Informaatioteknologian tiedekunnan julkaisuja
No. 72/2018

Riku Nyrhinen, Eija Hiekka, Minna Silvennoinen, Sivi Talvensola
ja Karoliina Talvitie-Lamberg

Sosiaalinen hyvinvointi tekoälyn luonnollisen kielen testiesimerkkinä

JYVÄSKYLÄN YLIOPISTO

Informaatioteknologian tiedekunnan julkaisuja
No. 72/2018

Editor: Pekka Neittaanmäki
Covers: Petri Vähäkainu ja Matti Savonen

ISBN 978-951-39-7663-7 (verkkoj.)
ISSN 2323-5004
Jyväskylä 2018

Sosiaalinen hyvinvointi tekoälyn luonnollisen kielen testiesimerkkinä

Riku Nyrhinen
Eija Hiekka
Minna Silvennoinen
Sivi Talvensola
Karoliina Talvitie-Lamberg

Tämä julkaisu on toteutettu osana WHC-hanketta, johon Jyväskylän yliopisto on saanut rahoituksen Business-Finlandilta.

Business Finland-hanke: WHC

SISÄLLYSLUETTELO

1 Johdanto

Sosiaaliasiamiestoiminnan yhtenä tavoitteena on lisätä asiakkaiden osallisuutta ja vahvistaa heidän itsemääräämisoikeuttaan, jolloin erilaisen kokemustiedon kerääminen ja analysoiminen toiminnan kehittämiseksi on keskeistä. Kokemustiedon analysoinnin avulla asiakkaan ja kansalaisen oikeus tulla kuulluksi voisi tulevaisuudessa toteutua aiempaa paremmin. Tällä hetkellä asiamiestoiminnan tiedonkeruumenetelmät eivät ole Suomessa vakiintuneet ja ne eroavat alueellisesti. Vaikka sosiaalihuolto on määrällisesti merkittävä palveluntarjoaja, siitä kerätään verrattain vähän asiakaspalautetta. Tilanne on kuitenkin muuttumassa, sillä sosiaalihuollossa edetään kohti määrämuotoista kirjaamista, mikä avaa uusia mahdollisuuksia asiakastiedon tehokkaampaan käsittelyyn. Tässä osahankkeessa Keski-Suomen sosiaaliasiamiehen vuosina 2010-2017 asiakasyhteydenotoista ja palvelupalautteesta koottua tekstimuotoista aineistoa käytiin läpi tekstianalytiikan keinoin. Tavoitteena oli selvittää, voisiko uusilla menetelmillä saada tekstimuotoisesta aineistosta lisätietoa sosiaalisesta hyvinvoinnista, palveluiden kehittämiskohteista tai lakimuutosten seurauksista. Hankkeessa toteutetun kokeilun tulokset osoittavat, että tekstianalytiikka tarjoaa runsaasti menetelmiä sosiaaliasiamiehelle kertyvän palautetiedon käsittelyn automatisointiin sekä sosiaalihuollon palveluiden parantamiseen ja ongelmakohtien havaitsemiseen. Esimerkiksi arkkityyppien, trendien ja tiivistelmien avulla voidaan tehdä tulkintoja palvelun laadusta ja siinä usein esiintyvistä ongelmista. Tällä hetkellä suurin haaste tekstianalytiikan menetelmien laajemmalle hyödyntämiselle ovat aineiston tallentamisen hajanaisuus ja kirjaamiskäytänteiden puuttuminen.

2 Sosiaaliasiamiesraportointi vuosina 2010-2017, asiamiestoiminnan ja raportoinnin nykytilan kuvaus

Sosiaaliasiamiehen työn perustana on laki sosiaalihuollon asiakkaan asemasta ja oikeuksista (812/2000). Jokaisen kunnan on nimettävä sosiaaliasiamies, jonka tehtäviin kuuluu sekä julkinen että yksityinen sosiaalihuolto. Sosiaaliasiamies neuvoo asiakkaita asiakaslain soveltamiseen liittyvissä asioissa; avustaa muistutuksen tekemisessä; tiedottaa asiakkaan oikeuksista; toimii muutenkin asiakkaan oikeuksien edistämiseksi ja toteuttamiseksi sekä seuraa asiakkaiden oikeuksien ja aseman kehitystä kunnassa ja antaa siitä selvityksen vuosittain kunnanhallitukselle. Kaikki Keski-Suomen 23 kuntaa ovat päätyneet hankkimaan sosiaaliasiamiestoiminnan Keski-Suomen sosiaalialan osaamiskeskukselta, KOSKE:lta (http://koskeverkko.fi/).

