Tekstianalyysin hyödyntäminen tutkittaessa terveydenhuollon digipalveluihin liittyvää kansalaismielipidettä

Arne Bäcklund

Haaga-Helia ammattikorkeakoulu
Tradenomi
Opinnäytetyö
2023

Tiivistelmä

Tekijä(t)
Arne Bäcklund

Tutkinto
Tradenomi

Raportin/Opinnäytetyön nimi
Tekstianalyysin hyödyntäminen tutkittaessa terveydenhuollon digipalveluihin liittyvää kansalaismielipidettä

Sivu- ja liitesivumäärä
36 + 2

Terveydenhuollon digitaalisten palvelujen käyttö Suomessa on lisääntynyt nopeasti. Palveluiden suuren suosion vuoksi on tärkeää ymmärtää, minkälaisia mielipiteitä palveluihin liittyy suomalaisten keskuudessa. Lisäksi palveluita pyritään kehittämään asiakaslähtöisesti, joka edellyttää ymmärrystä palveluiden käyttäjien kokemuksista ja tarpeista.

Palveluihin liittyvän kansalaismielipiteen kartoittamiseksi tarvitaan tiedonlähde, josta voidaan kerätä palveluihin liittyviä mielipiteitä riittävän suurissa määrin. Nykyään on alettu etenevissä määrin ymmärtää sosiaalisen median datamassojen potentiaali tietolähteenä, kun halutaan tutkia erilaisia ilmiöitä. Sosiaalisesta mediasta löytyy lähes kaikkia mahdollisia asioita käsitteleviä tekstejä ja suurin osa suomalaisista käyttää sosiaalista mediaa. Twitter eli nykyinen viestipalvelu X on erityisen houkutteleva tiedonlähde, sillä Twitterissä julkaistavat viestit eli twiitit ovat oletusarvoisesti julkisia.

Tämän tutkimuksen tarkoituksena oli tutkia, missä määrin tekstianalyysin avulla voidaan tuottaa tietoa terveydenhuollon digipalveluihin liittyvästä kansalaismielipiteestä Suomessa. Analysoitavat tekstit olivat suomenkielisiä twiitteja, jotka kerättiin Twitteristä palveluihin liittyvien hakutermien avulla. Tutkimuksessa käytetyt tekstianalyysin menetelmät olivat aihemallinnus ja tunneanalyysi. Aihemallinnuksella pyrittiin selvittämään, minkälaisia palveluihin liittyviä aiheita esiintyy kerätyissä twiiteissa ja arvioimaan aiheiden tärkeyttä suomalaisille. Tunneanalyysia käytettiin arvioimaan löydettyihin aiheisiin kuuluvien twiittien tunnelatauksia. Tarkoituksena oli vertailla mielipiteitä eri aiheita kohtaan tunnelatauksien perusteella ja arvioida, mitkä tekijät vaikuttavat aihetta kohtaan tunnettuihin mielipiteisiin.

Tutkimustulosten perusteella jouduttiin toteamaan, että tekstianalyysin kyky tuottaa tietoa kansalaismielipiteestä oli varsin rajallinen. Aihemallinnus ei kyennyt tuottamaan käytettyjen Twitterin hakutermien ulkopuolisia aiheita, eikä näin tuottanut uutta ennalta tuntematonta tietoa. Tunneanalyysin tulosten mukaan useimmissa aiheissa enemmistö twiiteista oli tunteiltaan neutraaleja, ja tunne-erot aiheiden välillä olivat vähäisiä. Valitettavasti käytettyjen tutkimusmenetelmien validiteetti osoittautui niin heikoksi, että saatuihin tuloksiin ei voitu luottaa. Kohdattujen ongelmien taustalla oli se, että kerätyt twiitit eivät sisältäneet halutunlaista tietoa, mikä heikensi merkittävästi menetelmien kykyä tuottaa luotettavaa tietoa kansalaismielipiteestä.

Asiasanat
tekstianalyysi, tekstinlouhinta, aihemallinnus, tunneanalyysi, terveydenhuollon digipalvelut, sosiaalinen media.

Sisällys

1 Johdanto ... 1
1.1 Tutkimuskysymykset ja rajaus ... 2
1.2 Keskeiset käsitteet ... 2
2 Terveydenhuollon digitalisaatio ... 4
3 Sosiaalinen media tiedonlähteenä ... 5
3.1 Twitter ... 5
3.2 Datan keräys ohjelmointirajapinnan kautta ... 6
4 Tekstianalyysi ... 7
4.1 Tekstin esiprosessointi ... 8
4.2 Tekstinlouhinta ... 10
4.2.1 Aihemallinnus ... 11
4.2.2 Tunneanalyysi ... 11
4.3 Analyysityökalut ... 12
4.3.1 Twitter API ... 12
4.3.2 DiscoverText ... 12
4.3.3 KNIME ... 13
4.3.4 Python ... 13
5 Käytetyt tutkimusmenetelmät ... 15
5.1 Kvantitatiivinen tutkimus ja otantatutkimus ... 15
5.2 Tutkimusmenetelmien validiteetti ... 16
5.3 Aihemallinnus epänegatiivisella matriisifaktorisoinnilla ... 16
5.4 Tunneanalyysi VADER-sanastolla ... 18
6 Tutkimusaineiston keräys ja esiprosessointi ... 19
6.1 Aineiston keräys ... 19
6.2 Aineiston esiprosessointi ... 20
6.2.1 Esiprosessointi aihemallinnusta varten ... 21
6.2.2 Esiprosessointi tunneanalyysia varten ... 22
7 Tekstinlouhinnan tulokset ... 23
8 Johtopäätökset ... 29
8.1 Vastaukset alatutkimuskysymyksiin ... 29
8.2 Validiteetin arviointia ... 31
8.3 Yhteenveto ... 32
9 Pohdinta ... 34
Lähteet ... 35
Liitteet ... 37

Liite 1. Aineistonhaun hakutermit ja kerättyjen twiittien lukumäärät 37
Liite 2. Terveydenhuollon digipalveluista aktiivisesti twiittaavat terveysalan toimijat 38

1 Johdanto

Terveydenhuollon digitaalisista palveluista on tullut lyhyessä ajassa merkittävä osa suomalaisten terveydenhuoltoa. Tämän vuoksi on tärkeää saada tietoa siitä, miten palveluiden käyttäjät ja suomalaiset yleensä kokevat nämä palvelut. Mielipiteitä voitaisiin selvittää kyselytutkimuksen avulla, mutta sellaisen toteuttaminen on työlästä ja aikaa vaativaa. Yhteiskunnan digitalisoitumisen ansiosta kaikenlaisista mieliä liikuttavista asioista – myös terveydenhuollon digipalveluista – tavataan kirjoittaa mielipiteitä Internetin keskusteluforumeille ja sosiaaliseen mediaan. Näille alustoille kirjoitetut tekstit tarjoavatkin runsaan tiedonlähteen aiheita tai ilmiöitä ymmärtämään pyrkiville tutkijoille.

Tekstien tutkimista varten on kehitetty ohjelmia, joilla voidaan automatisoidusti sekä kerätä tekstejä Internetistä että analysoida niitä. Tekstianalyysilla viitataan menetelmiin, jotka mahdollistavat tekstidatan analysoimisen automatisoidusti. Tekstianalyysistä tekee hyödyllisen se, että se on nopea toteuttaa ja ei juuri vaadi ihmistyötä. Tutkimuksessani käytetyt tekstianalyysin menetelmät ovat aihemallinnus ja tunneanalyysi. Aihemallinnusta käytetään paljastamaan tekstidokumenttien kokoelmista dokumenteille yhteisiä aiheita. Tunneanalyysia käytetään paljastamaan teksteihin sisältyviä tunteita. Käyttämällä tunneanalyysia tiettyä aihetta käsitteleviin teksteihin voidaan arvioida kirjoittajien mielipiteitä kyseistä aihetta kohtaan.

Tutkimukseni tarkoituksena on tekstianalyysia hyödyntäen kerätä ja analysoida sosiaalisen median tekstejä saadakseni tietoa suomalaisten mielipiteistä terveydenhuollon digitaalisia palveluita kohtaan. Tutkimusaineiston datanlähteenä on sosiaalisen median alusta Twitter eli nykyinen viestipalvelu X. Selvyyden vuoksi tässä tutkimuksessa alustaa kutsutaan kuitenkin edelleen Twitteriksi. Tutkimusaineistoon kuuluvat twiitit kerätään tekemällä Twitteriin tiedonhakuja terveydenhuollon digipalveluihin liittyvillä hakutermeillä. Aihemallinnuksella pyritään selvittämään, minkälaiset palveluihin liittyvät aiheet esiintyvät usein kerätyissä twiiteissa ja arvioimaan aiheiden tärkeyttä. Toivottavaa olisi löytää aiheita, jotka käsittelevät palveluiden saatavuutta, luotettavuutta ja eettisyyttä. Tunneanalyysilla pyritään puolestaan arvioimaan löydettyihin aiheisiin kuuluvien twiittien tunnelatauksia. Tarkoituksena on vertailla mielipiteitä eri aiheita kohtaan tunnelatauksien perusteella. Tällä tavoin pyritään tuottamaan tietoa siitä, missä aiheissa suomalaiset kokevat asioiden olevan hyvin, ja missä huonosti. Lisäksi pyritään selvittämään, onko tunteiden kehitykselle havaittavissa ajallista trendiä.

Opinnäytetyöni on osa Haaga-Helia ammattikorkeakoulun AI Forum-hanketta, jonka tavoitteena on vahvistaa asiantuntemusta ja tutkimusyhteistyötä tekoälyn saralla. Hankkeessa tutkitaan tekoälyn roolia digitaalisen muutoksen mahdollistajana ja sopeuttajana. Erityisesti keskitytään tekoälyn sovelluksiin digipalveluissa, ja sen mukanaan tuomaan työelämän murrokseen. (Haaga-Helia 2023.) Opinnäytetyöni tavoitteena on tuottaa Haaga-Helia ammattikorkeakoululle tietoa suomalaisten

kansalaismielipiteestä terveydenhuollon digipalveluita kohtaan sekä selvittää, missä määrin sosiaalisen median tekstien ja tekstianalyysin avulla ylipäätään voi hankkia tietoa palveluita koskevasta kansalaismielipiteestä. Tutkimuksen tuloksista voivat hyötyä terveydenhuollon digipalvelujen kehittäjät ja muut terveydenhuollon digitalisaation parissa työskentelevät.

1.1 Tutkimuskysymykset ja rajaus

Tutkimuksen päätutkimuskysymyksenä on selvittää, missä määrin tekstianalyysia käyttämällä voidaan tuottaa tietoa suomalaisten kansalaismielipiteestä terveydenhuollon digipalveluita kohtaan. Päätutkimuskysymykseen vastaaminen edellyttää vastaamista viiteen alatutkimuskysymykseen. Tekstianalyysin eri vaiheiden suorittamiseen on olemassa useita käyttökelpoisia ohjelmia, jonka vuoksi ensimmäiseksi haluankin selvittää, mitkä ohjelmat soveltuvat parhaiten tässä tutkimuksessa suoritettavaan tekstianalyysiin. Tutkimukselle olisi hyödyksi saada tietoa siitä, mitkä ovat suomalaisille tärkeitä terveydenhuollon digipalveluihin liittyviä aiheita. Toiseksi haluankin selvittää, minkälaisia aiheita esiintyy usein palveluita käsittelevissä twiiteissa. Tarpeellista on myös tutkia, mitkä löydetyistä aiheista ovat tärkeimpiä, eli mistä aiheista suomalaiset puhuvat eniten. Tämän vuoksi kolmanneksi haluan selvittää, minkälaisiin aiheisiin kuuluu eniten twiitteja. Seuraavaksi haluan tutkia, minkälaiset tekijät vaikuttavat twiittien sisältämiin tunnelatauksiin. Näin voidaan arvioida, mitkä palveluihin liittyvät asiat ovat kansalaisten mielestä hyvin, ja missä on kehityskohteita. Neljänneksi haluankin selvittää, miten twiittien aihe vaikuttaa niiden sisältämään tunnelataukseen. Lisäksi olisi hyödyllistä tietää, muuttuvatko mielipiteet palveluita kohtaan vuosien kuluessa. Tämän perusteella voidaan arvioida, onko palveluita onnistuttu kehittämään oikeaan suuntaan. Tämän vuoksi viimeiseksi haluan selvittää, löydetäänkö tunnelatauksien kehitykselle ajan mukana nouseva tai laskeva trendi.

Tutkimukseni on rajattu siten, että tutkin vain Twitterin suomenkielisiä twiitteja. Useista sosiaalisen median alustoista voisi löytyä aiheeseen liittyviä tekstejä, mutta Twitterin twiitit ovat oletusarvoisesti julkisia ja siten paremmin saatavilla. Rajaudun suomenkielisiin teksteihin, koska tekstianalyysia on vaikea käyttää, jos tutkimusaineisto on monikielistä. Jotta aineistosta tulisi riittävän suuri haen terveydenhuollon digipalveluihin liittyviä twiitteja Twitterin koko olemassaolon ajalta. Twitterissä on huomattavan paljon palveluihin liittyviä tekstejä, jotka ovat terveysalan ammatinharjoittajien tai organisaatioiden kirjoittamia. Tavoitteena on tutkia kansalaismielipidettä, joten pyrin rajaamaan aineistosta pois twiitit, joiden kirjoittajat ovat Twitteriin aktiivisesti kirjoittavia terveysalan toimijoita.

1.2 Keskeiset käsitteet

Aihemallinnus on menetelmä, joka pyrkii matemaattisten ja tilastollisten tekniikoiden avulla löytämään dokumenttikokoelmasta aiheita, teemoja tai käsitteitä (Sarkar 2019).

Luonnollisen kielen prosessointi on tietojenkäsittelyn alue, joka liittyy luonnollisten kielien – kuten englannin – koneelliseen tulkintaan. Prosessi pyrkii muuntamaan luonnollista kieltä sisältävät dokumentit muotoon, jota tietokoneet voivat hyödyntää. (Hapke, Howard & Hobson 2019.)

Poistosanat ovat kielessä usein esiintyviä, mutta tekstianalyysin kannalta merkityksettömiä sanoja, jotka pyritään rajaamaan pois analysoitavasta tekstistä (Sarkar 2019).

Sosiaalinen media on joukko Internet-pohjaisia sovelluksia, jotka mahdollistavat käyttäjien luoman sisällön välityksen ja vaihdon (Sloan ja Quan-Haase 2017).

Strukturoimaton data tarkoittaa dataa, joka on sellaisessa muodossa, että sitä ei voida analysoida tietokoneella ilman muokkausta (Witten, Frank, Hall & Pal 2017).

Tekstianalyysi viittaa menetelmiin, joilla kerätään halutunlaisia tekstidokumentteja, muunnetaan strukturoimatonta dataa luonnollisen kielen prosessoinnilla tai käytetään analytiikkaa erottamaan tietoa tekstipohjaisesta datasta (Sharda, Delen & Turban 2018, 277–278).

Tekstinlouhinta viittaa menetelmiin, joilla analysoidaan tekstiä tavoitteena paljastaa hyödyllistä tietoa (Witten ym. 2017).

Tunneanalyysi on menetelmä, joka pyrkii automatisoidusti havaitsemaan tekstidokumentissa olevat positiiviset ja negatiiviset näkemykset tiettyä aihetta kohtaan (Sharda ym. 2018).

Twitter on sosiaalisen median alusta ja mikroblogipalvelu, jonka käyttäjät voivat julkaista lyhyitä – enintään 280 merkin pituisia – viestejä eli twiitteja (Teodorowski ym. 2022).

Validiteetti kuvaa, missä määrin käytetty mittari kykenee mittamaan sitä, mitä pitikin mitata (Heikkilä 2014, 27).

