Avoimesti tieteestä!: Tomi Kinnunen – Sisäänrakennettua avoimuutta
Äänitteen kesto: 17 min

Litterointimerkinnät

Haastattelija:	Henriikka Mustajoki
Vastaaja:	Tomi Kinnunen
Puhuja	
sa-	sana jää kesken
(sana)	epävarmasti kuultu jakso puheessa tai epävarmasti tunnistettu puhuja
(-)	sana, josta ei ole saatu selvää
(--)	useampia sanoja, joista ei ole saatu selvää
, . ? :	kieliopin mukainen välimerkki tai alle 10 sekunnin tauko puheessa

---

Puhuja: Avoimesti tieteestä! – keskustelua tieteen, tutkimuksen ja oppimisen avoimuudesta ja vastuullisuudesta.

HENRIIKKA MUSTAJOKI: Tervetuloa Avoimesti tieteestä! -podcastin pariin! Minä olen Henriikka Mustajoki, ja tänään vieraanani on puheteknologian professori Tomi Kinnunen Itä-Suomen yliopistosta, tervetuloa!

TOMI KINNUNEN: Kiitos, kiitos!

HENRIIKKA MUSTAJOKI: Minkälaisen datan kanssa puheteknologian tutkimuksessa tehdään työtä?

TOMI KINNUNEN: Luonnollisesti nauhotetun puheen kanssa, elikkä se, mitä omassa tutkimusryhmässäni esimerkiksi tehdään, niin me ollaan keskitytty puhujantunnistukseen. Elikkä voi myöskin sanoa puheen biometriikka. Pyritään tunnistamaan henkilöitä puheäänen perusteella. Tää on niin kun menetelmäkehitystä, elikkä oon tietojenkäsittelytieteen taustalta tuleva, niin se mitä me käy käytännön tutkimustyössä tehdään, niin on näitä menetelmiä eli algoritmeja, ja kehitetään ne. Siinä on tietysti sitten nää datajoukot hyvin suuressa roolissa, eli puhutaan puhekorpuksista tai puheaineistoista tässä yhteydessä.

HENRIIKKA MUSTAJOKI: Sul on ilmeisesti aika isoja äänitetyn puheen aineistoja. Mistä ne kertyy? Keräättekö te niitä itse vai käytätteks te muiden keräämiä?

TOMI KINNUNEN: Itse asiassa hyvin monestakin eri lähteestä. Yksi lähde esimerkiks on tämmönen kansainväliset kilpailut, elikkä esimerkiks tämmönen

kuin NEST - National Institute of Standards and Technologies, tämmönen standardointijärjestö.

Niin he on 90-luvun puolivälistä jo tavallaan antaneet tutkijoiden vapaaseen käyttöön dataa. Sieltä on saanu ikään kuin ilmaiseksi datan tuloksia vastaan. Toisena on tilattu ihan tämmösiltä kielitieteen toimijoilta, esimerkiks LDC Linguistic Data Consortium -datajoukkoja. Mikä on nyt yleistyny ehkä viimeisen muutaman vuoden aikana, niin myöskin eri tutkimusryhmät sitten kerää esimerkiks ihan Internetistä dataa. Kun puhe on tavallaan tämmöstä ikään kuin julkista tietoa, esimerkiksi YouTubessa on ihan hirveesti keskustelupuhetta. Ja tällai automaattisesti tai puoliautomaattisesti tavallaan semmosia hyvin suuria aineistoja on eri tutkimusryhmät keränneet. Ja sit on tietysti ihan viimesenä myöskin ihan omia aineistojakin sillon tällön nauhotetaan, mutta tää viimeinen kategoria on oikeastaan aika harvinainen, koska jo me menetelmätietessä me ei semmosta mitään tiettyä ilmiötä ehkä tutkita niinkään, vaan pikemminkin sitä, että miten saahaan esimerkiksi nämä menetelmät algoritmit tunnistamaan paremmin henkilöltä. Ja siinä mielessä, niin ei oo oikeastaan aina edes relevantti kysymys, että onko se suomenkielistä materiaalia tai mitä tahansa, vaan se että onko sitä riittävästi.

