OULUN YLIOPISTO
Humanistinen tiedekunta
Informaatiotutkimus

Jouni Pitkäaho

SOSIAALISEN BIG DATAN MAHDOLLISUUDET TERVEYSVIESTINNÄN TUKENA JA TUNNEANALYYSIN KÄYTTÄMINEN MIELIPITEIDEN MITTAAMISESSA

Informaatiotutkimuksen
pro gradu -tutkielma
Oulu 2022

Jouni Pitkäaho TIIVISTELMÄ

Pro gradu -tutkielma, lokakuu 2022, 63 sivua
Oulun yliopisto, Humanistinen tiedekunta, Informaatiotutkimus

SOSIAALISEN BIG DATAN MAHDOLLISUUDET TERVEYSVIESTINNÄN TUKENA JA TUNNEANALYYSIN KÄYTTÄMINEN MIELIPITEIDEN MITTAAMISESSA

Pro gradu -tutkielman aihe käsittelee sosiaalista big dataa ja sen käyttömahdollisuuksia terveysviestinnässä. Sosiaalisella big datalla tarkoitetaan tämän tutkielman kontekstissa nettikeskusteluita, tarkemmin sanottuna Twitter-keskusteluita. Terveysviestinnän merkitys perustuu yksilön, tai ryhmän, elintapamuutoksiin, joilla on hyvinvointia lisääviä sekä taloudellisia vaikutuksia. Terveysviestintää käytetään ohjaamaan terveyskäyttäytymistä haluttuun suuntaan käyttäen apuna erilaisia strategioita, menetelmiä sekä työkaluja.

Tutkielman tarkoituksena on selvittää voiko sosiaalista big dataa hyödyntää terveysviestinnässä. Kirjallisuuskatsauksessa nostettiin esille useita käyttömahdollisuuksia, joista tutkielman kannalta oleellisin oli tunneanalyysi. Tunneanalyysi mittaa tutkittavaan kohteeseen liittyviä tunteita, asenteita ja mielipiteitä (Liu 2012, 1).

Ajatus tunneanalyysin hyödyntämisestä terveysviestinnällisiin tarkoituksiin pohjautuu Wilsonin (1983, 13–19) käsitykseen uskottavuuden vaikutuksesta kognitiivisiin auktoriteetteihin tiedonlähteinä sekä Euroopan tautienehkäisy ja -valvontakeskuksen (ECDC 2021) käsitykseen luotettavuuden ja viestien uskottavuuden tärkeydestä viestien vastaanottamisen kannalta. Auktoriteetin vaikutusvalta ja asema perustuu siihen, miten uskottavana sitä pidetään (Wilson 1983, 15).

Yksi tapa mitata tietyn tahon uskottavuutta on teettämällä mielipidemittaus. Mielipidemittauksen yhteydessä saadaan selville ketä auktoriteettia kansa pitää eniten uskottavana. Tätä tietoa voidaan hyödyntää terveysviestinnässä mm. valitsemalla

uskottavin auktoriteetti tiedonvälittäjäksi. Kuten tunneanalyysissä, myös mielipidemittauksessa pystytään mittaamaan tunteita, asenteita ja mielipiteitä. Tästä johtuen, tutkielman empiirinen osuus koskee tunneanalyysin tulosten vertaamista mielipidemittauksien kanssa.

Tutkimuskohteeksi valittiin Twitter-käyttäjät, ja aineisto koostuu tviiteistä. Kahta kognitiiviseksi auktoriteetiksi katsottua tahoa koskevat twiitit ladattiin ajalta 15.2–15.7.2022. Aineiston tunneanalyysi suoritettiin Python-koodilla. Tunneanalyysimalleiksi valittiin sanasto- ja sääntöpohjainen VADER-malli sekä tekoälypohjainen BERT-malli. Tunneanalyysien tuloksia verrattiin kolmansien osapuolten tuottamiin mielipidemittauksiin.

Tulosten analysoimisen yhteydessä huomattiin, että VADER-mallin vertautuivat parhaiten mielipidemittauksien kanssa. Yhdessä tapauksessa erot tunneanalyysin ja mielipidemittauksien välillä oli minimillään 5 ja maksimissaan 16 prosenttiyksikköä. Tuloksista kävi kuitenkin ilmi, etteivät mallit pystyneet tässä asetelmalla tuottamaan mielipidemittauksen kaltaisia tuloksia. VADER-malli sen sijaan vaikutti tutkimuksen perusteella osaavan arvioida auktoriteettien välisen sijoituksen suosion suhteen.

Tunneanalyysillä voidaan mahdollisesti selvittää auktoriteettien väliset uskottavuuteen liittyvät sijoitukset. Jatkotutkimuksen kohteena voisivat olla myös muunkieliset aineistot esim. tviitit. Lisäksi aineiston siivoaminen, eri tunneanalyysimallien vertaaminen keskenään sekä eri yhteisöpalveluihin keskittyminen voivat tuottaa laadukkaampia tutkimustuloksia.

Asiasanat: terveysviestintä, sosiaalinen media, tekstinlouhinta, tunneanalyysi

SISÄLLYS

1 JOHDANTO

Kiinnostuin terveysviestinnästä opiskellessani informaatiotutkimuksen Terveystieto ja -viestintä -kurssia varten. Kurssimateriaalina toimineet Enwaldin (2013) ja Hirvosen (2015) väitöskirjat sekä etenkin Johnsonin ja Casen (2012) teos *Health information seeking* avarsivat maailmankuvaani terveysviestinnän tärkeydestä ja potentiaalista. Terveysviestinnän merkitys on korostunut varsinkin viime aikana koronaviruspandemian johdosta, mikä tekee aiheesta ajankohtaisen.

Terveydenhuoltokustannuksien noustessa, netissä pyörivän dis- ja misinformaation yleistyessä sekä koronavirusopandemian yhteydessä terveysviestinnän tärkeys on korostunut. Terveysviestintä edistää terveyden ja hyvinvoinnin lisäksi myös yhteiskunnan toimintaa, sillä sairauksien ennaltaehkäisy tulee halvemmaksi kuin niiden hoito (Prevention institute 2012; tässä Case & Johnson 2012, 8).

Potilastietojen, kuluttajadatan ja muiden data-aineistojen avulla voidaan tehdä terveydenhuoltoa ja terveyttä edistäviä ratkaisuja. Tämä tutkielma keskittyy etsimään vastauksen siihen, voiko *sosiaalista big dataa* (SBD) hyödyntää terveysviestinnässä. Tutkielmaan SBD:ksi on valittu Twitter-keskustelut. SBD:n etu moneen muuhun aineistoon verrattuna on sen saavutettavuus. Sitä on paljon ja siihen pääsee helposti käsiksi.

Terveysviestintä on yksi monista tavoista vaikuttaa positiivisesti terveyteen ja hyvinvointiin, ja sosiaalinen big data on yksi monista aineistomuodoista, jota on hyödynnetty monien eri alojen tutkimuksissa. Tutkielmani tarkoitus on selvittää voiko näiden kahden yhdistelmällä vaikuttaa positiivisella tavalla ihmisten terveyteen ja hyvinvointiin terveysviestinnän kautta. Tutkielman perimmäinen tarkoitus on siis selvittää, voiko sosiaalista big dataa hyödyntää terveysviestinnän kehittämisessä.

Tutkielman tutkimuskysymykset ovat:

1. Millä tavoin SBD:a voitaisiin hyödyntää terveysviestinnässä?

2. Voiko Twitter-keskusteluista tuotettua tunneanalyysiä käyttää mielipidemittauksen korvikkeena?

Ensimmäiseen tutkimuskysymykseen pyritään löytämään vastaus kirjallisuuskatsauksen avulla sekä mahdollista lisävahvistusta tutkielman empiirisen osuuden kautta. Tutkielman toiseen tutkimuskysymykseen pyritään löytämään vastaus kirjallisuuskatsauksessa esille nousseen tutkimusmenetelmän *tunneanalyysin* avulla. Tunneanalyysi (myös sentimenttianalyysi) on yksinkertaisesti sanottuna menetelmä, jolla selvitetään tietyn sisällön, yleensä tekstin, tunnepitoisuutta. Joskus teksti voi olla negatiivisesti tai positiivisesti varautunut tai neutraali. Kun aiheen rajaa tiettyyn asiaan esim. presidenttiehdokkaaseen, saadaan selville usean nettikeskustelijan mielipide hänestä.

Tunneanalyysi muodostaa tutkielman empiirisen osuuden, ja siinä käytetään hyväksi nykyaikaisia vapaasti saatavilla olevaa tekoälymallia BERT ja sanasto- ja sääntöpohjaista VADER-mallia. Tunneanalyysissa pääaineistona toimvat Twitter-keskustelut, jotka liittyvät Yhdysvaltain tautikeskukseen CDC:hen (Centers for Disease Control) ja presidentti Joe Bideniin. CDC nousi Yhdysvalloissa keskeisemmäksi auktoriteetiksi taistelussa koronavirusta vastaan. Biden esiintyi usein kameroiden edessä, ja hän on vieläkin terveysviestinnällisesti ajatellen hyvin merkittävässä asemassa.

Sekä Wilson (1983, 13–19), että Euroopan tautienehkäisy- ja valvontakeskus (ECDC 2021) pitävät uskottavuutta tärkeänä elementtinä auktoriteettiaseman ylläpitämisessä ja viestien vastaanottamisessa. Wilsonin (1983, 13–19) mukaan auktoriteetin vaikutusvalta riippuu siitä, miten uskottavana pidämme sitä. ECDC (2021) pitää terveysviestinnän sisältöjen uskottavuutta yhtenä tehokkaan terveysviestinnän edellytyksenä.

Uskottavuutta ja luotettavuutta voidaan mitata mielipidemittauksien avulla, joten kysymys kuuluu: voiko sosiaalisesta mediasta (Twitteristä) saatuja keskusteluita ja mielipiteitä rinnastaa millään asteella mielipidemittauksien kanssa? Sosiaalisessa mediassa käydään paljon keskusteluita, mutta ovatko ne rinnastettavissa todellisuuden

kanssa vai eletäänkö somessa täysin omassa maailmassa? Tähän tutkielma pyrkii löytämään vastauksen tunneanalyysin avulla.

Mielipidemittaukset ovat yksi tapa mitata jonkin asian uskottavuutta ja luotettavuutta. Mielipidemittaukset antavat yleiskuvan kansalaisten mielipiteestä tietystä kohteesta esim. hallitusta, puoluetta, muutosta, johtohahmoa, organisaatiota yms. kohtaan Mielipidemittaukset toimivat työkaluina monille. Mielipidemittausten avulla eri tahot voivat tehdä tarvittavia muutoksia esim. puolue voi muuttaa kantaansa negatiivisen mielipiteen, eli suosion laskun, johdosta. Terveysviestinnästä vastaavia instituutioita ja johtohahmoja (joihin viitataan tästä edespäin tekstiä *kognitiivisina auktoriteetteina*) kohtaan olisi teoriassa otollista teettää mielipidemittauksia, sillä heidän uskottavuutensa on suoraan verrattavissa siihen, miten hyvin terveysviestintä toimii kansaan. Esimerkiksi Rokote-epäröinti on suurinta niissä maissa, joissa kansalaisten luottamus hallintoa kohtaan on matalimmillaan (Miyachi, Senoo, Takita & Yamamoto 2020, 31).

Tässä tutkielmassa terveystiedon kognitiivisiksi auktoriteeteiksi on valittu Yhdysvaltain tartuntatautivirasto CDC ja presidentti Joe Biden. Kummallakin auktoriteetilla oli, ja on vieläkin, iso rooli taistelussa koronaviruspandemiaa vastaan lähinnä viestinnällisin ja lainsäädännöllisin keinoin. Heistä on myös löydettävissä kolmannen osapuolten teettämiä mielipidemittauksia. Suomalaiset vastineet olisivat voineet olla esim. Terveyden ja hyvinvoinnin laitos (THL), Mika Salminen (entinen THL:n terveysturvallisuusosaston johtaja) sekä HUS:n ylilääkäri Asko Järvinen. He olivat usein mediassa esillä vuosien 2021-21 aikana, sekä he olivat osallisena pandemiaan liittyvissä päätöksenteoissa.

Syy miksi Yhdysvaltain eri viranomaiset valittiin tutkielmassa tutkimusaiheeksi, on se, että helposti saatavilla olevat tunneanalyysin mallit ovat yleensä tehty englanninkielisiä tekstejä varten. Lisäksi, Bidenista on saatavilla enemmän aineistoa, sillä hänestä keskustellaan paljon verkosta. Lisäksi molemmista tahosta on enemmän kolmannen osapuolten teettämiä mielipidemittauksia. Näistä syistä Yhdysvaltain viranomaiset ovat parempi tutkimusaihe, kuin suomalaiset vastineet. Joudun luottamaan kolmansien osapuolten teettämiin mielipidemittauksiin, sillä en itse kyennyt haastattelemaan

yhdysvaltalaisia, ja keräämään heidän mielipiteitänsä liittyen Yhdysvaltalaisiin terveystiedon kognitiivisiin auktoriteetteihin.

Tutkimuksen tarkoituksena on selvittää vastaavatko Twitter-keskustelujen ilmapiiri todellisuuden kanssa. Jos mielipidemittaukset ja Twitter-keskustelujen tunneanalyysit tuottavat yhdenmukaisia tuloksia, voidaan katsoa, että tunneanalyysia voidaan käyttää mielipidetutkimuksen korvikkeena. Julkinen mielipide on tärkeää, sillä yhteiskunnassa kansan luottamus auktoriteetteihin näkyy heidän terveyteensä liittyvissä valinnoissa.

Kansalaisten alhainen luottamus maan hallintoon näkyy esimerkiksi kasvuna rokote-epäröinnissä (Miyachi, Senoo, Takita & Yamamoto 2020, 31). Luottamus terveydenhuollon ammattilaisiin, auktoriteetteihin ja peräti paikallislehteen vaikutti saksalaisten aikomuksiin hankkia koronarokote (Blöbaum, Fujarski, Gehrau, Lorenz & Schieb 2021, 5–9). Mitä enemmän asenteet tiettyä tahoa kohtaan heikkenevät, sitä vähemmän kansa uskoo heidän tuottamiin viesteihinsä. Tämä epävarmuus heijastuu heidän käyttäytymiseensä.

Kansan mielipidettä voidaan hypoteettisesti pitää työkaluna, jolla terveysviestintää ohjataan. Yksinkertaisesti sanoen, jos asenteet jotain terveydenhuollon tahoa, hallitusta, johtohahmoa tai muuta kognitiivista auktoriteettia kohtaa heikkenevät, olisi syytä muuttaa tai tehostaa viestintää, jotta saadaan kansalaiset käyttäytymään halutulla tavalla. Jos nettikeskustelut ovat jollain tapaa linjassa kansalaisten mielipiteiden ja asenteiden kanssa (eli todellisuuden kanssa), mielipidemittausta paljon halvempi ja nopeampi vaihtoehto tunneanalyysi olisi tällöin terveysviranomaisille ja muille auktoriteeteille yksi tapa seurata kansalaisten mielipiteitä heitä kohtaan, sillä hyvä maine ja imago vaikuttavat heidän terveysviestintänsä toimivuuteen. Tunneanalyysin etuna on myös se, että sillä voidaan seurata yksittäisten viestien tuottamia tunteita ja reaktioita, jolloin viestintää voidaan seurata tviitti kerrallaan. Tunneanalyysillä voidaan myös selvittää, mitä terveysviestinnän kognitiivisista auktoriteettia kansa arvostaa eniten, ja käyttää häntä (tai sitä) tietynlaisena päähahmona median edessä.

Toivon, että pro gradu -tutkielmastani on hyötyä aiheesta kiinnostuneille. Toivon lisäksi, että tutkielmassa esille nousseista asioista on apua myös terveysviestinnän konkreettisella tasolla. Työtäni varten tekemäni koodit ovat saatavilla osoitteessa: https://github.com/jpitkaah/tunneanalyysit. Ne ovat vapaasti käytettävissä.

1.1 Keskeiset käsitteet

Tämän luvun tarkoituksena on esitellä tutkielmassa käytettyjä käsitteitä. Tutkielman kohdeyleisö on lähinnä informaatiotutkimuksesta ja viestinnästä kiinnostuneet, joten tutkielmassa käydyt tekniset asiat, kuten tekoäly, on pyritty selittämään mahdollisemman yksinkertaisesti ja käyttämällä varsin yleistäviä esimerkkejä.

1.1.1 Data ja sosiaalinen big data

Data -sanalle on olemassa monta eri määrittelyä. Cambridgen sanakirjan (Cambridge University Press) (2021) kaksi määritelmää datalle ovat mielestäni yleistajullisesti ajatellen parhaimmat eli dataa voidaan pitää informaationa, jota pystyy tutkimaan ja käyttämään apuna päätöksenteossa tai data on informaatiota elektronisessa muodossa, jota voi tallentaa ja käyttää tietokoneella. Perinteisesti dataa on usein kuvailtu eräänlaisina tiedon palasina (Kelleher & Tierney 2018, 39).

Big data (myös massadata) määritellään yleensä niin kutsutuilla V-malleilla. V-malleja yhdistää yleensä kolme ominaisuutta: tiedon suuruus (volume), nopeus (velocity) ja vaihtelevuus (variety). On olemassa myös 5V- ja 7V-malleja, sekä niiden variaatioita, mutta selkeyden vuoksi henkilökohtaisesti suosin 3V-mallia, jossa on ainoastaan kolme edellä mainittua ominaisuutta. Kolmesta ominaisuudesta tiedon suuruus viittaa yleensä valtaviin määriin dataa, muuttumisnopeus viittaa datan syntynopeuteen ja vaihtelevuus viittaa datan eri muotoihin ja tyyppeihin (Laney 2001).

Sosiaalisella big datalla (SBD) tarkoitetaan sosiaalisessa vuorovaikutuksessa syntynyttä dataa. Sosiaalisella vuorovaikutuksella viitataan pääasiassa sosiaalisen median käyttäjien keskeiseen kommunikointiin. Kommentit, päivitykset, tviitit, tykkäykset sekä äänien, linkkien, kuvien, videoiden jakaminen voidaan laskea osaksi sosiaalisessa mediassa tapahtuvaa kommunikointia. Sosiaalisesta mediasta käytetään joskus nimitystä *yhteisöpalvelut*.

Tässä tutkielmassa data-aineistona toimivat Twitteristä poimitut keskustelut. Analyysissa puhdistetaan kaikki mahdollinen metatieto pois niin, että jäljelle jäävät vain tekstiosiot eli tviitit (tweets) ja uudelleentviittaukset (retweets). Valitsin Twitterin, koska minulla on aikaisempaa kokemusta Twitteristä liittyen tiedonlouhintaan.