2 Terveydenhuollon digitalisaatio

Terveydenhuollon digitaalisilla palveluilla eli sähköisillä terveydenhuoltopalveluilla tarkoitetaan terveydenhuoltoalan välineitä ja palveluita, jotka hyödyntävät tieto- ja viestintätekniikkaa. Palvelut käsittävät tietojen vaihdon potilaiden ja terveydenhuollon palvelujentarjoajien välillä, sähköiset potilastietojärjestelmät, kannettavat potilaiden seurantalaitteet, lääketieteen etäpalvelut ja sähköisen asioinnin. Terveydenhuollon etäpalvelut ovat verrannollisia perinteisiin vastaanottokäynteihin, ja sähköinen asiointi tarkoittaa palveluiden käyttöä tieto- ja viestintätekniikan avulla. Tällaisia ovat esimerkiksi sähköinen ajanvaraus ja sähköisten lomakkeiden täyttäminen. Terveydenhuollon digipalvelujen tavoitteena on ennalta ehkäistä sairauksia sekä parantaa diagnosointia, hoitoa, seurantaa ja terveydenhuollon hallintoa. (Pirhonen 2016, 14–15.) Lisäksi niiden avulla voidaan lisätä kansalaisten kykyä ylläpitää itsenäisesti toimintakykyään ja terveyttään (Saranto, Kinnunen, Jylhä & Kivekäs 2020).

Sohlbergin (2021) mukaan sosiaali- ja terveysministeriön digitalisoitumiseen liittyvän strategian mukaan terveyspalveluja tulisi uudistaa asiakaslähtöisesti, jolloin suunnittelun perustana on palveluita käyttävien ihmisten tarpeet. Tätä varten tarvitaan ymmärrystä käyttäjien kokemuksista, käyttäytymisestä, toiveista ja tarpeista. Asiakasymmärryksen pohjalta voidaan sitten vastata asiakkaan tarpeisiin ja innovoida palveluratkaisuja, jotka perustuvat ymmärrykseen asiakkaiden tarpeista.

Sähköisten sosiaali- ja terveydenhuollon palveluiden käyttö Suomessa oli vuonna 2017 järjestetyn kyselytutkimuksen mukaan yleistä. Silloin ainakin jotain palvelua sähköisesti käyttäneitä oli kaikista vastaajista 68 %. Yleisimmin käytettiin tiedonhakua oman terveyden ja hyvinvoinnin edistämiseksi, jota oli käyttänyt 43 % vastaajista. Muita suosittuja sähköisiä palveluja olivat tiedonhaku palveluista, omien potilas/asiakastietojen tarkastelu, laboratoriotestien ja muiden tutkimustulosten vastaanotto, ajanvaraus ja lääkemääräyksen uusiminen. (Hyppönen, Pentala-Nikulainen & Aalto 2018, 30–31.)

Sähköisten palveluiden suosiosta huolimatta 54 % vuoden 2017 kyselytutkimuksen vastaajista oli täysin samaa mieltä ainakin yhdestä tutkimuksessa ehdotetusta potentiaalisesta esteestä. Yleisin este oli kokemus siitä, että henkilökohtaista tapaamista ei voi korvata sähköisellä yhteydenotolla, josta 35 % vastaajista oli täysin samaa mieltä ja 32 % osin samaa mieltä. Muita yleisesti koettuja esteitä olivat vaikeaselkoiset käyttöehdot, pelko virheistä, ei-lääketieteellisten seikkojen huomioimatta jättäminen etävastaanotossa sekä huoli tietosuojasta ja -turvasta. Muun muassa ilmeni, että osa vastaajista tuntee epäluottamusta palveluita kohtaan. Kolmannes vastaajista ei luottanut siihen, että henkilötiedot pysyvät salassa nimettömissä yhteydenotoissa. Lisäksi neljännes ei luottanut sähköisen palveluiden tuottajiin ja pelkäsivät huijatuksi joutumisen mahdollisuutta. (Hyppönen ym. 2018, 33–34.)

3 Sosiaalinen media tiedonlähteenä

Sosiaalista mediaa käytetään aktiivisesti ympäri maailman. Sosiaalisen median alustoilla yksityiset ihmiset, asiantuntijat ja organisaatiot keskustelevat ja jakavat tietoa. Näin alustoille on kasvanut valtava ja alati kasvava määrä tietoa, jonka merkitys tutkimukselle on alettu ymmärtämään. Sosiaalinen media sisältää kirjoituksia ja mielipiteitä liittyen melkein mihin tahansa asiaan. Näiden tietojen avulla voidaan saada syvällinen käsitys ihmisten ajatus- ja käyttäytymismalleista, ja tietoa voidaan edelleen hyödyntää esimerkiksi sosiologisten kysymysten ratkaisussa. Lisäksi yritykset ovat alkaneet käyttää sosiaalisesta mediasta kerättyjä tietoja liiketoiminnan edellytysten kehittämiseen.

Sosiaalisen median tavoittavuutta kuvaa se, että vuonna 85 % suomalaisista käytti aktiivisesti sosiaalista mediaa ja 59 % käytti sitä päivittäin tai melkein päivittäin. Nuoremmat suomalaiset käyttävät sosiaalista mediaa huomattavasti aktiivisemmin kuin vanhemmat – 84 % 16–24-vuotiasta käytti sitä päivittäin tai melkein päivittäin, kun taas 75–89-vuotiaiden kohdalla osuus oli vain 13 %. (Statista 2023.)

3.1 Twitter

Mikroblogipalvelu Twitter on merkittävä sosiaalisen median alusta. Twitterin käyttäjät voivat julkaista palveluun enintään 280 merkin pituisia viestejä eli twiitteja. Twitter on erityisen suosittu alusta viesteille, jotka liittyvät politiikkaan, journalismiin, yhteiskunnalliseen vaikuttamiseen, urheiluun ja viihteeseen (Isotalus, Jussila & Matikainen 2018, 9–10). Vuonna 2022 Twitteriä käytti 42 % suomalaisista ja 13 % käytti sitä päivittäin. Vertailun vuoksi mainittakoon, että noin 90 % suomalaisista käytti YouTubea ja WhatsAppia. (Statista 2023.)

Twitter ei siis ole suosituin sosiaalisen median alusta, mutta datan avoimuus Twitterissä tekee siitä houkuttelevan datalähteen tekstianalyysille. Monet sosiaalisen median alustat, kuten Facebook ja LinkedIn edellyttävät, että käyttäjät hyväksyvät toisensa kontakteiksi verkostoitumiskutsujen kautta. Vain kontaktit voivat nähdä toistensa viestit. Sen sijaan Twitterissä julkaistut twiitit ovat oletusarvoisesti julkisia, jolloin ne ovat muiden käyttäjien haettavissa ja luettavissa ilman suostumusta verkostoitumiseen. (Russell & Klassen 2019.)

Twitterin keskeinen elementti on *aihetunniste (hashtag)*. Twitterin käyttäjä voi tehdä mistä tahansa sanasta aihetunnisteen lisäämällä sen eteen #-merkin. Aihetunnisteet voivat liittyä esimerkiksi paikkaan, henkilöön, tapahtumaan tai asiaan. Aihetunnisteiden käyttö helpottaa tiedon löytämistä, sillä Twitteristä voi hakea tietoa aihetunnisteiden avulla. Aihetunnisteet toimivat hyperlinkkeinä, joita klikkaamalla saa nähtäväksi kaikki kyseisellä aihetunnisteella merkityt twiitit uutuusjärjestyksessä. (Isotalus ym. 2018, 10–11.)

3.2 Datan keräys ohjelmointirajapinnan kautta

Tutkimuksen edellyttäessä tutkimusaineistoa analyysien pohjaksi tutkijan on kerättävä data, josta muodostetaan tutkimusaineisto. Käyttämällä hakukoneita, kuten Googlea, toivottuja datoja ei välttämättä löydy, tai löydetyt valmiit datat voivat olla epäsopivia tutkimuksen tarkoitukseen nähden. Tällöin tutkijan on luotava oma tutkimusdatansa, ja kohdesovelluksen *ohjelmointirajapinta* API (*application program interface*) on yksi reitti datan hakemiseen hyödyntäen ohjelmointikoodia. Ohjelmointirajapinnat ovat liittymiä tai yhtymäkohtia, joiden kautta sovellukset voivat kommunikoida toistensa kanssa. API:n kautta sovelluksen tarjoamia palveluita voidaan käyttää ohjelmallisesti. (Albrecht, Ramachandran & Winkler 2020.)

Tutkijoille API:t ovat hyödyllisiä datan keräyksessä, sille ne ovat suunniteltu yhteyden muodostamista varten, ne ovat helppokäyttöisiä, ja niiden käyttö voidaan automatisoida. Lisäksi API:t ovat suositeltuja erityisesti silloin, kun tutkimuksessa käytetään dataa, joka muuttuu jatkuvasti tai, kun on kriittisen tärkeää, että tutkimus heijastelee viimeisintä tietoa. Ennen kuin aletaan hyödyntämään API:a on tärkeää tutustua käytön ohjeisiin, sillä käyttö voi edellyttää tunnistautumista, ja yhteydenottokutsulle tiettyä muotoa ja parametrejä. Lisäksi API:n käyttöön voi liittyä rajoituksia koskien esimerkiksi latausmääriä. (Albrecht ym. 2020.)

Useat suositut verkkoalustat, kuten GitHub ja Twitter tarjoavat mahdollisuuden alustalla olevan datan käsittelyyn API:n välityksellä. Verkkoalustat pyrkivät siihen, että pääsy dataan tapahtuu API kautta, sillä liiketoiminnan kannalta yritysten ei kannata tarjota rajoittamatonta pääsyoikeutta hallussaan olevaan dataan tunnistautumattomille datan käsittelijöille. Kun datan käsittely tapahtuu API:n kautta, datan omistavat yritykset voivat vaatia tunnistautumista ja laskuttaa datan käytöstä. Lisäksi yritykset tarjoavat usein erilaisia käyttöoikeuksia erilaisille asiakkaille. (Jürgens & Jungherr 2016.)

4 Tekstianalyysi

Yhteiskunnan tietoteknistymisen mukana on kehitetty suuri määrä työkaluja ja infrastruktuuria keräämään ja varastoimaan sähköisessä muodossa olevaa dataa. Suurin osa yritysten hallussa olevasta datasta on tekstimuotoista ja strukturoimatonta dataa eli sitä ei voida hyödyntää automatisoidusti ilman muokkausta. Arvioiden mukaan noin 80 % yritysten hallussa olevasta datasta on tekstimuotoista. Vastaavasti sosiaaliseen mediaan laitettu data on muodoltaan strukturoimatonta ja useimmiten tekstimuotoista. Ei olisi käytännöllistä turvautua yksittäisiin ihmisiin tällaisen datamassan prosessoinnissa ja analysoinnissa. Tämän vuoksi strukturoimattoman tekstimuotoisen datan tutkimista varten on kehitetty automaattisen tekstianalyysin menetelmiä. (Chakraborty, Pagolu & Garla 2014; Sharda ym. 2018.)

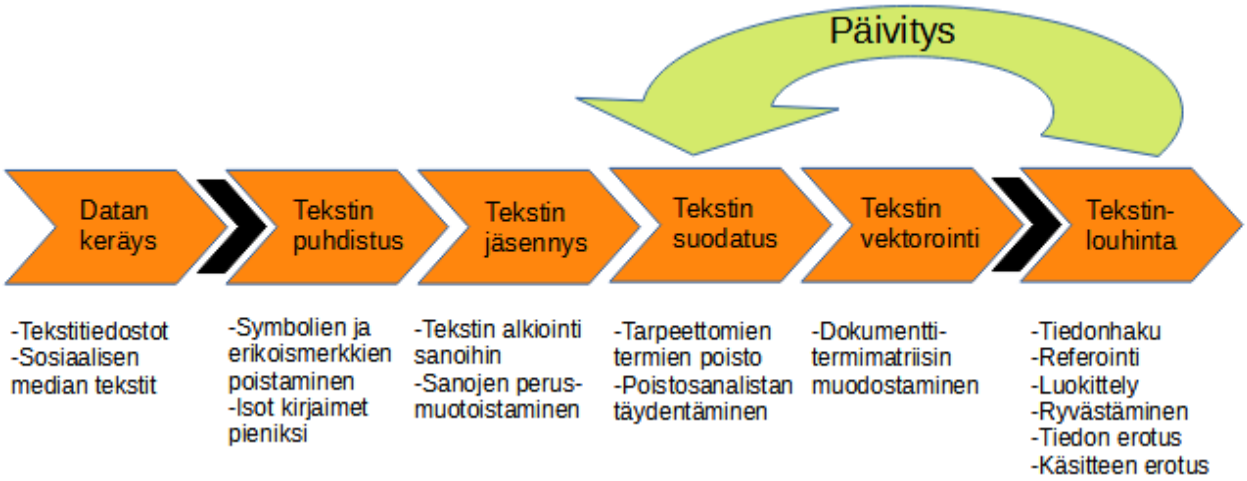
Tekstianalyysin mahdollistamiseksi käytettävä tekstimuotoinen data tulee muuntaa numeroiksi, jotta dataa voidaan käsitellä algoritmien avulla. Tekstin muuntamiseksi strukturoituun numeeriseen muotoon tarvitaan *luonnollisen kielen prosessointia* NLP *(natural language processing)*. Luonnolliset kielet, kuten suomen kieli ovat olemassa ihmisten välistä tiedonvälitystä varten. Luonnollinen kieli on strukturoimatonta dataa, jota tietokoneet eivät suoraan pysty hyödyntämään. Luonnollisen kielen prosessointi on menetelmä, jolla luonnollisella kielellä oleva data muokataan tietokoneiden ymmärtämään muotoon. (Hapke ym. 2019.)

Tekstianalyysin ajatellaan kattavan menetelmät, joita käytetään merkityksellisen tiedon löytämiseen strukturoimattomasta tekstidatasta. Halutun tiedon löytämiseksi käytetään luonnollisen kielen prosessointia ja analytiikkaa. Eristettäviä tietoja voivat olla esimerkiksi merkitykset, kaavat, aiheet ja tunteet. (Sharda ym. 2018.) Tekstianalyysilla on huomattava määrä erilaisia käyttöalueita, ja eri kirjoittajat määrittelevät nämä käyttöalueet eri tavoin. Kuitenkin yhteenvetona Minerin ja kumppaneiden (2012) sekä Shardan ja kumppaneiden (2018) aihetta koskevien ajatusten perusteella voidaan todeta, että tekstianalyysin tärkeimmät käyttöalueet ovat seuraavat:

- **Tiedonhaku** (information retrieval): tekstidokumenttien etsintä ja keräys kokoelmista hyödyntäen hakutermejä.
- **Referointi** (summarization): dokumenttien tiivistäminen sisältämiinsä pääteemoihin.
- **Luokittelu** (categorization): dokumentteja yhdistävien tekijöiden tunnistamien, ja dokumenttien luokitteleminen ennalta määrättyihin luokkiin.
- **Ryvästäminen** (clustering): dokumenttien ryhmittely samankaltaisuuden perusteella ennalta määräämättömiin rypäisiin.
- **Tiedon erotus** (information extraction): olennaisten termien ja yhteyksien tunnistaminen tekstistä.

- **Käsitteen erotus** (concept extraction): sanojen ja lauseiden ryhmittely merkitykseltään samankaltaisiin ryhmiin.