HENRIIKKA MUSTAJOKI: Kuulostaa siltä, että tämä on ala, joka on tehnyt avointa tiedettä jo ennen kuin muut oikein edes puhuivat avoimesta tieteestä. Mikä on sun ajatusten mukaan tällä alalla aloittanut tämmösen avoimuuden toimintakulttuuriin?

TOMI KINNUNEN: Uskon, että ihan vilpittömästi siitä, et on tavallaan haluttu teknologian kehitystä edistää. Elikkä tällainen termi ku evaluation driven resarch – siis tämä tavallaan ajatus, että datajoukkolähtösesti tavallaan pyritään niitä menetelmiä sitte parantamaan. Ymmärrykseni mukaan siis tosiaan siellä ihan alkutaipaleella, kun nämä standardointijärjestössä nämä (-) [03:10] ja muut rupes tutkimaan, niin siellä oli ihan kielitieteen osaajia taustalla tavallaan. Ja sitten yhtä aikaa näitä menetelmätieteiden osaajia. Ja jotkut tutkimusongelmat on niin hankalia, että siihen on tavallaan kaikille parasta jos sen tutkimusongelman avaa koko maailman ratkottavaks. Silloin tää joka kerää dataa, niin se saa tietenkin näitä tutkimustuloksia käyttöönsä. Ja nämä tutkijat saa sitten tutkimusdatan käyttöönsä. Elikkä vähän niin kuin kaikille jotain.

HENRIIKKA MUSTAJOKI: Ilmeisesti tämän avoimen tieteen perusajatus on sinne niin syvälle rakennettu, et on vaikea kuvitella edes sitä tutkimusta ilman tällaista aineiston jakamista ja avoimuutta. Kuulenko mä sitä oikein?

TOMI KINNUNEN: Juuri näin, silloin joskus 80-90 -luvulla vielä alkupuolella, niin tälläkin alalla oli vielä niin, että jokainen tutkimusryhmä keräsi ehkä oman datansa, otti mikrofonin kouraan ja pyysi vapaaehtoisia tulemaan laboratorioon ja näin poispäin. Mut sitten jossakin vaiheessa huomattiin, että ei tää mitenkään skaalaudu. Jos halutaan tietää vaikka, että miten joku tämmönen puhebiometrinen

järjestelmä toimii, niin me tarvitaan siihen 1000 tai 5000 puhujaa, että voidaan ylipäätään vastata koko kysymyksen. Niin semmosta ei enää ole. Pikemminkin niin päin, että julkaisussa tutkijat oikeastaan odottaa näkevänsä tulokset sitten aina näillä tietyillä

datajoukoilla. Ja näitä datajoukkojen vaihtoehtoja on toki hyvin monta. On toki myöskin sellaisia tilanteita, joissa voi tulla, että on jonkinlainen spesifisempi tutkimushypoteesi, josta me tarvitaan kerätä omaa dataa. Mutta lähtökohtaisest, jos tavoite on parantaa tätä tunnistusteknologiaa, niin sillon on oikeastaan de facto -standardi, että käytetään näitä datajoukkoja, mitä kaikki muutkin käyttää. Ni silloin kaikki on tavallaan samalla kartalla. Tässä yhteydessä vielä on hyvä ehkä mainita tämmösen termin kuin syväoppiminen – deep learning, ja myöskin syvät neuroverkot – deep neural networks. Nää on siis tämmösiä laskennallisia malleja, mitkä koulutetaan suuresta datamäärästä, nää on yleistynyt hyvin voimakkaasti tavallaan oikeastaan kaikilla menetelmätieteiden alalla väittäsin, että ne on jollakin tavalla alkaa olla käytössä. Tämmöset keinotekoset hermoverkkomallit, niiden suosioon on tavallaan vaihdellu vuosikymmenestä toiseen. Mut nyt noin 15 vuotta sitten tuli menetelmällisiä läpimurtoja, ja tämä nimenomaan tapahtu puheteknologian osalla. Elikkä tavallaan siihen mennessä parhaimman puheentunnistimen päälle onnistuttiin merkittävä parannus saamaan – tavallaan proof of conseption. Tämän jälkeen pikkuhiljaa eri tutkijat sitten on lähtenyt soveltamaan näitä hyvin erilaisiin ongelmiin. Ja se on tällä hetkellä se sanotaanko vallitseva paradigma, mitä tulee tämmöseen hyvin kompleksisten ongelmien ratkaisuun, missä halutaan tietokoneen tekevän jotain vähän samantapaista kuin ihminen. Vaikka tunnistamaan henkilöitä tai puhesisältöä tai muuten.