1.1.2 Terveysviestintä

Terveysviestintä (health communication) on viestintää, jolla pyritään vaikuttamaan yksilön ja yhteisön terveyskäyttäytymiseen erilaisilla viestinnällisillä keinoilla ja strategioilla. Tarkoitus on saada kohdeyleisö toimimaan halutulla tavalla, ja terveysviestinnässä tämä usein tarkoittaa terveyttä edistävää tapaa (ks. 1.1.3 Terveyskäyttäytyminen). Terveyttä ja hyvinvointia edistäviä tapoja ovat mm. asenteiden muuttaminen, tietoisuuden lisääminen, motivointi ja toimintatapojen muuttaminen esim. liikunnan aloittaminen tai ruokavalion vaihtaminen vähäkalorisemmaksi. (Thomas 2006, 1–2, 63, ECDC 2021.)

Terveystieto (health information) on internetin yleistyessä kasvanut huomattavasti muun tiedon ohella (Thomas 2006, 12). Terveystieto on yksinkertaisesti selitettynä terveyteen liittyvää ja terveyttä edistävää informaatiota. Terveysviestinnästä ja -tiedosta kerrotaan tarkemmin luvussa 3.

1.1.3 Terveyskäyttäytyminen

Terveyskäyttäytymisen (health behavior) muokkaaminen terveellisempään suuntaan on terveysviestinnän tavoite (Thomas 2006, 65). Terveyskäyttäytymisellä viitataan yksilön käyttäytymiseen ja valintoihin, joilla on terveyteen liittyviä vaikutuksia. Terveyskäyttäytymisellä on merkittävä vaikutus yksilön terveyteen. (Duodecim 2016.)

1.1.4 Tekstinlouhinta

Tekstinlouhinta (text mining) on joukko erilaisia tekstin analysoimiseen käytettyjä menetelmiä, joiden avulla teksteistä selvitetään relevanttia tietoa, trendejä ja tuntemattomia yhteyksiä asioiden välillä (LUC 2021, IBM 2020a). Tiedonlouhinnassa (tekstinlouhinnan yläkäsite) selvitettävät yhteydet voivat olla esimerkiksi yhteyksiä geenien ja tautien välillä tai yhteyksiä magnesiumin puutteen ja migreenin välillä (Case, D. & Johnson, J. 2012, 194–195, Smalheiser, N., Swanson, D. & Tovik, V. 2006, 1433–1438). Tekstinlouhinnalla on löydetty yhteyksiä sairauksiin, kuten tekonivelinfektioon, analysoimalla potilaskertomuksia (Hirviheimo, Kinnunen & Kivekäs 2015, 80). Yksi tekstinlouhintaan liittyvistä menetelmistä on tunneanalyysi.

Tekstinlouhinta on Finton eli Yleisen suomalaisen asiasanaston (2021) mukaan tiedonlouhinnan (data mining) alakäsite. Joskus tekstinlouhintaa ja tiedonlouhintaa pidetään toistensa synonyymeina ja joskus ne käsitetään toisistaan eriäviksi menetelmiksi. Joidenkin tutkijoiden mielestä tiedonlouhinta viittaa pääasiassa big datan prosessointiin, kun taas tekstinlouhinnalla viitataan pienten aineistomäärien tekstianalyysiin. Tekstinlouhinta on kuitenkin valtavirran mukaan aina tiedonlouhintaa, kun louhinnan kohteena on tekstiaineisto riippumatta aineiston koosta. Vanhemmissa suomalaisissa teoksissa saatetaan käyttää termiä *sisällön erittely* puhuttaessa verkkokeskusteluiden analysoimisesta (Laaksonen, Matikainen & Tikka 2013, 216–220). Tekstinlouhinta on kuitenkin terminä laajasti käytetty, minkä takia käytän kyseistä termiä myös tutkielmassani.

1.1.5 Tunneanalyysi

Tunneanalyysi (sentiment analysis) on yksi tekstinlouhinnan alitehtävistä (subtask) (IBM 2020a). Tunneanalyysia voidaan myös kutsua tekniikaksi tai menetelmäksi. Suomen kielessä alitehtävä -termi ei ole niin tunnettu, joten siksi onkin parempi kutsua tunneanalyysia tekstinlouhinnan tekniikaksi tai menetelmäksi. Tunneanalyysin tehtävä on mitata tiettyyn ilmiöön liittyviä tunteita, mielipiteitä ja asenteita (Liu 2012, 1).

Joskus törmää teksteihin, joissa käytetään mielipidelouhinta -termiä (opinion mining) synonyymina tunneanalyysin kanssa. Palpanas ja Tsytsarau (2011, 4) pitävät mielipidelouhinta -termin alkuperänä tiedonhaun tutkimuskenttää, kun taas tunneanalyysin juuret ovat heidän mukaansa *luonnollisen kielen käsittelyn* (natural language processing) puolelta. Molemmat pyrkivät löytämään vastauksen samoihin ongelmiin. Luonnollisen kielen käsittely on ala tai tutkimusalue, joka pyrkii kehittämään tekniikkaa, jolla tietokonetta voidaan opettaa ymmärtämään ja manipuloimaan luonnollista kieltä (Chowdhury 2005, 51). Liu (2012, 1) pitää molempia termejä periaatteessa toisiaan vastaavina. Tunneanalyysissä voidaan hyödyntää *tekoälyä* ja *koneoppimista*.

1.1.6 Tekoäly

Tunneanalyysissa hyödynnetään kahta erityyppistä tekoälyä. Tekoälyllä (artificial intelligence, AI) tarkoitetaan yleensä tietokonetta tai ohjelmaa, joka kykenee tekemään rationaalisia toimintoja. Tekoäly klassisesti ajatellen on ihmisen, tai ihmismielen, jäljittelyä. Tekoälyä voidaan myös pitää alana, joka tuottaa älykkäitä sovelluksia ja tietokoneita. (IBM 2020b, Nilsson 2010, 13.)

Tekoäly oppii *iteroinnin* kautta. Iterointi on menetelmä, jossa toistetaan samoja työvaiheita, kunnes saavutetaan haluttu lopputulos (Tilastokeskus 2022a). Voidaankin sanoa, että iterointi muistuttaa oppimista yritysten ja erehdysten kautta (trial and error). Toisin sanoen, tekoäly yrittää niin kauan, kunnes se tekee edes jotain oikein. Oikein

tehdystä asiasta tekoäly ”saa palautetta”, jonka jälkeen se tallentaa muistiin arvot, jotka johtivat onnistumiseen. Kun pieniä onnistumisia alkaa kertymään, tekoälyn voidaan katsoa ”viisastuvan”. Tämän vuoksi, aluksi tekoäly näyttää tekevän asioita hyvin sattumanvaraisesti.

Koneoppiminen on prosessi, jossa tietokone tai ohjelma kehittyy ja oppii itsenäisesti datasta ihmisen ohjastuksella tai ilman sitä (Merilehto 2018, 19). Kelleher (2020, 5) pitää koneoppimista kaksivaiheisena prosessina, jossa ensimmäinen vaihe on oppiminen ja toinen vaihe ennustaminen. Tekoälyä voidaan opettaa aineiston avulla esimerkiksi tunnistamaan käsin kirjoitettuja kirjaimia ja sanoja. Tekoälyä voidaan myös opettaa pelaamaan pelejä kuten shakkia, asettamalla se pelaamaan vastustajaa vastaan. Vastustajana voi toimia tietokonevastustaja, jossa on joko valmiiksi koodattu tekoäly, joka tuntee shakin säännöt, nappuloiden liikkeet ja taktiikat tai vastustajana voi toimia toinen tekoäly. Vastustajana voi toimia myös ihminen. Tietokone vastustajana mahdollistaa sen, että yksi shakkipeli saattaa olla ohi muutamassa sekunnissa riippuen mm. tietokoneen tehokkuudesta, tekoälyjen taitotasosta ja pelin kulusta. Ihmistä vastaan pelattaessa, tietokoneen joutuu odottamaan ihmistä. Tämän vuoksi yksi shakkipeli voi kestää useita minuutteja, mikä on tekoälyn oppimisen kannalta vähemmän tehokasta.

1.2 Tutkielman rakenne

Pro gradu -tutkielmani Johdanto-osassa esittelen tutkimuskysymykset ja tutkielman kannalta keskeisimmät termit. Seuraavassa luvussa käsittelen big dataa. Luvussa 3. keskitytään terveysviestintään ja siihen, mitkä tekijät vaikuttavat sen tehokkuuteen. Luku 4. on omistettu tutkimusmenetelmälle ja -aineistolle. Luku 5. avaa tutkimukseni tuloksia. Luvussa 6. Pohdin tuloksiani ja vertaan niitä aiempiin tutkimuksiin. Lisäksi pohdin tutkimukseni onnistumista ja esitän jatkotutkimusideoita.

2. BIG DATA

Maailmassa olevan datan määrä on massiivinen. Euroopan parlamentti (2021) arvioi, että datan kokonaismäärä tulee kasvamaan peräti 530 prosenttia vuoteen 2025 mennessä vuoteen 2018 verrattuna. Big datan tarjoamat mahdollisuudet eivät ole vielä täysin selviä sillä aihe on verrattain uusi. Big data luo uusia mahdollisuuksia monilla eri sektorilla. Big datan avulla teollisuus voi parantaa tuottavuutta ja tuotantotehokkuutta, joka voi kasvattaa innovaatiota ja johtaa kasvihuonekaasupäästöjen vähenemiseen. Datan avulla voidaan vähentää liikenneruuhkia ja sen myötä myös hiilidioksidipäästöjä. Potilastiedoista koostuvan datan analysointi voi johtaa parempiin diagnooseihin ja sen myötä hoitoihin. Big data voi myös auttaa lääkkeiden kehittämisessä ja kustannuksien vähentämisessä. Big data voi auttaa maataloutta tarjoamalla ajankohtaista tietoa viljelijöille olosuhteista. Lisäksi julkinen sektori voi parantaa palveluitaan big datan avulla lisäämällä tehokkuutta ja vaikuttavuutta. (Euroopan parlamentti 2021.)

2.1 Big datan hyödyntäminen terveydenhuollossa

Big dataa on jo hyödynnetty suomalaisessa lääketieteessä jo jonkin aikaa. Solutasolla, esim. geenitutkimuksissa, big data on tuottanut merkittäviä hyötyjä. Terveyteen liittyvän datan hyödyntäminen on Suomessa vasta alkamassa, ja suomalaisessa lääketieteessä big datan hyödyntäminen on lähtenyt liikkeelle hitaasti muihin aloihin verrattuna. (Tuomisto 2015, 2179.) Kuluttajadatan avulla voidaan seurata kuluttajien terveyteen liittyviä ruokatottumuksia, kuten alkoholin kulutusta, ja teettää niiden pohjalta kansanterveyttä edistäviä päätöksiä erilaisten rajoituksien ja verojen avulla (Erkkola, Fogelholm, Nevalainen, Saarijärvi & Uusitalo 2019, 82–83). Big datan avulla voidaan parantaa terveydenhuollon tehokkuutta (Binenbaum ym. 2019, 26). Big data-analytiikan avulla voidaan tunnistaa mm. yksilön riskiä sairastua tiettyyn sairauteen, parantaa hoidon laatua sekä potilaan turvallisuutta. Big dataa voidaan käyttää apuna vertailuun (benchmarking), laadun parantamiseen ja tutkimustyössä. (Bates, Heitmueller, Kakad & Saria 2018, 214.) Big datan avulla sairauksien diagnosoimista voidaan nopeuttaa ja tarkentaa, vähentää

vääriä diagnooseja, kehittää ja tarjota parempia hoitoja sekä vähentää readmissioneita eli potilaan hakeutumista takaisin hoitoon kotiuttamisen jälkeen (Krishnan 2016, 156–157).

Kuten edellä olevista esimerkeistä käy ilmi, big data tarjoaa paljon hyötyä yhteiskunnallisella tasolla. Varsinkin lääketieteen ja -teknologian saralla, big datan tulee osoittautumaan melko todennäköisesti hyvin arvokkaaksi. Suomalaisen terveysdatan, esim. potilas- ja genomitiedon, luovuttaminen edellyttää viranomaisilta riittävän hyvät perusteet (Genomikeskus 2021, Lääkäriliitto 2021). Tämä vaikeuttaa aineiston saatavuutta. Sosiaalisen big datan etu on se, että datasetit ovat kaikkien saatavilla, riippuen yhteisöpalvelun käytännöistä ja lainsäädännöstä.

2.2 SBD:n hyödyntäminen terveydenhuollossa

Sosiaalinen big data (SBD) on yksi big datan muoto. SBD:lla on monia muotoja, mutta ehkä yleisin ja tunnetuin muoto on nettikeskusteluista syntynyt tekstimuodossa oleva aineisto. SBD on vielä verrattain uusi tutkimusaineisto terveydenhuollon kontekstissa, sillä siitä löytyi todella niukasti kirjallisuutta. SBD:n analysointia on ehdotettu esim. henkilön itsetuhoisuuteen puuttumista skannaamalla avainsanoja hänen some-päivityksistään (Ryu & Song 2015, 7, Barros, Oliveira & Trifan 2021). Eräät tutkijat käyttivät Twitter-dataa apuna seuratakseen flunssan leviämistä (Kautz, Sadilek & Silenzio 2012, 322–329). Youngin (2014, 601–602) mukaan sosiaalinen media voi antaa kuvan tautien leviämisen seuraamisen lisäksi myös riskikäyttäytymisille.

Vuonna 2018 somessa levisi niin kutsuttu Tide pod challenge, jossa nuoret kuvasivat itseään ja toisiaan syömässä Tide pod-nimisiä pyykinpesukapseleita. Myös kanelihaaste (cinnamon challenge), jossa tarkoituksena on syödä jauhettua kanelia, on ollut muutamina vuosina trendinä. Molemmat haasteet ovat terveydelle vaarallisia. Näihin haasteisiin olisi periaatteessa voitu puuttua hyvissä ajoin netissä leviävien videoiden ja kommenttien perusteella, ja tiedottaa haasteiden aiheuttamista riskeistä. Eli näkisin, että SBD:taa voitaisiin näin hyödytää terveysviestinnän suuntaamisessa.

2.3 Tiedon- ja tekstinlouhinta

Tiedonlouhinnan voidaan kuvailla olevan prosessi, tekniikka tai menetelm(i)ä, jonka tarkoituksena on poimia tietoa, löytää kuvioita (patterns) ja muita tärkeitä tietoja datasta (IBM 2021, Galetsi, Katsaliaki & Kumar, 2020, 209, Ishikawa 2015, 86). Tiedonlouhinnan tarkoituksena on yksinkertaisesti sanottuna muodostaa kokonaiskuva datasta. Tiedonlouhinta on yksi askel KDD-prosessin (knowledge discovery in database process) kokonaisuudesta. (Ishikawa 2015, 91.) KDD kääntyy hieman huonosti suomen kieleen, mutta Moen (2021) kääntää sen muotoon *tietämyksen muodostuminen*. KDD-prosessi on moniaskeleinen tiedonlouhintaprosessin kokonaisuus, joka Ishikawan (2015, 91) mukaan koostuu seuraavista elementeistä:

1. Datan puhdistus.
2. Datan integroiminen eli datasettien yhdistely.
3. Datan vähentäminen ja valitseminen. Relevantin datan kanssa työskentely.
4. Datan muuttaminen tietorakenteeseen, joka sopii tiedonlouhintaan
5. Tiedonlouhinta. Datasta etsitään kuvioita erilaisten menetelmien avulla.
6. Kuvioiden evaluointi. Kuviot evaluoidaan ja niiden pohjalta muodostetaan käsitys.
7. Tulosten (käsityksen) visualisointi ja esittely.

Tekstinlouhinta on tyypillinen sosiaalisen big datan analyysimenetelmä, kun kyseessä on tekstimuodossa oleva aineisto. Tekstinlouhinnan avulla voidaan analysoida nimenomaan strukturoimatonta aineistoa esim. luonnollista kieltä, kuten tekstiä (Kaarakainen & Kaarakainen 2018, 26). Strukturoidulla aineistolla, tai datalla, viitataan siihen, että datalla on ennalta määritelty rakenne (Ashkpour ym. 2014, 545). Strukturoimattomalla datalla ei ole rakennetta, ja puolistrukturoitu data on näiden kahden välimaasto (Ashkpour ym. 2014, 545, Kaarakainen & Kaarakainen 2018, 26). Tyypillisiä tekstimuotoisia sosiaaliseen big dataan liittyviä strukturoimattomia aineistoja ovat esimerkiksi tviitit, arvostelut, kommentit ja Facebook-päivitykset. Tekstinlouhintaan voi käyttää monenlaisia sovelluksia. Internetissä löytyy paljon ilmaisia sekä maksullisia sovelluksia.

Käytän tässä tutkielmassa kuitenkin Python-ohjelmointikieltä, koska itse tehty ohjelma tarjoaa enemmän joustavuutta ja mahdollisuuksia.

Tekstinlouhinta koostuu useista eri menetelmistä ja tekniikoista. Analyysimenetelmän valinta riippuu tutkimuksen päämäärästä, luonteesta sekä siitä, mitä halutaan tutkia. Aineisto voidaan luokitella (classification), kun aineistoja halutaan organisoida. Yksi luokittelun esimerkki on erottaa sähköpostiliikenteestä roskaposteja niiden sisällön perusteella. Aineisto voidaan klusteroida (clustering), kun aineistoa halutaan ryhmittää tai erottaa toisistaan samankaltaisuuden perusteella. Klusteroinnin ansiosta aineistosta voidaan esim. löytää jotain uutta tai paikantaa erilaisia ongelmia. Aineiston avulla voidaan luoda ennusteita. (Damerau, Indurkhya, Weiss & Zhang 2005, 6–12.) Näiden lisäksi, tekstinlouhintateknologian avulla tekstiaineistoja voidaan tiivistää (IBM 2020a). Teknologiaa voidaan myös hyödyntää tiedonhankintajärjestelmissä dokumenttien noutamisessa eli tiedonhaussa (information retrieval) (Damerau, Indurkhya, Weiss & Zhang 2005, 8–9).

