitattu 26.5.2025. <https://www.brookings.edu/articles/ai-impact-elections-global-patterns>

Brin, S. & Page, L. 1998. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30 (1–7), 107–117. [https://doi.org/10.1016/S0169-7552\(98\)00110-X](https://doi.org/10.1016/S0169-7552(98)00110-X)

Britannica. 2025. Propaganda – definition, history, techniques, examples. Encyclopedia Britannica. Viitattu 26.5.2025. <https://www.britannica.com/topic/propaganda>

Chai, K.E.K., Lines, R.L.J., Gucciardi, D.F. & Ng, L. 2021. Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews. Systematic Reviews 10, 93. <https://doi.org/10.1186/s13643-021-01635-3>

Chandola, V., Banerjee, A. & Kumar, V. 2009. Anomaly detection: a survey. ACM Computing Surveys 41 (3), 15. <https://doi.org/10.1145/1541880.1541882>

Chen, F. & Neill, D.B. 2014. Non-parametric scan statistics for multivariate event detection and visualization. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1166–1175. <https://doi.org/10.1145/2623330.2623619>

Christou, L., Bompotas, A. & Makris, C. 2024. Document embeddings for long texts: A comparative analysis. Research Square Preprint. <https://doi.org/10.21203/rs.3.rs-5459822/v1>

Ciftci, U.A., Demir, I. & Yin, L. 2020. FakeCatcher: Detection of synthetic portrait videos using biological signals. IEEE Transactions on Pattern Analysis and

Machine Intelligence 45 (10), 11935–11956.
<https://doi.org/10.1109/TPAMI.2020.3009287>

Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D.S. 2020. SPECTER: Document-level representation learning using citation-informed transformers. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2270–2282. <https://doi.org/10.18653/v1/2020.acl-main.207>

Da San Martino, G., Yu, S., Barrón-Cedeño, A., Petrov, R. & Nakov, P. 2019. Fine-grained analysis of propaganda in news articles. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 5636–5646. <https://doi.org/10.18653/v1/D19-1565>

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R. & Nakov, P. 2020a. A survey on computational propaganda detection. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, 4826–4832. <https://doi.org/10.24963/ijcai.2020/672>

Da San Martino, G., Barrón-Cedeño, A., Wachsmuth, H., Petrov, R. & Nakov, P. 2020b. SemEval-2020 Task 11: Detection of propaganda techniques in news articles. Proceedings of the 14th International Workshop on Semantic Evaluation, 1377–1414. <https://doi.org/10.18653/v1/2020.semeval-1.186>

Dallison, P. 2024. The EU's first deepfakes election? Politico – EU Election Playbook 10.5.2024. Viitattu 26.5.2025. <https://www.politico.eu/article/eu-first-deepfakes-election-2024>

Das, B., Sharma, K., Chakraborty, T. & Goyal, P. 2025. Weakly supervised learning for textual propaganda detection. Communications in Computer and Information Science 2385. https://doi.org/10.1007/978-3-031-58547-0_15

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E. & Rizzo, C. 2023. ChatGPT and the rise of large language models: the new AI-

driven infodemic threat in public health. *Frontiers in Public Health* 11, 1134567. <https://doi.org/10.3389/fpubh.2023.1134567>

Delgado-Chaves, F.M., Bonada-Caparrós, J., Pareja-Flores, A. & Saez-Achaerandio, R. 2025. Transforming literature screening with artificial intelligence: Performance evaluation of ChatGPT and human experts. *Proceedings of the National Academy of Sciences* 122 (2), e2411962122. <https://doi.org/10.1073/pnas.2411962122>

Della Porta, D. & Keating, M. (toim.) 2008. *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511801938>

Della Vedova, M.L., Tacchini, E., Moret, S., Ballarin, G., DiPierro, M. & de Alfaro, L. 2018. Automatic online fake news detection combining content and social signals. 2018 22nd Conference of Open Innovations Association (FRUCT), 272–279. <https://doi.org/10.23919/FRUCT.2018.8468301>

Diakopoulos, N. 2015. Algorithmic accountability. *Digital Journalism* 3 (3), 398–415. <https://doi.org/10.1080/21670811.2014.976411>