Jotta sosiaaliasiamies pystyy seuraamaan sosiaalihuollon asiakkaiden oikeuksien ja aseman kehitystä kunnissa ja antamaan siitä vuosittain selvityksen, on ollut luotava seurantajärjestelmä. Sekä sosiaali- että potilasasiamiesten yhteydenottojen tilastointiin kehitettiin kymmenkunta vuotta sitten Sosiaali- ja terveysministeriön hankkeessa luokittelu, jonka useat asiamiehet ottivat käyttöön joko sellaisenaan tai paikallisiin tarpeisiin sovellettuna versiona. Sosiaaliasiamiehet toimivat kuntapohjaisesti ja terveydenhuollon potilasasiamiehet toimintayksikkökohtaisesti. Koska valtakunnallista ohjeistusta ei ole eikä tiedonmuodostusta edellytetä, lienee lähes yhtä monta asiakaspalautteen tilastointitapaa kuin asiamiestäkin. Kun ottaa huomioon sosiaali- ja terveydenhuollon integrointitavoitteen, hyvinvointitiedon tarpeen sekä pyrkimyksen asiakkaiden osallisuuden ja itsemääräämisoikeuden vahvistamiseen, paitsi sosiaalihuollon myös terveydenhuollon asiakaspalautetiedon nykyistä tehokkaampi, monipuolisempi ja ajantasaisempi hyödyntäminen olisi erittäin perusteltua.

Vuosiselvitys laaditaan asiakasyhteydenottotilastoinnin, asiamiehen muuten saamien tietojen ja havaintojen sekä niitä täydentävän kuntakyselyn pohjalta. Keski-Suomessa yhteydenotoista tilastoidaan verkkopohjaista tilastointi- ja kyselyohjelmaa käyttäen päivämäärä, yhteydenottaja, kehen tapahtuma kohdistuu, kunta, onko kyseessä julkinen vai yksityinen palvelu, yhteydenoton tehtäväalue, luonne ja syy sekä mihin toimenpiteisiin asiointi antoi aihetta. Lopuksi laaditaan lyhyt ***sanallinen kuvaus yhteydenoton sisällöstä***. Samasta asiasta voi tulla useita yhteydenottoja, mutta pääsääntöisesti sama asia tilastoidaan vain kerran vuodessa. Tällä tilastointitavalla yhteydenottoja on 275 000 asukkaan maakunnasta kertynyt vuosittain 400–660. Tilastoon ei merkitä asiakkaan eikä yhteydenottajan nimeä kuten ei muitakaan yksilöintitietoja. Tilastoaineistosta on suodatettu kunta- ja sosiaalipalvelukohtaisia tietoja tarvittavia ehtoja lisäämällä, mutta ***tekstin sisällöstä ei ole aiemmin pystytty saamaan tilastollista tietoa***.

Sosiaalihuollon suureen palveluvolyymiin nähden asiakaspalautetta on toistaiseksi kerätty varsin vähän ja silloinkin satunnaisesti ja työläästi. Kun **tavoitteena on lisätä asiakkaiden osallisuutta ja vahvistaa heidän itsemääräämisoikeuttaan, erilaisen kokemustiedon merkitys kasvaa entisestään.** Sosiaalisen hyvinvointitiedon tarve lisääntyy kaikilla tasoilla, niin paikallisesti, alueellisesti ja kansallisesti kuin myös kansainvälisesti. Nykyisin koottava kansallinen hyvinvointitieto kerätään oikeastaan pahoinvoinnin indikaattoreilla ja hyvinvointitietoa saa käänteisellä päättelyllä. Sama hyvinvointitiedon käänteisyys toteutuu myös sosiaaliasiamiehen tilastoissa: noin 30 % yhteydenotoista on tiedustelua ja muulloin syynä on tyytymättömyys sosiaalihuoltoon. Asiakkaan oikeuksien ja aseman kehitystä on seurattu ensisijaisesti tilastojen avulla. Sen sijaan **satojen sanallisten yhteydenottokuvausten läpikäyminen on jäänyt valitettavan pintapuoliseksi jopa vuositasolla, puhumattakaan useiden vuosien kehityksen tarkastelusta**. Asiakaspalaute on alihyödynnettyä, vaikka sitä onkin suodatettu yksittäisten kuntien tarpeisiin, sosiaalihuollon tehtäväalueiden koulutustilanteisiin ja kehittämishankkeiden suunnitteluun.

3 Raakadatan ja menetelmän kuvaus ja tekoälyn opettaminen sosiaaliasiamiehen kokoamalla rakenteettomalla datalla

Aineisto koostui sosiaaliasiamiehen kahdeksan vuoden aikana tilastointi- ja kyselyohjelmalla (Webropol) kerätyn aineiston avoimista kysymyksistä. Tekoälykokeilusta toivottiin vastauksia siihen, miten nykyistä laajempia tekstejä pystytään tarkastelemaan systemaattisesti. Yhteistyö aineiston analysoinnissa eteni sosiaaliasiamiehen kanssa säännöllisillä tapaamisilla, joissa rakennettiin yhteistä ymmärrystä aineistosta ja mahdollisuuksista löytää sopivia käyttötapoja tekstianalytiikalle seuraavien ongelmien/kysymysten selvittämisessä:

a) Löydetäänkö tekstianalytiikalla sosiaaliasiamiehelle vuosina 2010–2017 kertyneestä sosiaalihuollon asiakkaiden palvelupalautteesta uusia hyvinvointivajeista kertovia ilmiöitä?

b) Antaako aineisto viitteitä sosiaalista hyvinvointia edistävien palvelujen kehittämiskohteista?

c) Voiko aineistosta löytää yhteyksiä toteutuneisiin lakimuutoksiin tai perusteita valmisteilla olevien lakien ja uudistusten arviointiin?