3. TERVEYSVIESTINTÄ

Viime vuosisadalla kansanterveyttä uhanneet pandemiat, kuten espanjantauti sai aikaan sen, että eri maiden viranomaiset julkaisivat sanomalehdissä ja lentolehtisissä hoito-ohjeita, ohjeita hygieniasta huolehtimiseen ja kehotuksia välttämään väkijoukkoja (Linnanmäki 2006, 2028, Stephens 2020). Sata vuotta sitten tapahtunut terveysviestintä saattoi olla periaatteeltaan samankaltainen nykyaikaisen terveysviestinnän kanssa, mutta viestintäteknologia ja -laitteet olivat tuohon aikaan eri luokkaa. Tuolloin painatteet olivat joukkoviestinnän tehokkain työkalu. Painatteiden rinnalle yleistyi myöhemmin radio ja televisio. Näiden kahden elektronisen keksinnön yleistyessä kuluttajien keskuudessa tietoa pystyttiin siirtämään kansalaisille aikaisempaa nopeammin ja laajemmin. Lehdistö, radio ja televisio olivat joukkoviestinnän tukipilareita vuosikymmeniä, kunnes kuvioon ilmestyi digitalisaation myötä uusi osallistuja, nimittäin internet.

Internetin välityksellä voidaan tavoittaa kattava yleisö. Internetissä on saatavilla hurjat määrät tietoa aina salaliittoteorioista tieteellisiin julkaisuihin. Sosiaalisen median alustoilla käyttäjät voivat jakaa sisältöä tai tuottaa sitä itse, jolloin vaarana on valheellisen tai haitallisen tiedon leviäminen. Terveysviestinnän näkökulmasta ajatellen, haitallisella tiedolla voi olla terveyttä vaarantavia vaikutuksia. Viranomaiset joutuvatkin taistelemaan huomiosta ja siitä, että terveydelle hyväksi olevaa tietoa saadaan sitä tarvitseville. Hukka (2014, 118) pitää asiantuntijoiden hyviä verkostoitumistaitoja (kuuntelu ja osallistuminen) verkkohuomion ja uskottavuuden kannalta tärkeinä.

Terveysviestinnässä yhdistyy useita eri informaatiotutkimuksen osa-alueita: tiedonhaku ja -hankinta, informaatiolukutaito (etenkin terveyslukutaito), informaatioympäristöt ja informaatiokäyttäytyminen. Tutkielman tarkoituksena ei ole lähestyä aihetta pelkästään yhden tai kahden em. osa-alueen näkökulmasta. Toisin sanoen, tutkielmassa nousee esille monipuolisesti eri informaatiotutkimuksen osa-alueita, joitakin enemmän ja joitakin vähemmän.

Tutkielmassa on lisäksi eri tieteenalojen teoksia, jotka ovat lähellä informaatiotutkimusta ja sen osa-alueita. Esimerkiksi viestinnän tohtorien Johnson ja Casen (2013) teos, johon

tulen viittaamaan useasti tässä tutkielmassa, käsittelee terveysviestintää viestinnän näkökulmasta. Teoksessa on paljon yhtäläisyyksiä informaatiotutkimuksen kanssa, ja se on ollutkin meillä kurssikirjana. Tulen käyttämään myös eri tieteenalojen teoksia esim. Richard Thomasin teos *Health Communication* (2006) käsittelee terveysviestintää terveystieteen (health science) ja terveyssosiologian (medical sociology) näkökulmasta.

Tieteenalojen sisäiset termit voivat erottua toisistaan, mutta niiden sisältö ja tutkimustuloksien tulkinta ja tutkimuksien johtopäätökset ovat pääasiassa samoja. Esimerkiksi termi *terveyskäyttäytyminen* on periaatteeltaan hyvin paljon samankaltainen kuin informaatiotutkimuksesta tuttu termi *informaatiokäyttäytyminen*. Pyrin kuvaaman tutkielmassa käytyjä ilmiöitä pääasiassa informaatiotutkimuksen termeillä aina kun mahdollista. Poikkeuksena edellä mainittu termi terveyskäyttäytyminen, sillä se on terminä mielestäni tarkempi kuin monen asian kattava informaatiokäyttäytyminen.

3.1 Terveysviestinnän tärkeys

Terveydenhuollon menot kasvoivat tasaista vauhtia Suomessa jo ennen koronakriisiä. Terveydenhuollon menot yltivät 22 miljardiin euroon vuonna 2019, mikä tekee asukasta kohden 3983 euroa. Kasvua edellisvuoteen oli reaalisesti 3 prosenttia. (THL 2021.)

Terveydenhuollon kustannuksien vähentämiseksi on ehdotettu terveysviestinnän kehittämistä ja lisäämistä. Yksi terveysviestinnän tehtävistä on lisätä tietoisuutta eri taudeista ja niiden ennaltaehkäisykeinoista (Case & Johnson 2012, 3–8.) Terveysviestinnällä voidaan kansanterveyden ja hyvinvoinnin parantamisen lisäksi saada aikaan taloudellista hyötyä, kustannuksien alenemisen muodossa. Jokaista dollaria kohden, joka investoidaan ennaltaehkäisyyn (prevention), saadaan viiden dollarin hyödyt (return) viiden vuoden sisällä (Prevention institute 2012; tässä Case & Johnson 2012, 8).

3.2 Onnistunut terveysviestintä

Terveysviestintä voidaan katsoa onnistuneeksi silloin kun viestintä saa yksilöissä aikaan halutun muutoksen. Terveysviestinnällä pyritään ohjaamaan yksilön terveyskäyttäytymistä niin, että sillä on positiivinen vaikutus terveyteen. Yleensä tällainen muutos tarkoittaa toimintatapojen muuttamiseen niin, että muutoksilla on terveyttä edistäviä vaikutuksia esim. liikunnan lisäämisellä. WHO (2022) näkee tehokkaan terveysviestinnän koostuvan useista ominaisuuksista, esim. viestin on oltava: ymmärrettävä (understandable), relevantti (relevant) ja uskottava (credible). Thomas (2006, 99–100) pitää tärkeänä myös viestien merkityksellisyyttä sekä viestien synnyttämiä tunneperäisiä stimulaatioita. Tutkimukset ovat osoittaneet, että terveysviestinnällä on merkittävä positiivinen vaikutus terveyteen liittyviin uskomuksiin, asenteisiin ja käyttäytymiseen. Euroopan tautiehkäisy ja -valvontakeskus (ECDC 2021) uskoo toimivan terveysviestinnän koostuvan useista ominaisuuksista (ks. Taulukko 1).

Taulukko 1. ECDC:n luettelemat ominaisuudet edellytyksenä tehokkaaseen terveysviestintään ja viestien kehittämiseen suomennettuna (ECDC 2021).

suomennos	englanniksi	tarkennus
tarkkuus	accuracy	Sisältö on virheetön
saavutettavuus	availability	Sisältö toimitetaan tai sijoitetaan niin, että se on yleisön saatavilla
tasapaino/puolueetto-muus	balance	Sekä hyödyt että riskit esillä. Esitetään eri näkökulmia
johdonmukaisuus	consistency	Sisältö pysyy ajan myötä yhtenäisenä, ja on johdonmukainen muiden muista lähteistä saatujen tietojen kanssa
kulttuurinen kelpoisuus	cultural competence	Otetaan huomioon tiettyjen ryhmien ongelmat, koulutustaso ja mahdolliset kyvyttömyydet
näyttöön perustuvuus	evidence base	Laadukkaat ja kriittisesti arvioidut tieteelliset julkaisut
kattavuus	reach	Sisältö tavoittaa tai on saatavilla mahdollisemman monelle kohderyhmän jäsenille
luotettavuus	reliability	Sisältö on ajan tasalla ja on uskottavaa
toistettavuus	repetition	Sisältöä toistetaan vaikutuksen vahvistamiseksi
oikea-aikaisuus	timeliness	Sisältö on saatavilla heille, jotka sitä tietoa sinä hetkenä eniten tarvitsevat
ymmärrettävyys	understandability	Sisältö on oltava kieli- ja lukutasollisesti sellaisessa muodossa, että se on helposti ymmärrettävä

Viestien kohdentaminen (targetin) on tehokas viestinnän ja markkinoinnin keino. Ihmisten tarpeet ja muut ominaisuudet poikkeavat toisistaan niin paljon, että yksi sama viesti ei toimi kaikille. Yleisö jaetaan ryhmiin, mitä kutsutaan segmentoinniksi. Ryhmät segmentoidaan niin, että ryhmän jäsenillä on jotain yhteistä keskenään. Ryhmät voi jakaa esim. sukupuolen ja iän perusteella tai niiden yhdistelmällä. (Case & Johnson 2012, 202–203, Armstrong, Kotler, Saunders & Wong 2005, 391–393, 399.)

Terveysviestinnän toimivuutta voidaan myös parantaa viestien räätälöinnillä (tailoring). Räätälöinnillä tarkoitetaan viestien sanoman muuttamista siihen muotoon, että uppoaa mahdollisemman tehokkaasti tiettyyn yksilöön. Tarkoitus on saada yksilö toimimaan halutulla tavalla. Pelko on yksi käytetyimmistä terveysviestinnän rakennuspalikoista. Yksilöiden väliset erot, sekä koetun uhan vakavuus, ovat tekijöitä, jotka vaikuttavat elintapamuutoksiin. (Enwald 2013, 121–122.) Viestien kohdentaminen eroaa räätälöinnistä siinä, että kohdennuksen kohteena on usein ryhmä, kun taas viestien räätälöinnissä keskitytään tiettyyn yksilöön (Aldrich, Harrington & Noar 2009, 75–76). Glassman, Kreuter & Strecher (1999, 277) määrittelevät räätälöinnin olevan:

> …any combination of strategies and information intended to reach one specific person, based on characteristics that are unique to that person, related to the outcome of interest, and derived from an individual assessment. (Strecher 1999, 277.)

Verrattuna kohdentamiseen, räätälöinnin etuna on sen henkilökohtaisuus. Esimerkiksi terveydenhoidon ammattilainen voi seurata potilastaan ja mukauttaa viestiä vastaamaan hänen sen hetkisiä tiedontarpeitaan. (Case & Johnson 2012, 202–203.) Viestien räätälöinti voi tosin olla enemmän aikaa ja resursseja vievää viestien kohdentamiseen verrattuna, kun kohteena on vain yksi henkilö.

Myös palautteella (feedback) voi olla positiivinen vaikutus henkilön terveyskäyttäytymiseen. Yli 70 prosenttia nuorista miehistä piti ipsatiivista, n. 60 prosenttia normatiivista ja noin 40 prosenttia teoreettista palautetta liikkumista kannustavana tekijänä (Hirvonen 2015, 105). Ipsatiivinen palaute tarkoittaa nykytilan vertaamista aikaisempaan. Normatiivinen palaute vertaa yksilön käyttäytymistä tai tilaa

hänen vertaisiinsa. Teoreettinen palaute viittaa fakta- ja teoriapohjaisiin argumentteihin, jotka esitetään palautteen saajalle. (Hirvonen 2015, 33.)

3.3 Terveyskäyttäytymiseen vaikuttavat tekijät

Tässä kappaleessa käyn läpi lyhyesti terveyskäyttäytymiseen, eli käytännössä terveysviestinnän toimivuuteen, vaikuttavia tekijöitä pääasiallisesti informaatiotutkimuksen näkökulmasta. On olemassa monia eri tekijöitä, joilla on vaikutusta terveysviestinnän tehokkuuteen, jotka ovat suljettu ulos tämän opinnäytetyön käsittelystä.

ECDC (2021) pitää viestien ymmärrettävyyttä terveysviestinnän toimivuuden kannalta olennaisena. Ymmärrettävyys voidaan yhdistää informaatiolukutaitoon (information literacy) ja etenkin terveyslukutaitoon (health literacy). Terveyslukutaidolla tarkoitetaan kykyä ymmärtää, hakea, arvioida ja soveltaa oppimaansa terveystietoa terveyttä edistävällä tavalla (Brand & Sørensen 2013, 634). Tällä on vaikutusta terveyskäyttäytymiseen; liian vaikea aihe saattaa pakottaa yksilöä lopettamaan tiedonetsinnän kokonaan tai tyytymään helppolukuisempiin epätieteellisiin lähteisiin. Lopulta, ”huonon” tiedon soveltaminen voi johtaa tehottomaan tai jopa henkeä uhkaavaan lopputulokseen.

Erilaiset tekijät tiedonhakuun ja -hankintaan liittyen yksilöiden ja ryhmien välillä vaikuttavat jossain määrin terveyskäyttäytymiseen. Esimerkiksi naiset etsivät aktiivisemmin terveyteen liittyvää tietoa kuin miehet. Naiset ovat myös miehiä aktiivisempia tiedonhakijoita, kun haetaan tietoa sairaudesta, joka koskettaa lähipiiriläistä. Myös iällä on vaikutusta. Tutkimuksissa on myös huomattu, että iäkkäämmät ovat myöntyväisempiä tottelemaan lääkärien neuvoja eivätkä ole niin aktiivisia etsimään lisää tietoa. Tämä ei välttämättä ole hyvä asia, sillä aktiiviset tiedonhakijat ovat parempia varautumaan ongelmiin, harjoittamaan ennaltaehkäisevää käytöstä ja hakeutumaan nopeammin hoitoon. (Case & Johnson 2013, 39–52, 154) Motivointia tiedonhakuun tulisi pitää varteenotettavana terveysviestinnän menetelmänä.

Demografisten tekijöiden lisäksi terveyskäyttäytymiseen vaikuttaa erilaisia sosiokognitiivisia ja -psykologisia tekijöitä. Omakohtaiset kokemukset, varsinkin sairastuminen, on yksi tekijöistä, joka laukaisee tiedontarpeen. Yksilön sosiaaliset verkostot vaikuttavat myös terveyskäyttäytymiseen. Henkilöt, joilla no vahvat sosiaaliset verkostot, ovat myös luultavammin vahvemmin terveyteen orientoituneita. Sosiaaliset verkostot myös lisäävät yksilön tiedonlähteitä ja monipuolistavat niitä. Lisäksi yksilön uskomukset vaikuttavat hänen terveyskäyttäytymiseensä. (Case & Johnson 2013, 51–59.) Esimerkiksi nuoret miehet, jotka omaavat vahvan pystyvyyden sekä hyötyperäisen ja positiivisen tunneperäisen asenteen huolehtivat paremmin hygieniastaan. Myös muiden taholta koetut odotukset vaikuttivat positiivisesti hygieniasta huolta pitämiseen eli sosiaalinen paine. (Hankonen, Haukkala, Jallinoja & Laine 2013, 227.)

Negatiiviset tunteet, kuten pelko ja ahdistus, voivat johtaa lopulta siihen, ettei yksilö hakeudu esim. seulontaan. Toisaalta negatiiviset tunteet voivat toimia joillekin yksilöille motivoivana tekijänä. (Hawkings, Hwang, Lee & Pingree 2008, 359–361.) Tosin, pelon käyttäminen terveysviestinnässä, ja sen tutkimuksissa, on eettisesti arveluttavaa. Positiivisesti ajattelevien ihmisten ajatusprosessi on laajempaa ja monipuolisempaa, mistä voi olla terveyden kannalta hyötyä pitkällä aikatähtäimellä, kun taas negatiiviset tunteet voivat olla hyödyllisiä lyhyellä aikatähtäimellä (Branigan, Fredrickson, Mancuso & Tugade 2000, 239).

Kuten aiemmin sanottu, terveyskäyttäytymiseen vaikuttavia tekijöitä on paljon ja nämä tekijät olisi syytä ottaa huomioon terveysviestinnän suunnittelussa. Paljon jäi mainitsematta, näistä esimerkkinä informaatioähky, ympäristölliset, biologiset ja kulttuurilliset tekijät. Myös preferenssit ja psykologiset vinoumat jäävät tämän opinnäytteen ulkopuolelle. Relevanttiudella ja ajankohtaisuudella on myös merkittävä rooli riippuen taustatekijöistä ja kontekstista.

3.4 Tiedonlähteet terveysviestinnän kontekstissa

Internetin kautta voidaan helposti tavoitella suurta yleisöä. Sosiaalisen median käyttäjät usein jakavat muiden sisältöä tai luovat niitä itse. Huono puoli sisällöntuottamisessa ja -jakamisessa on se, että joskus tieto on osittain tai kokonaan väärää tai harhaanjohtavaa. Tällöin puhutaan dis- ja misinformaatiosta. Misinformaatio on tahatonta väärää, tai puutteellista, tietoa, kun taas disinformaatio on tahallisesti tuotettua väärää tietoa. Malinformaatiosta on kyse, kun ”oikeaa” tietoa jaetaan tarkoituksenmukaisesti haittamielessä. (Pace University 2022.) Tulen viittaamaan tästä lähtien tutkielmassa dis- ja misinformaatioon termillä *väärä tieto*. Puolueettomaan, luotettavaan ja tutkimuspohjaiseen tietoon tulen viittaamaan termillä *oikea tieto*.

Yhä useammat meistä saavat ja hankkivat tietonsa internetistä, joten riski altistua väärälle tiedolle on läsnä. Väärää tietoa voi sisältää blogikirjoitus, Youtube-video, asiantuntijan lausunto, luokkakaverin Facebook-kommentti tai melkein mikä tahansa sisältö. Netissä surffailevan voi olla vaikea välttää väärää tietoa. Väärän tieto voi löytää tiensä myös kahvikeskusteluihin tai sukujuhliin. Tämän vuoksi luotettavat terveydenhuoltoon liittyvät instituutiot, sekä alan ammattilaiset, ovat nyky-yhteiskunnassa avainasemassa oikean tiedon saattamisessa ihmisten keskuuteen. Sananvapaus ja tieteellinen argumentointi ovat tärkeitä elementtejä demokraattisessa yhteiskunnassa, joten sensuuri ei ole välttämättä paras ratkaisu.

Internetin tarjoama sananvapaus (tai joskus keskustelupalstojen tai yhteisöpalveluiden sensuuri) voi välillä toimia yhteiskuntaa heikentävällä tavalla mitä tulee terveyskäyttäytymiseen. Ihmiset tarvitsevatkin luotettavan ja vaikutusvaltaisen auktoriteetin, joka pystyy tuottamaan puolueetonta ja tutkimuksiin pohjautuvaa tietoa ja viestiä siitä eteenpäin heille, jotka sitä tietoa eniten kaipaavat. Moraalisesti väärä tieto on sidoksissa henkilön asenteisiin ja arvoihin. Tiettyjä arvoja edustavat yksilöt ja ryhmän suosivat tietoa, jotka puoltavat heidän arvokäsityksiään, ja suhtautuvat kriittisesti tietoon, jotka uhkaavat sitä. (Haasio, Mattila & Ojaranta 2018, 32.)