Dimitrov, D., Bin Ali, B., Shaar, S., Alam, F., Silvestri, F., Firooz, H., Nakov, P. & Da San Martino, G. 2021. Detecting propaganda techniques in memes. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 6603–6613. <https://doi.org/10.18653/v1/2021.acl-long.516>

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M. & Ferrer, C.C. 2019. The Deepfake Detection Challenge preview dataset. arXiv 1910.08854. <https://doi.org/10.48550/arXiv.1910.08854>

Dubé, E., Dionne, M., Rochette, L., Désilets, J., Tchoubi, S. & Sauvageau, C. 2022. Understanding the influence of web-based information on COVID-19 vaccine uptake in Quebec: Web-based questionnaire study. *JMIR Research Protocols* 11 (10), e41012. <https://doi.org/10.2196/41012>

Dunn, A.G., Leask, J., Zhou, X., Mandl, K.D. & Coiera, E. 2015. HPV vaccine negativity on the Web: Amplification of risks in online social media. *Journal of Medical Internet Research* 17 (6), e144. <https://doi.org/10.2196/jmir.4343>

Euroopan parlamentin ja neuvoston asetus (EU) 2024/1689, annettu 13.6.2024, tekoälyä koskevien yhdenmukaistettujen sääntöjen vahvistamisesta ja eräiden unionin säädösten muuttamisesta (*tekoälyasetus*, Artificial Intelligence Act). *Euroopan unionin virallinen lehti L*, 12.7.2024. Viitattu 26.5.2025. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

Fazio, L.K., Brashier, N.M., Payne, B.K. & Marsh, E.J. 2015. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General* 144 (5), 993–1002. <https://doi.org/10.1037/xge0000098>

Ferrara, E., Varol, O., Menczer, F. & Flammini, A. 2017. Detection of promoted social media campaigns. *Proceedings of the 10th International AAAI Conference on Web and Social Media*, 420–426. <https://doi.org/10.1609/icwsm.v10i1.14889>

Ferrer, X., van Nuenen, T., Such, J.M. & Criado, N. 2020. Language biases in Reddit. arXiv 2008.02754. <https://doi.org/10.48550/arXiv.2008.02754>

Fisch, C. & Block, J. 2018. Six tips for your systematic literature review in business and management research. *Management Review Quarterly* 68, 103–106. <https://doi.org/10.1007/s11301-018-0142-x>

Gao, S., Alawad, M., Young, M.T., Gounley, J., Schaefferkoetter, N., Yoon, H.J., Wu, X-C., Durbin, E.B., Doherty, J., Stroup, A., Coyle, L. & Tourassi, G. 2021. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics* 25 (9), 3596–3607. <https://doi.org/10.1109/JBHI.2021.3062322>

Gentzkow, M., Kelly, B. & Taddy, M. 2022. Algorithmic content and polarization: Evidence from social media. *Proceedings of the National Academy of Sciences* 119 (48), e2208352119. <https://doi.org/10.1073/pnas.2208352119>

George, A.L. & Bennett, A. 2005. Case Studies and Theory Development in the Social Sciences. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/1668.001.0001>

Ghenai, A. & Mejova, Y. 2018. Fake cures: user-centric modeling of health misinformation in social media. Proceedings of the ACM on Human-Computer Interaction 2 (CSCW), 1–20. <https://doi.org/10.1145/3274327>

Glenski, M., Johnston, T. & Weninger, T. 2015. Random voting effects in social media. Proceedings of the 26th ACM Conference on Hypertext & Social Media, 290–293. <https://doi.org/10.1145/2700171.2791050>

Glenski, M., Johnston, T. & Weninger, T. 2017. Rating effects on social news posts and comments. ACM Transactions on Intelligent Systems and Technology 8 (6), 68. <https://doi.org/10.1145/3106367>

Gordon, C.S., Reeds, K., Vanderbilt, K., Lyles, D., Allicock, M. & Ribisl, K.M. 2021. Outcomes of the SoMe social media literacy program on adolescent tobacco and nicotine use. Nutrients 13 (11), 3825. <https://doi.org/10.3390/nu13113825>

Gorwa, R., Binns, R. & Katzenbach, C. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 7 (1). <https://doi.org/10.1177/2053951720933991>

Graves, L. 2018. Understanding the Promise and Limits of Automated Fact-Checking. Oxford: Reuters Institute for the Study of Journalism.