Tekstianalyysi ja tekstinlouhinta samaistetaan usein, mutta nykyään etenevissä määrin ajatellaan, että tekstianalyysi on laajempi menetelmäkokonaisuus, johon tekstinlouhinta sisältyy. Chakraborty ja kumppanit (2014) esittävät tekstianalyysin kolmivaiheisena prosessina. Heidän mukaansa prosessin ensimmäisenä vaiheena on tutkittavan tekstidatan keräys. Data voi olla esimerkiksi tietokannoissa olevia tekstidokumentteja tai sosiaalisen mediaan ladattuja kirjoituksia. Halutunlaisen datan löytämiseksi tarvitaan usein hakutermejä, joilla määritellään, minkälaisia tekstejä halutaan kerätä. Seuraavana vaiheena on tekstin esiprosessointi, jossa teksti muunnetaan luonnollisen kielen prosessoinnilla numeeriseen muotoon. Esiprosessointi voidaan puolestaan jakaa neljään myöhemmin kuvattavaan vaiheeseen. Viimeisenä prosessissa on tekstinlouhinta, jossa muokatulle datalle suoritetaan haluttu analyysi. Usein tekstinlouhinnan tulokset osoittavat, että tekstin esiprosessoinnissa oli puutteita, jolloin esiprosessoinnissa käytettäviä asetuksia päivitetään iteratiivisesti tekstinlouhinnan tulosten perusteella, kunnes ollaan tyytyväisiä saatuihin tuloksiin. Tekstianalyysi voidaan esittää kuvan 1 kaltaisena prosessina.



Kuva 1: Tekstianalyysin esitys prosessina

4.1 Tekstin esiprosessointi

Tekstin esiprosessointi alkaa tekstin puhdistamisella, jossa tunnistetaan ja poistetaan tekstistä analyysin kannalta tarpeettomat tai haitalliset osat, kuten erikoismerkit, symbolit ja HTML-osoitteet. Lisäksi puhdistuksessa suoritetaan merkkien normalisointi, jossa isot kirjaimet muunnetaan pieniksi, ja epätavanomaisessa muodossa olevat merkit, kuten aksenttimerkillä varustetut kirjaimet muunnetaan tavanomaiseen muotoonsa. (Albrecht ym. 2020.)

Jäsennysvaiheessa tekstille tehdään *alkiointi (tokenization)*. Alkiointi on tekstidokumenttien segmentointia, jossa teksti jaetaan osuuksiin eli segmentteihin. Tavallisesti tekstit jaetaan sanoihin. Tekstissä olevat sanat esiintyvät usein jossain taivutusmuodossaan. Kun tekstit ovat alkioitu sanoiksi, sanat pyritään yhdenmukaistamaan palauttamalla ne perusmuotoonsa. *Perusmuotoistamisella (lemmatization)* tarkoitetaan sanan muuntamista sen taivuttamattomaan perusmuotoon. Perusmuotoistamista suorittavat ohjelmat kykenevät huomioimaan eri sanojen saman merkityksen hyödyntämällä listoja sanojen synonyymeista. Perusmuotoistamista käytetään, jotta analysointivaiheessa kyetään käsittelemään samalla tavalla saman sanan eri taivutusmuotoja. Lisäksi sanat esiintyvät erilaisissa sanastoissa perusmuodossaan, jonka vuoksi sanojen perusmuotoistaminen mahdollistaa niiden linkityksen sanastoihin. (Hapke ym. 2018; Sarkar 2019.)

Tekstin suodatusvaiheessa pyritään tunnistamaan ja rajaamaan analyysin ulkopuolelle termit (sanat), jotka eivät tarjoa haluttua tietoa dokumenttien sisällöstä eivätkä auta erottelemaan dokumentteja toisistaan. Tämänlaisten termien etsintä dokumenteista on työlästä ja vaatii ymmärrystä tutkittavasta ilmiöstä. Kielessä usein esiintyviä, mutta tekstianalyysissa merkityksettömiä sanoja kutsutaan *poistosanoiksi (stopwords)*. Sanat, kuten *koko*, *kun* ja *mitkä* ovat yleensä poistosanoja. Universaalia listaa poistosanoista ei ole olemassa, vaan eri ohjelmilla on omat poistosanalistat, jotka riippuvat käytetystä kielestä. Poistosanalistalla olevia sanoja ei käytetä esiprosessoinnin vektorointivaiheessa. Löydettäessä tutkittavan ilmiön kannalta tarpeettomia termejä ne voidaan lisätä poistosanalistaan sen sijaan, että ne poistettaisiin tekstistä. (Chakraborty ym. 2014; Sarkar 2019.)

Tarpeettomien termien suodatuksessa hyödyllisiä työkaluja *sanaluokkakoodit (part-of-speech tags)*, joiden avulla analyysi voidaan rajata perustumaan vain niihin datan sanoihin, jotka kuuluvat tiettyihin sanaluokkiin, kuten substantiiveihin tai verbeihin. Kielessä pronominit, prepositiot, konjunktiot ilmaisevat sanojen välisiä suhteita lauseessa, mutta ne eivät anna tietoa lauseen merkityksestä. Sen sijaan substantiivit, verbit, adjektiivit ja adverbit määrittelevät lauseen merkityksen. Tämän vuoksi analyysissa usein riittää, että käsitellään vain näihin sanaluokkiin kuuluvia sanoja. (Albrecht ym. 2020.)

Vektorointivaiheessa teksti muunnetaan numeeriseen muotoon. Dokumentin tekstin kuvaamiseen kattavasti ei riitä yksi luku, vaan tarvitaan useita lukuja vektorin muodossa. Vektoroinnissa teksti muunnetaan lukuja sisältäviksi vektoreiksi. (Albrecht ym. 2020.)

Erittäin hyödyllinen malli tekstidatan käsittelyssä on termivektorimalli, joka esittää tekstidokumentin eri termit eli sanat numerovektorin ulottuvuuksina. Voidaksemme vertailla vektoreita ja laskea yhtäläisyyksiä, eri dokumenteille muodostetuilla vektoreilla täytyy olla samat ulottuvuudet. Vektorien ulottuvuudeksi tulee erillisten termien lukumäärä koko aineistossa lukuun ottamatta poistosanoja.

Tämän ymmärtämiseksi ajatellaan, että d on yhdelle aineiston dokumenteista muodostettu vektori, ja n on vektorin ulottuvuus. Nyt voidaan kirjoittaa, että

$$d = \{w_{D1}, \dots, w_{Dk}, \dots, w_{Dn}\},$$

missä w_{Dk} on termille k laskettu paino kyseisessä dokumentissa. Paino on numeerinen arvo, joka riippuu painojen laskemiseen käytetystä mallista ja voi olla esimerkiksi sanan frekvenssi eli esiintymismäärä dokumentissa. Järjestämällä jokaiselle dokumentille lasketut vektorit riveiksi matriisiin, saadaan *dokumentti-termimatriisi*, joka siten sisältää vektoriesityksen kaikista aineiston dokumenteista. Dokumentti-termimatriisi toimii analyysien perustana tekstinlouhinnassa. (Albrecht ym. 2020; Sarkar 2019.)

Dokumentin vektorointiin on olemassa useita malleja, joista tunnetuimmat ovat *sanakassi (bag of words)* ja TF-IDF *(term frequency times inverse document frequency)*. Sanakassi-mallissa vektorin painot ovat sanojen frekvenssit dokumentissa. Malliin sisältyvä oletus on, että mitä useammin sana esiintyy, sitä tärkeämpi se on dokumentille. Tällainen oletus voi kuitenkin johtaa harhaan, sillä sanan esiintymistiheys dokumentissa ei itsessään välttämättä kerro paljoa sanan tärkeydestä kyseisessä dokumentissa. Tiettyyn aiheeseen liittyvässä aineistossa voi olla paljon aiheeseen liittyviä sanoja halki kaikkien aineiston dokumenttien. Tämänkaltaiset runsaslukuiset sanat eivät tuo uutta tietoa eivätkä auta erittelemään dokumentteja. Usein voisi olla hyödyllistä määritellä tällaiset sanat poistosanoiksi. (Albrecht ym. 2020; Hapke ym. 2018.)

Parempi lähtökohta olisi se, että malli vähentäisi liian usein aineistossa esiintyvien sanojen merkitystä. Tällainen malli on TF-IDF, joka laskee dokumenttien lukumäärän, joissa tietty sana esiintyy. Mallissa vektorin painot ovat sanojen frekvenssit dokumentissa jaettuna aineiston dokumenttien määrällä, joissa kyseinen sana esiintyy. Näin ollen malli vähentää aineistossa usein esiintyvien sanojen painoa ja lisää aineistossa harvinaisempien sanojen painoa. Malliin sisältyvä oletus on, että sanat, jotka eivät esiinny runsaasti koko aineistossa, mutta ovat runsaita tietyissä dokumenteissa, saattavat kertoa jotain oleellista näiden dokumenttien luonteesta. Vastaavasti, jos sana ilmenee yhdessä dokumentissa useita kertoja, mutta harvoin muualla, niin sana kertoo jotain olennaista kyseisestä dokumentista. (Albrecht ym. 2020; Hapke ym. 2018.)

4.2 Tekstinlouhinta

Kun tekstidata on esiprosessoitu strukturoituun muotoon, voidaan aloittaa tekstinlouhinta, joka on datan varsinainen analyysivaihe. Tässä vaiheessa vektoroitua tekstiä käsitellään tekstinlouhinnan algoritmeilla, jotta saadaan tehtyä haluttu analyysi, kuten esimerkiksi dokumenttien ryhmittely ryppäisiin. Tekstinlouhintaa tehdään usein iteratiivisesti, koska tekstinlouhinnan antamien tulosten perusteella tekstin esiprosessointi tulisi tehdä toisin, jolloin esiprosessointi tehdään uudelleen

muuttamalla asetuksia sekä suodattamalla aineistosta pois analyysin kannalta haitallisia termejä. (Chakraborty ym. 2014.)

4.2.1 Aihemallinnus

Käytettäessä tutkimusaineistoa, joka sisältää suuren määrän tekstidokumentteja voidaan haluta selvittää, mitä aiheita dokumentit käsittelevät. Tällöin voidaan käyttää tekstinlouhinnan menetelmää *aihemallinnus (topic modelling)*, joka käyttää tilastollisia ja matemaattisia menetelmiä erottamaan aiheita tai käsitteitä laajoista tekstidokumenttien kokoelmista. Mitä monimuotoisempi dokumenttikokoelma on, sitä enemmän aiheita tai käsitteitä siitä voidaan erottaa. (Sarkar 2019.)

Aiheet koostuvat jakaumasta sanoja, jotka esiintyvät usein yhdessä aineiston dokumenteissa. Valittu vektorimalli voi vaikuttaa aihemallinnuksen tuottamiin aiheisiin. Aiheet tavallisesti jossain määrin limittyvät toistensa kanssa, ja dokumentteja ei ole mahdollista luokitella vain yhteen aiheeseen kuuluvaksi. Sen sijaan dokumentit sisältävät sekoituksen eri aiheita. Aihemallinnuksen tavoitteena ei olekaan määritellä dokumenteille yhtä aiheitta, vaan paljastaa aineiston aiherakenne. (Albrecht ym. 2020.)

4.2.2 Tunneanalyysi

Haluttaessa arvioida tekstidokumenttien sisältämiä tunteita tai mielipiteitä tekstien sisällön perusteella, voidaan turvautua *tunneanalyysiin (sentiment analysis)*. Tunneanalyysi on tekstinlouhinnan menetelmä, jonka tarkoituksena selvittää tekstin kirjoittajan tuntemia tunteita tai mielialaa kirjoitettujen tekstien perusteella. Tunneanalyysin avulla voidaan selvittää, mitä ihmiset tuntevat tiettyä aihetta kohtaan. Yritykset voivat edistää liiketoimintaansa menetelmän avulla, sillä ne voivat hankkia automatisoidusti tietoa asiakkaidensa mielipiteistä tekemällä tunneanalyysin asiakkaiden tuotteista tai palveluista kirjoittamille arvosteluille. (Albrecht ym. 2020; Chakraborty ym. 2014.)

Tunneanalyysin tavallinen käyttötarkoitus on tekstidokumenttien luokittelu niiden sisältämien tunteiden mukaan. Dokumentit voidaan luokitella esimerkiksi positiivisiksi, negatiivisiksi tai neutraaleiksi. Luokittelun kriteerinä käytetään dokumenttien tunteiden *polariteettia (polarity)*. Polariteetilla viitataan siihen, sisältääkö dokumentti positiivisen, negatiivisen vai neutraalin tunteen. Lisäksi polariteetille lasketaan yleensä jonkinlainen pistemäärä, joka kuvaa dokumentin positiivisen tai negatiivisen tunteen voimakkuutta. (Chakraborty ym. 2014; Sarkar 2019.)

Tavallisesti tunneanalyysissa käytetään *sanastopohjaista (lexicon-based)* menetelmää. Sanastopohjainen tunneanalyysi hyödyntää tunnesanastoja, jotka on laadittu erityisesti tunneanalyysia varten. Sanastot sisältävät sanoja, jotka liitetään positiiviseen tai negatiiviseen tunteeseen. Sanoille on määritelty polariteetti ja polariteetin voimakkuuden ilmaiseva pistemäärä. Dokumentin

sisältämän tunteen voi määritellä antamalla sanaston avulla jokaiselle dokumentin sanalle tunnepistemäärä. Summaamalla pistemäärät saadaan dokumentin kokonaistunne. Tunnesanastot normaalisti sisältävät useita taivutusmuotoja samasta sanasta, ja ne eivät sisällä tavallisia poistosanoja. Ainoastaan sanat, jotka löytyvät sanastosta saavat pistemäärän. Sanastopohjaisen menetelmän heikkous onkin siinä, että analyysissa joudutaan rajautumaan sanaston kokoon. Jos tunnesisältöä omaava sana ei ole sanastossa, niin sanan sisältämää tietoa ei lasketa mukaan tunneanalyysissa. (Albrecht ym. 2020; Sarkar 2019.)

4.3 Analyysityökalut

Tekstianalyysin eri vaiheiden suorittamiseen tarvitaan yksi tai useampi ohjelmointityökalu. Tätä varten tutustuin muutamiin tarjolla oleviin tekstianalyysiin pystyviin ohjelmiin. Tarkoituksenani oli löytää ja valita paras ohjelma kuhunkin vaiheeseen. Twitter tarjoaa ohjelmointirajapinnan (API), jonka kautta voidaan tehdä tiedonhakuja, mutta on myös olemassa kolmannen osapuolen tarjoamia ohjelmia, jotka ovat – tai ainakin olivat vuoteen 2023 asti – lisensoituja tekemään tiedonhakuja Twitterin API:n kautta. Nämä edellyttävät käyttäjän tunnistautumista. Lisäksi on ohjelmia, jotka mahdollistavat tiedonhaun Twitteristä ilman, että käyttäjän täytyy tunnistautua. Tässä on kuitenkin otettava huomioon, että Twitter on uuden pääomistajansa Elon Muskin aikana pyrkinyt rajoittamaan mahdollisuuksia tehdä tiedonkeruuta Twitteristä.

4.3.1 Twitter API

Twitterin API mahdollistaa käyttäjille pääsyn Twitterin dataan ohjelmallisesti eli luomalla ohjelmakoodin, jolla määritellään, mitä tietoa haetaan. API:n kautta on mahdollista muun muassa hakea ja luoda twiitteja sekä etsiä käyttäjiä ja mainintoja käyttäjistä. Twitterin API.n käyttö on kuitenkin huomattavan rajoitettua ja datan käyttömahdollisuuden laajuus riippuu valitusta pääsyoikeustasosta. Tällä hetkellä tarjoaa pääsyoikeustasot *Free*, *Basic* ja *Enterprise*. Näistä ensimmäinen on ilmaiseksi tarjolla kaikille, joilla on Twitterin kehittäjätili. Ilmainen taso ei tosin mahdollista twiittien keruuta, ja kaksi muuta ovat maksullisia. (Twitter 2023.)