Media on internetin ohella myös yksi tärkeimmistä tiedonlähteistämme. suurin osa meistä sai koronaan liittyvää tietoa televisio- ja radiouutisten kautta. Vuosien 2020–2022 aikana Hanna Nohynek, Mika Salminen sekä Asko Järvinen ovat tulleet monille televisiota seuranneelle suomalaiselle tutuksi. Perinteisen median (tv, radio ja sanomalehdet) vahvuuksiin kuuluu levikki; perinteisen median avulla tietoa voidaan levittää laajalle yleisölle. Tämä tarkoittaa myös valitettavasti sitä, että tietoa ei voida räätälöidä tai kohdentaa yksilö- tai ryhmäkohtaisesti. Perinteisen median viestintä perustuu massaviestintään, jossa tietoa ”pusketaan” kansalle ”samassa muotissa.”

3.4.1 Kognitiivinen auktoriteetti

Wilson (1983, 14–16) kutsuu ihmisten pitämiä luotettavia tiedonlähteitä heidän *kognitiivisiksi auktoriteeteikseen*. Wilsonin (1983, 15) mielestä kognitiivinen auktoriteetti liittyy selvästi uskottavuuteen (credibility). Yhdysvallassa tapahtuneessa terveysviestinnän kontekstissa, kognitiivisiksi auktoriteeteiksi voidaan Wilsonin (1983, 81) mukaan lukea erilaisia instituutioita tai organisaatioita, kuten Yhdysvaltain tautikeskus CDC, eikä pelkästään yksilöitä.

Tässä tutkimuksessa kognitiivisiksi auktoriteeteiksi on valittu yhdysvaltalaiset toimijat, sillä heistä keskustellaan enemmän Twitterissä eli heistä on saatavilla enemmän dataa. Kognitiivisiksi auktoriteeteiksi on valittu CDC sekä Joe Biden. Tarkastelun kohteena on tarkemmin sanottuna Twitter-käyttäjien näkemys ja mielipide liittyen edellä mainittuihin kognitiivisiin auktoriteetteihin.

Wilsonin lisäksi ECDC (2021) sekä WHO (2022) pitää luotettavuutta, ja viestien uskottavuutta, viestinnän kannalta tärkeinä elementteinä. On sanomattakin selvää, ettei viestintuojaa (saatikka viestiä) voi ottaa tosissaan, ellei häneen luoteta. Luottamuksen parantaminen tiettyyn auktoriteettiin loisi periaatteessa lisää uskoa tämän tuottamiin viesteihin. Jotta kansalaisten sen hetkinen mielipide kognitiivisiin auktoriteetteihin saataisiin tiedettyä, on teetättävä tutkimus, yleensä mielipidetutkimus. Seuraavassa

kappaleessa perehdytään lyhyesti mielipidetutkimukseen, sen hyötyihin ja haittoihin, sekä pohditaan sille mahdollisia korvaajia.

3.5 Mielipidetutkimus ja tunneanalyysi sen korvaajana

Mielipidetutkimus on työkalu, jonka tavoitteena on selvittää ihmisten asenteita ja mielipiteitä liittyen tiettyyn aiheeseen. Mielipidetutkimuksia tehdään kyselyinä, joista yleisimpiä ovat puhelinalla ja kasvotusten toteutettavat haastattelut. Muita keinoja ovat mm. posti- ja internetkyselyt. Kyselyihin valitaan osallistujia laaja-alaisesti ja sattumanvaraisesti. (Gallup 2007, 2.)

Tunneanalyysi on tutkimusmenetelmä tai työkalu, jolla selvitetään sisältöjen heijastamia tunteita. Sisällöt voivat olla monessa eri muodossa, mutta yleensä tunneanalyysin avulla tutkitaan tekstimuodossa olevaa sisältöä. Tietokoneen teettämässä tunneanalyysissä sanasto- (lexicon) tai tekoälypohjainen (artificial intelligence) ohjelma laskee sanojen arvoja tekstissä, ja ilmoittaa onko teksti esim. positiivinen, neutraali tai negatiivinen. (Taboada 2006, 325–333, Abhinandan 2021, Boldenthusiast 2019.)

Sanastopohjainen tunneanalyysi vertaa tekstin sanoja sanastoon, jossa jokaiselle sanalle on annettu ennalta määritelty tietty arvo riippuen sanan sentimentistä. Yksinkertaisesti sanottuna: positiiviset sanat ja virkkeet saavat positiiviset arvot, ja negatiiviset saavat negatiiviset arvot. Virke: ”vihaan maanantaita” sisältää sanan ”viha”, jolla on sanastossa tietty arvo (esim. negatiivinen arvo), ja taas sana ”maanantai” on luultavasti arvoltaan neutraali. Nämä kaksi arvoa ohjelma laskee yhteen tietyllä funktiolla, ja ilmoittaa virkkeen olevan sentimentiltään negatiivinen. Tekoälypohjainen ohjelma taas on oppinut itse useista tuhansista tekstinpätkistä laskemaan, mikä sentimentti milläkin tekstillä on. Molempien ohjelmamallien prosessit ovat todellisuudessa paljon teknisempiä ja monimutkaisempia, mutta niihin ei tässä kohtaa tutkielmassa tätä syvemmälle uppouduta. (Abhinandan 2021).

Eräässä tutkimuksessa Trumpiin liittyviä tviittejä analysoitiin tunneanalyysin avulla 2016 Yhdysvaltain presidentinvaalin aikana. Aineistoa verrattiin sinä aikana tehtyihin mielipidemittauksiin ja niiden välillä havaittiin peräti 94 prosenttinen positiivinen korrelaatio sanastopohjaisella ja 85 prosenttinen korrelaatio koneoppineella algoritmilla. Hillary Clinton kohdalla, korrelaatio oli heikompi, välillä negatiivinen. (Deng & Joyce 2017, 4.)

Myös useampi tutkimus on osoittanut korrelaatiota Twitter-keskusteluista tehtyjen tunneanalyysien ja politiikkaan liittyvien mielipidemittauksien välillä. Presidentti Obaman kannatusluku vuonna 2009 (job approval rating) oli eräässä tutkimuksessa niin myötäilevä Twitteristä saadun datan kanssa, että tutkijat sanoivat tekniikan näyttävän lupaavalta vaihtoehdolta mielipidemittauksien rinnalle tai peräti sen korvaajaksi (Balasubramanyan, O'Connor, Routledge & Smith 2010, 128). Toisessa tutkimuksessa analysoitiin yli 100 000 tviittiä koskien Saksan 2009 liittopäivävaaleja. Tutkijoiden mielestä tviitit reflektoivat perinteisiä mielipidemittauksien tuloksia. (Tumasjan, Sandner, Sprenger & Welpe 2010, 183.) Kaikki tutkijat eivät kuitenkaan ole vakuuttuneita. Bermingham ja Smeaton (2011, 9) pitävät Twitter-datan käytettävyyttä ennustamistarkoituksiin epäselvänä.

3.6 Muut mahdolliset työkalut ja keinot terveysviestinnän apuna

Internet tarjoaa paljon aineistoa ja vaikuttamiskeinoja tehokkaamman terveysviestinnän kannalta. Keräsin tähän kappaleeseen hypoteettisia terveysviestinnällisiä keinoja. Alla mainitut hypoteettiset työkalut eivät suoranaisesti liity tutkielmaan, vaan ne on kirjoitettu ajatellen lukijaa, joka saattaa olla kiinnostunut terveysviestinnästä tai työskentelee sen saralla. Kappaleessa mainitut työkalut ja keinot perustuvat omakohtaisiin kokemuksiin HTML:stä, JavaScriptistä, tunneanalyysistä, hakukoneoptimoinnista ja -markkinoinnista, joita olen itse vuosien aikana hankkinut. Näistä hypoteettisista keinoista ei yksinkertaisesti ole olemassa puolueettomia tutkimuksia, jonka vuoksi viittauksia tieteellisiin teoksiin ei ole.

Terveydenhuollon auktoriteetit voivat ottaa käyttöönsä muitakin työkaluja terveysviestinnän seuraamiseen ja tehostamiseen. Kävijälaskuri (web counter) laskee, kuinka moni klikkasi annettua linkkiä tai kuinka moni kävijä ylipäätään löysi tiensä tietylle nettisivulle. Kävijälaskurin avulla voidaan seurata linkkien jakamisen tai nettisivun mainostamisen tehokkuutta.

Evästeet (cookies) tuovat monipuolisesti erilaisia mahdollisuuksia. Evästeiden avulla nettisivustojen ylläpitäjät mm. saavat tietoonsa, mitä kautta kävijä löysi tiensä nettisivustolle, miten kauan he sivustolla viipyivät ja miten he navigoivat sivustolla. Näistä tiedoista on hyötyä itse nettisivun käytettävyyden takia. Helppokäyttöiset nettisivut voivat tarkoittaa sitä, että kävijät viihtyvät nettisivustolla kauemmin.

Evästeet voivat myös kerätä käyttäjistä tietoa, jonka avulla ne voivat tarjota käyttäjälle kohdennettuja mainoksia tai sisältöä. Jos suomalainen etsii Youtubesta tietoa tietystä viruksesta, heille voitaisiin tarjota esim. THL:n teettämiä Youtube-videoita tai mainoksia (linkkejä), jotka tarjoavat lisätietoa aiheesta. Tämä voidaan toistaa sairauksien, riippuvuuksien, mielenterveysongelmien ja monien muiden terveyteen liittyvien asioiden kanssa.

Hakukoneoptimointi ja -markkinointi ovat myös keinoja, jolla nettisivustojen ylläpitäjät voivat lisätä näkyvyyttä hakukoneissa. Hakukoneoptimointi tarkoittaa erilaisia teknisiä ja sisällöllisiä tapoja, joita nettisivustolle tehdään, jotka nostattavat sivuston näkyvyyttä eli sijaa (ranking) esim. Googlen hakukoneessa. Jos tiedonhakija googlettaa koronarokotteen sivuoireista, olisi parempi, että hän saisi hakutuloksen kärkeen linkin luotettavalle sivustolle, eikä sivustolle, jossa levitetään perättömiä väitteitä. Hakukonemarkkinointi tarkoittaa sijan ostamista. Jossain hakusanoissa ja -lausekkeissa ensimmäiset Googlen hakutulokset ovat mainoksia, ja sen huomaa siitä, että linkkien edessä lukee ad tai mainos.

Google rankingin merkityksestä verkkoliikenteeseen, eli kävijämäärän prosentuaalisesta jakautumisesta sijasta riippuen (ensimmäisenä näytetyt linkit saavat eniten klikkauksia), on olemassa useita eri yritysten tuottamia tutkimuksia. Nämä tutkimukset ovat pääasiassa hakukoneoptimointia tuottavien yritysten teettämiä, joten tutkimuksien luotettavuus on

kyseenalaista. Tieteellisiä artikkeleita asiaan liittyen ei ollut, mutta kun huomioidaan Zipfin (1949) *vähemmän vaivan laki* (the principle of least effort), niin tiedetään, että ihmisillä on tapana hakea tietoa mahdollisemman vaivattomasti. Tämä tarkoittaa käytännössä sitä, että ensimmäiset hakutulokset saavat eniten klikkauksia, minkä vuoksi ensimmäinen sija Googlen hakutuloksissa on tärkeää, kun halutaan taistella huomiosta.

Tekstinlouhintaa voidaan teoriassa käyttää myös eri tavoin hyödyksi. Yksi tapa on käyttää tekstinlouhintaa tunnistamaan verkossa esim. rokotevastainen henkilö. Henkilöä voidaan sitten lähestyä yksityisviestien kautta, yrittäen muuttaa hänen näkökantaansa erilaisten keinojen, esim. todisteiden avulla. Hänen julkisesti esillä oleviin väittämiinsä voidaan vastata vasta-argumentteja avulla. Tämä esimerkki on eettisesti arveluttava, sillä se muistuttaa dystooppista ”isoveli valvoo”-tilannetta. Toisaalta väärän tiedon levittäjät, botit tai muut ongelmalliset tahot, ja heidän levittämänsä väärät tiedot, voidaan paljastaa yleisölle ennen kuin ne leviävät tai niistä koituu enemmän harmia.

Tunneanalyysin kenties optimaalisin esimerkki olisi seurata yksittäisten viestien, esim. tviittien, tuottamaa reaktiota. Tietty taho, esim. THL, voisi tviitata viestinsä ja seurata sen siihen liittyviä kommentteja ja tunteita. Jos tviitti saa paljon negatiivista palautetta, voidaan tviitti poistaa, korvata tai siihen voi liittää lisätietoa ennen kuin se aiheuttaa liian paljon tuhoa organisaation maineelle, narratiiville, johdonmukaisuudelle tai mille tahansa muulle elementille, jolla on vaikutusta organisaation luotettavuuteen ja uskottavuuteen negatiivisella tavalla.

3.6 Internet informaatioympäristönä

Internetistä on tullut tärkeä terveyteen liittyvän tiedon hankintakanava useille ihmisille. Tilastokeskuksen (2020) tutkimuksessa käy ilmi, että vuonna 2020 72 prosenttia suomalaisista etsi tietoa liittyen sairauksiin, ravitsemuksiin tai terveyteen internetistä. Nettisivustojen, tiedon, datan ja sisällöntuotannon kasvanut määrä tarkoittaa sitä, että tietoa on todella paljon tarjolla, välillä liikaakin.

Samalla mis- ja disinformaation määrä on kasvanut. Internetin käyttäjien nauttima anonymiteetti tarjoaa eräänlaisen suojan, jonka varjossa osa ihmisistä voi käyttäytyä epäeettisesti esim. jakamalla väärää tietoa, valehdella tai johdatella muita harhaan. Botit ja trollit voivat toimia ”terveysviestinnän aseena” levittämällä väärää tietoa ja aiheuttamalla eripuraisuutta osapuolten välillä. Tämä käy ilmi 2014–2017 vuosina tehdyssä tutkimuksessa, jossa tutkijat antoivat ilmiölle nimen *Weaponized Health Communication* eli aseellinen terveysviestintä. (Broniatowski ym. 2022, e1–e6.)

Sosiaaliseen mediaan lukeutuvat yhteisöpalvelut ovat alustoja, jotka tarjoavat käyttäjille mahdollisuutta luoda sisältöä ja jakaa niitä (Twitter, Youtube jne.). Lisäksi voidaan tarjota mahdollisuutta kommunikoimiseen. Sisältö voi olla tekstiä, kuvia, videoita, ääniä yms. Yhteisöpalveluiden käyttäjillä on periaatteessa aina mahdollisuus olla sosiaalisessa vuorovaikutuksessa toisten käyttäjien kanssa.

Yhteisöpalveluiden heikkouksiin kuuluu väärän ja vahingollisen tiedon leviäminen. Joskus yhteisöpalveluiden jäsenet jakaantuvat ideologisesti, poliittisesti tai ikäryhmittäin eri alustojen alle, jolloin vaarana on se, että yhteisöistä muodostuu eräänlainen kaikukammio (echo chamber), joissa tietty näkökulma tai maailmankuva on hyväksytty ja muut taas ei. Kaikukammio (myös sosiaalinen kupla) on tilanne, jossa syystä tai toisesta, verkkoyhteisön jäsenistä on muodostunut samaa maailmankuvaa kannattava joukko (pieniä poikkeuksia lukuun ottamatta). (Cinelli, Galeazzi, Morales, Starnini & Quattrociocchi 2021, 1–3.)

Yhteisöpalveluissa vaarana on sen muuttuminen ns. kaikukammioksi. Kaikukammiolla tarkoitetaan sitä, että keskustelupalvelusta tai sosiaalisesta mediasta tulee poliittisesti tai ideologisesti sellainen ympäristö, jossa hyväksytään lähinnä tietyt näkökulmat. Tämä tarkoittanee pahimmassa tapauksessa sitä, että vastapuolen ajatukset, eli viestit, ja käyttäjätunnukset voidaan poistaa tai algoritmi voi piilottaa käyttäjän ja hänen tuottamansa sisällöt (shadow banning).

Blank ja Dubois (2018, 737–741) uskovat kaikukammion olevan liioiteltu ilmiö. Blank ja Dubois (2018, 730) kuitenkin totesivat tekstissään, että aiemmissa tutkimuksissa on

kyllä huomattu todisteita kaikukammioilmiöstä Twitterissä, kun taas Facebookissa (Meta) ei ole havaittu samankaltaista trendiä. McClainin, Riveron, Smithin ja Widjayan (2021, 2) huomiota herättänyt tutkimus osoitti, että Twitterin käyttäjät ovat pääasiassa ideologisesti lähellä Yhdysvaltain demokraattista puoluetta (45 demokraatit, 27 republikaanit), ja aktiivisin käyttäjäryhmä (päivittäisten vierailuiden perusteella) oli liberaalit demokraatit.

Arviddsonin, Colleonin ja Rozzan (2014) data-analytiikka Twitterin big datasta osoitti, että demokraatit suosivat ystävikseen muita demokraatteja paljon useammin kuin republikaanit. Tällä on siten vaikutusta mm. Twitterin tarjoamaan sisältöön, jota käyttäjille näytetään ja näkökantoihin, joihin hän siellä törmää. Kaikukammioilmiö voi osoittautua ongelmalliseksi sillä se heikentää tiedon, tiedonlähteiden ja näkökulmien monipuolisuutta rajaamalla, ja joskus jopa sensuroimalla, vastapuolen argumentteja ja näkemyksiä riippumatta siitä ovatko ne faktuaalisia tosiasioita taikka ei.

Tutkielman aineisto on ladattu Twitteristä. Kuten mainittiin, Twitterin käyttäjäkunta on poliittisesti ajatellen enemmän vasemmalla kuin oikealla. Käyttäjäryhmän näkemykset voivat siis erota paljonkin siitä, mitä internetin ulkopuolella eli todellisuudessa, ihmiset ajattelevat. Konservatiivit yhdysvaltalaiset olivat enemmän koronarokotteita vastaan kuin liberaalit (Cowan, Mark & Reich 2021, 1–2). Euroopassa ei ollut samanlaista oikeiston ja vasemmiston kahtiajakoa, vaan lähinnä populismin kannattajat olivat enemmän koronarokotteita vastaan (Carter, Lyons, Reifler & Stoeckel 2022, 4–5, Kennedy 2019, 515–516).