HENRIIKKA MUSTAJOKI: Monella muulla alalla ollaan huolissaan sen datan kerryttämisestä saatavasta hyödystä. Mä mietin, että onko puheteknologian puolella sit olemassa jo hyvin vakiintuneet käytännöt esimerkiksi dataviittauksille tai jotenkin, että meritoituvatko ne, jotka keräävät hyvän aineistokokonaisuuden? Vai millä tavalla tämä jakamisesta ja sen aineiston kerryttämisen työstä saatava meriitti, niin miten se näkyy?

TOMI KINNUNEN: Se mikä on hirveästi yleistyny viime vuosina, ni on tämmöset niin kun kilpailut, elikkä eri tutkimusryhmät on ruvennu tavallaan avaamaan niitä datoja sillä tavalla, että se pistetään niin kun kaikille tämmönen avoin kilpailu – challenge pystyyn. Ja tässä saatte datan, ja teillä on niin ja niin paljon aikaa sitten prosessoida tätä mainontaa. Henkilökohtaisesti oon siis nähnyt tällaisen erittäin vahvana urabuusterina, elikkä se homma menee oikeestaan niin päin, että kun avaat semmosen kilpailun, pistät sen datajoukon saataville, ja jos siihen liittyy vaikka tämmönen paperi. Elikkä yleensä on kirjoitettu tämmönen tiivistelmäpaperi siitä datajoukosta, miten se on kerätty, määritelty ja yleisesti ottaen tämmösistä kilpailun tuloksista, niin kyllä se vaan tämmönen magiikka tapahtuu, että ne muut tutkijat sitten rupee viittamaan niihin. Tää on jossakin määrin oikeastaan ongelma tällä hetkellä. Eli esimerkiksi kun on julkaisufoorumia ja muuta, niin näitähän

oikeastaan millään tavalla huomioida. Mutta oma vankka mielipiteeni on se, että datajoukko on julkaisu siinä missä muutkin julkaisut. Tää on niinku tämmönen potentiaali, mikä minusta tavallaan missataan. Tai joka on hieman harmi, että se tavallaan. Se ei mitenkään sitten näy myöskään tämmösessä teknisessä tuloslaskennassa niin sanotusti sitten.

HENRIIKKA MUSTAJOKI: Jotenkin vois kuvitella, että siihen suuntaan ollaan nyt menossa. Puheteknologia on ehkä alana kehittyneempi kuin moni muu. Että alan sisällä tunnistetaan hyvien aineistojen kerryttäjät, ja teil on alan sisäiset viittaukset kunnossa, mutta tutkimusyhteisö laajemmin ei vielä osaa käsitellä tätä.

TOMI KINNUNEN: Tämmöstä avoimen lähdekoodin ajattelua on paljon. Ihmiset haluu pistää jakoon koodit ja muut (-) [07:36] tapaan, että no sitten kun tutkimus on valmis, niin pistetään jakoon koodit ja muut niin kun tavallaan koko maailma saataville mahdollisimman avoimesti. Ja tää on niinku minun mielestä tämmönen toimintakulttuuri, mikä on laajemminkin koneoppimisen ja tietojenkäsittelytieteen puolella muodostunu. Ja se on hyvin hyvä kulttuuri minun mielestä.

HENRIIKKA MUSTAJOKI: Onko teidän oman tutkimusalan sisällä. Tämmösen jakamiseen ja avoimuuteen liittyviä huolia tai haasteita?

TOMI KINNUNEN: En ihan suoraan nyt keksi, että mikä olis. Se on jotenkin niin selkärangassa tämä ajatus. Mutta mikäänhän ei oo koskaan täydellinen järjestelmä. Ja siis tällä hetkellä muun muassa koneoppimisen puolella… Siellä on myöskin huomattu muun muassa tämä, että vaikka se data laitetaan jakoon, niin ei se vielä takaa esimerkiksi toistettavuutta välttämättä. Aina voi toki olla eri mieltä, että onko minkälainen datajoukko tai kilpailutettu tutkimusongelma kiinnostava tai relevantti. Mut tietysti ihmiset äänestää niissäkin. Että ne ketkä sitten rupee käyttämään jotain dataa, niin kyllähän se siinä tavallaan nähään sitte sitä kautta.