4.3.2 DiscoverText

DiscoverText on tekstianalyysin ohjelma, jolla voidaan kerätä ja analysoida tekstimuotoista dataa. DiscoverText on selainpohjainen ja sitä käytetään graafisen käyttöliittymän kautta. Graafinen käyttöliittymä mahdollistaa ohjelman käytön ilman koodausosaamista. DiscoverText:n tarjoamiin ominaisuuksiin sisältyy monikielisuus, tekstinlouhinta, data-analyysi ja koneoppiminen. (DiscoverText 2023.) Mielestäni DiscoverText on yksinkertainen käyttää, mutta ongelmana on kunnollisten käyttöohjeiden puute. DiscoverText oli aiemmin lisensoitu keräämään dataa Twitterin API:n kautta, mutta menetti tämän oikeuden vuoden 2023 aikana.

4.3.3 KNIME

KNIME on avoimen lähdekoodin ohjelma, jolla on graafinen käyttöliittymä. Ohjelman avulla voidaan kerätä, analysoida ja visualisoida dataa. Ohjelmisto mahdollistaa myös koneoppimisen mallien käytön luokittelussa, ryhmittelyssä ja regressiossa hyödyntäen kehittyneitä algoritmeja. (Knime 2023.) Samoin kuin DiscoverText, myös KNIME menetti oikeuden kerätä dataa Twitterin API:n kautta.

4.3.4 Python

Python on avoimen lähdekoodin ohjelmointikieli, joka on ominaisuuksiltaan erittäin monipuolinen. Python on suunniteltu olemaan mahdollisimman helppokäyttöinen tekemällä koodista yksinkertaista ja kaunista. Pythonin ominaisuuksia voi helposti laajentaa lataamalla ilmaiseksi saatavilla olevia kirjastoja, joiden sisältä otetaan käyttöön kulloinkin tarvittu moduuli eli kooditiedosto. Python soveltuu hyvin tekstidatan analysointiin, sillä Pythonilla on käytettävissään useita luonnollisen kielen prosessointiin ja tekstianalyysiin soveltuvia kirjastoja. Hyödyntämällä näitä kirjastoja voidaan säästää paljon aikaa, joka kuluisi tekstin käsittelyyn ja analysointiin tarvittavan koodin kirjoittamiseen. (Sarkar 2019.)

NLTK *(Natural language toolkit)* on kirjasto, joka käsittää yli lukuisia tekstin prosessointiin ja analysointiin tarvittavia työkaluja. Lisäksi NLTK sisältää moduuleita tekstin luokitteluun, alkiointiin, perusmuotoistamiseen ja puhdistamiseen. Useimmiten NLTK valitaan työkaluksi, kun tehtävä edellyttää luonnollisen kielen prosessointia. (Sarkar 2019.)

Eräs varsin uusi kirjasto on spaCy, mutta se on kenties yksi parhaista kirjastoista luonnollisen kielen prosessintiin. Sen toiminnallisuuksien laajuus ei ole yhtä suuri kuin NLTK:n, mutta se suoriutuu silti tärkeimmistä tekstin käsittelyyn liittyvistä tehtävistä. Tutkimus on osoittanut, että spaCy on erittäin nopea algoritmien suorittamisessa ja sopiikin erityisen hyvin suuren mittakaavan tiedonerotuksen tehtäviin. Yksi tärkeimmistä ominaisuuksista on se, että spaCy tukee useita kieliä. (Sarkar 2019.)

Pythonilla voidaan tehdä tunneanalyysi hyödyntäen useita eri tunnesanastoja. Pythonin kirjastojen sisältämiä tunnesanastoja ovat mm. Bing Liun sanasto, AFINN, SentiWordNet ja VADER. Jokainen näistä sanastoista sisältää listan sanoista, jotka liitetään positiiviseen ja negatiiviseen tunteeseen sekä sanojen tunteiden voimakkuuden ja suunnan. Valitettavasti useimmat Pythonin tunnesanastoista on saatavilla vain englanninkielisinä. (Sarkar 2019.)

Pythonilla on myös tarjota kirjastoja datan keräykseen Twitteristä. Usein tutkija tarvitsee dataa pitkältä aikaväliltä, jolloin API:sta saatavat viimeisinä seitsemänä päivänä julkaistut twiitit

muodostaisivat kattavuudeltaan ja usein myös määrältään puutteellisen tutkimusaineiston. Ratkaisuna ongelmaan Pythonille on kehitetty kirjastoja, jotka kykenevät keräämään dataa Twitteristä ohittaen sen API:n. Kirjasto nimeltä Snscrape mahdollistaa Twitterin historiallisen datan keräyksen ilman latausmäärien rajoituksia. Snscrape toimii melkein samalla tavalla, kuin Twitterin hakupalkki. Snscrapen Twitter-moduuliin kirjoitetaan haluttu hakutermi samoin kuin se kirjoitettaisiin Twitterin hakupalkkiin. Haetut twiitit voidaan sitten tallentaa csv- tai json-tiedostoihin. (Bettenbuk 2022.)

5 Käytetyt tutkimusmenetelmät

Tässä osuudessa esitellään opinnäytetyössä käytetyt tutkimusmenetelmät ja perustellaan niiden valinta. Lisäksi esitellään muutamia asioita, jotka voivat vinouttaa menetelmien tuottamia tuloksia, ja ne pitää sen vuoksi huomioida arvioidessa tutkimuksen tuottamien tuloksien luotettavuutta.

5.1 Kvantitatiivinen tutkimus ja otantatutkimus

Kvantitatiivinen eli määrällinen tutkimus pyrkii selvittämään lukumääriin ja prosenttiosuuksiin liittyviä kysymyksiä. Tutkittavaa asiaa kuvataan numeerisesti, ja tuloksia havainnollistetaan taulukoin tai kuvioin. Tavallista on myös, että halutaan selvittää asioiden välisiä riippuvuuksia tai tutkittavassa asiassa tapahtuneita muutoksia. Kvantitatiivisella tutkimuksella pystytään yleensä kartoittamaan olemassa oleva tilanne, mutta sillä ei pystytä riittävästi selvittämään asioiden syitä. (Heikkilä 2014, 15.)

Tutkimuksen perusjoukko on tutkimuksen kohteena oleva joukko, josta halutaan kerätä tietoa tutkittavan ilmiön kuvaamiseksi. Tutkimus voi olla joko kokonaistutkimus, jolloin tutkittava aineisto kattaa koko perusjoukon tai otantatutkimus, jolloin aineistona on perusjoukosta poimittu otos. Käyttäessä otantatutkimusta otoksen tulee edustaa perusjoukkoa mahdollisimman hyvin. Otoksen edustavuus merkitsee sitä, että perusjoukosta poimitussa otoksessa on samoja ominaisuuksia samassa suhteessa kuin perusjoukossa. Poimitun otoksen on vastattava perusjoukkoa tutkittavien asioiden suhteen. Kvantitatiivinen tutkimus edellyttää lukumäärältään suurta ja edustavaa otosta, jotta voidaan luottaa siihen, että otoksesta laskettujen numeeristen tunnuslukujen avulla voidaan estimoida eli tilastollisen päättelyn keinoin arvioida perusjoukon vastaavia tunnuslukuja. (Heikkilä 2014, 12–32.)

Otantatutkimuksen edustavuuden saavuttamiseksi olennaista on otannan satunnaistaminen. Otokseen tulevien yksiköiden tulee määräytyä sattumanvaraisesti. Käytettävä otantamenetelmä harkitaan tarkasti, sillä virheellinen otantamenetelmä saattaa vinouttaa tuloksia. *Todennäköisyysotannassa* jokaisella perusjoukon alkiolla on otosta kerätessä tunnettu poimintatodennäköisyys. Käytännössä otos kuitenkin joudutaan keräämään enemmän tai vähemmän harkinnanvaraisesti. Tällöin puhutaan *ei-todennäköisyysotannasta*. Tällaisen otantamenetelmän perusteella ei saada edustavia otoksia eikä voida johtaa yleistyksiä perusjoukkoon. Jos otosaineisto valitaan jollain muulla tavalla, kuin todennäköisyysotantaa käyttäen on otosta syytä kutsua *näytteeksi*. Tilastollista päättelyä ja tarkkojen estimaattien laskemista perusjoukon suureille ei voi näytteen perusteella tehdä luotettavasti. Näytettä voidaan kuitenkin käyttää esimerkiksi ideoiden tuottamisessa varsinaiselle kyselylle. (Heikkilä 2014, 32–39.)

Tämä opinnäytetyö on kvantitatiivinen tutkimus, sillä tutkittavaa ilmiötä eli terveydenhuollon digipalveluihin liittyvää kansalaispidettä pyritään kuvaamaan nimenomaan numeeristen tunnuslukujen avulla. Tunnuslukuja ovat esimerkiksi twiittiaineistosta lasketut tunteiden keskiarvot ja osuudet, joiden avulla pyritään arvioimaan palveluihin liittyviä mielipiteitä. Lisäksi opinnäytetyö on otantatutkimus, jonka perusjoukkona ovat suomalaiset, joiden mielipiteitä halutaan tutkia. Tutkimukseni aineisto kerätään Twitteristä, jonka vuoksi todennäköisyysotanta ei ole mahdollinen. Tämän vuoksi otosta tulee pitää näytteenä.

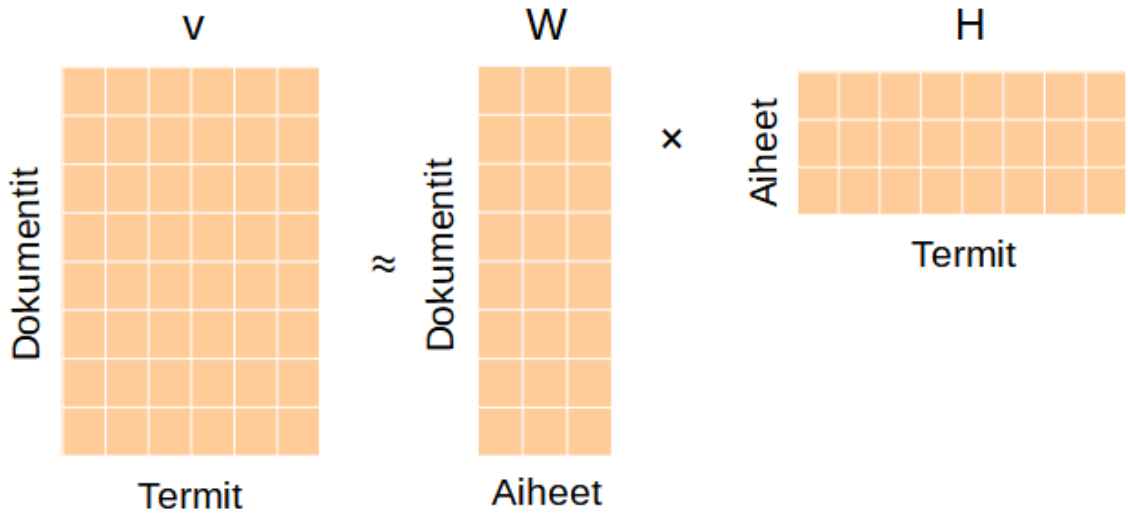
5.2 Tutkimusmenetelmien validiteetti

Validiteetti eli pätevyys kuvaa sitä, missä määrin tutkimuksessa on onnistuttu mittaamaan sitä, mitä oli tarkoitus mitata. Validius tarkoittaa karkeasti systemaattisen virheen puuttumista mittaustuloksista, jolloin validilla mittarilla tehdyt mittaukset ovat keskimäärin oikeita. Systemaattisen virhe vääristää tuloksia, sillä virheen vaikutus mittaustuloksiin ei vähene otoskoon kasvaessa, ja sen suuruutta on vaikea arvioida. (Heikkilä 2014, 27.)

Opinnäytetyössäni tutkimuksen validiteettia heikentää se, että aineisto ei ole edustava otos tutkittavasta perusjoukosta eli suomalaisista. Twitteristä voidaan kerätä vain niiden ihmisten tekstejä, jotka ovat kirjoittaneet twiitteja liittyen terveydenhuollon digipalveluihin. Oletettavasti vain pieni osa suomalaisista on kirjoittanut tällaisia twiitteja. Lisäksi toiset voivat olla Twitterissä hyvinkin aktiivisia kirjoittamaan palveluista. Tämän vuoksi suuri osa suomalaisista ei voi valikoitua otokseen, kun taas aktiiviset twiittaajat voivat tulla useaan kertaan otokseen. Lisäksi suurin osa sosiaaliseen mediaan jaetusta sisällöstä on positiivista, ja eniten sisältöä jakavat miehet ja nuoret ihmiset (Strengell & Sigg 2018). Näin ollen aineiston edustavuus on niin heikko, että sen perusteella ei voi luotettavasti estimoida Suomessa vallitsevaa kansalaismielipidettä palveluita kohtaan. Aineiston pohjalta saatuja tuloksia voi pitää lähinnä suuntaa antavina.

5.3 Aihemallinnus epänegatiivisella matriisifaktorisoinnilla

Aihemallinnusta varten on kehitetty useita menetelmiä. Yksi näistä menetelmistä on *epänegatiivinen matriisifaktorisointi* NMF *(non-negative matrix factorization)*. Tämä valittiin tutkimuksen aihemallinnuksen menetelmäksi, sillä Sarkarin (2019) mukaan se antaa erinomaisia tuloksia, vaikka onkin suhteellisen uusi menetelmä. Menetelmässä faktorisoidaan TF-IDF-mallilla saatu dokumentti-termimatriisi. Hyödyntämällä lineaarialgebraa tämä matriisi voidaan esittää kahden epänegatiivisen matriisin *(faktorin)* tulona. Olkoon dokumentti-termimatriisi V, ja faktorit W ja H. Nyt V voidaan esittää kuvan 2 mukaisena matriisitulona. (Albrecht ym. 2020.)



Kuva 2: Epänegatiivisen matriisifaktorisoinnin laskukaava

Kuvan *dokumentti-aihematriisi* W määrittää dokumenttien ja aiheiden välisen yhteyden. Tämän matriisin solujen sisältämät luvut – *aihearvot (topic dominance scores)* – ilmoittavat, missä määrin aiheet ilmenevät dokumenteissa. Aihearvojen avulla voidaan määrittää kullekin dokumentille hallitseva aihe, mutta ne eivät määritä aiheiden muodostamia prosenttiosuuksia dokumenteista, vaan ne ovat absoluuttisia lukuja. (Sarkar 2019.)

Epänegatiivisessa matriisifaktorisoinnissa aiheet ovat jakauma aineiston termejä, ja *aihe-termimatriisi* H määrittää, minkälainen jakauma on eri aiheilla. Matriisin solujen sisältämät luvut ilmoittavat, kuinka suuren prosenttiosuudet termit eli sanat muodostavat aiheista. Tältä pohjalta voidaan antaa tulkinta aiheille. Analysoitavat dokumentit sisältävät tavallisesti runsaasti tekstiä, jolloin useimpien termien osuudet jäävät pieneksi. Kuitenkin, jos löydetään termejä, jotka muodostava suuria osuuksia aiheesta, niin nämä termit ovat aiheelle tärkeitä ja määrittelevät aiheen olemuksen. Lisäksi osuudet ovat mittari mallin laadulle. Jos osuudet laskevat nopeasti kuljettaessa tärkeimmistä termeistä vähemmän tärkeisiin, niin aihe on hyvin määritelty. Vastaavasti, jos osuudet ovat hyvin samankaltaisia, niin aihe on vähemmän selväpiirteinen. (Albrecht ym. 2020.)

Käytettäessä epänegatiivista matriisifaktorisointia aiheiden määrä – matriisin W sarakkeet ja matriisin H rivit – voidaan valita mielivaltaisesti. Tällöin tutkija voi kokeilemalla päätellä, millä aiheiden määrällä saadaan laadukkain aihemalli. (Albrecht ym. 2020.)