3.1. Menetelmän kuvaus

Aineisto koostuu litteroiduista yhteydenotoista vuosilta 2010-2017 sekä yhteydenottojen ominaisuuksia kuvaavista kategorisoivista muuttujista, joita ovat muun muassa yhteydenoton päivämäärä, yhteydenottajan tyyppi (*Asiakas, Edustaja/muu, Henkilöstö*), yhteydenottajan kieli, asiatapahtuman kohde, yhteydenottajan kotikunta, yhteydenoton tehtäväalue ja luonne sekä käsittelyn lopputulos. Yhteydenottoja käsiteltiin yksittäisinä dokumentteina, jotka jaettiin kahteen osaan: litteroituun yhteydenoton sisältöön ja oleellisiksi koettuihin kategoriamuuttujiin. Tarkastelun ulkopuolelle jätettiin yhteydenottajan sukupuoli sekä yhteydenottajan kieltä koskevat kaksi muuttujaa.

Yhteydenoton sisältö ajettiin TurkuNLP Group'n *Finnish Dependency Parser* -työkalun (TurkuNLP Group) komentoketjun läpi saaden tulosteena CoNLL-U-standardin mukaiset versiot alkuperäisistä yhteydenotoista. CoNLL-U on standardisoitu tapa kuvata kielen sanojen ominaisuuksia sekä sanojen välisiä syntaktisia relaatioita. Kirjoitettua kieltä CoNLL-U-standardiin muuntavat ohjelmat tuottavat CoNLL-U:n määritelmän mukaisen tulosteen riippumatta kielestä; tulosteessa esiintyvät termit ja tunnukset löytyvät standardin tietokannoista. CoNLL-U-muotoisessa tiedostossa jokainen lähdeaineiston sana on jaettu omalle rivilleen. Jokaista sanaa varten on kymmenen solua, jotka on eroteltu toisistaan tabulaattorilla. Solut sisältävät muun muassa sanan lemman, taipumistavan (esimerkiksi *sijamuoto, aikamuoto*), sanaluokan ja syntaktisen relaation muihin sanoihin nähden (esimerkiksi *subjekti, objekti*) (Universal Dependencies (CoNLL-U)). Kaikki aineistoon kohdistetut menetelmät hyödyntävät jollain tavalla CoNLL-U-standardin mukaisia soluja.

3.2. Case-esimerkkejä tekstianalytiikkakokeiluista

Tekstimuotoisesta asiakasyhteydenottojen tiedostosta haettiin muun muassa sekä tyytymättömyyteen, että tiedusteluun liittyviä sosiaalihuollon palvelujen frekvenssejä ja kategorioita. Aineistoa tarkasteltiin esimerkiksi sen kautta, mitkä sanat liittyivät toisiinsa ja kuinka lähekkäin nämä sanat toisiaan esiintyivät. Asiakaskertomuksista pyrittiin muodostamaan automaattisia tiivistelmiä ja tarkasteltiin sitä, kuinka hyvin tekstianalytiikan työkalut näitä pystyivät tuottamaan. Teemallisesti tarkastelun kohteena olivat aluksi lastensuojelun, huoltoriitojen ja lähisuhdeväkivallan mainintoja sisältävät asiointimerkinnät. Yhteydenotoista pyrittiin myös muodostamaan tyypillisiä esimerkkitapauksia. Sen jälkeen analysoitiin vielä muutamia muita sosiaalihuollon osa-alueita päätyen sosiaalihuollon ja mielenterveyspalvelujen sekä päihdehuollon yhtymäkohtiin etsien muun muassa tyyppiasiakkaita eri yhteydenottosyiden takaa. Seuraavissa kappaleissa kerrotaan tarkemmin kokeilujen toteutuksesta esimerkkien avulla.

3.2.1. Aineiston hahmottaminen tekstianalytiikan kautta

Aineistoa lähdettiin hahmottelemaan ***yhteydenottojen kategorioiden kautta*** esimerkiksi laskemalla yleisimmät sanat, jotka esiintyivät "Vammaispalvelut"-kategoriassa. ***Sanafrekvenssien avulla on mahdollista luoda parempi kuva, mistä yhteydenotoissa tyypillisesti puhutaan***. Oletuksena on, että kategoriassa yleisimmin esiintyvät sanat liittyvät yhteydenottojen yleisimpiin puheenaiheisiin. Keskeisten puheenaiheiden avulla saadaan selville, mitkä ongelmat, tilanteet, jne. aiheuttavat eniten yhteydenottoja, mikä taas voi toimia hyödyllisenä apuvälineenä palveluiden parantamiseksi tai ongelmien syntymisen ehkäisyyn yhteiskunnallisella tasolla. Lähdeaineistolle laskettiin sekä raakafrekvenssit että TF-IDF-painotetut frekvenssit.