HENRIIKKA MUSTAJOKI: Miten sä näkisit tän avoimuuden roolin lähitulevaisuudessa? Tuottaako se teille jotain uusia mahdollisuuksia, jos muut tutkimusalat esimerkiksi pääsee vauhtiin. Tai näätkö sellasia mahdollisuuksia, mitä sun omalla alalla on? Et tutkimus voisi vielä saada tästä jotain lisäpotkua ja vetovoimaa?

TOMI KINNUNEN: Kyl mä ihan aidosti uskon tavallaan, että riippumatta tieteenalasta, niin mitä avoimemmaksi nää datat ja tämmöset mallit, ja mitä sitten ikinä onkaan, niin mitä avoimempia ne on, niin sitä parempi. Koska silloin päästään aina siihen, sanotaanko saman keskustelupöydän ääreen, että kun kaikilla on sama ymmärrys. Muuten on tämmönen riski, että jos on vaikka julkaisussa kuvataan joku tämmönen mahollisesti hyvinkin kiinnostava skenaario, tai joku tämmönen mielenkiintonen data. Mut se on pelkästään tekstiä, että sie et oikeastaan, kukaan muka tutkiva ryhmä, ettei oo nähny sitä aineistoa tai dataa.

Mut siinä vaiheessa, jos muutkin pääsee kattelemaan sitä aineistoa. Niin siinä voi käydä niin, että jollakin tasolla saattaa olla uudenlainen idea. Mitä sitä koko ongelmaa kannattaisi miettiä ylipäätään, hankala sanoa. Että voisiko siinä sitten jotain vielä paremmin pystyy tekemään. Mutta ehkä se on tavallaan, vaikka nyt tossa mainihinkin, että on

tämmönen avoimen tieteen kulttuuri, niin se on edelleenkin myöskin hyvin nopeasti muuttuvaa. Elikkä ei oo mitää tämmöstä yhteistä blueprinttiä, että ensin tehdään tällä tavalla, sitten tällä tavalla. Se on pikemminkin alakohtaista, ja vielä alan sisällä olevien kentille ominaista, että miten, tyylin minne repositoryihin laitetaan dataa, ja mitä lisenssejä mitä ohjelmakoodeja. Tavallaan on ehkä se tutkimusyhteisö ite löytää sieltä ne käytännöt, mitä sitten muut tutkijat samalla alalla käyttää.

HENRIIKKA MUSTAJOKI: Me avoimen tieteen koordinaation puolella pohditaan kansallisia linjauksia. Miten tämmöset ulkoapäin tulevat, tämmöselle itseohjautuvalle nopeasti kehittyvälle… Niin mimmosen roolin sä näät tämmösellä kansainvälisillä ja kansallisilla isoilla linjauksilla siihen oman tutkimusalan käytäntöjen kanssa?

TOMI KINNUNEN: Kun ne on tavallaan niin yleisiä sitten. Niin niistä on aina toisinaan hankala sitten konkretiaa saada, että miten tämä sitten soveltuu. Se miten itte tavallaan nään asiat, niin hyvin pitkälti just esimerkiks nää (-) [10:36] - periaatteet, niin näitä kyllä nähdäkseni noudatetaan. Mut sit on tavallaan just, että miten siinä vois tehdä enempi, niin se ehkä just, että sitä konkretiaa vois olla enempi. Että kenties siinä pitäis miettiä, että olisiko vaikka jokaiselle tieteenalalle tai vaikka luonnontieteelle laajemmin, tai jotenkin tämmönen tutkija tai joku asiantuntija, mikä vois hieman tulla vastaan. Ja minun mielestä myöskin hieman toiseen suuntaan, elikkä välillä ainaki ittellä on tavallaan mielikuva, että kun näistä periaatteista puhutaan, niin siinä on jotenki sammonen - sanotaanko tämmönen top down -ajattelu, että tässä on nyt nää ohjeet, seuratkaa näitä ohjeita. Ja tavallaan tutkijan vastuulla on sitten ottaa selville, mitä se tarkottaa. Mut kyl se menee minun mielestä hieman molempiin suuntiin, elikkä tavallaan se mitä tutkijat jo tekee, niin kenties sieltä vois siirtää sitä konkreettisiks käyttöesimerkeiks muillekin tavallaan.