Syy siihen, miksi Yhdysvalloissa nähtiin rokotteen osalta oikeiston ja vasemmiston kahtiajako piilee osittain siinä, että oikeistokonservatiivit (republikaanit) olivat koronarajoituksia, maskipakkoja ja pakollisia rokotuksia vastaan, kun taas vasemmisto, sekä maltilliset demokraatit, olivat niiden puolesta. CDC sekä Joe Biden olivat pandemian alkuvuosina koronarajoituksien puolestapuhujia. Tästä syystä vasemmistovoittoisen Twitterin käyttäjät ei välttämättä ajattele yhtä negatiivisesti näitä kolmea kognitiivista auktoriteettia kohtaan mitä käyttäjät muissa yhteisöpalveluissa, tai mitä ihmiset todellisuudessa ajattelevat. Twitter saattaa olla siis tässä tapauksessa

alustana huono vaihtoehto mielipiteen mittaamiseen tunneanalyysin avulla, sillä sen käyttäjäkunta ei edusta Yhdysvaltain kansalaisia realistisesti.

Yhteisöpalvelut tarjoavat myös paljon hyvää. Käyttäjät pystyvät puuttumaan väärinymmärryksiin ja valheelliseen tietoon hyvinkin nopeaa. Käyttäjät voivat myös jakaa terveyttä edistävää sisältöä ja tietoa. Kenties yksi parhaimmista yhteisöpalveluiden käyttömahdollisuuksista on vertaistuki. Sairastuneille vertaistuki on henkisen hyvinvoinnin kannalta tärkeää, sillä se auttaa selviämään sairauden aiheuttamasta psyykkisestä rasituksesta. Sairautta kärsivä voi lukea muiden kokemuksista, kommentoida niitä, jakaa ja saada uutta tietoa, sekä antaa ja vastaanottaa henkistä tukea (Case & Johnson 2013, 179–182). Yhteisöpalvelut auttavat samaa sairautta kärsiviä tarjoamalla heille alustan, eli ympäristön, jossa he voivat keskustella, jakaa kokemuksia ja auttaa toinen toistaan.

3.7 Yhteenveto

Internetin käyttäjämäärä on kasvanut vuosi vuodelta, ja kasvu tulee jatkumaan. Monet ihmiset käyttävät internetiä terveyteen liittyvän tiedonlähteenä (Tilastokeskus 2020b). Internetiin voi periaatteessa kuka vain tuottaa tietoa, myös väärää tietoa. Moni törmää sattumalta väärään tietoon, josta voi koitua negatiivisia seuraamuksia. Koska tietoa voi tuottaa kuka tahansa, on auktoriteeteilla tärkeä tehtävä säilyttää ja pitää huolta omasta imagostaan harjoittamalla toimivaa viestintää, tiedottamalla ammattitaitoisesti, sekä tuottamalla laadukasta ja uskottavaa sisältöä. Terveystiedon kognitiivisten auktoriteettien on taisteltava nykyaikaisessa maailmassa asemastaan pääasiallisena tiedonlähteenä, sillä vaihtoehtoista tietoa pystyvät tuottamaan muutkin kuin pelkästään auktoriteetit itse.

Tehokas ja onnistunut terveysviestintä vaatii ennen kaikkea kohdeyleisön huomioon ottamista. Kulttuurilliset sekä ryhmä- ja yksilökohtaiset tekijät on hyvä ottaa huomioon. Viestinnän toimivuutta voidaan tehostaa kohdentamalla, räätälöinnillä ja palautteen avulla (Enwald 2013, 141, Hirvonen 2015, 160). Viestien ymmärrettävyys sekä kohdeyleisön kulttuurilliset ja sosiologiset tekijät on hyvä myös ottaa huomioon. Viestin

tuojaan ja itse viestiin liittyvä yleisön luottamus ja uskottavuus on tärkeää, jotta kohdeyleisö saadaan toimimaan halutulla tavalla.

Pro gradu -tutkielmani ensimmäinen tutkimuskysymys kuului: ”Millä tavoin SBD:a voitaisiin hyödyntää terveysviestinnässä?”. Kirjallisuuskatsauksen perusteella, tunneanalyysillä voidaan selvittää mielipiteitä, ainakin usean politiikkaan ja äänestämiseen liittyvän tutkimuksen perusteella. Tekstinlouhinnalla voidaan löytää terveyskäyttäytymiseen liittyviä tekijöitä ja esimerkiksi puuttua itsetuhoiseen käytökseen. Viestinnän onnistuvuutta voidaan seurata kävijälaskurin avulla, ja nettisivun näkyvyyttä hakukoneissa voidaan parantaa hakukoneoptimoinnilla ja -markkinoinnilla.

Kognitiivisten auktoriteettien luotettavuutta ja uskottavuutta voi mahdollisesti pystyä seuraamaan sosiaalisen median keskusteluista saatujen aineistojen perusteella. Näiden aineistojen tutkimiseen voidaan hyödyntää tekoäly- ja sanastopohjaisia tunneanalyysin menetelmiä. Tunneanalyysiä on tutkittu poliittisiin tarkoitusperiin esim. vaalitulosten vertaamista mielipidemittauksiin ja jopa vaalitulosten ennustamiseen, mutta tulokset ovat olleet ristiriitaisia. Tutkielman empiirinen osuus voi ”onnistuessaan” tarjota vahvistuksen tunneanalyysin käyttöön osana terveysviestintää.

Tässä tutkielmassa pyritään löytämään lisäksi vastaus siihen, voiko Twitteristä saatua SBD:aa käyttää perinteisten mielipidemittauksien korvikkeena. Jos näyttää siltä, että voi, niin silloin tunneanalyysiä voidaan potentiaalisesti käyttää terveysviestinnässä työkaluna. Tunneanalyysillä voidaan hypoteettisesti selvittää esimerkiksi sitä, millä kognitiivisella auktoriteetilla on paras maine ja kenellä ei, jotta heidän väliltään voidaan valita paras edustaja median eteen. Tunneanalyysillä voidaan selvittää myös, miten paljon kansan luottamusta tietyllä kognitiivisella auktoriteetilla, esim. terveydenhuollon organisaatiolla, on. Tutkielman empiirinen osuus (luku 5.) voi tarjota vastauksen myös ensimmäiseen tutkimuskysymykseen.

4. TUTKIMUSMENETELMÄ JA -AINEISTO

Tutkimukseni käsittelee big datan hyödyntämistä terveysviestinnässä ja sen mahdollisuukssista korvata mielipidemittaus. Tässä luvussa kerron lyhyesti tunneanalyysin työkaluista eli sanastopohjaisesta VADER- ja tekoälypohjaisesta BERT-mallista. Kuvaan aineistonkeruutani, tutkimusprosessiani sekä aineistoani eli tutkimuksen kohteeksi valittuja tviittejä. Kappaleessa pohdin myös tutkimusetiikkaan liittyviä asioita.

4.1 Tutkimusmenetelmä

Molemmat tutkielmaani valikoituneet tunneanalyysityökalut on koodattu Python-ohjelmointikielellä. Ensimmäinen tunneanalyysimalli pohjautuu Googlen kehittämää koneoppimismalliin nimeltä BERT (Bidirectional Encoder Representations from Transformers). BERT-koneoppimismalleja voidaan hyödyntää myös muuhun kuin tunneanalyysiin, sen voi esimerkiksi opettaa tuottamaan tekstiä. Jos haluaa käyttää tekoälyä tutkimukseen, tutkijalla on edessä kaksi vaihtoehtoa: joko hänen pitää kouluttaa se itse, tai käyttää valmiiksi opetettua mallia (pretrained model). Valmiiksi opetettu malli on useimmissa tapauksissa paras vaihtoehto. Käyttämäkseni malliksi valikoitui NLP Town-yhtiön tuottama valmiiksi opetettu BERT-tekoälymalli *bert-base-multilingual-uncased-sentiment,* sillä se suoriutui muihin malleihin verrattuna parhaiten. BERT on tällä hetkellä yksi käytetymmistä ja tunnetuimmista tekoälymalleista. NLP Townin malli osaa tuottaa tunneanalyysin kuudella eri kielellä, ja se on alun perin tehty arvioimaan tekstin sentimentin arvolla yhdestä viiteen, jossa arvo yksi on erittäin huono, arvo kolme neutraali ja arvo viisi erittäin hyvä. Malli on koulutettu lukemalla arvosteluita asiakkaiden ostamista palveluista ja tuotteista, sekä niihin liitetyistä arvosanoista (NLP Town 2022.) Mallin saa kuitenkin muutettua siten, että se luettelee tekstin joko positiiviseksi, neutraaliksi tai negatiiviseksi. Tämä onnistuu, kun arvot 1–2 tulkitaan negatiiviseksi, arvo 3 neutraaliksi ja arvot 4–5 positiiviseksi. NLP Townin valmiiksi opetettu malli suoriutui parhaiten verrattuna muutamaan muuhun kokeilemaani malleihin. Tämän vuoksi valitsin heidän tekoälymallinsa tutkielman empiiriseen osuuteen.

Toinen opinnäytetyössäni käyttämä tunneanalyysisovellus hyödyntää avoimeen lähdekoodiin perustuvaa VADER-työkalua. VADER (Valence Aware Dictionary and sEntiment Reasoner) on nimenomaan tarkoitettu sosiaalisen mediasta saadun aineiston tunneanalyysiin (Gilbert & Hutto 2014). VADER on melko yleisesti käytetty tunneanalyysin työkalu, varsinkin tutkimuskäytössä. VADER perustuu sanasto- ja sääntöpohjaiseen malliin, jossa tekoäly muuttaa sanat perusmuotoon, ja etsii ne sanastosta, johon sanoille on merkitty tietty arvo. Arvo määrittää, ovatko tekstin sanat kokonaisuudessaan positiivisia, neutraaleita vai negatiivisia.

Testasin Python-ohjelmien tunneanalyysimallien tarkkuuksia teksteistä koostuvien datasettien avulla. Testin aikana ohjelmalle syötettiin pelkästään positiivisia tai negatiivisia tekstejä, jotta saatiin selville mallien tarkkuus. Neutraaleita tekstejä ei testattu, sillä viestien neutraalisuus on subjektiivista ja tulkinnanvaraista. Yleensä mallien tekijät ilmoittavat tarkkuuden, mutta objektiivisuuden nimissä, päätin testata ne itse. Testissä käytettyjä datasettejä kokeiltiin yksittäin ja yhdistelemällä. VADER:n tarkkuus positiivisten viestien kohdalla oli 83–89 prosenttia ja BERT:n (NLP Townin valmiiksi opetettu malli) 75–79 prosenttia. Negatiivisten viestien kohdalla VADER:n tarkkuus oli 80–85 prosenttia ja BERT:n 87–91 prosenttia.

Testien perusteella sanasto- ja sääntöpohjainen VADER osasi luetella positiiviset tekstit paremmin, kun taas tekoälyyn pohjautuva BERT luokitteli paremmin negatiiviset tekstit. Testatut datasetit koostuivat hyvin yksinkertaisista, selvästi positiivisista tai negatiivisista, oikeinkirjoitetuista tviiteistä. On syytä muistaa, että tviitit ovat usein todellisuudessa kaikkea paitsi näitä. Testiaineistot olivat usein kooltaan alle 100 tekstiä, ja ne olivat yleensä lyhyitä ja yksinkertaisia.

4.2 Tutkimusprosessi (KDD-PROSESSI)

Tutkimus noudatti suurilta osin tiedonlouhinnan KDD-prosessia (katso luku 2.3). Datan puhdistus ja datasettien yhdistely eivät olleet olennaisia askeleita, joten ne jätettiin

suorittamatta. Datasta karsittiin epäolennainen data, ja se muutettiin tietorakenteeseen, joka sopii tiedonlouhintaan (tekstinlouhintaan). Itse tiedonlouhinta suoritettiin Python-koodilla. Tulokset listattiin Excel-taulukkoon. Lopuksi tulokset visualisoitiin Pythonilla.

4.3 Tutkimuskohde

Pro gradu -tutkielmani tutkimuskohteena ovat Twitter-käyttäjien keskustelut, ja niihin liittyvät asenteet ja mielipiteet. Twitter on yhdistelmä sosiaalista mediaa sekä mikroblogia. Tviitit (tweets) ovat 280 merkin pituisia viestejä, joita voi kommentoida tai jakaa (retweet) eteenpäin. Tviiteistä voidaan myös tykätä, ja niitä voi lisäksi jakaa (share).

Twitterin ilmoittama päivittäinen kävijämäärä koostuu käyttäjistä, joita voi rahallistaa (Average monetizable daily active usage). Vuoden 2022 ensimmäisellä neljänneksellä päivittäinen kävijämäärä oli Twitterin mukaan keskiarvoltaan 229 miljoonaa (Twitter 2022). Twitter ilmoitti, että bottien määrä sivustolla on alle viisi prosenttia (Dang 2022). Twitterin ulkopuolinen lähde ilmoitti bottien määrän olevan aktiivisten käyttäjien keskuudessa yhdeksän ja viidentoista prosentin välillä vuonna 2017 (Davis, Ferrara, Flammini, Menczer & Varol 2017, 288).

4.4 Tutkimusaineisto

Tutkimusaineistonani toimivat Twitter-keskustelut liittyen kahteen ajankohtaiseen yhdysvaltalaiseen terveysviestinnän kognitiiviseen auktoriteettiin: yhdysvaltalainen sairauksista ja valvonnasta vastaava viranomaisorganisaatio CDC:hen sekä istuvaan presidentti Joe Bideniin. Tutkimusaineisto ladattiin aikavälillä 15.2-15.7.2022. Twitter-keskusteluista koostuvat datasetit ladattiin kerran kuukaudessa, noin 30 päivän välein (15.2., 15.3., 15.4. jne.).

Tutkimusaineisto koostuu lähinnä englanninkielisistä tviiteistä. CDC:hen liittyvää aineistoa ei puhdistettu, sillä puhdistaminen hidastaisi prosessia merkittävästi, ja

tutkielman tarkoituksena on tutkia nimenomaan nopeaa ja vähän resursseja kuluttavaa tapaa mielipidetutkimuksen korvikkeeksi. Halutessaan tutkija voi puhdistaa aineiston linkeistä ja Pythonin lisäämistä rivienvaihdoista (”\n”), mutta molemmilla oli mitätön vaikutus tunneanalyysien tuloksiin.

Joe Bideniin liittyvä aineisto jaettiin kahteen osaan: raakaan ja puhdistettuun aineistoon. Raaka aineisto sisälsi kaikki, mukaan lukien uudelleentviittaukset (retweetit) ja puhdistettu data sisälsi pelkästään tviitit. Lisäksi aineistosta poistettiin spämmitviitit. Spämmitviitit olivat pääasiassa aiheeseen kuulumattomia promootioita ja kylkiäisten (giveaways) mainostamista. Puhdistus tehtiin, koska halusin nähdä, oliko sillä mitään vaikutusta tuloksiin. Puhdistus tehtiin manuaalisesti, mikä oli aineiston suuresta koosta johtuen aikaa vievää.

Aineisto koostui yhteensä noin 233 000 tviitistä. Yksittäisten datasettien koko vaihteli 2780 tviitistä vajaaseen 18 000 tviittiin. Datasettien sisältö vaihteli laadultaan merkittävästi. Esimerkiksi helmikuun #CDC-datasetti sisälsi 2 spämmäävää käyttäjää, josta molemmat joko lähettivät n. tuhat identtistä tviittiä ja jonka seuraajat uudelleentviittasivat heidän spämmiviestejänsä (datasetti koostui 8643 tviittistä). Toisen käyttäjän viesti ei edes liittynyt millään tavalla CDC:hen, vaan hän käytti kyseistä hashtagia tuntemattomasta syystä, luultavasti huomion saamiseen.

Sama datasetti sisälsi lisäksi paljon ulkomaalaisia tviittejä, lähinnä espanjaksi. Maaliskuun #CDC-datasetti taas sisälsi hyvin vähän spämmiä ja ulkomaankielisiä tviittejä. VADER-ohjelma luokitteli ulkomaalaiset tekstit usein neutraaliksi, ja monikielinen BERT-malliin pohjautuva ohjelma taas antoi niille jonkin arvon. En osaa espanjaa, joten en ole täysin varma BERT-mallin espanjankielisen tekstin tarkkuudesta. Alle on koottu lista aineisosta ja sen koosta (ks. Taulukko 2.). Jokainen aineisto koostuu kuudesta datasetistä.

Taulukko 2. Lista ladatuista aineistosta, sekä niiden yhteenlasketusta koosta, joka on ilmaistu tviittien määränä.

Datasetit:	#CDC	CDC	Joe Biden	Joe Biden (s)*
tviitit (noin)	39 000	79 000	95 000	20 000

* Siivotut Joe Biden-datasetit

4.4.1 Kolmansien osapuolien teettämät mielipidemittaukset

Tutkimuksessani Twitter-datasta saatuja tunneanalyysejä verrataan kolmansien osapuolien teettämiin mielipidemittauksiin, sillä Suomesta käsin on vaikea mitata yhdysvaltalaisten mielipiteitä. Tästä syystä jouduin luottamaan muiden tahojen teettämiin mielipidemittauksiin. Ennen datasettien lataamista, kognitiivisista terveydentiedon auktoriteeteista, etenkin Joe Bidenista ja CDC:stä löytyi hyvin mielipidemittauksia, mutta aineiston lataamisen aikana, niitä ei valitettavasti julkaistu yhtä tiheästi. Tämä tosiasia oli tutkielman isoin heikkous, sillä jouduin luottamaan siihen, että mielipidemittauksia järjestettäisiin jatkossa.

Löysin kaksi eri mielipidemittausta, joita vertaan tunneanalyysin tuloksiin. Ensimmäinen on Morning Consultin (2022) viikoittainen mielipidemittaus ajalta 15. helmikuuta – 15. kesäkuuta (4 kuukautta). Morning Consultin (2022) mukaan heidän viikottaisiin mielipidemittauksiinsa osallistuu noin 1900 rekisteröitynyttä äänestäjää, ja virhemarginaaliksi he ilmoittavat ±2 prosenttiyksikköä. Tulokset ilmoitetaan nettona, eli tyytyvyys (*approval)* miinus tyytymättömyys (*disapproval)*. Mielipidemittauksessa näkyvät molemmat tutkitut auktoriteettiasemassa olevat toimijat eli Joe Biden, sekä CDC.

Toinen mielipidemittaus, jota käytän tunneanalyysin tulosten vertaamiseen, on FiveThirtyEightin (2022) mittaus kuuden kuukauden ajalta. He esittävät useiden eri arvostettujen lähteiden teettämiä mielipidemittauksia yhdistelemällä niitä. Tässä tutkielmassa käytin vaihtoehtoa: ”All Polls”. Heidän mielipidemittauksensa mittaa kannatuslukua (approval rating). (FiveThirtyEight 2022.)