Sanan raakafrekvenssi kertoo, kuinka monta kertaa se esiintyy dokumentissa tai kokoelmassa dokumentteja. Sitä ei normalisoida eikä painoteta mitenkään, joten yleisimmät sanat -- riippumatta ovatko ne merkitykseltään keskeisiä -- nousevat frekvenssilistan kärkeen. Suomen kielessä on monia sanoja, jotka esiintyvät tekstissä kuin tekstissä riippumatta aiheesta. Tällaisia ovat muun muassa *olla, että, koska, ja, kun, sekä, ei, mutta*. Suodattamalla tuloksista nämä yleisiksi tunnetut, vähän merkitystä kantavat sanat saadaan paremmin aihetta kuvaava lista. Frekvenssien normalisointi ottaa huomioon sanan esiintymismäärän lisäksi sanojen kokonaismäärän dokumentissa, johon sana kuuluu, ja sitä kautta tekee eri pituisten dokumenttien vertailemisen helpommaksi.

TF-IDF painotus huomioi sanan dokumenttikohtaisen esiintymismäärän lisäksi sen yleisyyden koko dokumenttikannassa. TF-IDF-painotukselle on lukuisia laskentatapoja, mutta tässä tutkimuksessa käytettiin versiota, jossa sanan yleisyys dokumenttikannassa esitetään dokumenttien (eli yhteydenottojen) kokonaismäärän ja sanan sisältävien dokumenttien osamäärän luonnollisena logaritmina. Tämä tarkoittaa, että sanat, jotka esiintyvät kaikissa yhteydenotoissa (kuten esimerkiksi "olla"-verbi), painottuvat tärkeysjärjestyksessä hyvin alas. Tällä yksinkertaisella metodilla saatiin korostettua eri kategorioissa esiintyviä puheenaihetta indikoivia sanoja paremmin kuin

raakafrekvensseillä tai normalisoiduilla frekvensseillä. Using TF-IDF to Determine Word Relevance in Document Queries

Yksittäisten sanojen lisäksi kategorioita hahmoteltiin lingvististen relaatioiden kautta esimerkiksi selvittämällä yleisimmät subjekti-predikaatti-objekti-tripletit. Menetelmällä yritettiin selvittää, kuka tekee mitäkin ja kenelle yhteydenotoissa. Toinen kokeiltu relaatioihin liittyvä menetelmä koski substantiivien attribuutteja eli sanoja, jotka tarkentavat, millainen substantiivi on tai mitä ominaisuuksia siihen liittyy. Esimerkiksi "punainen omena" kertoo, että "omena" on "punainen", jossa "omena" on substantiivi ja "punainen" sen välitön määrite. Tällöin aineistosta nousi esille muun muassa "huono kohtelu" -yhdistelmä ("huono" on attribuutti), mikä toimii loistavana hakuterminä ja mittarina päätelmiä tehdessä.

Edellä mainitut metodit eivät kuitenkaan suoraan paljasta päällekkäisyyksiä yhteydenottojen aiheissa. Jos yhteydenotossa puhutaan kahdesta eri asiasta, frekvenssien ja relaatioiden avulla on hankala päätellä, mitkä aiheet esiintyvät yleensä yhdessä. Tätä ongelmaa selvitettiin hyödyntäen word2vec-tekniikoita, jotka muodostavat teksteissä esiintyvistä sanoista listan, jonka alkiot ovat moniulotteisia vektoreita. Listan vektorit ovat suoraan yhteydessä yksittäisiin sanoihin ja niiden järjestys listan sisällä kertoo, ***kuinka lähellä sanat keskimäärin esiintyvät tekstissä***. Lähekkäisyyden määritelmä ja "keskimääräisyys" riippuvat käytetystä w2v-tekniikasta. Kun frekvenssimenetelmillä löydetyt aihesanoja verrattiin listan indekseihin, saatiin selville, mitkä aihesanat esiintyvät yhdessä kaikkein useimmiten. (Word2Vec)

3.2.2. Asiakaskertomusten automaattiset tiivistelmät

Tekstin automaattinen tiivistäminen on prosessi, jossa tyypillisesti kookkaasta, rönsyilevästä ja informaatioharvasta lähdeaineistosta (joka voi koostua yhdestä tai useammasta dokumentista) muodostetaan lyhyt ja ytimekäs lyhennelmä. Lyhennelmä voi sisältää yleiset pääkohdat tai keskittyä tiettyyn aiheeseen. Sosiaaliasiamiehen aineiston tapauksessa yritettiin kategorioittain antaa ***paras kuva siitä, millaista informaatiota yhteydenotossa välitettiin: kuka ottaa yhteyttä, kenen puolesta, millä asialla ja mihin tulokseen päädyttiin.***

Tiivistämisessä hyödynnettiin aikaisemmin esiteltyjä tapoja aihesanojen ja relaatioiden havaitsemiselle. Aihesanat indikoivat, mikä yhteydenotossa on olennaista. Sanojen väliset syntaktiset relaatiot taas kertovat, miten paljon tekstiä voidaan typistää. Lauseesta "punainen omena putosi puusta" voidaan huoletta poistaa attribuutti "punainen", ellei sanalla koeta olevan suurta merkitystä. Jäljelle jäävä "omena putosi puusta" -lause on kuvaava, mutta jos omena putoilee milloin mistäkin, voidaan myös sana "puu" jättää pois lopullisesta tiivistelmästä. Tällöin jäljelle jää lauseke "omena putosi", joka tunnetaan myös lauseen *neksuksena* (subjekti-predikaatti-kaksikko).