Grimmelikhuijsen, S., & Meijer, A. (2022). Legitimacy of Algorithmic Decision-Making: Six Threats and the Need for a Calibrated Institutional Response. Perspectives on Public Management and Governance, 5(3), 232–242. <https://doi.org/10.1093/ppmgov/gvac008>

Groh, M., Epstein, Z., Firestone, C. & Picard, R. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119 (1), e2110013119. <https://doi.org/10.1073/pnas.2110013119>

Haddaway, N.R., Bethel, A., Dicks, L.V., Koricheva, J., Macura, B., Petrokofsky, G., Pullin, A.S., Savilaakso, S. & Stewart, G.B. 2020. Eight problems with literature reviews and how to fix them. *Nature Ecology & Evolution* 4, 1582–1589. <https://doi.org/10.1038/s41559-020-01295-x>

Hahn, U. & Oaksford, M. 2007. The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review* 114 (3), 704–732. <https://doi.org/10.1037/0033-295X.114.3.704>

Hammer, R., Park, S., Kwon, O., Lee, J. & Shin, H. 2023. Deepfake video detection: a comprehensive review. *IEEE Access* 11, 37210–37235. <https://doi.org/10.1109/ACCESS.2023.3245688>

Horne, B.D. & Adalı, S. 2017. This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 759–766. <https://doi.org/10.1609/icwsm.v11i1.14976>

Huang, J.T., Choi, J. & Wan, Y. 2024. Politically biased moderation drives echo-chamber formation in social media. SSRN 17.10.2024. <https://doi.org/10.2139/ssrn.4990127>

Ichihashi, S. & Meng, D. 2021. The Design and Interpretation of Information. SSRN Working Paper. <https://doi.org/10.2139/ssrn.3922025>

Janze, C. & Risius, M. 2017. Automatic detection of fake news on social media platforms. *Proceedings of the 21st Pacific Asia Conference on Information Systems Paper* 261.

Jawahar, G., Abdul-Mageed, M. & Lakshmanan, L.V.S. 2020. Automatic detection of machine-generated text: A critical survey. arXiv 2011.01314. <https://doi.org/10.48550/arXiv.2011.01314>

Jia, S., Zhang, T., Li, X. & Zhang, R. 2024. Can ChatGPT detect deepfakes? A study of using multimodal large language models for media forensics. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 4325–4334. <https://doi.org/10.1109/CVPRW63382.2024.00438>

Jiang, T., Li, J.P., Haq, A.U., Saboor, A. & Ali, A. 2021. A novel stacking approach for accurate detection of fake news. IEEE Access 9, 22626–22639. <https://doi.org/10.1109/ACCESS.2021.3056079>

Johnson, M.K. & Raye, C.L. 1981. Reality monitoring. Psychological Review 88 (1), 67–85. <https://doi.org/10.1037/0033-295X.88.1.67>

Jolliffe, I.T. & Cadima, J. 2016. Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A 374 (2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>

Joos, L., Keim, D.A. & Fischer, M.T. 2024. Cutting through the clutter: A novel approach to text embeddings for enhanced clustering and retrieval. arXiv 2407.10652. <https://doi.org/10.48550/arXiv.2407.10652>

Jowett, G.S. & O'Donnell, V. 2015. Propaganda & Persuasion. 6. painos. Thousand Oaks: SAGE Publications.

Jurafsky, D. & Martin, J.H. 2025. Speech and Language Processing. 3. painos. Viitattu 26.5.2025. <https://web.stanford.edu/~jurafsky/slp3/>

Kaati, L., Omer, E., Prucha, N. & Shrestha, A. 2015. Detecting multipliers of jihadism on Twitter. Proceedings of the 2015 IEEE International Conference on Data Mining Workshops (ICDMW 2015), 954–960. <https://doi.org/10.1109/ICDMW.2015.9>

Kamenica, E. & Gentzkow, M. 2011. Bayesian persuasion. *American Economic Review* 101 (6), 2590–2615. <https://doi.org/10.1257/aer.101.6.2590>

Karras, T., Aila, T., Laine, S. & Lehtinen, J. 2018. Progressive growing of GANs for improved quality, stability, and variation. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*. <https://openreview.net/forum?id=Hk99zCeAb>