Tässä tutkielmassa epänegatiivista matriisifaktorisointia käytetään löytämään aiheita twiittiaineistosta. Tavoitteena on löytää mahdollisimman tarkasti määriteltyjä terveydenhuollon digipalveluihin liittyviä aiheita sekä määrittää jokaiselle twiitille hallitseva aihe, jotta twiitit voidaan luokitella aiheisiin. Epänegatiivinen matriisifaktorisointi tehdään Pythonin Scikit-Learn-kirjaston NMF-moduulilla.

Aiheiden määrä valitaan tekemällä aihemallinnusta iteratiivisesti katsoen, millaisia tärkeitä termejä saadaan aiheisiin eri aihemäärillä ja miten hyvin määriteltyjä aiheet ovat.

5.4 Tunneanalyysi VADER-sanastolla

Tutkielman aineistona on twiittidata, joka luonnollisesti sisältää emojeja ja lyhenteitä (esim. *LOL* ja *WTF*), joilla on usein merkityksellinen tunnelataus. Tällöin on perustelua pohjata tunneanalyysi VADER *(Valence Aware Dictionary and sEntiment Reasoner)* -tunnesanastoon. VADER on muokattu soveltumaan erityisesti sosiaalisen median tekstien analysointiin. VADER kykenee huomioimaan lyhenteet ja emojit, ja sen tarkkuus arvioitaessa sosiaalisen median tekstejä on huomattavasti parempi kuin muilla tunneanalyysin menetelmillä. VADER:in sanasto antaa tunnearvon 7500 englanninkieliselle sanalle. Tunnearvot on asetettu asteikolle [-4, 4], jossa -4 on erittäin negatiivinen, 4 on erittäin positiivinen ja 0 on neutraali. Kokonaisen lauseen tai tekstin sisältämä tunne lasketaan yhdistelmänä näistä sanoista, jolloin saadaan *yhdistetty tunnearvo (aggregated sentiment score)*, joka asetetaan asteikolle [-1, 1], jossa -1 on erittäin negatiinen, 1 on erittäin positiivinen, ja 0 on neutraali. (Sarkar 2019; Strengell & Sigg 2018.)

Twiitit halutaan lisäksi luokitella niiden sisältämien tunteiden mukaan. Tarkoituksena on jakaa twiitit negatiivisiin, positiivisiin ja neutraaleihin tunneluokkiin. VADER suosittelee määrittelemään positiiviseksi dokumentin, jolla yhdistetty tunnearvo > 0,5, neutraaliksi tunnearvolla [-0,5, 0,5], ja negatiiviseksi tunnearvolla < -0,5. Sarkar (2019.) Tässä tutkielmassa toimimme kuitenkin Sarkarin (2019) ehdottamalla tavalla ja valitsemme kynnysarvoiksi -0,4 ja 0,4.

Tutkielmassa halutaan määrittää yhdistetty tunnearvo ja tunneluokka jokaiselle aineiston twiitille. Pythonissa VADER löytyy NLTK-kirjastosta. Koska aihemallinnuksen ansiosta tiedetään jokaisen twiitin hallitseva aihe, twiitit voidaan luokitella aiheisiin ja laskea kunkin aiheen twiiteille keskimääräinen yhdistetty tunnearvo sekä selvittää, miten twiitit jakautuvat kolmeen tunneluokkaan aiheittain.

6 Tutkimusaineiston keräys ja esiprosessointi

Kuten aiemmin mainittiin, tekstianalyysia voidaan pitää kolmivaiheisena prosessina. Tässä luvussa esitellään, kuinka tutkimuksessa suoritettiin kaksi ensimmäistä vaihetta eli tutkimusaineiston keräys Twitteristä ja aineiston tekstin esiprosessointi. Esiprosessointi tehtiin valittujen tekstinlouhinnan menetelmien – aihemallinnus ja tunneanalyysi – tarpeita silmälläpitäen.

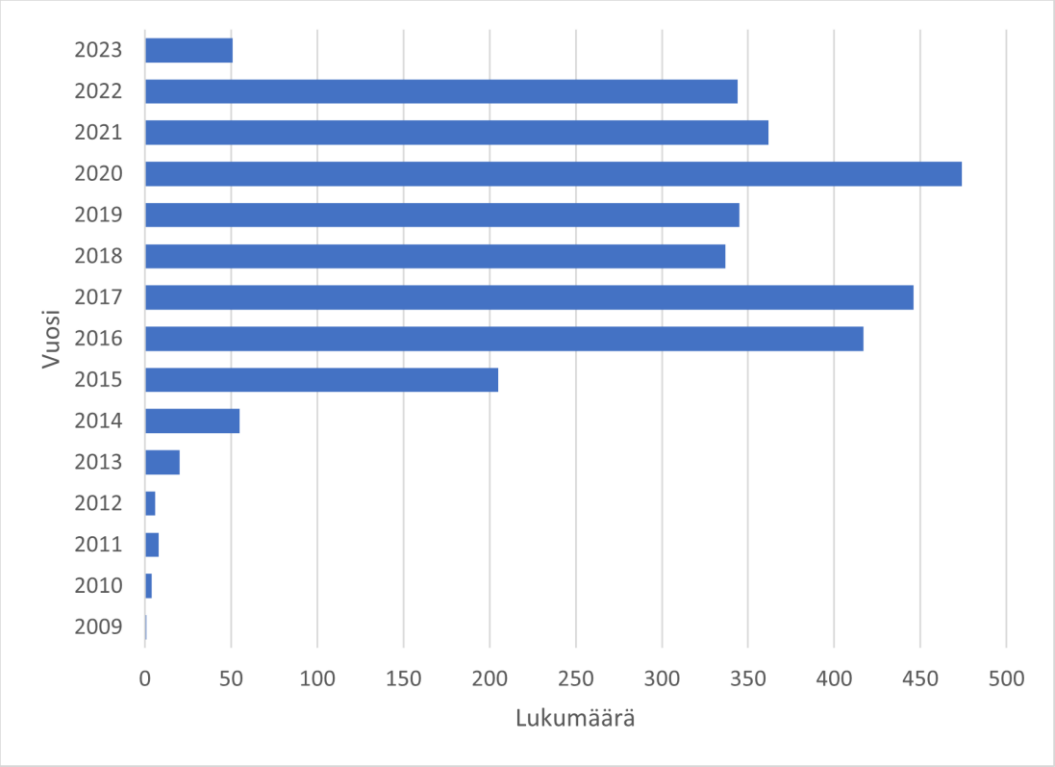
Valitsin opinnäytetyöni ainoaksi ohjelmointityövälineeksi Pythonin, sillä sen kirjastojen avulla voidaan suorittaa kaikki tekstianalyysiin liittyvät tehtävät. Lisäksi Pythonin tekstipohjainen käyttöliittymä mahdollistaa komentojen yksityiskohtaisen määrittelyn. Pythonin käyttöä tekstianalyysissa on myös käsitelty runsaasti kirjallisuudessa. Tutkimuksessani käytetyt Python-koodit ovat katsottavissa Github-sivullani https://github.com/bgj377/Tutkimus-Tekstinalalyysi.

6.1 Aineiston keräys

Tutkimuksen twiittiaineisto kerättiin Twitteristä Pythonin Snscrape-työkalun avulla 20.4.2023. Ennen aineiston keräystä täytyi kuitenkin ensin valita hakutermit, joilla twiitit haetaan. Twitterin verkkosivuilla on hakupalkki, johon syötetään hakutermejä. Haku palauttaa vain ne twiitit, jotka sisältävät jokaisen hakutermiin kuuluvan sanan. Potentiaalisia hakutermejä etsittiin syöttämällä hakupalkkiin suomenkielisiä terveydenhuollon digipalveluihin liittyviä sanoja. Hakutermien tuottamia twiitteja läpikäymällä voitiin arvioida hakutermien toimivuutta. Lisäksi nämä twiitit ja niihin liittyvät aihetunnisteet auttoivat löytämään lisää potentiaalisia hakutermejä. Tämän prosessin kautta valittiin hakutermit, joita käytettiin Snscrape-hauissa. Valitut hakutermit ja löydettyjen twiittien lukumäärät ovat liitteessä 1.

Tutkielman tavoitteena on tutkia suomalaisten kansalaismielipidettä terveydenhuollon digipalveluita kohtaan, jonka vuoksi tutkimusaineistoon halutaan vain twiitteja, jotka ovat yksityisten kansalaisten kirjoittamia. Merkittävä osa aihetta käsittelevistä twiiteista on terveysalan organisaatioiden ja ammatinharjoittajien kirjoittamia. Jotta aineisto kuvaisi nimenomaan yksityisten kansalaisten mielipiteitä, terveysalan toimijoiden kirjoittamat twiitit tulisi mahdollisuuksien mukaan poistaa. Tosiasia on kuitenkin se, että kaikkien näiden twiittien poistaminen aineistosta vaatisi sitä, että haetut twiitit käydään yksitellen läpi käsityönä, joka olisi työlästä suuren aineiston kanssa. Tässä tutkimuksessa tyydytään siihen, että syötetään valittuja hakutermejä Twitterin hakupalkkiin ja tunnistetaan terveysalan toimijoita, joiden kirjoittamat twiitit esiintyivät usein hakutuloksissa. Näin tunnistettiin yhteensä kaksikymmentä liitteessä 2 olevaa terveysalan organisaatiota ja ammatinharjoittajaa, joiden kirjoittamat twiitit poistetaan aineistosta. Aineiston kooksi tuli 3075 twiittia, kun näiden toimijoiden twiitit ja kaksoiskappaleet oli poistettu aineistosta.

Terveydenhuollon digipalveluiden yleistyessä niiden saatavuus ja kattavuus lisääntyvät, jolloin on syytä olettaa, että myös niihin liittyvät mielipiteet voivat muuttua. Tutkimuksen tavoitteena on kerätä tietoa mielipiteistä nykyhetken palveluita kohtaan, jolloin olisi toivottavaa, että aineiston twiitit eivät ole liian kaukaa menneisyydestä. Aineisto olisi voitu rajata vain twiitteihin, jotka on kirjoitettu parin viimeisen vuoden aikana, mutta tämä olisi supistanut aineistoa liikaa. Automaattisen tekstianalyysin käyttö on perusteltua vain, kun aineiston tekstimäärä on niin suuri, että sen läpikäynti manuaalisesti olisi liian työlästä. Kuvassa 3 on kerättyjen twiittien lukumäärät niiden luontivuoden mukaan. Kuvasta nähdään, että selkeä enemmistö twiiteista on alle kymmenen vuotta vanhoja.



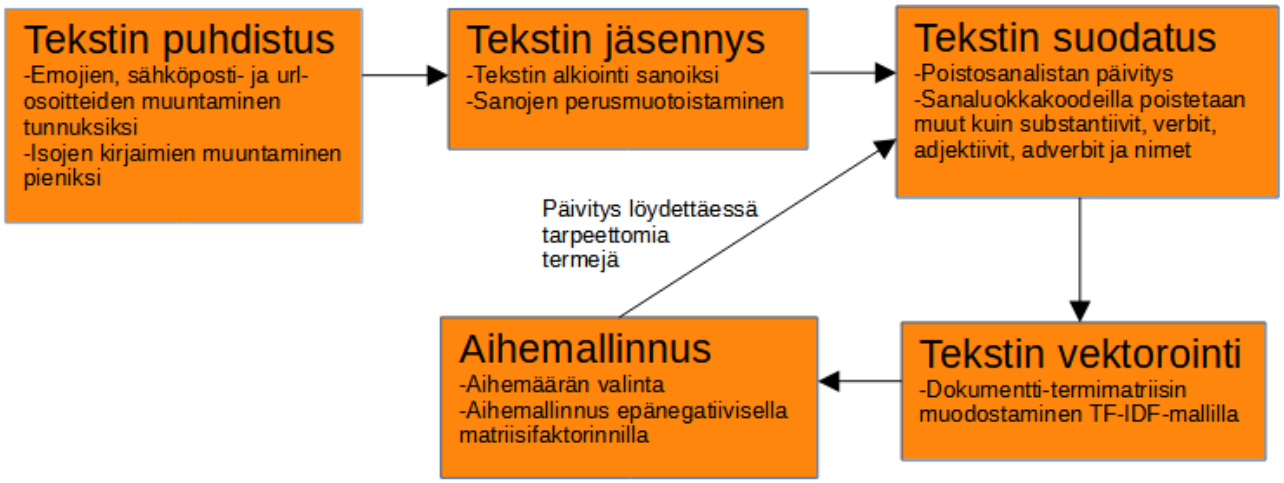
Kuva 3: Aineiston twiittien lukumäärät luontivuoden mukaan

6.2 Aineiston esiprosessointi

Aihemallinnus ja tunneanalyysi ovat erilaisia tekstinlouhinnan menetelmiä, ja niillä on hyvin erilaiset tekstidatan esiprosessoinnin tarpeet. Aihemallinnusta varten tehtävä esiprosessointi sisältää kaikki aiemmin mainitut esiprosessoinnin vaiheet, kun taas tunneanalyysia varten tehtävä esiprosessointi on huomattavasti kevyempi ja useimmat esiprosessoinnin vaiheet voidaan ohittaa. Tämän vuoksi aineisto esiprosessoitiin erikseen kumpaakin menetelmää varten.

6.2.1 Esiprosessointi aihemallinnusta varten

Kuvassa 4 on esitelty aihemallinnusta edeltävät tekstin esiprosessoinnin vaiheet. Esiprosessointi aihemallinnusta varten noudattaa luvussa 4.1 kuvattuja esiprosessoinnin vaiheita. Twiitteja tarkastelemalla havaittiin, että ne sisältävät usein URL-osoitteita, sähköpostiosoitteita ja emojeja, jotka ovat aihemallinnukselle rasitteita. Tekstin puhdistuksessa nämä muunnettiin yhtenäiseen tunnusmuotoon _URL_, _EMAIL_ ja _EMOJI_. Tämä tehtiin Pythonin textacy-kirjaston avulla, joka kykenee suorittamaan useita luonnollisen kielen prosessointiin liittyviä tehtäviä ja on suunniteltu toimimaan yhdessä spaCyn kanssa (Albrecht ym. 2020). Teksti jäsennettiin – alkiointi ja perusmuotoistaminen – spaCyn avulla. Tekstissä haluttiin säilyttää vain sanoja, jotka ovat hyödyllisiä aihemallinnukselle. Tämän vuoksi tekstistä suodatettiin pois sanaluokkakoodeja käyttämällä sanat, jotka eivät olleet substantiiveja, verbejä, adjektiiveja, adverbeja tai nimiä. Lisäksi suodatettiin sanat, jotka olivat spaCyn suomenkielisellä poistosanalistalla, ja listaan lisättiin sähköpostien, emojien ja URL-osoitteiden tunnukset. Jäljelle jäänyt teksti vektoroitiin ja hyödynnettiin aihemallinnuksessa.



Kuva 4: Aihemallinnusta edeltävät tekstin esiprosessoinnin vaiheet

Aihemallinnuksen suorituksen jälkeen havaittiin, että saatujen aiheiden tärkeiksi termeiksi tuli terveydenhuollon digipalveluihin liittymättömiä sanoja. Tämä on ymmärrettävää, sillä aiheiden tärkeiksi termeiksi voi valikoitua mitä tahansa twiiteissa usein yhdessä esiintyviä sanoja. Ongelman korjaamiseksi tekstin esiprosessointi tehtiin uudestaan lisäämällä ei-toivotut tärkeät termit poistosanalistaan. Aihemallinnusta tehtiin iteratiivisesti päivittämällä poistosanalistaa, kunnes löydettyjen aiheiden tärkeät termit olivat aiheeseen kuuluvia.