Tiivistelmien lopputulokset olivat yleensä varsin koherentteja kokonaisuuksia, mutta yhteydenottojen litteroinnin informaatiotiheys teki tuloksista vähemmän luotettavia. Tiivistelmäalgoritmi saattoi pudottaa lopullisesta lyhennelmästä tärkeitäkin osia pois, sillä algoritmin pyrkimyksenä oli tuottaa kaikissa tapauksissa alkuperäistä tekstiä lyhyempi versio. Jos alkuperäinen teksti oli jo lyhyt ja ytimekäs, oleellista informaatiota katosi väistämättä, mikä ei tee valvomattomasta tiivistämisestä toimivaa reaalimaailman työkalua.

3.2.3. Sosiaaliasiamiehen työprosessin automaattinen hahmottaminen – “polun kulku”

Työprosessin hahmottaminen noudatti pitkälti jo esiteltyjä menetelmiä. ***Tutkimusta tarkennettiin sosiaalitointa koskevilla erityistiedoilla***, kuten esimerkiksi lapsen huostaanottoa koskevilla termeillä ja tyypillisellä vaiheittaisuudella. Pyrittiin esimerkiksi selvittämään, millaisen matkan lapsi kulkee "systeemissä" ennen lopullista sijoituspaikkaa. Yritettiin myös selvittää, millaiset tekijät ovat mukana huostaanottoprosessissa (koulu, koti, vanhemmat, naapurit) tutkimalla lastensuojelutapauksissa usein esiintyviä sanoja ja niiden välisiä relaatioita ("opettaja teki ilmoituksen", "äidillä päihdeongelma"). Parhaissa tapauksissa pystyttiin muodostamaan kulkukaavioita, jotka tiivistävät käsittelyprosessin askeleisiin, kuten: “huume” -> “huostaanotto” -> “sijaisperhe”. Esimerkin mukaisten ***ketjujen muodostaminen vaatii kuitenkin runsaasti lähdeaineistoa*** jotta sama kuvio toistuu tarpeeksi monta kertaa ja ylittää näin kynnysarvon, kiinnittääkseen algoritmin huomion.

3.2.4. Tekstimuotoinen tilastollinen luokittelu ja päättely

Edellä mainituilla metodeilla selvitettiin, ***mihin yhteydenottajat ovat tyytymättömiä***. Sen jälkeen opittua sovellettiin tilastollisesti kategorioittain ja pääteltiin, missä yhteydenottokategorioissa ja missä asioissa esiintyy eniten tyytymättömyyttä. Esimerkiksi vanhuspalveluja koskevissa yhteydenotoissa omaiset tunsivat usein tyytymättömyyttä hoitokodissa asuvien vanhusten epäreilua kohtelua kohtaan. Vammaispalveluissa itsemääräämisoikeus nousi esiin. Tukia koskevat yhteydenotot toivat esiin ihmisten tyytymättömyyden päätösten koetun mielivaltaisuuden tai epäreilun kohtelun tai joustamattomuuden suhteen.

3.2.5. Arkkityypit / trendit

Erilaisille yhteydenottojen kategorioille yritettiin hakea tyypillisiä esimerkkitapauksia. Nämä arkkityypeiksi tai trendeiksi kutsutut tulokset tiivistettiin kokoelmaan sanoja tai ilmaisuja, jotka esiintyvät samassa kontekstissa usein saman yhteydenottokategorian sisällä. ***Arkkityypit pyrkivät kuvaamaan yleisesti, millaisista asioista kategoriassa puhutaan***. Esimerkkinä vanhuspalvelut, joista nousi esiin trendi: [sairaus] vanhalla ihmisellä; ollut sairaalassa; kotiutettu; omaiset kysyvät. Toisiinsa liittyvät sanat tai ilmaisut on eroteltu puolipisteellä. Ne esiintyvät usein samassa kontekstissa, mistä voidaan päätellä, että vanhuspalveluihin otetaan usein yhteyttä koskien sairasta vanhaa ihmistä, joka on kotiutettu sairaalasta ja jonka omaiset tarvitsevat lisää ohjeistusta.