HENRIIKKA MUSTAJOKI: Ehdottomasti sen dialogin pitäisi toimia paremmin kuin mitä se toimii nyt. Ja mul on sit viel semmonen kysymys, en oo varma yhtään, mihinkä tämä keskustelua vie. Mutta koska ääni on henkilötieto, niin millä tavalla nämä henkilötietosuojakysymykset näkyy silloin kun teillä on valtavia korpuksia?

TOMI KINNUNEN: Tää on siis tärkee kysymys ilman muuta. Näissä korpuksessa, mitä yleensä käytetään, niin näissä pseudonymisoitu nämä. Elikkä ei oo mitään henkilöiden nimiä suoraan. On kuitenkin muuta metadataa. Voi olla tieto sukupuolesta tai iästä. Tässä on itse asiassa yksi tämmönen juttu, missä… tai hyvin

paljon kollegoiden kanssa ollut puhetta. Elikkä, että voidaanko henkilö tunnistaa puheesta. Ja se riippuu myöskin siitä, mitä teknologia käytetään tunnistamiseen. Mut se et tavallaan, miten tätä jotenki tuntuu, että se miten sitä on katottu… Niinku kategorisestihan se, että tämä on yksikäsitteisesti yksilöivä henkilötieto, joten se on sensitiivinen, joten se pitää anonymisoida tai poistaa nämä tiedostot. Tämä tuntuu olevan tämä sentimentti. Ja se yksinkertaisesti ei ole näin, elikkä siis se on aina niin kuin todennäköisyys, onko sama henkilö vai ei. Ja itte mitä aattelen, just tätä, että

kun käsitellään tämmösiä aineistoja, missä on vaikka, sanotaan nyt vaikka 1000 henkilöä. Ja jos meillä ei oo mitään identifioivia tunnuksia siellä mukana, niin jos haluttais tietää, että no voidaanko täältä nyt sitten tunnistaa joku, niin tää vaatis joka tapauksessa jonkinlaisen referenssiäänen. Sanotaanko jostakin muualta, että okei satun tietämään, että täältä löytyy tämä henkilö. Sitte pystyttäs vastaamaan kysymykseen, että löytykö tämä henkilö nyt sitten siitä tuhannen henkilön joukosta. Oliko se joku näistä tai joku tietty henkilö näistä. Mutta itte en nää sitä mitenkään realistisena, että kukaan lähtisi tällaista tekemään. Ja vaikka lähtiskin, niin edelleenkään sitä ei taattas, että sieltä löytyisi tämän vastine. Ihan poikkeuksen tekee kuitenkin, niin on sitten esimerkiks tämmönen just näitä YouTubesta kerättyjä korpuksia. Elikkä siinä on tämmönen meidän alalla paljon käytetty aineisto nimeltä VoxCeleb, nimensä mukaisesti siinä on niin kuin julkkisten ääniä kerätty YouTubesta. Ja siitä on se metadata saatavilla, elikkä kuka julkisuuden henkilö sitten on tässä äänitiedostossa saatavilla. Mutta tässäkin ainakin minulla on se mielipide, että se on siellä YouTubessa joka tapauksessa saatavilla, joten se ei oo välttämättä sen sensitiivisempää, että se kerätään sitte tämmöseen korpuksen metadataan.

HENRIIKKA MUSTAJOKI: Nää kysymyksethän on kaikilla tutkimusaloilla monimutkaisia, ja on ilmeisen selvää, että te ootte paljon sitä pohtineet ja puhuneet ja etsineet siihen yhdessä ratkaisua. Ja niin kuin tiedämme, GDPR jättää meille aika paljon tulkittavaa ja päätöksentekoa. Se mitä mä tästä saan, niin voimakkaasti on se, et sä edustat sellaista tutkimusalaa, jossa se mitä kaikkialla muualla puhutaan, siihen mihin halutaan mennä, niin on täällä jo nyt. Ja ois tietysti kauhean kiinnostavaa ehkä lähemminkin tutkia sitä, et miten se kulttuuri on muodostunu, ja miltä se näyttää. Se että jakaminen on perusoletus. Ehkä mä vielä kysyisin sulta, et miltä se tuntuu?