6.2.2 Esiprosessointi tunneanalyysia varten

Tunneanalyysin vaatima esiprosessointi alkaa twiittien kääntämisellä englanniksi, sillä VADER-tunnesanasto sisältää vain englanninkielisiä sanoja. Twiitit käännettiin englanniksi käyttäen Googlen tarjoamaa maksutonta Google Translate-käännöspalvelua Pythonin Googletrans-kirjaston kautta. Tarkastelemalla käännettyjä twiitteja voitiin havaita, että palvelun tekemä käännöstyö on virheetöntä. Seuraavaksi isot kirjaimet muunnettiin pieniksi, ja twiitit alkioitiin Pythonin NLTK-kirjaston avulla. Aineiston sanojen perusmuotoistaminen voitiin ohittaa, sillä VADER sisältää useita taivutusmuotoja sanoista. Tunneanalyysia varten ei tarvitse laatia poistosanalistoja tai suodattaa tekstiä muilla tavoin, sillä vain tunnesanastossa olevat sanat saavat tunnearvon. Tekstin vektorointi voidaan myös ohittaa, sillä twiitin saama yhdistetty tunnearvo riippuu vain kyseisen twiitin sisältämistä sanoista, eikä sen laskeminen näin ollen edellytä dokumentti-termimatriisin luomista.

7 Tekstinlouhinnan tulokset

Esiprosessoinnin jälkeen voitiin siirtyä varsinaiseen analyysivaiheeseen tekemällä aihemallinnus epänegatiivisella matriisifaktorisoinnilla. Ensimmäisenä tehtävänä oli valita aiheiden määrä tutkittavaan malliin. Mallinnus päädyttiin tekemään kuudella aiheella, sillä suurempi määrä aiheita olisi tuottanut vain tutkimuksen kannalta merkityksettömiä aiheita. Taulukossa 1 on lueteltu saatujen aiheiden viisi tärkeintä termiä. Aihenumeron perässä suluissa oleva luku kertoo, kuinka monella twiitilla kyseinen aihe oli hallitseva. Tärkeiden termien perässä suluissa oleva luku ilmoittaa termit prosenttiosuuden aiheesta – aihe-termimatriisin H solu kyseiselle aihe-termiparille. Osuuksia tarkastellessa on muistettava, että ne eivät liity alkuperäisiin twiitteihin, vaan suodatettuihin twiitteihin, joita on typistetty poistosanalistan ja sanaluokkakoodien avulla.

Taulukko 1: Aihemallinnukset tuottamat aiheet ja aiheiden viisi tärkeintä termiä

Aihe 0 (504)	Aihe 1 (485)	Aihe 2 (465)
terveydenhuolto (13,53)	sähköinen (14,73)	etälääkäri (26,49)
digitalisaatio (12,17)	ajanvaraus (14,22)	terveyspalvelu (2,90)
digitaalinen (1,75)	lääkemääräys (0,94)	hoito (2,66)
tekoäly (1,32)	resepti (0,85)	lääkäri (2,19)
lääkäri (0,83)	asiointi (0,85)	potilas (0,95)

Aihe 3 (778)	Aihe 4 (440)	Aihe 5 (403)
etävastaanotto (17,00)	digiterveys (23,58)	terveys (12,18)
lääkäri (3,83)	digitalheath (2,97)	digi (9,32)
potilas (1,21)	terveysteknologia (1,35)	hyvinvointi (2,37)
vastaanotto (1,07)	ehealth (1,15)	digitaalinen (1,40)
kännykkälääkäri (0,72)	digitaalinen (0,98)	hoito (1,08)

Taulukosta huomataan, että aiheet ovat hyvin määriteltyjä, sillä termien prosenttiosuuksia tarkastelemalla havaitaan, että aiheilla on yksi tai kaksi selvästi tärkeintä termiä, joiden jälkeen osuudet laskevat nopeasti. Lisäksi voidaan huomata, että nämä aiheen määrittelevät termit ovat samoja kuin ne twiittien haussa käytetyt hakutermit, joilla kerättiin eniten twiitteja – ks. liite 1. Taulukosta

löytyy myös termejä, jotka eivät olleet hakutermejä, kuten *tekoäly* ja *digitalhealth*, mutta niiden osuudet aiheissa ovat varsin pieniä.

Aiheille voidaan antaa tulkinta tärkeiden termien avulla. Aihe 0 on *terveydenhuollon digitalisaatio*, sillä sanat terveydenhuolto ja digitalisaatio ovat prosenttiosuuksien perusteella aiheen määrittelevät termit. Tämä merkitsee sitä, että twiitit, joiden hallitseva aihe on aihe 0 sisältävät runsaasti näitä kahta sanaa. Vastaavasti aihe 1 on *sähköinen ajanvaraus*, aihe 2 on *etälääkäri*, ja aihe 3 on *etävastaanotto*. Aiheet 4 ja 5 ovat käytännössä samat, ja aiheelle voitaisiin antaa nimi *digitaalisuus terveysasioissa*. Aiheiden 4 ja 5 yhdistelmä on twiittimäärältään suurin aihe sen ollessa hallitseva aihe 843 twiitille. Seuraavaksi suurin aihe oli etävastaanotto, johon kuului vajaat kahdeksansataa twiittia.

Taulukossa 2 on yksi esimerkkitwiitti jokaisesta aiheesta. Taulukkoon valittiin twiitteja, joilla oli suuria aihearvoja – dokumentti-aihematriisin W solu kyseiselle dokumentti-aiheparille.

Taulukko 2: Esimerkkitwiitit jokaisesta aiheesta

Aihe	Aihearvo	Twiitti
0	0.238	#Digitalisaatio muokkaa voimakkaasti tulevaisuuden osaamisvaatimuksia terveydenhuollossa https://t.co/EkWneOU2Q7 #digiosaaminen #digitaidot #terveydenhuolto
1	0.359	Sähköinen ajanvaraus laajenee terveyskeskuksessa vähitellen. Käytössä jo osin ehkäisyneuvolassa. #nkkyselytunti
2	0.361	Rakentuuko etälääkäri palvelut langattomuuden ja kännykän varaan? https://t.co/UkLT1rga0l
3	0.370	Lääkärin 24h etävastaanotto on nykyaikaa. 😷
4	0.361	Mona Mannevuon kolumni: Soten myötä saat omahoitaa itsesi kunnon kansalaiseksi https://t.co/RQBAWLnSkn #sote #digiterveys #tietosuoja
5	0.379	Älyrollaattorit avuksi kuntoutumisen seuraamiseen #digi #terveys #kuntoutuminen https://t.co/Hb11QGFM7j

Esimerkkitwiitteja katsomalla voidaan havaita ilmiö, joka esiintyi laajemminkin aineistossa. Useat taulukon twiiteista eivät olleet mielipiteitä tietyistä terveydenhuollon digipalveluista, vaan luonteeltaan hyvin toisenlaisia. Aineiston twiitteja läpikäymällä havaittiin, että huomattava osuus – ei kuitenkaan kaikkia – twiiteista voitaisiin karkeasti jakaa *latentteihin* (piileviin) *ulottuvuuksiin*. Ensimmäinen ulottuvuus kattaa twiitit, jotka ovat kirjoittajien mielipiteitä palveluista. Mielipiteet voivat

pohjautua kirjoittajien omiin kokemuksiin tai ei. Toinen ulottuvuus kattaa twiitit, jotka ovat käsittelevät terveysteknologian mahdollisuuksia. Kolmas ulottuvuus kattaa twiitit, jotka ovat mainoksia tai tiedonantoja liittyen terveyspalveluihin. Usein kolmannen ulottuvuuden twiitit ovat terveysalan toimijoiden kirjoittamia.

Tarkastelemalla twiitteja edelleen havaitaan, että aiheisiin terveydenhuollon digitalisaatio ja digitaalisuus terveysasioissa kuuluvien twiittien sisällöt viittaavat enimmäkseen toiseen ja kolmanteen latenttiin ulottuvuuteen. Aiheisiin sähköinen ajanvaraus, etälääkäri ja etävastaanotto kuuluvat twiitit viittaavat näiden lisäksi myös ensimmäiseen ulottuvuuteen. Tämä on ymmärrettävää, sillä jälkimmäiset aiheet viittaavat tiettyihin olemassa oleviin terveydenhuollon palveluihin, kun taas ensimmäisenä mainitut aiheet liittyvät yksittäisiä palveluja laajempiin ilmiöihin.

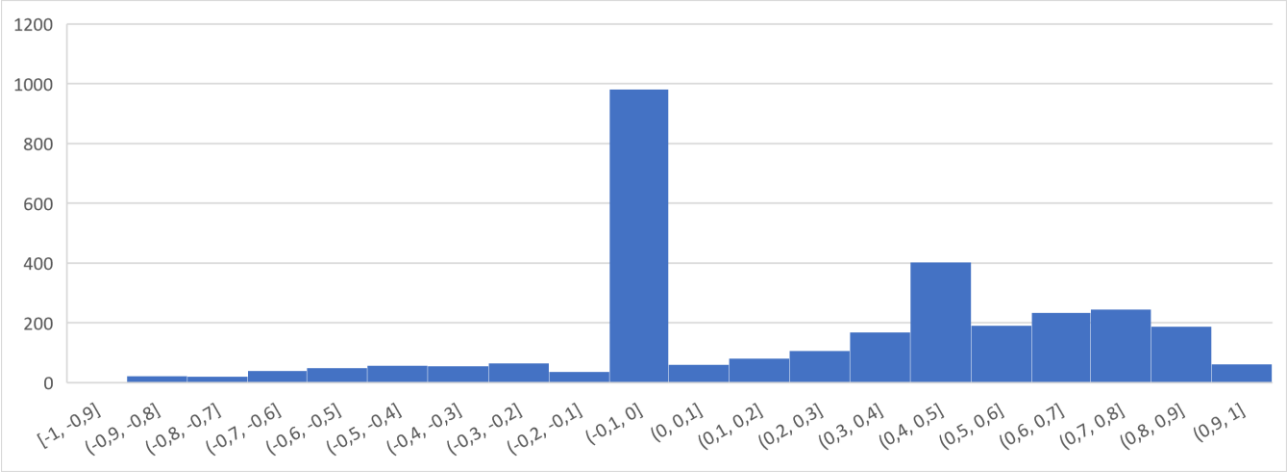
Seuraavaksi tehdään tunneanalyysi VADER-tunnesanaston pohjalta. Tunneanalyysi tuottaa jokaiselle twiitille yhdistetyn tunnearvon, jonka vaihteluväli on [-1, 1]. Löydettyihin aiheisiin liittyvien mielipiteiden arvioimiseksi twiiteille lasketaan aiheittain tiettyjä tilastollisia tunnuslukuja yhdistettyjen tunnearvojen pohjalta. Taulukossa 3 ovat yhdistettyjen tunnearvojen keskiarvo, keskihajonta, suurin ja pienin arvo. Tunnusluvut on laskettu sekä kaikille aineiston twiiteille että aiheittain, kun twiitit on luokiteltu hallitsevan aiheen mukaan.

Taulukko 3: Yhdistetyistä tunnearvoista lasketut tilastolliset tunnusluvut

Aihe	Keskiarvo	Keskihajonta	Minimi	Maksimi
Kaikki	0,256	0,397	-0,949	0,980
0	0,363	0,368	-0,859	0,980
1	0,209	0,389	-0,898	0,955
2	0,168	0,396	-0,913	0,976
3	0,234	0,413	-0,949	0,957
4	0,287	0,355	-0,761	0,976
5	0,287	0,418	-0,866	0,968

Taulukosta nähdään, että kaikissa aiheissa twiittien keskimääräinen tunnearvo on positiivinen. Aiheen 0 (terveydenhuollon digitalisaatio) keskiarvo on kaikkein suurin, ja aiheen 2 (etälääkäri) kaikkein pienin. Nähdään myös, että jokaiseen aiheeseen liittyy tunnearvoltaan hyvin negatiivisia ja positiivisia twiitteja. Aiheet 4 ja 5 (digitaalisuus terveysasioissa) saavat saman keskiarvon, joka oli odotettavaa. Keskihajonta kertoo kuinka kaukana muuttujan arvot keskimäärin ovat keskiarvosta.

Yhdistetyn tunnearvon vaihteluvälin huomioon ottaen voidaan todeta, että twiittien tunnearvojen hajonta on melko suurta. Tunnearvojen jakauman havainnollistamiseksi kuvassa 5 on histogrammi, joka esittää kaikkien aineiston twiittien tunnearvojen jakaumaa. Kuvasta havaitaan, että lähes tuhat twiittia saa tunnearvon nolla, eli niissä ei ole sanoja, jotka VADER-sanaston mukaan sisältäisivät tunteita. Lisäksi havaitaan, että tunnearvoltaan negatiivisia twiitteja on suhteellisen vähän.



Kuva 5: Twiittien yhdistettyjen tunnearvojen jakauma

Twiitit luokiteltiin tunnearvon perusteella positiivisiin, neutraaleihin ja negatiivisiin twiitteihin. Taulukossa 4 ovat näihin tunneluokkiin kuuluvien twiittien prosenttiosuudet koko aineistossa ja luokiteltuna twiittien hallitsevan aiheen mukaan. Huomataan, että tunnelatauksellaan negatiiviset twiitit ovat vähemmistönä kaikissa aiheissa, mikä oli jo kuvan 5 perusteella odotettavissa. Eniten oli neutraaliksi luokiteltuja twiitteja, ja kaikista twiiteista puolet oli neutraaleja. Myös tämä oli kuvan 5 perusteella odotettavissa, sillä tunnearvon nolla saavien twiittien suuri määrä kasvattaa neutraalia luokkaa. Taulukosta voidaan kuitenkin havaita, että aiheessa 0 (terveydenhuollon digitalisaatio) enemmistö twiiteista positiivisia. Vähiten positiivisia ja eniten negatiivisia twiitteja oli aiheessa 2 (etälääkäri). Huomattavaa on kuitenkin se, että edes tässä aiheessa negatiivisten twiittien osuus ei ole poikkeuksellisen suuri. Hieman yllättäen aiheiden 4 ja 5 (digitaalisuus terveysasioissa) saamat osuudet poikkeavat toisistaan.

Taulukko 4: Tunneluokkien prosenttiosuudet twiiteista

Tunneluokka	Kaikki	Aihe 0	Aihe 1	Aihe 2	Aihe 3	Aihe 4	Aihe 5
Negatiivinen	6,3	3,6	6,8	8,0	7,8	3,4	7,2
Neutraali	50,7	40,9	56,1	58,5	51,4	52,3	44,4
Positiivinen	43,0	55,6	37,1	33,5	40,7	44,3	48,4

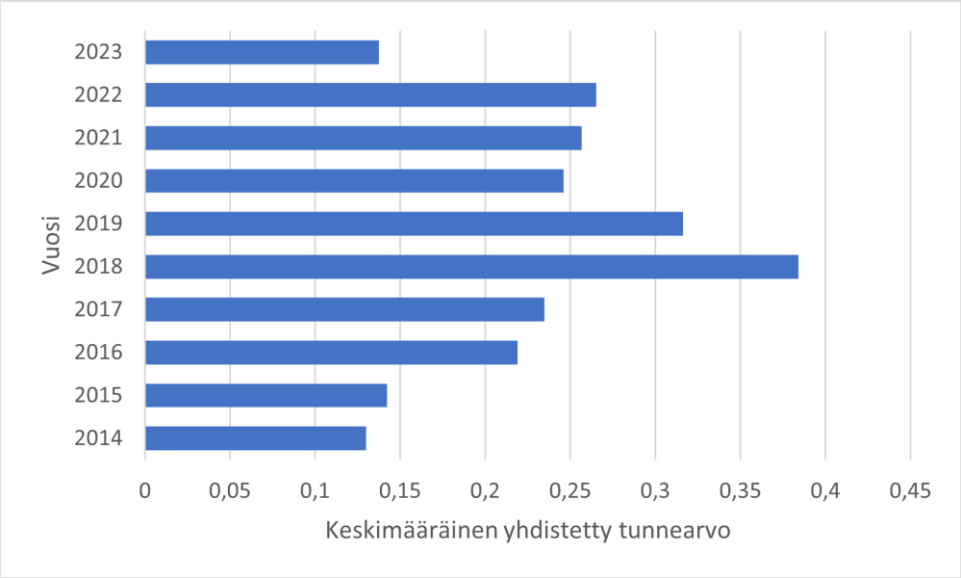
Taulukossa 5 on esimerkkitwiitteja, joilla on suuria positiivisia tai negatiivisia yhdistettyjä tunnearvoja. Havaitaan, että positiivisia tunnearvoja saavat twiitit eivät ole mielipiteitä terveydenhuollon digipalveluista, vaan twiitit korostavat terveysteknologian mahdollisuuksia tai palveluntarjoajien mainoksenomaisia tiedotteita. Positiiviset esimerkkitwiitit kuuluvat siten toiseen ja kolmanteen latenttiin ulottuvuuteen. Tarkastelemalla twiitteja manuaalisesti voidaan havaita, että tämä positiivisia twiitteja koskeva ilmiö esiintyy laajemminkin aineistossa. Negatiivisissa esimerkkitwiiteissa on sen sijaan mielipiteitä palveluista. Lisäksi neljäs esimerkkitwiitti on palvelun käyttäjän kertomus siitä, miten hän koki etälääkärin palvelun. Tämän kaltaiset twiitit ovat erityisen tärkeitä, kun halutaan arvioida kansalaismielipidettä näitä palveluita kohtaan. Aineiston twiittien perusteella voidaan väittää, että tunnearvoltaan negatiiviset twiitit ovat usein mielipiteitä palveluista, eli ne kuuluvat ensimmäiseen latenttiin ulottuvuuteen.