4.5 Työkalut

Tviittiaineisto ladattiin NCapturella, joka on selaimelle asennettava NVivo-ohjelmistoon liittyvä lisäosa. Se pystyy lataamaan tviittejä tietyn avainsanan perusteella. Se, millä ehdoin ja millä logiikalla Twitter antaa tviittejään ladattavaksi, ei ole yleisessä tiedossa. NVivo-ohjelmaa käytettiin pelkästään viemään (export) eli muuttamaan NCapturen NVCX-tiedostomuodot XLSX-muotoon (Excel). Tämän jälkeen aineisto muutettiin Excelillä CSV-muotoon, sillä kyseistä tiedostomuotoa on helpompi käsitellä Pythonilla. Exceliä käytettiin datan pyyhkimiseen kaikesta tarpeettomasta tiedosta, esim. tviittien kirjoittaja, aika, spatiaalinen data sekä muut metatiedot poistettiin. Excelin käyttö ei ole välttämätöntä, sillä isoja tietomääriä käsitellessä, se toimii Pythonia hitaammin. Excelillä on tosin huomattavasti helpompi pyyhkiä sarakkeet pois verrattuna siihen, jos puhdistuksen olisi tehnyt koodaamalla.

Tunneanalyysit on koodattu Pythonilla (versio 3.8.8). Python on yksi yleisimmistä ohjelmointikielistä, ja se on suosittu etenkin sen helppokäyttöisyyden vuoksi. Python ladattiin osana Anaconda-nimistä pakettia, jossa on paljon data-analytiikkaan, koneoppimiseen ja muihin tieteellisiin tarkoituksiin käytettyjä ominaisuuksia. Koodit tehtiin Jupyter Notebook-ympäristössä. Anakonda-paketti sisältää myös Orange -nimisen työkalun, joka on aloittelijaystävällinen tiedonlouhinnan ohjelma.

Koodeissa on käytetty kolmea eri kirjastoa. BERT-tekoälymalli käyttää apunaan Transformers- ja Torch-kirjastoja, ja sanastopohjainen malli käyttää VaderSentiment-nimistä kirjastoa. Pythonin kirjastoilla viitataan joukkoihin moduuleita. Moduulit ovat kooditiedostoja. Kirjastojen ansiosta, samaa koodinpätkää ei tarvitse kirjoittaa uudelleen vaan kirjastoja voi tuoda (import) osaksi koodia. (Geeksforgeeks 2021.) Kirjastojen voidaan karkeasti ajatellen olevan lisäosia, jotka tuovat lisäfunktioita koodiin, ilman että niitä tarvitsee itse kirjoittaa.

Transformers- ja Torch-kirjastot liittyvät koneoppimiseen, neuroverkkoihin ja tekoälyyn (Wolf ym. 2020, Chanan, Chintala, Gross & Paszke 2022). VaderSentiment-kirjasto sisältää sanasto- ja sääntöpohjaisia työkaluja tunneanalyysin tekemistä varten (Gilbert &

Hutto 2014). Pandas on myös yleisesti käytetty kirjasto, kun käsitellään isoja määriä dataa, mutta tässä tutkielmassa se ei ollut hyödyllinen.

4.6 Tutkimusetiikka

katso edellisestä kommenttiversiostani tähän antamani kommentit ja ota ne huomioon.

Tutkielmassa tarkastellaan ainoastaan itse tviittejä. Muut metatiedot pyyhitään pois, sillä ne ovat analyysin kannalta tarpeettomia. Osa tviiteistä voi olla yhdistettävissä sen kirjoittajaan pelkkää Twitterin omaa hakukonetta käyttämällä, jonka vuoksi tviittejä on pidettävä turvassa. Työssäni en kuitenkaan tuo esille yksittäisten kirjoittajien tviittejä. Aineisto säilytetään muistikortilla sekä tietokoneen kovalevyllä siihen asti, kunnes niitä ei tarvita. EU:n tietosuoja-asetus (GDPR) sallii henkilötietojen käytön tutkimustarkoituksiin. Käytän työssäni Tietosuojavaltuutetun toimiston (2022) nimeämää tietosuojapolkua, joka on käytännössä ohjelista tutkimusaineiston huolehtimiseen koko sen elinkaaren ajan.

4.7 Odotukset ja tulosten tulkinta

Tutkimuksen tarkoituksena on tutkia tunneanalyysin vertautuvuutta kolmansien osapuolten teettämiin mielipidemittauksiin. Molemmat tunneanalyysit luokittelevat tekstin joko positiiviseksi, neutraaliksi tai negatiiviseksi. Tunneanalyysin tulokset muutettiin sellaiseen muotoon, jotta niitä voitiin verrata kolmansien osapuolten tekemiin mielipidemittauksiin esim. Morning Consultin (2022) mittaukset esitettiin nettona, joten tunneanalyysin tulokset muutettiin samankaltaiseen muotoon.

Odotan aineistojen vertailun tuottavan vähintään yhden johtopäätöksen seuraavista neljästä skenaariosta: 1. Tunneanalyysin ja mielipidemittausten tulokset eivät ole verrattavissa, 2. Tunneanalyysin tulokset ennustavat mielipiteisiin liittyviä muutoksia ja trendejä (eli ovat lähinnä suuntaa antavia), 3. Tunneanalyysin ja mielipidemittausten

tulokset ovat korrelaatiossa keskenään, 4. Tunneanalyysistä käy ilmi se, millä kognitiivisella auktoriteetilla on paras maine.

Kaksi ensimmäistä skenaariota ovat todennäköisempiä. On mahdollista, että tunneanalyysin tulokset eivät vastaa mielipidemittauksien tuloksia ollenkaan, jolloin ensimmäinen skenaario käy toteen. Näen mahdolliseksi myös sen, että tunneanalyysi ennustaa trendit oikein, eli negatiivinen diskurssi (tyytymättömyys) on havaittavissa sosiaalisen median sisällä sekä sen ulkopuolella. Kolmannen skenaarion toteutumiseen on vaikea uskoa jo pelkästään kirjallisuuskatsauksen perusteella. On myös mahdollista, että tunneanalyysistä käy ilmi kummalla kognitiivisella auktoriteetilla on parempi maine.

5. TULOKSET

Tässä luvussa esittelen tutkimustulokset ja löydökset. Tutkimuksessani toteutin tunneanalyysit koodaamillani ohjelmilla. Tulokset visualisoitiin helpottamaan niiden tulkitsemista sekä helpottamaan niiden vertaamista mielipidemittauksiin. Tulosten visualisointi ja taulukointi olivat riittäviä menetelmiä tulosten tulkitsemiseen. Taulukot 3. ja 4. on värikoodattu siten, että vihreä edustaa positiivisia, punainen negatiivisia ja harmaa neutraaleita tviittien sentimenttejä.

5.1 Tunneanalyysin tulokset

Tunneanalyysin tulokset, eli tviittien sentimentit, on esitetty taulukoissa 3. ja 4. Aineistot on nimetty sen hakusanan mukaan, millä niitä ladattiin, paitsi Joe Bidenin kohdalla on käytetty lyhennettä JB. Tulokset (prosentit) on pyöristetty lähimpään kokonaislukuun. Taulukoissa näkyy VADER- ja BERT-mallin erot. BERT on selvästi luokitellut tviitit paljon negatiivisemmaksi kuin VADER. BERT on myös luokitellut selvästi vähemmän tviittejä neutraaliksi.

Taulukko 3. VADER-mallin tuottamat tunneanalyysin tulokset aineistoittain. Symbolit +, - ja kirjain N viittaavat positiivisiin, negatiivisiin ja neutraaleihin tviitteihin. Lyhenne JB:llä viitataan Joe Bideniin liittyvään aineistoon. S-kirjain (s) viittaa siivottuun dataan.

VADER	#CDC+	#CDC-	#CDC N	CDC+	CDC-	CDC N	JB+	JB-	JB N	JB+ (s)	JB- (s)	JB N (s)
helmikuu	36 %	29 %	35 %	37 %	27 %	36 %	26 %	51 %	22 %	29 %	42 %	29 %
maaliskuu	31 %	25 %	43 %	35 %	29 %	36 %	41 %	31 %	28 %	27 %	37 %	36 %
huhtikuu	36 %	28 %	36 %	33 %	31 %	36 %	32 %	36 %	32 %	34 %	38 %	28 %
toukokuu	35 %	29 %	36 %	23 %	38 %	39 %	24 %	57 %	19 %	29 %	46 %	25 %
kesäkuu	21 %	42 %	37 %	31 %	27 %	42 %	32 %	49 %	18 %	33 %	43 %	24 %
heinäkuu	20 %	47 %	33 %	38 %	36 %	27 %	31 %	43 %	26 %	35 %	33 %	32 %
Keskiarvo:	30 %	33 %	37 %	33 %	31 %	36 %	31 %	45 %	24 %	31 %	40 %	29 %

Taulukko 4. BERT-mallin tuottamat tunneanalyysin tulokset aineistoittain. Symbolit +, - ja kirjain N viittaavat positiivisiin, negatiivisiin ja neutraaleihin tviitteihin. Lyhenne JB:llä viitataan Joe Bideniin liittyvään aineistoon. S-kirjain (s) viittaa siivottuun dataan.

BERT	#CDC+	#CDC-	#CDC N	CDC+	CDC-	CDC N	JB+	JB-	JB N	JB+ (s)	JB- (s)	JB N (s)
helmikuu	23 %	74 %	2 %	26 %	68 %	6 %	16 %	80 %	4 %	24 %	72 %	4 %
maaliskuu	16 %	81 %	3 %	13 %	82 %	4 %	9 %	88 %	2 %	20 %	76 %	5 %
huhtikuu	18 %	77 %	4 %	5 %	92 %	3 %	35 %	63 %	2 %	23 %	71 %	6 %
toukokuu	16 %	80 %	4 %	14 %	86 %	1 %	13 %	79 %	8 %	19 %	76 %	5 %
kesäkuu	10 %	86 %	4 %	23 %	73 %	4 %	24 %	71 %	5 %	22 %	72 %	6 %
heinäkuu	10 %	88 %	2 %	20 %	77 %	4 %	17 %	78 %	4 %	23 %	72 %	5 %
keskiarvo:	16 %	81 %	3 %	17 %	80 %	4 %	19 %	77 %	4 %	22 %	73 %	5 %

Tuloksista käy ilmi, että VADER-mallin mukaan Joe Biden on keskiarvoltaan CDC:tä epäsuositumpi: JB- (eli Joe Bidenin aineiston negatiivinen sentimentti) on keskiarvoltaan 45 prosenttia ja JB- (s) 40 prosenttia, kun taas #CDC- on 33 prosenttia ja CDC- 31 prosenttia (ks. Taulukko 3.). Morning Consultin (2022) mielipidemittaukset ovat samaa mieltä VADER-tunneanalyysin tuloksien keskiarvojen kanssa eli CDC on mielipidemittauksissa korkeammalla mitä Joe Biden. Myös Pew Research Centerin tutkimus on samoilla linjoilla. Funkin ja Tysonin (2022) kirjoittamassa Pew Research Centerin artikkelissa olevassa mielipidemittauksessa mitattiin tosin ainoastaan tammikuun lopun mielipiteet Joe Bidenin ja CDC:n teoista liittyen koronapandemiaan. Vaikka tammikuun loppu on mittausajankohdan ulkopuolella kaksi viikkoa (ensimmäinen datasetti ladattiin 15.2.2022), niin tätä tietoa voi myös käyttää tukena arvioimaan, kummalla auktoriteetilla on parempi maine. Heidän mielipidemittauksestansa käy ilmi, että CDC oli tammikuun lopulla arvioitu korkeammalle mielipidemittauksessa kuin Joe Biden. (Funk & Tyson 2022.)

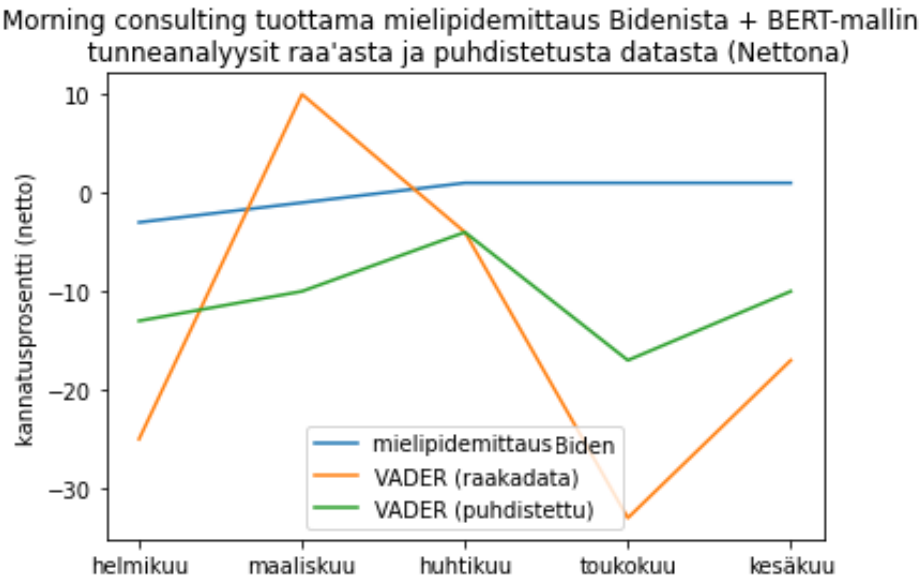
Vastaava ilmiö ei toistunut BERT-mallin kohdalla, vaan tulos oli itseasiassa päinvastainen; CDC oli BERT-mallin tunneanalyysin tulosten perusteella epäsuositumpi (ks. Taulukko 4.). Seuraavissa kappaleissa selviää, että VADER-mallin tulokset ovat huomattavasti tarkempia kuin BERT-mallin tuottamat tulokset verrattaessa niitä kolmansien osapuolten teettämiin mielipidemittauksiin. On kuitenkin vaikea arvioida, voiko tutkimukseen luotua VADER-mallia käyttää selvittämään auktoriteettien välinen

hierarkia liittyen kansalaisten mielipiteisiin pelkän Twitter-keskusteluiden perusteella. Vaikka VADER-malli piti Joe Bidenia negatiivisten viestien perusteella vähemmän suosittuna, niin tästä asiasta on vaikea tehdä johtopäätöksiä tutkielmassa tehdyn tutkimuksen perusteella.

5.2 Tunneanalyysin tulosten verrattavuus Morning Consultin mielipidemittauksen kanssa

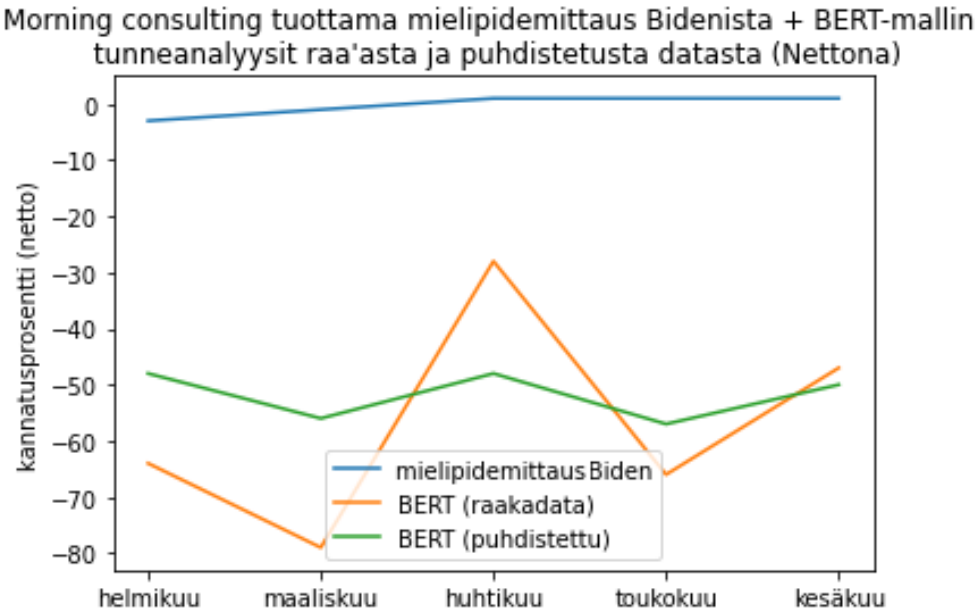
Ensimmäisenä vertasin Morning Consultin (2022) mielipidemittauksia ajalta 15. helmikuuta – 15. kesäkuuta (4 kuukautta) VADER-ohjelmiston tuottamiin tuloksiin. Bidenin kannatus (nettona) muuttui viiden kuukauden aikana hyvin vähän: miinus kolmesta (-3) plus yhteen (1) (ks. Kuvio 1.). VADER-mallin tuottama tunneanalyysi raa'asta datasta erosi hieman mielipidemittauksesta. Puhdistettu aineisto pärjäsi paremmin, sillä se erosi mielipidemittauksesta minimillään 5 (huhtikuu) ja maksimissaan 16 prosenttiyksikköä (toukokuu), kun taas raakadata-aineisto tuotti radikaalimpia tuloksia (ks. Kuvio 1). Puhdistettu aineisto näytti tuottavan tasaisempia tuloksia.

Molemmissa tapauksissa, VADER tulkitsi Twitter-keskustelujen olevan negatiivisempia, kuin mielipiteet todellisuudessa olivat. Morning Consultin (2022) mielipidemittaukset liittyivät nimenomaan Joe Bidenin ja CDC:n kykyihin pandemian hoidossa. Lisäksi on huomioitava, että Joe Bidenia koskevat Twitter-keskustelut liittyvät moneen muuhun aiheeseen, kuin pelkästään pandemian hoitoon. Pandemian hoito kuitenkin osittain vaikuttaa Joe Bidenin kannatukseen, kuten moni muukin asia. Yhdysvaltalaisten usko Bidenin kykyyn pandemian hoidossa saa varmasti vaikutteita hänen aiemmistaan pandemiaan liittymättömistä teoistaan. Huomautuksena lukijalle, että kuviossa 1–4 vasemmalla puolella olevat prosentit näkyvät skaalautuvasti, eivät staattisesti.