Kekkonen, K. 2025. Kaiutin vai suodatin? Suomalainen media, Katalonian itsenäisyysliike ja informaatiovaikuttaminen. *Tampere University Dissertations No. 1218*. Tampereen yliopisto. Väitöskirja. <http://urn.fi/URN:ISBN:978-952-03-3887-9>

Keraghel, I., Morbieu, S. & Nadif, M. 2024. Beyond words: Comparative analysis of LLM embeddings for effective clustering. *Lecture Notes in Computer Science* 14641, 278–291. https://doi.org/10.1007/978-3-031-58547-0_17

Klinger, U., Kreiss, D. & Mutsvairo, B. 2024. *Platforms, Power, and Politics: An Introduction to Political Communication in the Digital Age*. Cambridge: Polity Press.

Kreps, S., McCain, M. & Brundage, M. 2020. All the News That's Fit to Fabricate: AI-generated text as a tool of media misinformation. *SSRN Working Paper No. 3525002*. <https://doi.org/10.2139/ssrn.3525002>

Lenin, V.I. 1961. What Is to Be Done? *Teoksessa Collected Works* 5, 347–530. Moscow: Foreign Languages Publishing House.

Li, Y., Yang, X., Wu, P., Li, H., Li, X. & Zhao, J. 2020. MM-COVID: A multilingual and multimodal COVID-19 misinformation dataset. *arXiv* 2011.04088. <https://doi.org/10.48550/arXiv.2011.04088>

Li, Y. & Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv* 1811.00656. <https://doi.org/10.48550/arXiv.1811.00656>

Li, Q., Zhang, X., Chen, J., Wu, Y. & Xiao, J. 2024. Towards multimodal disinformation detection by vision-language interaction. *Information Fusion* 102, 102037. <https://doi.org/10.1016/j.inffus.2025.102037>

Li, W., Gao, W., Ma, R., Li, J. & Li, Y. 2022. Span identification and technique classification of propaganda in news articles. *Complex & Intelligent Systems* 8 (5), 3603–3612. <https://doi.org/10.1007/s40747-021-00393-y>

Lin, J. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37 (1), 145–151. <https://doi.org/10.1109/18.61115>

Lindstrom, M. R., Jung, H., & Larocque, D. 2020. Functional Kernel Density Estimation: Point and Fourier Approaches to Time Series Anomaly Detection. *Entropy*, 22(12), 1363. <https://doi.org/10.3390/e22121363>

Liu, Y. & Wu, Y-F.B. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 354–361. <https://doi.org/10.1609/aaai.v32i1.11268>

Maier, A. & Riess, C. (2024). Reliable Out-of-Distribution Recognition of Synthetic Images. *Journal of Imaging*, 10 (5), 110. <https://doi.org/10.3390/jimaging10050110>.

Marwick, A.E. & Lewis, R. 2017. *Media Manipulation and Disinformation Online*. New York: Data & Society Research Institute.

Miles, M.B., Huberman, A.M. & Saldaña, J. 2014. *Qualitative Data Analysis: A Methods Sourcebook*. 3. painos. Thousand Oaks: SAGE Publications.

Miller, D., Benson, R. & Fowler, B. (toim.) 2015. *Sociology, Propaganda and Psychological Operations*. London: Routledge.

Miller, J.K. & Alexander, T.J. 2025. Moving past single metrics: A comprehensive framework for evaluating scientific impact. arXiv 2502.17020. <https://doi.org/10.48550/arXiv.2502.17020>

Min, C., Bu, Y., Wu, D., Ding, Y. & Zhang, Y. 2020. Citation cascade and topic relevance: Understanding the interaction between content and citation networks. arXiv 2004.12275. <https://doi.org/10.48550/arXiv.2004.12275>

Mirsky, Y. & Lee, W. 2021. The creation and detection of deepfakes: A survey. ACM Computing Surveys 54 (1), 1–41. <https://doi.org/10.1145/3425780>

Mishra, A. & Sadia, H. 2023. A comprehensive analysis of fake-news detection models. Engineering Proceedings 59 (1), 28. <https://doi.org/10.3390/engproc2023059028>

Moed, H.F. 2005. Citation Analysis in Research Evaluation. Dordrecht: Springer.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D.G. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLOS Medicine 6 (7), e1000097. <https://doi.org/10.1371/journal.pmed.1000097>

Monti, F., Frasca, F., Eynard, D., Mannion, D. & Bronstein, M.M. 2019. Fake news detection on social media using geometric deep learning. arXiv 1902.06673. <https://doi.org/10.48550/arXiv.1902.06673>

Mouton, C.A., Lucas, C. & Ee, S. 2025. Defending American Interests Abroad: Early Detection of Foreign Malign Information Operations. Santa Monica: RAND Corporation.