Taulukko 5: Esimerkkejä tunnearvoltaan positiivisista ja negatiivisista twiiteista

Aihe	Tunnearvo	Twiitti
0	0,980	Onnittelut @helsinginsote suunhoito -hienoa työtä👍"Helsingin kaupungin suun terveydenhuolto ja kotihoito haluavat parantaa kotihoidon asiakkaiden suun terveyttä ja ravitsemustilaa. Päätimme toteuttaa kotihoidon henkilöstön koulutuksen verkkokurssina."#digitalisaatio #kotihoito
4	0,976	Onnea @Terveystalo'n tiimi Parhaan käyttökokemuksen palkinnosta @grandonefinland-kisassa! Huikeaa työtä ajanvarauspalvelun kehittämiseksi - asiakkaat palkitsevat työn kehumalla ja käyttämällä palvelua. #digiterveys #kohtiparempaa https://t.co/i99h6F3B9G
5	0,955	Terveys on tärkein voimavaramme ja siitä on pidettävä kiinni! Digitalisoimalla saamme enemmän irti ja vastaamme paremmin vanhenevan väestön tuomiin haasteisiin. Datan ja digin avulla pidämme huolta, että suomalaisten terveys pysyy hyvällä tolalla jatkossakin. #sotedigi #digi https://t.co/oKThU1IjP8
2	-0,858	2020 kesällä en päässyt vastaanotolle kovan yskän vuoksi, alla 3 negatiivista koronatestiä. Etälääkäri diagnosoi 2 min chat-ajan perusteella astman. Ei auttanut astmalääkkeet ja yskin lopulta kylkiluuhun hiusmurtuman.
0	-0,859	Koronapandemia vauhditti maailmanlaajuisesti digitaalisten palveluiden käyttöönottoa syöpäpotilaiden hoidossa. Kukapa meistä enää haluaisi luopua asioinnista verkkopankissa? #digitalisaatio #terveydenhuolto #syöpä
2	-0,875	@akuuttigeriatri @IPiikki Ne kuolee palvelutaloissa, ei avausta. Etälääkäri kirjaa kuolinsyyn. Hautausurakoitsija hakee.

Tunneanalyysi voi joskus tuottaa odottamattomia tuloksia. Viides esimerkkitwiitti on tunnearvoltaan negatiivinen, vaikka twiitissa korostetaan digitaalisten palveluiden mahdollisuuksia tehostaa syöpäpotilaiden hoitoa. Twiitti ei siis ole sisältämältään mielipiteeltä negatiivinen saadusta tunnearvosta huolimatta. Tämän voi olettaa johtuvan siitä, että sana *syöpä (cancer)* saa VADER-sanastossa hyvin negatiivisen arvon.

Tutkimuksessani pyrittiin myös selvittämään, kehittyykö keskimääräinen yhdistetty tunnearvo johonkin suuntaan, eli ovatko uudempien twiittien tunnearvot positiivisempia tai negatiivisempia kuin vanhojen twiittien. Jotta twiitteja voitaisiin analysoida vuosittain, jokaiselle vuodelle pitää olla niin paljon twiitteja, että yksittäiset twiitit eivät vaikuta merkittävästi tunneanalyysin tulokseen. Kuten kuvasta 3 huomattiin, aineistossa on vain vähän twiitteja vuosilta 2009–2013. Varhaisimpina vuosinaan Twitter ei ollut Suomessa erityisen suosittu ja terveydenhuollon digipalvelut eivät olleet suuresti käytettyjä, joten aineistossa on vasta vuodesta 2014 alkaen niin runsaasti twiitteja, että voidaan tehdä vuosittainen tunneanalyysi ilman, että sattumalla olisi merkittävä vaikutus. Vuosittaisessa tunneanalyysissa ei oteta huomioon twiittien aiheita. Kuvassa 6 on esitelty twiittien keskimääräiset yhdistetyt tunnearvot luokiteltuna twiittien luontivuoden mukaan. Kuvasta nähdään, että tunnearvojen kehitykselle ei ole selkeää trendiä. Tunnearvot nousivat trendinomaisesti vuoteen 2018 asti, jolloin keskimääräinen tunnearvo oli melkein 0,4. Sen jälkeen nouseva kehitys katkesi.



Kuva 6: Keskimääräiset yhdistetyt tunnearvot twiittien luontivuoden mukaan

8 Johtopäätökset

Tämän tutkimuksen tavoitteena oli löytää vastaus tutkimuskysymykseen siitä, missä määrin tekstianalyysilla voidaan tuottaa tietoa terveydenhuollon digitaalisiin palveluihin liittyvästä kansalaismielipiteestä. Vastauksen löytämiseksi tähän päätutkimuskysymykseen vastattiin viiteen alatutkimuskysymykseen. Seuraavaksi pohdittiin käytettyjen menetelmien validiteettia eli arvioitiin, mitataanko tutkimusmenetelmillä haluttuja asioita. Saatujen vastausten pohjalta voitiin tehdä yhteenveto, joka vastasi päätutkimuskysymykseen.

8.1 Vastaukset alatutkimuskysymyksiin

Ensimmäisenä alatutkimuskysymyksenä oli selvittää, mitkä ohjelmat kannattaa valita tekstianalyysin eri vaiheiden suorittamiseen. Tähän tarkoitukseen on saatavilla useita ohjelmia, mutta päädyin valitsemaan Pythonin ainoaksi ohjelmointityövälineeksi. Ensinnäkin datan keräystä ei enää voitu suorittaa ohjelmilla, jotka keräävät twiitteja Twitterin API:n kautta, sillä tämä mahdollisuus lakkautettiin Twitterin toimesta vuonna 2023. Pythonin käyttöä puolsivat myös ne seikat, että Python tarjoaa lukuisia tekstianalyysia varten tehtyjä kirjastoja, Pythonin tekstipohjainen käyttöliittymä mahdollistaan komentojen yksityiskohtaisen hallinnan, ja Pythonin käyttöä tekstianalyysissa käsitellään runsaasti kirjallisuudessa, mikä tarjoaa runsaasti ohjemateriaalia.

Toisena alatutkimuskysymyksenä oli selvittää, minkälaiset aiheet esiintyvät usein terveydenhuollon digipalveluita käsittelevissä twiiteissa. Tarkoituksena oli tuottaa tietoa siitä, mitkä ovat suomalaisille tärkeitä palveluihin liittyviä aiheita. Erityisenä toiveena oli löytää palveluiden luotettavuuteen ja eettisyyteen liittyviä aiheita. Valitettavasti kaikki aihemallinnuksen tuottamat aiheet vastasivat twiittien haussa käytettyjä hakutermejä. Kaikki hakutermit, joilla kerättiin Twitteristä useita satoja twiitteja, näkyivät aihemallinnuksessa aiheina, joita olivat *terveydenhuollon digitalisaatio, sähköinen ajanvaraus, etälääkäri, etävastaanotto ja digitaalisuus terveysasioissa*. Näin ollen aihemallinnus ei kyennyt tuottamaan hakutermien ulkopuolisia aiheita tai uutta tietoa kansalaisille tärkeistä palveluihin liittyvistä aiheista. Tiedetään, että aihemallinnuksen tuottamat aiheet ovat termijakaumia, joiden tärkeiksi termeiksi valikoituu aineistossa usein yhdessä esiintyviä sanoja. Koska aihemallinnus ei löytänyt uusia tutkimuksen kannalta hyödyllisiä aiheita, voidaan olettaa, että aineisto ei sisältänyt tarpeeksi suuria määriä palveluita koskevia sanoja muodostamaan tällaisia aiheita.

Kolmantena alatutkimuskysymyksenä oli selvittää, minkälaisista terveydenhuollon digipalveluihin liittyvistä aiheista viestitään eniten Twitterissä. Twiittimääriltään suurimmat aiheet olivat etävastaanotto ja digitaalisuus terveysasioissa. Nämä aiheet ovat hyvin erilaisia ja sisältävät erityyppisiä twiitteja. Digitaalisuus terveysasioissa sisältää paljon twiitteja, jotka käsittelevät yleisellä tasolla terveyttä ja digitaalisuutta. Näin ollen aihe on twiittisisältönsä puolesta hyvin laaja ja täsmentymätön.

Etävastaanoton merkitystä aiheena sen sijaan selittää kyseisen palvelun suosio ja tärkeys kansalaisten keskuudessa. Etävastaanotto on virtuaalinen vaihtoehto perinteiselle lääkärin vastaanotolle. Etävastaanoton aikana asiakas vastaanottaa ohjeita ja diagnooseja lääkäriltä netin Internetin välityksellä. Olisi perusteltua olettaa, että asiakkaat kommentoivat saamaansa palvelua Twitterin välityksellä.

Neljäntenä alatutkimuskysymyksenä oli selvittää, miten twiittien aihe vaikuttaa niiden sisältämään tunnelataukseen. Tarkoituksena oli vertailla aiheita ja tuottaa tietoa siitä, minkälaiset tekijät vaikuttavat twiitteihin sisältyviin tunteisiin. Kysymykseen vastaamiseksi twiiteille laskettiin aiheittain keskimääräinen yhdistetty tunnearvo sekä luokiteltiin twiitit aiheittain positiivisiin, negatiivisiin ja neutraaleihin luokkiin. Tämän jälkeen vertailtiin keskimääräisiä tunnearvoja ja twiittien jakautumista luokkiin aiheiden välillä. Havaittiin, että aiheella terveydenhuollon digitalisaatio on merkittävästi muita suurempi yhdistetty tunnearvo, ja se oli ainoa aihe, jonka twiiteista suurin osa luokiteltiin positiivisiksi. Muiden aiheiden väliset erot olivat pienemmät, ja enemmistö twiiteista luokiteltiin neutraaleiksi.

Aineiston twiittien manuaalisen tarkastelun kautta huomattiin, että twiitit voitiin tietyissä määrin jakaa *latentteihin ulottuvuuksiin*. Ensimmäiseen ulottuvuuteen kuuluivat twiitit, jotka olivat twiittaajien mielipiteitä palveluista, toiseen ulottuvuuteen kuuluivat terveysteknologian mahdollisuuksia käsittelevät twiitit, ja kolmanteen ulottuvuuteen kuuluivat twiitit, jotka olivat palveluihin liittyviä tiedonantoja tai mainoksia. Näiden latenttien ulottuvuuksien avulla pystyttiin tulkitsemaan tunneanalyysin tuottamia tuloksia.

Latentit ulottuvuudet ja aiheen luonne saattoivat tietyssä määrin selittää aiheen sisältämien twiittien tunnearvoja. Aiheet voitaisiin jakaa palveluita käsitteleviin aiheisiin sähköinen ajanvaraus, etälääkäri ja etävastaanotto sekä ilmiöitä käsitteleviin aiheisiin terveydenhuollon digitalisaatio ja digitaalisuus terveysasioissa. Huomattiin, että ilmiöitä käsittelevillä aiheilla keskimääräiset tunnearvot ja positiivisten twiittien osuudet olivat suurempia kuin palveluita käsittelevillä aiheilla. Tämän voi ajatella johtuvan siitä, että ilmiöitä käsitteleviin aiheisiin kuuluvat twiitit voidaan luokitella enimmäkseen toiseen ja kolmanteen latenttiin ulottuvuuteen, sillä nämä twiitit harvoin käsittelevät jotain tiettyä palvelua. Näiden ulottuvuuksien luonteen vuoksi niihin kuuluvien twiittien voi olettaa olevan usein tunteiltaan positiivisia. Sen sijaan palveluita käsitteleviin aiheisiin kuuluvista twiiteista monet voidaan luokitella ensimmäiseen ulottuvuuteen, vaikka kaksi muuta ulottuvuutta ovat myös näissä aiheissa hyvin edustettuina. Ensimmäinen ulottuvuus kattaa twiitit, jotka sisältävät mielipiteitä palveluista. Luonnollisesti mukana on positiivisia ja negatiivisia mielipiteitä. Aineistoa tarkastelemalla havaittiin, että negatiivisia twiitteja esiintyi varsinkin silloin, kun kirjoittaja oli kokenut saaneensa

huonoa palvelua. Negatiiviset mielipiteet palveluita kohtaan voivat selittää, miksi palveluita käsittelevät twiitit olivat keskimäärin vähemmän positiivisia kuin ilmiöitä käsittelevät twiitit.

Latenttien ulottuvuuksien avulla voidaan myös selittää, miksi aihe terveydenhuollon digitalisaatio sain merkittävästi positiivisempia tunnearvoja kuin muut aiheet. Aiheen voi olettaa sisältävän paljon twiitteja, jotka kuuluvat toiseen latenttiin ulottuvuuteen eli käsittelevät terveysteknologian mahdollisuuksia. Tarkemmin sanottuna twiitit käsittelevät digitalisaation tarjoamia mahdollisuuksia terveydenhuollon kehittämisessä. Tämänkaltaiset twiitit ovat oletusarvoisesti tunteiltaan hyvin positiivisia, joka näkyy saaduissa tunnearvoissa.

Viidentenä alatutkimuskysymyksenä oli selvittää, liittyykö tunteiden kehitykseen nouseva tai laskeva trendi. Yhdenmukaista trendiä ei kuitenkaan löytynyt. Havaittiin, että keskimääräinen tunnearvo oli korkeimmillaan vuonna 2018, mutta huomattavasti matalampi vuodesta 2020 eteenpäin. Koronapandemialla saattoi olla vaikutuksensa tunnearvojen kehitykseen. Pandemian alettua Suomessa vuonna 2020 digitaalisten terveyspalvelujen tarve nousi yllättäen, jolloin palveluiden mahdolliset puutteet tulivat entistä näkyvämmiksi.

8.2 Validiteetin arviointia

Lähdettäessä tulkitsemaan tuloksia on tärkeää omata käsitys käytettyjen tutkimusmenetelmien validiteetista eli siitä, missä määrin valituilla menetelmillä kyetään mittaamaan sitä, mitä oli tarkoitus mitata. Tuloksista tehtävien johtopäätösten luotettavuus riippuu menetelmien validiteetista. Aiemmin todettiin, että käytetty tutkimusaineiston keräysmenetelmä ei voi tuottaa edustavaa otosta tutkimuksen kohdejoukosta eli suomalaisista. Lisäksi tarkastelemalla aineiston twiitteja voitiin havaita, että twiittien laatu saattaa heikentää tunneanalyysin validiteettia.