Arkkityypit muodostettiin kokoamalla kategorian yhteydenotoista mahdollisimman usein esiintyviä yksittäisiä sanoja tai sanojen yhdistelmiä. Sanoja yhdistävä tekijä saattoi olla peräkkäisyys, syntaktinen relaatio, yhteiset naapurisanat, aihe tai yhteinen konteksti. Toisiinsa liittyvät sanat tai sanayhdistelmät järjestettiin niiden yleisyyden mukaan ja koottiin arkkityypiksi, jos yleisyys ylitti ennalta määritetyn kynnysarvon.

4. Pohdinta - millaista tukea tekoälyn analytiikka voi tarjota sosiaaliasiamiehen työhön tulevaisuudessa?

Sosiaaliala ja tekoäly voi tuntua yllättävältä ja ristiriitaisiakin ajatuksia herättävältä yhdistelmältä. Jo lyhyt sosiaaliasiamiehen yhteydenottoaineistoon kohdistunut tekoälyn soveltamiskokeilu osoitti kuitenkin lukuisia potentiaalisia tekstianalytiikan soveltamispaikkoja, kuin myös uusia kehittämiskohteita.

Erityisen mielenkiintoiseksi kokeilun teki uusi tapa tarkastella usealta vuodelta Keski-Suomesta kertynyttä asiakaspalautetta. Aineiston analysoinnissa edettiin kokeillen vaiheesta toiseen. Yhden vaiheen löydökset herättivät uusia kysymyksiä, joihin etsittiin vastauksia – välillä siinä onnistumatta. Esimerkiksi aineistosta etsittiin merkittäväksi hyvinvointivajeeksi tunnistettuun yksinäisyyteen liittyvää tematiikkaa, mutta se ei tullut tässä aineistossa esiin. Toisaalta ***aineistosta löytyi eri vuosien tekstidataa yhdistämällä tyypillisiä esimerkkitapauksia,*** joita ei olla tilastollisiin frekvensseihin pohjaavassa analysointitavassa pystytty tähän mennessä nostamaan esiin. ***Palveluiden kehittämisen näkökulmasta hyödyllisinä nähtiin, että aineistosta saatiin koottua esimerkiksi syitä asiakkaiden tyytymättömyydelle, yhteydenottojen keskeisiä aiheita tai tehtyä automaattisia tiivistelmiä yhteydenotoista.*** Liitteestä raportin lopusta löytyy lisäksi muutamia esimerkkihakuja ja niiden tuloksia.

Tekstianalytiikka tarjoaa runsaasti menetelmiä sosiaaliasiamiehen työn automatisointiin sekä palveluiden parantamiseen ja ongelmakohtien havaitsemiseen. Aineiston analysointi pohjautui litteroituihin yhteydenottoihin, mikä tarkoittaa, että tulokset olivat riippuvaisia siitä, miten asia oli kirjattu aineistoon. Jos tulevaisuudessa pyritään integroimaan tekstianalytiikka- ja tekoälymenetelmiä mukaan prosessiin, olisi hyvä, jos litteroinnille olisi tarkemmat säännöt tai standardi. Rakenteellista tekstiä on helpompi analysoida kuin täysin rakenteetonta, sillä rakenteellisen tekstin analysoimiselle voi asettaa kokoelman empiirisesti havaittuja aineistolle ja tarkoitukseen sopivia sääntöjä. Käytetyt metodit kohtasivat ongelmia erisnimien ja sosiaalihuollon ammattitermien kanssa. Erisnimet, kuten hoitopaikkojen nimet, voisi korvata aineistossa [hoitopaikka]-tunnuksella, jolloin analytiikkaohjelmisto osaisi tämän säännön pohjalta käsitellä [hoitopaikka]-esiintymät järkevämmin. Monet ammattitermit, kuten “pd” ja “mt”, löytyivät tekstistä sellaisenaan ja aiheuttivat tulkinnallisia ongelmia niin koneelle kuin ihmisellekin. Vasta termien tarkoituksen selvittämisen ja sääntöpohjan luomisen jälkeen termeihin saatiin tolkkua.

Arkkityyppien, trendien ja tiivistelmien avulla voidaan tehdä tulkintoja palvelun laadusta ja siinä usein esiintyvistä ongelmista. Esimerkkinä kappaleessa 3.2.5 esitelty vanhuspalveluja koskeva tapaus, jossa sairas vanhus on ollut sairaalassa, tullut kotiutetuksi ja omaiset eivät tiedä, miten toimia, vaan kaipaavat lisää ohjeistusta. Tällöin voisi tehdä päätelmän, että omaisille suunnattua ohjeistusta, mitä ikinä se tarkoittaakaan, kuuluisi lisätä tapauksissa, joissa vanhempi, isovanhempi tai muu iäkäs sukulainen on sairas tai muuten tuen ja avun tarpeessa.