Muchnik, L., Aral, S. & Taylor, S.J. 2013. Social influence bias: A randomized experiment. Science 341 (6146), 647–651. <https://doi.org/10.1126/science.1240466>

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T.L., et al. 2022. Crosslingual generalization through multitask finetuning. arXiv 2211.01786. <https://doi.org/10.48550/arXiv.2211.01786>

Muennighoff, N., Tazi, N., Magne, L. & Reimers, N. 2023. MTEB: Massive Text Embedding Benchmark. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2014–2037. <https://doi.org/10.18653/v1/2023.eacl-main.148>

Nakamura, K., Levy, S. & Wang, W.Y. 2020. r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. Proceedings of the 12th Language Resources and Evaluation Conference, 1617–1626. <https://doi.org/10.48550/arXiv.1911.03854>

Nakamura, Y.K., Levy, S. & Wang, W.Y. 2020. Emotion, echo chambers, and virality: A deep learning approach to understanding the spread of misinformation. arXiv 2012.01740. <https://doi.org/10.48550/arXiv.2012.01740>

Nguyen, V-H., Sugiyama, K., Nakov, P. & Kan, M-Y. 2020. FANG: Leveraging social context for fake news detection using graph representation learning. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 1615–1624. <https://doi.org/10.1145/3340531.3412046>

Nielsen, D.S. & McConville, R. 2022. MuMiN: A large-scale multilingual multimodal fact-checked misinformation social network dataset. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 3141–3153. <https://doi.org/10.1145/3477495.3531744>

Nightingale, S.J. & Farid, H. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences 119 (8), e2120481119. <https://doi.org/10.1073/pnas.2120481119>

OpenAI. 2024. OpenAI Platform Documentation – Embeddings. Viitattu 26.5.2025. <https://platform.openai.com/docs/guides/embeddings>

Oshikawa, R., Qian, J. & Wang, W.Y. 2018. A survey on natural language processing for fake news detection. arXiv 1811.00770. <https://doi.org/10.48550/arXiv.1811.00770>

Ozbay, F.A. & Alatas, B. 2019. *Fake news detection within online social media using supervised artificial intelligence algorithms*. *Physica A: Statistical Mechanics and its Applications* 540, 123271. <https://doi.org/10.1016/j.physa.2019.123271>

Pennycook, G. & Rand, D.G. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>

Qiao, J., Feng, X., Wang, Z., Mao, K. & Jiang, J. 2025. Improving multimodal fake news detection with cross-modal consistency and external knowledge. <https://doi.org/10.1016/j.ipm.2025.104120>

Rakhi, B., Gupta, B. & Lamba, S.S. 2024. Local outlier detection using kernel density estimation. *Franklin Open* 8, 100162. <https://doi.org/10.1016/j.fraope.2024.100162>

Rakholia, N. & Bhargava, S. 2017. "Is it true?" – Deep learning for stance detection in fake news. Stanford University. Projektiraportti.

Ratkiewicz, J., Conover, M.D., Meiss, M., Gonçalves, B., Flammini, A. & Menczer, F. 2011. Detecting and tracking political abuse in social media. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 297–304. <https://doi.org/10.1609/icwsm.v5i1.14127>

Raza, S., Garg, M., Reji, D.J., Bashir, S.R. & Ding, C. 2025. VLDBench: A comprehensive benchmark for vision-language disinformation detection. arXiv 2402.01999. <https://doi.org/10.48550/arXiv.2402.01999>

Ricardo Campello, R.J.G.B., Moulavi, D., Zimek, A. & Sander, J. 2015. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data* 10 (1), 5. <https://doi.org/10.1145/2733381>

Rodríguez de Las Heras Ballell, T. 2025. Mapping generative AI rules: The regulatory landscape in the EU and China. *Cambridge Forum on AI* 1, e5. <https://doi.org/10.1017/cfa.2025.5>