Twiittien tarkastelu osoitti, että huomattava osa twiiteista kuului toiseen tai kolmanteen latenttiin ulottuvuuteen, eli ne käsittelivät terveysteknologian mahdollisuuksia tai olivat palveluihin liittyviä tiedonantoja tai mainoksia. Tutkimuksen validiteetille olisi tärkeää, että mahdollisimman suuri osa twiiteista kuuluisi ensimmäiseen ulottuvuuteen, eli ne olisivat mielipiteitä palveluista. Toiseen ja kolmanteen ulottuvuuteen kuuluvat twiitit ovat usein tunteiltaan positiivisia tai neutraaleja, kun taas ensimmäiseen ulottuvuuteen kuuluvat twiitit voivat mielipiteinä palveluista olla hyvinkin negatiivisia.

Validiteetin kannalta olisi toivottavaa, että twiitit olisivat enimmäkseen tavallisten kansalaisten kirjoittamia, jotta tulokset antaisivat tietoa vallitsevasta kansalaismielipiteestä. Tätä varten aineistosta poistettiin twiitteja, jotka olivat terveysalan ammatinharjoittajien ja organisaatioiden kirjoittamia. Kyseessä olivat kuitenkin vain Twitterissä aktiivisimpien terveysalan toimijoiden kirjoittamat twiitit. Aineistoon jäi poistojen jälkeenkin runsaasti terveysalan toimijoiden kirjoittamia twiitteja. Tämä heikentää validiteettia, sillä terveysalan toimijoiden twiitit ovat sisällöltään erilaisia kuin kansalaisten

kirjoittamat twiitit. Twiitteja tarkastelemalla voitiin havaita, että terveysalan toimijoiden kirjoittamat twiitit kuuluivat usein kolmanteen ulottuvuuteen, joka vinouttaa tunneanalyysin tuloksia positiiviseen suuntaan.

Vertailemalla yhden esimerkkitwiitin sisältöä ja yhdistettyä tunnearvoa havaittiin, että twiitin yhdistetty tunnearvo ei vastannut kirjoittajan twiitille tarkoittamaa tunnetta. Sanastopohjainen tunneanalyysi laskee teksteille yhdistetyn tunnearvon yksittäisten sanojen tunnearvojen yhdistelmänä. Tämä tuottaa sen ongelman, että sanastopohjainen tunneanalyysi ei kykene huomioimaan sanojen asiayhteyttä. Esimerkiksi tunteiltaan positiiviseksi tarkoitetussa twiitissa voi olla runsaasti negatiivisia sanoja, kuten *syöpä* ja *kuolema*. Tällöin twiitti saa negatiivisen yhdistetyn tunnearvon, vaikka kirjoittajan tarkoituksena oli kertoa hoitojen tehokkuudesta syöpäkuolemien ehkäisyssä. Näin ollen tunneanalyysi voi pahimmillaan tuottaa yhdistettyjä tunnearvoja, jotka ovat päinvastaisia verrattuna twiittien todelliseen tunnesisältöön.

Kaikkiaan voitiin havaita, että tunneanalyysin validiutta heikentävät monenlaiset asiat, jotka voivat tuottaa tuloksiin systemaattista virhettä. Havaittujen ongelmien perusteella voidaan olettaa, että virhe näkyy siten, että negatiiviset mielipiteet ovat aineistossa aliedustettuina, kun taas neutraalit ja positiiviset ovat yliedustettuina. Virheen suuruuden luotettava arviointi on kuitenkin mahdotonta.

8.3 Yhteenveto

Tehtyjen havaintojen pohjalta voidaan vastata päätutkimuskysymykseen eli, missä määrin tekstianalyysin avulla voidaan tuottaa tietoa terveydenhuollon digipalveluihin liittyvästä kansalaismielipiteestä. Toiveena oli löytää tietoa, jonka avulla voitaisiin kehittää palveluita. Mielestäni tekstianalyysi ei soveltunut hyvin tähän tarkoitukseen. Aihemallinnus ei löytänyt aiheita käytettyjen Twitterin hakutermien ulkopuolelta riippumatta käytetyistä aihemääristä. Tunneanalyysin avulla saatiin tietoa palveluihin liittyvästä kansalaismielipiteestä, mutta käytettyjen tutkimusmenetelmien validiteetti asetettiin siinä määrin kyseenalaiseksi, että tuloksiin ei voi luottaa. Tuloksissa oli melko varmasti systemaattista virhettä, jonka suuruutta ei voida laskea.

Tekstianalyysin kohtaamia ongelmia voi pitkälti selittää Twitteristä kerätyn aineiston laadulla. Tutkimuksessa haluttiin tutkia kansalaismielipidettä eli yksityisten kansalaisten mielipiteitä palveluista, mutta on syytä uskoa, että suomalaiset eivät yleensä halua kertoa kokemuksistaan julkisesti Twitterissä. Isotaluksen ja kumppaneiden (2018) mukaan toisin kuin Yhdysvalloissa, Suomessa henkilökohtaisista terveysasioista twiittaaminen on harvinaista. Sen sijaan terveysalan toimijat suosivat Twitteriä viestinnässä. Työssä havaittiinkin, että huomattava osuus aineiston twiiteista oli terveysalan toimijoiden kirjoittamia, ja heidän twiittinsa poikkeavat sisältönsä puolesta yksityisten kansalaisten twiiteista. Näin ollen ongelmien ytimessä luultavasti oli se, että Twitteristä kerätty aineisto ei

yksinkertaisesti sisältänyt tarpeeksi suuria määriä yksityisten kansalaisten terveydenhuollon digipalveluista kirjoittamia mielipiteitä.

9 Pohdinta

Työni kuluessa pohdin, minkälaisessa tarkoituksessa voidaan menestyksekkäästi hyödyntää tekstianalyysia ja sosiaalista mediaa datalähteenä. Sosiaalinen media on huono tietolähde, kun tutkitaan asioita, jotka ovat luonteeltaan yksityisiä. Varsinkaan suomalaiset eivät tapaa kertoa henkilökohtaisista asioistaan julkisesti, vaan niistä kerrotaan yksityisesti. Näin ollen sosiaalinen media sopii paremmin datanlähteeksi, kun tutkitaan mielipiteitä, jotka liittyvät vähemmän arkaluonteisiin asioihin. Tämänkaltaisiin asioihin kansalaisten on helpompi ottaa kantaa julkisesti. Esimerkkinä voisivat olla ensi-iltaelokuviin liittyvät aiheet.

Tutkimuksen aikana esille tuli mahdollisuus lisätä aineiston kokoa käyttämällä useampia hakutermejä Twitterissä. Tämä voisi hyödyttää etenkin aihemallinnusta, jossa aineiston lisääntynyt sanamäärä voisi paljastaa uusia aiheita. Varjopuolena olisi kuitenkin se, että jouduttaisiin turvautumaan hakutermeihin, jotka liittyvät löyhemmin terveydenhuollon digipalveluihin – ainakin, jos valittaisiin termejä, joilla voitaisiin kerätä merkittäviä määriä twiitteja. Tällaisten hakutermien käyttö tuottaisi aineistoon twiitteja, joista suurin osa ei käsittelisi terveydenhuollon digipalveluja. Esimerkiksi, jos hakutermiksi valittaisiin *terveydenhuolto*, saataisiin varmasti runsaasti twiitteja, mutta harva sisältäisi tutkimukselle arvokasta tietoa. Pidän todennäköisempänä, että lisäämällä hakutermejä ei saataisi hyödyllistä tietoa, vaan ennemminkin lisää poistosanoja. Tämän lisäksi tutkittavaan ilmiöön liittymättömät twiitit heikentäisivät entisestään tutkimuksen validiteettia.

Miljardööri Elon Musk osti Twitterin syksyllä 2022, jonka jälkeen sen nimi muuttui viestipalvelu X:ksi ja tapahtui monia muita muutoksia, joilla on vaikutusta sen käytettävyyteen tekstianalyysin datanlähteenä. Twitter on tehnyt parhaansa, jotta palvelun omistaman datan käytöstä joutuu maksamaan. Mahdollisuus kerätä dataa Twitterin API:n kautta hyödyntäen kolmansien osapuolien ohjelmia, kuten DiscoverText:iä on lopetettu. Lisäksi muuta datankeräystä on hankaloitettu asettamalla kirjautuminen palvelun käytön ehdoksi ja rajoittamalla twiittimäärää, jonka käyttäjä saa katsoa päivittäin. Näin ollen Twitterin avoimuus on tämän opinnäytetyön teon aikana heikentynyt hyvin merkityksellisellä tavalla. Uskonkin, että vastaisuudessa mietittäessä työni kaltaisia tutkimuksia on otettava huomioon se, että datan kerääminen Twitteristä on maksullista, ja keräys tapahtuu suoraan Twitterin API:n kautta, mikä edellyttää tutkijalta ohjelmointitaitoja.

Lähteet

Albrecht, J., Ramachandran, S. & Winkler, C. 2020. Blueprints for Text Analytics Using Python. O'Reilly Media. E-kirja. Luettu: 16.5.2023.

Bettenbuk, Z. 2022. How to Scrape Twitter Data Using Python Without Using Twitter's API. ScraperAPI blogi. Luettavissa: https://www.scraperapi.com/blog/scrape-twitter-data/. Luettu: 22.5.2023.

Chakraborty, G., Pagolu, M. & Garla, M. 2014. Text mining and analysis. SAS Institute. E-kirja. Luettu: 7.4.2023.

DiscoverText 2023. About our text analysis data science software. Luettavissa: https://discovertext.com/. Luettu: 20.4.2023.

Haaga-Helia 2023. AI Forum. Luettavissa: https://www.haaga-helia.fi/fi/hankkeet/ai-forum. Luettu: 9.6.2023.

Hapke, H., Howard, C. & Lane, H. 2019. Natural Language Processing in Action. Manning Publications. E-kirja. Luettu: 23.4.2023.

Heikkilä, T. 2014. Tilastollinen tutkimus. 9. painos. Edita Publishing Oy. Helsinki.

Hyppönen, H., Pentala-Nikulainen, O. & Aalto, A. 2018. Sosiaali- ja terveydenhuollon sähköinen asiointi 2017 – Kansalaisten kokemukset ja tarpeet. Juvenes Print – Suomen Yliopistopaino Oy. Helsinki.

Isotalus, P., Jussila, J. & Matikainen, J. 2018. Twitter viestintänä – ilmiöt ja verkostot. Vastapaino. Tampere.

Jürgens, P. & Jungherr, A. 2016. A tutorial for using Twitter data in the social sciences: data collection, preparation, and analysis. Luettavissa: https://ssrn.com/abstract=2710146. Luettu: 1.4.1023.

Knime 2023. KNIME Analytics Platform. Luettavissa: https://www.knime.com/knime-analytics-platform.

Miner, G., Balakrishnan, K., Delen, D., Elder, J., Fast, A., Foley, R., Hill, T., Thompson, J., Wner, A., Winters-Miner, L., Andrew, F., Nisbet, R. 2012. Practical text mining and statistical analysis for non-structured text data applications. Academic Press, 2012. E-kirja. Luettu: 19.6.2023.

Pirhonen, K. 2016. Teknologia sosiaali- ja terveydenhuollossa. Fioca. Helsinki.

Russell, M. & Klassen, M. 2019. Mining the Social Web. 3. painos. O'Reilly Media. E-kirja. Luettu: 31.10.2023.

Saranto, K., Kinnunen, U., Jylhä, V. & Kivekäs, E. 2020. Digitalisaatio ja sähköiset palvelut uudistuvassa sosiaali- ja terveydenhuollossa. Tampere. Luettavissa: http://urn.fi/URN:ISBN:978-952-359-022-9. Luettu: 29.3.2023.

Sarkar, D. 2019. Text Analytics with Python. Apress. E-kirja. Luettu: 16.4.2023.

Sharda, R., Delen, D. & Turban, E. 2018. Business intelligence, analytics, and data science: a managerial perspective. Pearson Education. E-kirja. Luettu: 8.4.2023.

Sloan, L. & Quan-Haase, A. 2017. The SAGE handbook of social media research methods. SAGE Publications Ltd. Lontoo. E-kirja. Luettu: 26.4.2023.

Sohlberg, S. 2021. Asiakasymmärrys digitaalisten sosiaali- ja terveyspalveluiden kehittämiseen asiakaslähtöisesti – Kohderyhmänä erityisen tuen tarpeessa olevat nuoret aikuiset. Ylempi AMK-opinnäytetyö. Laurea-ammattikorkeakoulu, restonomin koulutusohjelma. Luettavissa: https://urn.fi/URN:NBN:fi:amk-202104064266. Luettu: 3.4.2023.

Statista. 2023. Social media usage in Finland. Luettavissa: https://www-statista-com.ezproxy.haaga-helia.fi/study/37855/social-media-usage-in-finland-statista-dossier/. Luettu: 23.4.2023.

Strengell, N. & Sigg, S. 2018. Local emotions - Using social media to understand human-environment interaction in cities. In 2018 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2018 (pp. 615-620). [8480364] IEEE. Luettavissa: https://doi.org/10.1109/PERCOMW.2018.8480364. Luettu: 9.6.2023.

Teodorowski, P., Rodgers, S., Fleming, K. & Frith, L. 2022. Use of the Hashtag #DataSavesLives on Twitter: Exploratory and Thematic Analysis. Journal of Medical Internet Research, 24. Luettavissa: https://doi.org/10.2196/38232. Luettu: 30.10.2023.

Twitter 2023. Getting started. Luettavissa: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api. Luettu: 21.5.2023.

Witten, I., Frank, E., Hall, H. & Pal, C. 2017. Data Mining – Practical Machine Learning Tools and Techniques. 4. painos. Morgan Kaufmann. E-kirja. Luettu: 23.4.2023.

Liitteet

Liite 1. Aineistonhaun hakutermit ja kerättyjen twiittien lukumäärät

Hakutermi	Määrä
digi terveys	988
digitaalinen terveydenhuolto	61
digitalisaatio terveydenhuolto	500
etälääkäri	507
etäterveydenhuolto	19
etävastaanotto	828
sähköinen ajanvaraus	410
sähköinen hoito	14
sähköinen lääke	22
sähköinen lääkemääräys	49
sähköinen potilas	22
sähköinen terveydenhuolto	39
sähköinen terveys	58
terveys verkkopalvelu	21

Liite 2. Terveydenhuollon digipalveluista aktiivisesti twiittaavat terveysalan toimijat

Toimijatyyppi	Nimi	Käyttäjätunnus
Organisaatio	Kanta	@Kantapalvelut
Organisaatio	Lääkärilehti	@Laakarilehti
Organisaatio	Potilaan Lääkärilehti	@Potlaakarilehti
Organisaatio	Suomen Lääkäriliitto	@Laakariliitto
Organisaatio	Apotti	@Apotti_OyAb
Organisaatio	THL	@THLorg
Organisaatio	Sosiaali- ja terveysministeriö	@STM_Uutiset
Organisaatio	Keusote	@KU_Sote
Organisaatio	Sosiaali- ja terveyspalvelut Hel.	@helsinginsote
Organisaatio	Päijät-Hämeen hyvinvointialue	@paijatha
Ammattilainen	Sanni Isometsä	@sannimaria_i
Ammattilainen	Mikko Huovila	@mikkohuo
Ammattilainen	Dr Jouni Laurila	@drjounilaurila
Ammattilainen	Vesa Jormanainen	@VJormanainen
Ammattilainen	Tuomo Oikarainen	@tuomOikarainen
Ammattilainen	Mikko Huovila	@mikkohuo
Ammattilainen	Anna Karjalainen	@AnnaKarjalainen
Ammattilainen	Miia Turpeinen	@miia_turpeinen
Ammattilainen	Juha Wahlstedt	@Sote_laatu
Ammattilainen	Mika Salminen	@mika_salminen