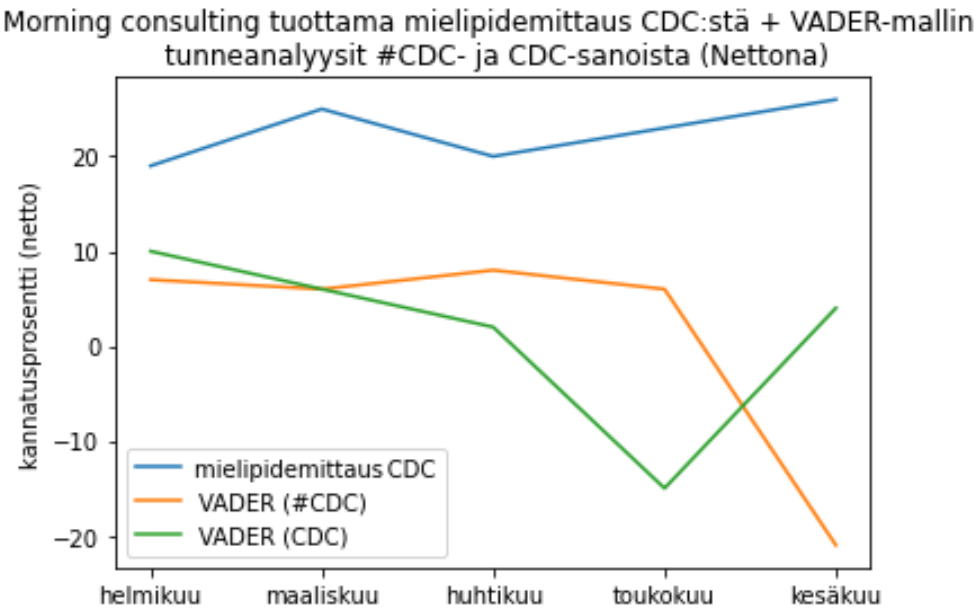
Kuvio 1. Morning Consultin (2022) mielipidemittaus (sinisellä), käsittelemättömän (oranssilla) ja puhdistetun (vihreällä) aineiston tunneanalyysi VADER:lla. Mielipidemittauksen virhemarginaali ±2 prosenttiyksikköä.

Toisessa mittauksessa saman mielipidemittauksen tuloksia verrattiin BERT-ohjelman tuottamiin tuloksiin. BERT:n tuottamat tulokset olivat paljon negatiivisimpia kuin VADER:n tulokset (ks. Kuvio 2.). Tässäkin tapauksessa tunneanalyysi puhdistetusta datasta tarjosi tasaisemman tuloksen. Erot mielipidemittauksiin olivat molemmissa BERT:n tapauksessa useana kuukautena yli 50 prosenttiyksikköä. Joe Bideniin liittyvä aineisto oli ainoa, jossa oli sekä raakaa, että puhdistettua, dataa.



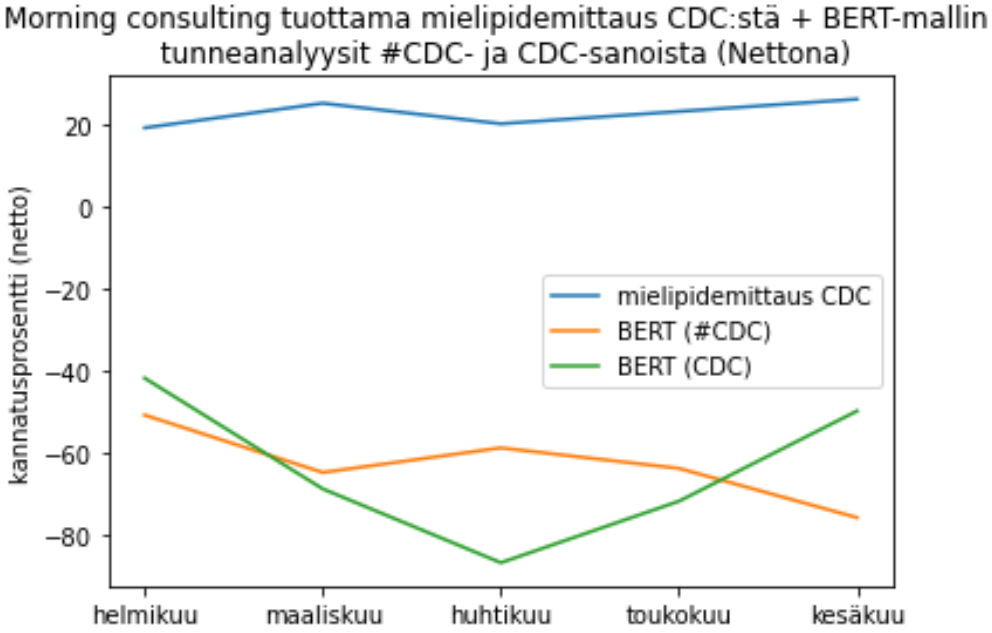
Kuvio 2. Morning Consultin (2022) mielipidemittaus (sinisellä), käsittelemättömän (oranssi) ja puhdistetun (vihreä) aineiston tunneanalyysi BERT:llä. Mielipidemittauksen virhemarginaali on ±2 prosenttiyksikköä.

Kolmannessa vertailussa aineistojen #CDC ja CDC välillä oli pieniä eroja (ks. Kuvio 3.). Pienimmillään aineistojen ero oli mielipidemittauksiin 9 prosenttiyksikköä (helmikuu CDC) ja enimmillään 47 prosenttiyksikköä (kesäkuu #CDC). Kummassakin (ilman, ja hashtagin kanssa) ladatussa aineistossa oli eroja, mutta tulosten pohjalta ei pysty päättelemään, kumpi aineisto tuottaa laadukkaampia tutkimustuloksia.



Kuvio 3. Morning Consultin (2022) mielipidemittaus (sinisellä), #CDC- (oranssi) ja CDC-aineiston (vihreä) tunneanalyysi VADER:lla. Mielipidemittauksen virhemarginaali ±2 prosenttiyksikköä.

Viimeisenä mitattiin BERT:n tuottamia tuloksia samoihin Morning Consultin (2022) mielipidemittauksiin. Kuten Joe Biden -hakusanaan pohjautuneessa aineistossa (ks. Kuvio 2.), tässäkin tapauksessa BERT:n tulokset olivat erittäin negatiivisia (ks. Kuvio 4.). Tässäkin tapauksessa hashtagin ja normaalin hakusanoin ladatun aineiston välillä ei ollut merkittäviä eroja.



Kuvio 4. Morning Consultin (2022) mielipidemittaus (sinisellä), #CDC- (oranssi) ja CDC-aineiston (vihreä) tunneanalyysi BERT:llä. Mielipidemittauksen virhemarginaali ±2 prosenttiyksikköä.

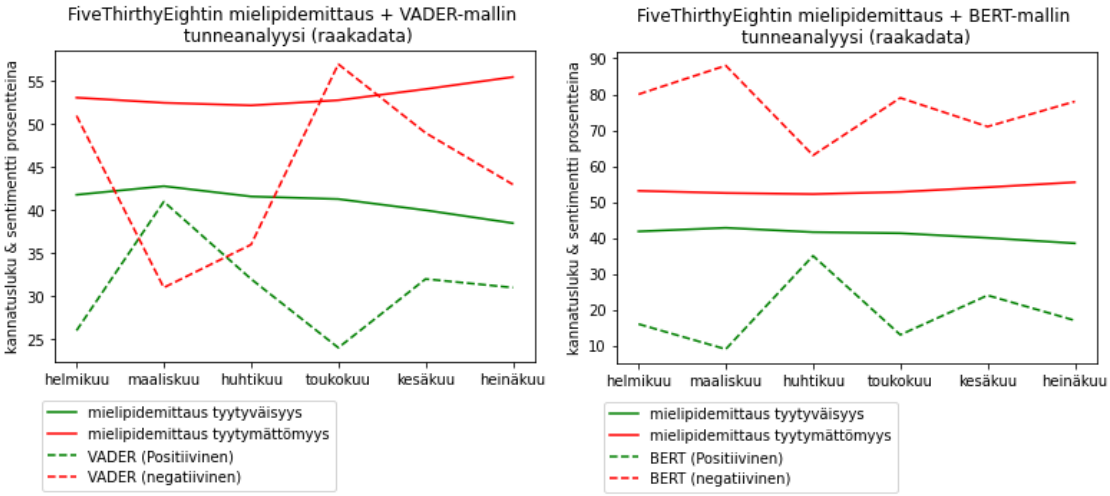
Kuten aiemmin mainittiin, testausten aikana (NLP Townin) BERT- ja VADER-mallien erot eivät olleet kovinkaan suuret. Testauksen aikana, BERT-mallin huomattiin kuitenkin luokittelevan sentimentiltään negatiiviset tekstit tarkemmin, mikä voi osittain selittää negatiivisvoittoiset tulokset. Jos verrataan raakaa ja puhdistettua aineistoa, NLP Townin BERT-malli näyttää toimivan oletettua heikommin kaikkien aineistojen kohdalla.

Kaiken kaikkiaan, VADER-tunneanalyysi suoriutui paljon paremmin, kuin BERT-koneoppimismalliin pohjautuva NLP Townin malli. Kumpikaan tunneanalyysin malleista ei näytä pystyvän tuottamaan mielipidemittauksen kaltaisia tuloksia, eivätkä ne myöskään näytä pystyvän ennustamaan muutoksia. Toisaalta täytyy muistaa, että Morning Consultin (2022) mielipidemittauksien mukaan Joe Bidenin ja CDC:n muutokset tuona ajanjaksona, olivat melko vähäiset, joten muutosten ennustaminen heidän aineistollansa on vaikeaa. Se, ladattiinko aineistoa hashtagin kanssa tai ilman (#CDC ja CDC), ei näyttänyt tuottavan merkittäviä eroja tunneanalyysin tuloksiin (ks. Kuvio 3 ja Kuvio 4). Puhdistettu data tarjosi tasaisemmat ja hieman tarkemmat tulokset (ks. Kuvio 1 ja Kuvio 2).

5.3 Tunneanalyysin tulosten verrattavuus FiveThirtyEightin mielipidemittauksen kanssa

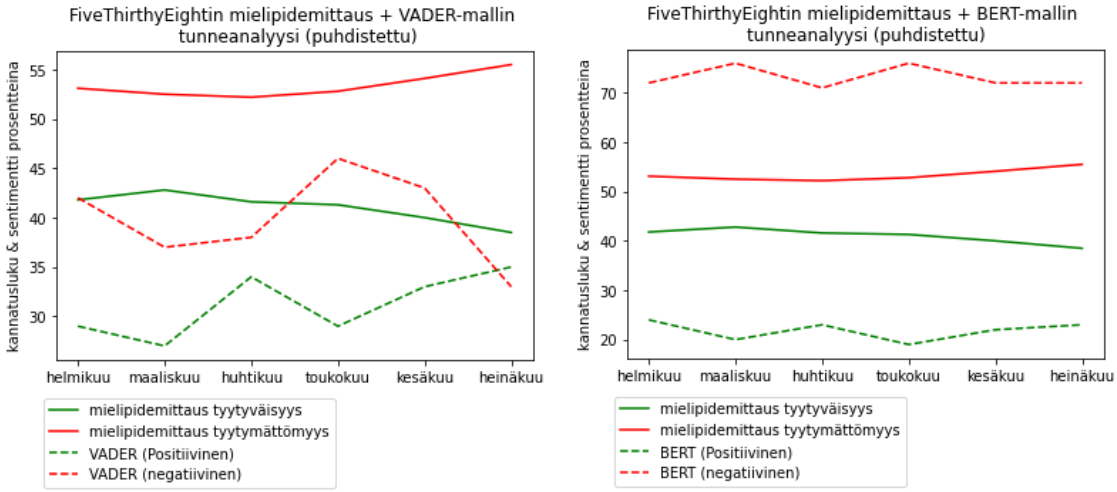
Tässä tapauksessa tunneanalyysien tuloksia (raakadatasta ja puhdistetusta datasta) verrataan Joe Bidenin kannatuslukuihin. Kannatusluvut on mitannut yritys nimeltä FiveThirtyEight, joka on erikoistunut mielipidemittauksiin. FiveThirtyEight (2022) laskee mukaan muiden osapuolten mielipidemittaukset, ja he ottavat huomioon niiden laadun (quality), tuoreuden (recency), otoskoon (sample size) sekä puoluepoliittisen suuntauksen (partisan lean). Huomiona taas, kuvioiden 5. ja 6. vasemmalla puolella olevat prosentit näkyvät skaalautuvasti, ei staattisesti.

Kuten aiemmin, sekä VADER- että BERT-malli epäonnistuivat tuottamaan mielipidemittauksen vertaisia tuloksia. Mallit eivät myöskään pystyneet ennustamaan muutoksia. VADER-malli suoriutui tässäkin tapauksessa paremmin (ks. Kuvio 5. ja Kuvio 6.). VADER-mallin positiivisten viestien ero Joe Bidenin kannatukseen (tyytyväisyyteen) oli raakadatan kanssa minimissään 1,8 (maaliskuu) ja maksimissaan 28,2 (toukokuu) prosenttiyksikköä. Tyytymättömyyden, ja negatiivisten viestien erot olivat suuremmat (ks. Kuvio 5.)



Kuvio 5. FiveThirtyEightin mielipidemittaus verrattuna VADER- ja BERT-mallien tunneanalyysin tuloksiin raakadatasta. Tyytyväisyys (approval) Bidenia kohtaan on esitetty vihreänä ja tyytymättömyys (disapproval) punaisena viivana. Tunneanalyysien (VADER:n ja BERT:n) positiiviset sentimentit on esitetty vihreinä ja negatiiviset punaisina katkoviivoina.

Puhdistettu data ei näyttänyt antavan tässä tapauksessa tarkempia tuloksia mitä puhdistettu data (ks. Kuvio 6). Puhdistetun datan tulokset olivat tasaisempia, mutta lisäksi myös negatiivisempia kuin raakadatasta tuotetut tulokset (ks. Kuvio 5. ja Kuvio 6.). Toisin sanoen, tyytyväisyys oli alhaisempaa, ja tyytymättömyys korkeampaa puhdistetun datan kanssa.



Kuvio 6. FiveThirtyEightin mielipidemittaus verrattuna VADER- ja BERT-mallien tunneanalyysin tuloksiin puhdistetusta datasta. Tyytyväisyys (approval) Bidenia kohtaan on esitetty vihreänä ja tyytymättömyys (disapproval) punaisena viivana. Tunneanalyysien (VADER:n ja BERT:n) positiiviset sentimentit on esitetty vihreinä ja negatiiviset punaisina katkoviivoina.

VADER-malli näytti tuottavan tarkempia tuloksia verrattavaan aineistoon mitä BERT-malli. Kummatkaan mallit eivät tarjonneet mielipidemittauksen tasoista kuvaa. Ne eivät myöskään osanneet ennakoida muutoksia oikein. Molempien mallien tunneanalyysin tulokset olivat kuitenkin verrattavissa osittain FiveThirtyEightin mielipidemittauksen kanssa. Yhdysvaltalaiset olivat enemmän tyytymättömiä Presidentti Bideniin, ja taas Twitter-käyttäjät puhuivat hänestä enemmän negatiiviseen sävyyn (ks. Kuvio 5 ja Kuvio 6).

6. JOHTOPÄÄTÖKSET JA POHDINTA

Informaatiotutkimuksen tutkimuskentällä pro gradu -tutkimukseni sijoittuu sosiaalisen median keskusteluiden ja terveysviestinnän tutkimukseen. Terveysviestinnän onnistumiseen ja terveyskäyttäytymisen muuttamiseen vaaditaan ennakkotietoa kohdeyleisöstä (kohderyhmä- tai yksilötasolla). Viestinnän suunnittelijan pitää olla tietoinen mahdollisista käyttäytymiseen, ennakkoluuloihin ja kulttuuriin vaikuttavista tekijöistä, jotka poikkeavat yksilöiden ja ryhmien välillä.

Erilaiset viestinnän vastaanottamiseen vaikuttavat tekijät tulisi tunnistaa, ja viestintästrategia muovata niiden mukaan. Myös eri viestintätavat sopivat eri tilanteisiin. Viestintästrategian suunnittelijan olisi hyvä tietää pitääkö hänen lähestyä yleisöä massaviestinnän keinoin vai käyttää yksilöllisempiä keinoja kuten kohdentamista ja räätälöintiä. Yksilöllisemmät lähestymiskeinot vaikuttavat olevan paljon tehokkaampia kuin perinteinen massaviestintä, mutta ne vaativat enemmän resursseja. Myös palautteen annolla on vaikutusta terveyskäyttäytymiseen. (Enwald 2013, 141, Hirvonen 2015, 160.) Kirjallisuudesta nousi esille myös se, miten yleisön luottamus kognitiivista auktoriteettia ja viestien sisältöä kohtaan ovat tärkeitä viestin vastaanottamisen kannalta. Jos arvostus viestin tuojaa kohtaan on alhainen, viestejä on hankala ottaa tosissaan (Wilson 1983, 13–19.) Lähteeseen kohdistuva luotettavuus sekä uskottavuus nousivat esille monissa eri kirjallisuudessa.

Ensimmäinen tutkimuskysymykseni kuului: ”Millä tavoin SBD:a voitaisiin hyödyntää terveysviestinnässä?” Tutkielmassa tuotiin kirjallisuuden kautta esille muutamia mahdollisia tapoja käyttää SBD hyödyksi, joista (tutkielman kannalta) potentiaalisin ehdokas oli tunneanalyysi. Spekuloitiin, että tunneanalyysin avulla voisi mahdollisesti selvittää eri auktoriteetteihin liittyviä mielipiteitä ja asenteita sekä mielipiteeseen liittyviä muutoksia, jotka olisivat verrattavissa perinteisten mielipidemittauksien kanssa.

Tekstinlouhinnan avulla voidaan saada selville terveydelle haitallisia viraaleja trendejä tai viestejä, jotka viittaavat itsetuhoiseen käytökseen, mikä mahdollistaa niihin ajoissa puuttumisen (Ryu & Song 2015, 7, Barros, Oliveira & Trifan 2021). Viestinnän

onnistuneisuuden seuraamiseksi voidaan käyttää kävijälaskuria, evästeitä, hakukoneoptimointia ja -markkinointia (ks. luku 3.6). SBD:aa ei kuitenkaan ole vielä toistaiseksi hyödynnetty terveydenhuollon ja terveysviestinnän alalla muutamia poikkeuksia lukuun ottamatta.

Toinen tutkimuskysymys liittyi juuri mielipiteiden mittaamiseen SBD:sta, ja siihen liittyvä tutkimuskysymys oli seuraavanlainen: ”Voiko Twitter-keskusteluista tuotettua tunneanalyysiä käyttää mielipidemittauksen korvikkeena?”. Tässä tutkielmassa käytetyllä aineistolla ja työkaluilla ei kuitenkaan voitu tuottaa tuloksia, jotka olisivat verrattavissa mielipidemittauksien kanssa. Tunneanalyysi ei ollut verrattavissa mielipidemittauksiin, eikä ennustaa oikein siihen liittyviä trendejä tai muutoksia.