Rogers, R. 2021. *Mainstreaming the Fringe: How Misinformation Propagates on Social Media*. Amsterdam: Amsterdam University Press.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. & Nießner, M. 2019. FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11. <https://doi.org/10.1109/ICCV.2019.00009>

Roy, A., Basak, K., Ekbal, A. & Bhattacharyya, P. 2018. A deep ensemble framework for fake news detection and classification. *arXiv* 1811.04670. <https://doi.org/10.48550/arXiv.1811.04670>

Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I. & Natarajan, P. 2019. Recurrent convolutional strategies for face manipulation detection in videos. *arXiv* 1905.00582. <https://doi.org/10.48550/arXiv.1905.00582>

Sahin, U., Senaratne, H., Winkler, A. & Riedl, M. 2023. ARC-NLP at multimodal hate-speech event detection 2023: A multimodal approach for hate-speech detection in memes. *arXiv* 2310.01234. <https://doi.org/10.48550/arXiv.2310.01234>

Santos Jr., E. 2006. *Deception detection in expert source information: A computational approach*. Air Force Research Laboratory. Tutkimusraportti. Semantic Scholar. 2025. Frequently Asked Questions. Viitattu 26.5.2025. <https://www.semanticscholar.org/faq>

Shu, K., Wang, S., & Liu, H. 2019a. Beyond News Contents: The Role of Social Context for Fake News Detection. Proceedings of the 12th ACM International Conference on Web Search and Data Mining (WSDM 2019), 312–320. <https://doi.org/10.1145/3289600.3290994>

Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. 2019b. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. arXiv (Version 3), arXiv:1809.01286. <https://doi.org/10.48550/arXiv.1809.01286>

Shu, K. & Liu, H. 2022. Detecting Fake News on Social Media. Cham: Springer. <https://doi.org/10.1007/978-3-031-01915-9>

Singh, I.P., Goyal, S., Kumari, A. & Malhotra, D. 2023. Multi-type deepfake detection. 2023 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP) Proceedings, 1–6. <https://doi.org/10.1109/MMSP59012.2023.10337635>

Slovic, P. 2010. The Feeling of Risk: New Perspectives on Risk Perception. London: Routledge. <https://doi.org/10.4324/9781849776677>

Solaiman, I., Brundage, M., Clark, J., Askeel, A., Herbert-Voss, A., Wu, J., ym. 2019. Release strategies and the social impacts of language models. arXiv 1908.09203. <https://doi.org/10.48550/arXiv.1908.09203>

Spitale, G., Biller-Andorno, N. & Germani, F. 2023. AI model GPT-3 (dis)informs us better than humans. Science Advances 9 (26), eadh1850. <https://doi.org/10.1126/sciadv.adh1850>

Stanley, J. 2015. How Propaganda Works. Princeton: Princeton University Press. <https://doi.org/10.1515/9781400865802>

SuthanthiraDevi, P., Rajalakshmi, S., Kannan, K. & Devi, A. 2020. Detection of propaganda from news articles using deep learning. *International Journal of Advanced Trends in Computer Science and Engineering* 9 (3), 3500–3505. <https://doi.org/10.30534/ijatcse/2020/166932020>

Tacchini, E., Ballarin, G., Della Vedova, M.L., Moret, S. & de Alfaro, L. 2017. Some like it hoax: Automated fake news detection in social networks. arXiv 1704.07506. <https://doi.org/10.48550/arXiv.1704.07506>

Thorne, J. & Vlachos, A. 2018. Automated fact checking: Task formulations, methods and future directions. arXiv 1806.07687. <https://doi.org/10.48550/arXiv.1806.07687>

Tondo, L. 2025. Italian opposition complaint over racist AI images. *The Guardian* 18.4.2025. Viitattu 26.5.2025. <https://www.theguardian.com/world/2025/apr/18/italian-opposition-complaint-racist-ai-images>

Trielli, D. & Diakopoulos, N. 2022. Algorithmic agenda setting: How search engines and social media platforms influence journalistic practices. *International Symposium on Online Journalism Journal* 12 (1), 45–70.