Sosiaaliasiamiehen aineistolla toteutetut tekstianalytiikan kokeilut tuottivat myös ***paljon ideoita menetelmän jatkojalostamiseksi koskemaan muita sosiaalityön ja sosiaalihuollon aineistoja***. Kyseinen aineisto on ns. välittynyttä asiakkaan palautetta, sillä sen on sosiaaliasiamies kirjannut tilastomerkinnäksi mm. asiakkaiden sähköposteista ja puheluista saadusta tiedosta. Siksi ***entistä parempi hyöty asiakaspalautteesta saataisiin, jos asiakkaiden ja heidän edustajiensa luvalla voitaisiin jatkossa tarkastella suoria palautteita***, esimerkiksi sähköposteja, nauhoitettuja puheluja ja keskusteluja. Tuolloin päästäisiin käsittelemään autenttista aineistoa ja subjektiivisia kokemuksia sen sijaan, että työntekijä niitä poimii, suodattaa, tiivistää ja välittää, ja samalla väistämättä muokkaa.

Sosiaalihuollon asiakaspalautteesta olisi saatavissa nykyistä enemmän tietoa kansalaisten hyvinvointivajeista, palvelutarpeista ja niihin vastaamisasteesta. Täsmällisemmällä tiedolla tuotetuissa palveluissa toteutuu paremmin laatu, riittävyys, oikea-aikaisuus, vaikuttavuus ja myös taloudellisuus. Luonnollista kieltä analysoimalla kohtelupalaute saataisiin esiin elävämpänä ja puhuttelevampana. Erityisen hyödyllistä olisi saada tietää, mitä tietoa kansalaiset sosiaalihuollosta tarvitsevat ja kysyvät. Tuolloin olisi mahdollista kehittää tehokkaampia itseapumenettelyjä ja suorempia palveluprosesseja.

Sosiaalihuollossa edetään kohti määrämuotoista kirjaamista ja se avaa uusia näkymiä asiakasrekisteritiedon tehokkaampaan käsittelyyn, olipa kyse dokumenttien laadusta ja oikeellisuudesta, tilastoista, palvelutarpeen aikaisemmasta tunnistamisesta tai vaikuttavuuden arvioinnista. Sosiaalihuollossa on runsaasti mahdollisuuksia tekoälyn hyödyntämiseen siten, että teknologia palvelee sosiaalista hyvinvointia ja resurssit riittävät paremmin yksilölliseen inhimilliseen vuorovaikutukseen perustuvaan asiakastyöhön.

Huomattavaa on, että tekoälypohjaisen tekstianalytiikan tuottamien tulosten tulkitsemiseen tarvitaan kuitenkin aina alan asiantuntija arvioimaan niiden pätevyyttä. ***Tekstianalytiikka toimii siis pikemminkin asiantuntijoiden tukena tekstipohjaisen informaatiotulvan ymmärtämiseksi ja hallitsemiseksi.*** Näin esimerkiksi sosiaalitoimen alueella tekoälypohjaista tekstianalytiikkaa voisi hyödyntää samaan tapaan kuin tekoälypohjaisia lääkärien työn tueksi kehitettyjä diagnosointityökaluja, automatisoituna tukena syy-seuraussuhteiden ymmärtämisessä, lopullisen ymmärryksen tuottaminen ja konkreettiset toimet ovat kuitenkin ihmistoimijan vastuulla.

5. Tekstianalytiikan tulevaisuuden kuvat

Suomen kieli on pieni marginaalikieli, joka kuuluu suomalais-ugrilaisiin kieliin, mikä tekee esimerkiksi englanninkielisten tekstianalytiikan teorioiden, työkalujen ja tietokantojen suoran hyödyntämisen suomen kielen analysoinnissa vaikeaksi tai jopa mahdottomaksi kielten erilaisuuden vuoksi. Suomenkielinen tekstianalytiikka onkin jäänyt jälkeen isompien kielien tahdista, mutta nykyisillä työkaluilla ja resursseilla saa silti aikaan kelvollisia tuloksia. Tulevaisuudessa on odotettavaa, että hyödynnettävien resurssien määrä ja suomenkielisen tekstianalytiikan tutkijakanta tulee kasvamaan, kuten myös tekoälyjen koulutuksessa käytettävät tekstimassat. Internetresurssit, kuten sosiaalinen media, ja avoimet kannat, kuten FinnWordNet, ovat loistavia esimerkkejä lähes vapaasti käytettävistä aineistoista tekstianalytiikan kehittämiseen. Suomenkielinen tekstianalytiikka on kovaa vauhtia matkalla sinne, missä isommat kielet jo ovat.