TOMI KINNUNEN: Sanotaanko näin, että se on hieman jopa tämmönen, minun mielestäni sosiaalinen kynnys. Elikkä tavote on aina edistää tiedettä ja tietämystä. Mut tavallaan siinä vaiheessa kun näät kuitenkin, että hetkinen tällä mitä mie teen on väliä. Elikkä ihmiset oikeasti viittaa siihen aineistoon. Niin kyllä sen oon huomannut, että ihan eri tavalla se fiilis on, mitä sillon joskus hyvin alkuaikoina. Kun tavallaan ei oikein tiennyt, onko tässä mitään järkeä, onko tässä meidän tutkimuksessa mitään pointtia. Mut sitten tavallaan kun näkee, että sielä sitten muutkin tutkijat on niitä datoja käyttäny ja näin poispäin, niin siin on tämmönen,

ittelle tulee tämmönen fiilis, että on jotenkin ollut mahdollisuus tämmösessä interaktiossa olla muihinkin kun omien kollegoiden kanssa vaan tälleen. On tullut keskustelua hyvin laajalti sitten tavallaan näistä datajoukoista ja kilpailuista, mitä me on esimerkiksi oltu tekemässä.

HENRIIKKA MUSTAJOKI: Saaks sä itelles semmoisen kokemuksen siitä, et tää on paras tapa tehdä tiedettä ja edistää tiedettä?

TOMI KINNUNEN: Ehkä siinä on se pointti, että kun tavallaan nämä ainakin oma näkemys on just semmonen, että nää mallinnusongelmat, mitä tällä hetkellä on käytössä. Elikkä just kun mainitsin esimerkiks näistä syvistä neuroverkoista, niin näissä on miljoonia parametreja ja (nämä luvat) [15:11] ja paljon laskentatehoa. Ja nää ongelmat, elikkä halutaan esimerkiksi tunnistaa henkilöitä hyvin meluisissa olosuhteissa tai kun henkilö puhuu eri tavalla ja näin poispäin, niin se on uskomattoman monimutkaista sitten mallinnuksen kannalta. Ja sillon tavallaan se on, kun en itse ainakaan usko, että tällä omalla alalla kukaan yksittäinen henkilö tai kukaan mikään yksittäinen tutkimusryhmä pystys ratkaisemaan näitä ongelmia. Et se vaatii sen dialogin, myöskin tämmöset datajoukot käyttöönsä. Kun näitä kilpailuita ja näitä benchmark settejä tavallaan on julkastu, niin näissä se logiikka oikeastaan on mennyt sieltä 90-luvun puolivälistä lähtien niin, että datajoukot kasvaa kasvamistaan koko ajan, ja muuttuu monimutkaisemmiksi monimutkaisemmiksi. Ajatus on siis se, että kun teknologia kehittyy, elikkä laskentateho on lisääntynyt. Ja nämä menetelmäpuolen kehitys on kehittynyt, niin näillä pystytään nyt paljon hankalampia tehtäviä ratkasemaan. Tavallaan sitten kun ihmiset on (saanu) okei, tämä ongelma on nyt ratkottu, niin sitten mietitäänki nexti leveli niin sanotusti, elikkä voisiko tätä vielä haastavammaksi tehdä. Aina löytyy niitä avoimia tutkimusongelmia. Ja siinä minun mielestä tää avoin tiede on ihan hyvin tärkeässä roolissa kyllä.

HENRIIKKA MUSTAJOKI: Kiitoksia Tommi Kinnunen, puheteknologian professori Itä-Suomen yliopistosta. Tää on ollu oikein tämmöst mieltä ylentävää keskustelua siitä, mitä avoimuus näyttää, sitten kun siitä on tullut tieteen tekemisen arkipäivää. Kiitos.

TOMI KINNUNEN: Paljon kiitoksia.