Trubey, P. 2025. Exploring Multivariate Extreme Value Theory with Applications to Anomaly Detection. UC eScholarship. <https://escholarship.org/uc/item/2r43f9zn>

Vaccari, C. & Chadwick, A. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society* 6 (1). <https://doi.org/10.1177/2056305120903408>

Vignotto, E. & Engelke, S. 2020. Extreme value theory for anomaly detection—the GPD classifier. *Extremes* 23, 501–520. <https://doi.org/10.1007/s10687-020-00393-0>

Vijjali, R., Potluri, P., Kumar, S. & Teki, S. 2020. Two-stage transformer model for COVID-19 fake news detection and fact checking. arXiv 2011.13253. <https://doi.org/10.48550/arXiv.2011.13253>

Vosoughi, S., Roy, D. & Aral, S. 2018. The spread of true and false news online. *Science* 359 (6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>

Walton, D. 2007. *Media Argumentation: Dialectic, Persuasion, and Rhetoric*. Cambridge: Cambridge University Press.

Wang, W.Y. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 422–426. <https://doi.org/10.18653/v1/P17-2067>

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L. & Gao, J. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 849–857. <https://doi.org/10.1145/3219819.3219903>

Wang, Y., Yang, W., Ma, F., Xu, J., Zhong, B., Deng, Q. & Gao, J. 2019. Weak Supervision for Fake News Detection via Reinforcement Learning. arXiv 1912.12520. Viitattu 26.5.2025. <https://arxiv.org/abs/1912.12520>

Wang, R., Ma, F., Zhang, X., Liu, H. & Li, J. 2020. FakeSpotter: A multimodal framework for fake news detection. Proceedings of the 29th International Joint Conference on Artificial Intelligence, 3445–3451. <https://doi.org/10.24963/ijcai.2020/476>

Wang, C., Neill, D.B. & Chen, F. 2022. Calibrated non-parametric scan statistics for anomaly detection in graphs. Proceedings of the 36th AAAI Conference on Artificial Intelligence 36 (4), 4201–4209. <https://doi.org/10.1609/aaai.v36i4.20339>

Wang, Y., Sun, Y., Huang, H. & Rudin, C. 2024. Dimension reduction with locally adjusted graphs. Proceedings of the 38th AAAI Conference on Artificial Intelligence, 15556–15564. <https://doi.org/10.1609/aaai.v38i14.29481>

World Health Organization (WHO). 2020. How to report misinformation online. Viitattu 26.5.2025. <https://www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/how-to-report-misinformation-online>

Yang, K., Zhou, S., Dong, Z., Yang, J. & Wang, W. 2023. Detecting multimodal AI-generated content with limited training data. arXiv 2303.10115. <https://doi.org/10.48550/arXiv.2303.10115>

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. & Choi, Y. 2019. Defending against neural fake news. Advances in Neural Information Processing Systems 32, 9054–9065. <https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html>

Zhao, Z., Resnick, P. & Mei, Q. 2018. Fake news propagates differently from real news even at early stages of spreading. EPJ Data Science 7, 1–14. <https://doi.org/10.1140/epjds/s13688-018-0131-2>

Zhou, X. & Zafarani, R. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys 53 (5), 109. <https://doi.org/10.1145/3395046>

Zhou, X., Zafarani, R., Shu, K. & Liu, H. 2019. Fake news: Fundamental theories, detection methods, and opportunities. Proceedings of the 12th ACM International Conference on Web Search and Data Mining, 836–837. <https://doi.org/10.1145/3289600.3291382>

Zuckerman, M., DePaulo, B.M. & Rosenthal, R. 1981. Verbal and nonverbal communication of deception. Advances in Experimental Social Psychology 14, 1–59. [https://doi.org/10.1016/S0065-2601\(08\)60369-X](https://doi.org/10.1016/S0065-2601(08)60369-X)

Zurstiege, G. 2016. Propaganda. Teoksessa Heesen, J. (toim.) Handbuch Medien- und Informationsethik, 146–153. Stuttgart: J.B. Metzler.
https://doi.org/10.1007/978-3-476-05394-7_20

Zhao, J., Cao, N., Wen, Z., Song, Y., Lin, Y.-R. & Collins, C. 2014. #FluxFlow: Visual analysis of anomalous information spreading on social media. IEEE Transactions on Visualization and Computer Graphics 20 (12), 1773–1782.
<https://doi.org/10.1109/TVCG.2014.2346922>