Nykyinen tutkimus on keskittynyt pääasiassa tekstianalytiikan mahdollistamiseen eli esimerkiksi sanojen morfologian päättelemiseen koulutusaineiston pohjalta. Jotta tekoäly oppisi tunnistamaan suomen kielen lukuisat taivutusmuodot, on ihmisen käytävä opetusaineisto läpi ja annotoitava siihen oleellinen sisältö, jonka perusteella tekoäly tekee päätelmiä sille syötetystä uudesta materiaalista. Prosessi on kumuloituva, joten kaikki lisäaineisto, jonka tekoäly saa käyttöönsä, vie suomenkielistä tekstianalytiikkaa eteenpäin. Vaikka työkalut, kuten TurkuNLP Group'n *Finnish dependency parser,* ovat vielä puutteellisia, ne eivät ole käyttökelvottomia. Työkalujen suurimmat ongelmat liittyvät erisnimien, puhekielisyyden ja kirjoitusvirheiden tulkitsemiseen. Työkalujen koulutusmateriaali ei kata väärin kirjoitettuja sanoja, puhekielisyyttä tai slangi-ilmaisuja, joten sellaisten ilmaisujen käsitteleminen tuottaa automaattisesti virheellisiä tuloksia. Kirjoitusvirheet voidaan yrittää korjata vertaamalla kirjoitusvirheen sisältämää sanaa kokoelmaan sitä muistuttavia sanoja ja yrittää valita sopivin kirjoitusasu tilalle. Sama ei kuitenkaan päde puhekielisyyteen, joka yleensä rikkoo kielen sääntöjä enemmän kuin satunnainen kirjoitusvirhe. Erisnimissä tulkintavirheitä aiheuttavat etenkin ulkomaiset nimet ja nimiä sisältävät nimet: onko “Juhani” “minun Juha” vai erisnimi “Juhani”?

Suurin haaste tulevaisuudessa on lingvistisen ymmärryksen tuottaminen koneellisesti. Robotti, jonka sisällä toimiva tekoäly on kuin papukaija, ei ole erityisen joustava tai hyödyllinen ihmisten keskuuteen päästettynä. Tulevaisuuden palvelurobottien on kyettävä tekemään tulkintoja niille puhutusta kielestä lähes yhtä hyvin -- tai ellei jopa paremmin -- kuin ihminen vastaavassa tilanteessa. Jo pelkästään homofonisten (samalta kuulostavien) sanojen eri merkityksien tulkitseminen on koneille vielä haasteellista puhumattakaan aidonkuuloisen tekstin tai puheen tuottamisesta.

LIITTEET

Liite 1: Esimerkkejä tuloksista

Toistuvat narratiivit ja arkkityypit:

1. [isä/äiti] ottaa yhteyttä tiedustellakseen; huosta, jälkihuolto
2. Lastensuojelu-kategoriassa: [isä/äiti] muodostunut päihdeongelma
3. Sijaisperheestä ls-ilmoitus; [lapsi] siirretty
4. huoltajuuskiista pitkittynyt
5. synnytyksen jälkeinen masennus (vuosina 2015, 2016)

Eri yhteydenottajien osuus yhteydenotoissa:

Kaikki yhteydenotot [Asiakas, Edustaja/muu, Henkilöstö] Edustajat %

Kaikki kategoriat					
Vuosi	Kaikki yhteydenotot	Asiakas	Edustaja/muu	Henkilöstö	Edustajat %
2010	517	359	126	31	30,4 %
2011	592	421	135	35	28,7 %
2012	621	422	170	28	31,9 %
2013	664	438	178	47	33,9 %
2014	577	408	138	30	29,1 %
2015	595	412	145	37	30,6 %
2016	560	381	142	36	31,8 %

Lastensuojelu ja perheasiat					
Vuosi	Kaikki yhteydenotot	Asiakas	Edustaja/ muu	Henkilöstö	Edustajat %
2010	127	98	20	8	22,2
2011	144	113	24	6	20,9
2012	170	135	28	6	20,1
2013	180	125	37	17	30,1
2014	148	119	23	5	19,0
2015	133	100	20	12	24,2
2016	113	79	26	7	29,4

Lastensuojelu					
Vuosi	Kaikki yhteydenotot	Asiakas	Edustaja/ muu	Henkilöstö	Edustajat %
2010	98	71	18	8	26,8
2011	105	80	19	5	23,0
2012	135	104	24	6	22,3
2013	145	96	32	16	33,3
2014	95	69	20	5	26,5
2015	95	66	18	10	29,7
2016	89	59	22	7	32,9

Vammaispalvelut					
Vuosi	Kaikki yhteydenotot	Asiakas	Edustaja/ muu	Henkilöstö	Edustajat %
2010	40	25	11	3	35,8
2011	54	29	19	5	45,2
2012	65	33	27	4	48,4
2013	81	51	27	2	36,2
2014	55	27	25	2	50,0
2015	67	51	13	2	22,7
2016	66	38	24	3	41,5

Palvelukomponentit / prosessin kulku:

1. asiakas tyytymätön -> päätös -> hallinto-oikeus
2. oikaisuvaatimus -> hallinto-oikeus
3. lääkärintodistus -> vammaispalvelut
4. oikeudenkäynti -> lääkärintodistus

Trendit ja arkkityypit (ikääntyneiden palvelut -kategoria; tyytymättömyys-alakategoria):

1. ei yöhoitoa; ei ulkoiluteta
2. hammashoito & lääkkeet
3. kodin realisointi
4. puoliso hoitanut
5. henkilökunta + vähentäminen / henkilökunta + vähäinen + määrä / henkilökunta + lisää
6. hoitopaikka + vuokra

Informaatioteknologian tiedekunnan julkaisuja
No. 72/2018

ISBN 978-951-39-7663-7 (verkkoj.)
ISSN 2323-5004

JYVÄSKYLÄN YLIOPISTO