SBD voidaan mahdollisesti hyödyntää siten, että tunneanalyysiä käyttämällä voidaan selvittää *millä auktoriteetilla on paras tämänhetkinen maine.* CDC:hen luotetaan mielipidemittauksen mukaan enemmän mitä Bideniin, ja vastaavasti VADER-tunneanalyysin mukaan CDC:stä puhuttiin Twitterissä positiivisempaan sävyyn (ks. Taulukko 3.) Tunneanalyysiä voidaan siis käyttää antamaan jonkinlainen karkea arvio auktoriteettien välisistä sijoituksista arvostuksen ja luotettavuuden suhteen. Tätä tietoa voidaan mahdollisesti hyödyntää terveysviestinnässä esim. valitsemalla paras mahdollinen auktoriteetti tiedonvälittäjäksi. Tämä mahdollisuus on tutkielman empiirisen tutkimuksen perusteella mahdollinen.

Täytyy muistaa, että Joe Bideniin liittyvä aineisto, eli Twitter-keskustelut, keskittyvät moneen muuhunkin häneen liittyvään aiheisiin, kun taas CDC:n liittyvä aineiston sisältö keskittyi enimmäkseen COVID 19 -pandemiaan. Toisaalta on vaikea arvioida, miten paljon vaikutusta kansaa puhuttavilla tapahtumilla, kuten inflaatiolla, Ukrainan tilanteella, energiapoliittisilla päätöksillä, skandaaleilla yms. on kansan näkemykseen hänen kykyihinsä pandemian hoidossa, jota Morning Consultin (2022) mielipidemittaus mittasi. Wilsonin (1983, 15) mielestä auktoriteetin vaikutusvalta on sidottu uskottavuuteen yleisellä tasolla. Olipa kyse sitten energia, talous- tai terveyssektorin hoitamisesta (jotka kuuluvat Yhdysvaltain presidentin tehtäviin), ja Wilsonin näkemyksen myötä, näkisin, että presidentti Bidenin pandemian hoidon ulkopuolisilla

teoilla on vaikutusta uskottavuuteen, ja lopulta kansan näkemykseen hänen kykyihinsä pandemian hoitamisessa. Aineiston manuaalisella puhdistuksella yksinkertaisin kriteerein, sekä aineiston lataamisella hashtagin kanssa tai ilman, ei ollut nähtävästi merkittävää vaikutusta tutkielman lopputulokseen.

Jatkotutkimukset voisivat valottaa tunneanalyysin soveltuvuutta lisää. Ehdottaisin koodien, mallien ja laskentafunktioiden testaamista siten, että testaus suoritetaan oikean elämän esimerkeillä eli nettikeskusteluilla. Useat netistä tunneanalyysin testaamiseen tarkoitetut datasetit, jotka sisältävät tekstinpätkiä, ovat ehkä liian helppoja ohjelmille, minkä vuoksi ohjelman tehokkuudesta voi saada väärän kuvan. Tämän vuoksi olisi parempi, että ohjelmat testattaisiin ”oikealla” datalla. Ehdottaisin myös hankkimaan aineistoa muualtakin kuin Twitteristä. Saattaisi olla myös hyvä kerätä aineistoa tiheämpään esim. kerran viikossa tai päivittäin. Aineiston lataus kannattaa ajoittaa aikavyöhykkeitä ajatellen, jolloin maan ulkopuolisia käyttäjiä on mahdollisemman vähän ns. online-tilassa. Tällä saattaa olla vaikutusta aineiston laatuun. Aineiston siivoaminen manuaalisesti tai automaattisesti voi osoittautua positiiviseksi asiaksi, vaikka se onkin aikaa kuluttava prosessi. Olisi hyvä pyrkiä keräämään esimerkiksi suomenkielisiä nettikeskusteluita englanninkielisten sijaan, sillä englannin kieli on maailmanlaajuisesti puhuttu, joka houkuttelee huijareita, mainostajia, Yhdysvaltain ulkopuolisia käyttäjiä yms., jotka täyttävät datasetit irrelevanteilla viesteillä, roskalla ja spämmiviesteillä. Lisäksi olisi hyvä pyrkiä joko itse, tai yhteistyön tuumin, keräämään mielipidemittaukset, jotta ei tarvitsisi luottaa muiden teettämiin mielipidemittauksiin. Se, että mielipidemittauksia teetetään jatkossakin, on vaikea ennustaa etukäteen.

Sosiaalista big dataa on runsaasti saatavilla ja uskon, että sille löytyy paljon käyttöä, myös terveysviestinnällisiin tarkoituksiin. Sen tutkiminen saattaa vaikuttaa kankealta siihen liittyvän kirjallisuuden loistaessa poissaolollaan, mutta niinhän monen muunkin uuden aiheen kohdalla, ainakin sen alkutaipaleella. Toivon, että pro gradu -tutkielmasta on muille hyötyä.

LÄHTEET

Abhinandan, R. (2021). Sentiment Analysis- Lexicon Models vs Machine Learning. https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746 (käytetty 29.8.2022)

Aldrich, R., Harrington, N., & Noar, S. (2009). The Role of Message Tailoring in the Development of Persuasive Health Communication Messages. *Annals of the International Communication Association, 33(1),* 73–133. DOI: 10.1080/23808985.2009.11679085

Alkulaib, L., Benton, A., Broniatowski D., Chen, T., Dredze, M., Jamison, A., & Qi, S. (2018). *Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. American Journal of Public Health, e1–e7.* DOI: 10.2105/ajph.2018.304567

Ashkpour, A., Breure, L., Harmelen, F., Erp, M., Mandemakers, K., Meroño-Peñuela, A., Scharnhorst, A., & Schlobach, S. (2014). *Semantic technologies for historical research: A survey. Semantic Web, 6(6)*, 539–564. DOI: 10.3233/sw-140158

Arvidsson, A., Colleoni, E., & Rozza, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication, 64(2)*, 317–332. DOI: 10.1111/jcom.12084

Balasubramanyan, R., O'Connor, B., Routledge B. & Smith N., (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media,4(1) 122–129.* https://ojs.aaai.org/index.php/ICWSM/article/view/14031 (käytetty 23.6.2022)

Barros, L., Oliveira, J., & Trifan, A. (2021). VADER meets BERT: sentiment analysis for early detection of signs of self-harm through social mining. *Conference and Labs of the Evaluation Forum 2021*. Romania, Budapest.

Bates, D., Heitmueller, A., Kakad, M., & Saria, S. (2018). Why policymakers should care about “big data” in healthcare. *Health Policy and Technology, 7(2)*, 211–216. DOI: 10.1016/j.hlpt.2018.04.006

Bermingham, A. & Smeaton, A. (2011). On using Twitter to monitor political sentiment and predict election results. *IJCNLP 2011, 2–10*, Thaimaa, Chiang Mai.

Binenbaum, I., Boccia, S., De Vito, C., Glocker, K., Migliara, G., Pastorino, R., & Ricciardi, W. (2019). Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. *European Journal of Public Health, 29(Supplement_3)*, 23–27. DOI: 10.1093/eurpub/ckz168

Blank, G. & Dubois E. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society, 21(5), 729-745*, DOI: 10.1080/1369118X.2018.1428656

Blöbaum, B., Fujarski, S., Gehrau, V., Lorenz, H., & Schieb, C. (2021). The Impact of Health Information Exposure and Source Credibility on COVID-19 Vaccination Intention in Germany. *International Journal of Environmental Research and Public Health, 18(9), 4678.* doi:10.3390/ijerph18094678

Boldenthusiast (2019). Sentiment Analysis – The Lexicon Based Approach. https://www.alphabold.com/sentiment-analysis-the-lexicon-based-approach/ (käytetty 29.8.2022)

Brand, H. & Sørensen, K. (2013). Health literacy lost in translations? Introducing the European Health Literacy Glossary. *Health Promotion International, 29(4)*, 634–644. DOI: 10.1093/heapro/dat013

Branigan, C., Fredrickson B., Mancuso, R. & Tugade M., (2000). The undoing effect of positive emotions. *Motivation and Emotion, 24(4)*, 237–258. DOI: 10.1023/a:1010796329158

Cambridge University Press (2021). Meaning of data in english. https://dictionary.cambridge.org/dictionary/english/data (käytetty 27.10.2021)

Carter, C., Lyons, B., Reifler, J., & Stoeckel, F. (2022). The politics of vaccine hesitancy in Europe. *European Journal of Public Health, ckac041*, https://doi.org/10.1093/eurpub/ckac041

Case, D. & Johnson, J. (2012). *Health information seeking.* New York: Peter Lang Publishing.

Chanan, G., Chintala, S., Gross, S., & Paszke, A (2022). Torch (versio 1.12.0) [Python-kirjasto]. https://pypi.org/project/torch/

Chowdhury, G. (2005). Natural language processing. *Annual Review of Information Science and Technology*, 37(1), 51–89. DOI:10.1002/aris.1440370103

Cinelli, M., Galeazzi, A., Morales, G., Starnini, M., & Quattrociocchi, W. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences, 118(9), 1-8*. DOI: 10.1073/pnas.2023301118

Cowan, S., Mark, N., & Reich, J. (2021). COVID-19 Vaccine Hesitancy Is the New Terrain for Political Division among Americans. Socius: *Sociological Research for a Dynamic World, 7, 237802312110236*. https://doi.org/10.1177/23780231211023657

Dang, S. (2022). Twitter estimates spam, fake accounts comprise less than 5% of users - filing. Reuters. https://www.reuters.com/technology/twitter-estimates-spam-fake-accounts-represent-less-than-5-users-filing-2022-05-02/ (käytetty 14.7.2022)

Damerau, F., Indurkhya, N., Weiss, S., & Zhang T. (2005). *Text mining: Predictive methods for analyzing unstructured information.* New York: Springer.

Davidson, J., Delangue, C., Debut, L., Chaumond, J., Cistac, P., Funtowicz, M., Gugger, S., Jernite, Y., Louf, R., Ma, C., Moi, A., Platen, P., Plu, J., Rault, T., Scao, T., Sanh, V., Shleifer, S., Wolf, T., & Xu, C. (2020). Transformers: State-of-the-Art Natural Language Processing. *The 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45*. DOI: 10.18653/v1/2020.emnlp-demos.6

Davis, C., Ferrara, E., Flammini, A., Menczer, F., & Varol, O. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017), 280-289. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

Deng, J. & Joyce, B. (2017). *Sentiment analysis of tweets for the 2016 US presidential election. 2017 IEEE MIT Undergraduate Research Technology Conference (URTC), 2017*, 1-4, DOI: 10.1109/urtc.2017.8284176

Duodecim (2016). *Terveyskäyttäytyminen.* https://www.terveyskirjasto.fi/ltt03441 (käytetty 23.11.2021)

ECDC (2021). *What is health communication?* https://www.ecdc.europa.eu/en/health-communication/facts (käytetty 14.9.2021).

Enwald, H. (2013). *Tailoring health communication: the perspective of information users' health information behaviour in relation to their physical health status.* (Acta Universitatis Ouluensis. B, Humaniora 118) [väitöskirja, Oulun yliopisto] JULTIKA Oulun yliopiston julkaisuarkisto http://jultika.oulu.fi/Record/isbn978-952-62-0279-2

Erkkola, M., Fogelholm, M., Saarijärvi, H., Nevalainen, J., & Uusitalo L. (2019). Kuluttajadatan mahdollisuudet ja haasteet kansanterveystutkimuksessa; case LoCard.

Sotilaslääketieteellinen aikakausilehti, 2019(56), 76–87. DOI: https://doi.org/10.23990/sa.70960

Euroopan parlamentti (2021). *Massadata: määritelmä, hyödyt, haasteet (infografiikka).* https://www.europarl.europa.eu/news/fi/headlines/priorities/tekoaly-eu-ssa/20210211STO97614/massadata-maaritelma-hyodyt-haasteet-infografiikka (käytetty 29.11.2021)

FiveThirtyEight (2022). *How popular is Joe Biden?* https://projects.fivethirtyeight.com/biden-approval-rating/ (käytetty 24.7.2022)
Funk, C. & Tyson, A. (2022). *Increasing Public Criticism, Confusion Over COVID-19 Response in U.S.* https://www.pewresearch.org/science/2022/02/09/increasing-public-criticism-confusion-over-covid-19-response-in-u-s/ (käytetty 24.7.2022)

Galetsi, P., Katsaliaki, K., & Kumar, S. (2020). Big data analytics in health sector: Theoretical framework, techniques and prospects. *International Journal of Information Management, 50*, 206–216. DOI:10.1016/j.ijinfomgt.2019.05.003

Gallup (2007). *What Is Public Opinion Polling and Why Is It Important?* http://media.gallup.com/muslimwestfacts/pdf/pollingandhowtouseitr1dreveng.pdf (käytetty 19.6.2022)

Geeksforgeeks (2021*). Libraries in Python.* https://www.geeksforgeeks.org/libraries-in-python/ (käytetty 15.7.2022)

Genomikeskus (2021). *5. KUKA VOISI KÄYTTÄÄ GENOMIKESKUKSEN TIETOJA?* https://www.genomikeskus.fi/meilta-kysyttya.html (käytetty 1.12.2021)

Gilbert, E. & Hutto, C. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14).* Ann Arbor, Michigan.

Hankonen, N., Haukkala, A., Jallinoja, P., & Laine, H. (2013). Mitkä sosiaaliskognitiiviset tekijät selittävät varusmiesten käsihygieniaa? *Sosiaalilääketieteellinen aikakausilehti 2013(50),* 221–233.

Hawkings, R., Hwang, H., Lee S., & Pingree, S. (2008). Interplay of Negative Emotion and Health Self-Efficacy on the Use of Health Information and Its Outcomes. *Communication Research, 35(3)*, 358–381. DOI:10.1177/0093650208315962

Hirvonen, N. (2015). *Health information matters: everyday health information literacy and behaviour in relation to health behaviour and physical health among young men.* (Acta Universitatis Ouluensis. B, Humaniora 133) [väitöskirja, Oulun yliopisto] JULTIKA Oulun yliopiston julkaisuarkisto http://jultika.oulu.fi/Record/isbn978-952-62-1040-7

Hukka, E. (2014). Potilaasta partneriksi – sosiaalinen media haastaa terveydenhuollon. Teoksessa: Järvi, U. (toim.) *Tautinen media.* Helsinki: Kustannus Oy Duodecim.

Ishikawa H. (2015). *Social big data mining*. Boca Raton: CRC press.

Kaarakainen, M.-T. & Kaarakainen, S.-S. (2018). Tulevaisuuden kansalaisia rakentamassa: Uudet lukutaidot koulutuksen ja opetuksen digitalisaation kehityksessä. *AFinLAn vuosikirja 2018 (Suomen soveltavan kielitieteen yhdistyksen julkaisuja 76)*. 22–40. DOI: https://doi.org/10.30661/afinlavk.69269

Kautz, H., Sadilek, A., & Silenzio, V. (2021). Modeling Spread of Disease from Social Interactions. *Proceedings of the International AAAI Conference on Web and Social Media, 6(1), 322–329.* https://ojs.aaai.org/index.php/ICWSM/article/view/14235

Kennedy, J. (2019). Populist politics and vaccine hesitancy in Western Europe: an analysis of national-level data. *European Journal of Public Health, 29(3)*, 512–516. https://doi.org/10.1093/eurpub/ckz004

Kreuter, M., Strecher, V., & Glassman, B. (1999). One size does not fit all: The case for tailoring print materials. *Annals of Behavioral Medicine, 21(4),* 276–283. DOI:10.1007/bf02895958

Krishnan, S. (2016). Application of Analytics to Big Data in Healthcare. *2016 32nd Southern Biomedical Engineering Conference (SBEC)*. 156–157. doi:10.1109/sbec.2016.88

Lääkäriliitto (2021). *Potilastietojen hyödyntäminen.* https://www.laakariliitto.fi/laakarin-etiikka/potilas-laakarisuhde/potilastietojen-hyodyntaminen/ (käytetty 1.12.2021)

Hirviheimo, M., Kinnunen, U-M., & Kivekäs, E. (2015). Tekonivelinfektioita aiheuttavien riskitekijöiden selittäminen tai ennustaminen potilaskertomukseen tallennetun tiedon avulla. *Finnish Journal of eHealth and eWelfare* 7(2-3), 75–82. https://journal.fi/finjehew/article/view/50894

IBM (2021). Data mining. https://www.ibm.com/cloud/learn/data-mining (käytetty 1.12.2021)

IBM (2020a). *Text mining.* https://www.ibm.com/cloud/learn/text-mining (käytetty 14.9.2021)

IBM (2020b). *Artificial Intelligence (AI).* https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (käytetty 27.10.2021)

Kelleher, J. & Tierney, B. (2018). *Data Science*. Cambridge, MA: The MIT Press.

Laaksonen, S-M., Matikainen, J., & Tikka, M. (2013). *Otteita verkosta: Verkon ja sosiaalisen median tutkimusmenetelmät*. Tampere: Vastapaino.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *Application Delivery Strategies by META Group Inc 2001(949)*.

Linnanmäki, E. (2006). Historian influenssapandemiat. Duodecim, 2006(122), 2028. http://urn.fi/URN:NBN:fi-fe201211029597

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5(1), 1–167. DOI:*10.2200/S00416ED1V01Y201204HLT016*

LUC (2021). *Mitä on tekstinlouhinta?* https://lib.luc.fi/tekstinlouhinta (käytetty 14.9.2021)

Merilehto, A. (2018). *Tekoäly: Matkaopas johtajalle*. Helsinki: Alma talent.

McClain, C., Rivero, G., Smith., A., & Widjaya, R. (2021). *The behaviors and attitudes of U.S. adults on Twitter.* https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/ (käytetty 28.11.2021)

Miyachi, T., Senoo, Y., Takita, M., & Yamamoto, K. (2020). Lower trust in national government links to no history of vaccination*. The Lancet, 395(10217), 31–32.* DOI:10.1016/s0140-6736(19)32686-8

Moen, P. (2021). TIEDON LOUHINNAN MENETELMÄT: Kurssin keskeisten käsitteiden englanti-suomi-sanasto. https://www.cs.helsinki.fi/u/ronkaine/tilome/sanasto.html (käytetty 1.12.2021)

Morning Consult (2022). *Net Approval for Leaders' Handling of Coronavirus.* https://morningconsult.com/views-on-the-pandemic/ (käytetty 24.7.2022)

Nilsson, N. (2010). *The quest for artificial intelligence: a history of ideas and achievements.* Cambridge, NY: Cambridge University Press.

NLP Town (2022). *bert-base-multilingual-uncased-sentiment.* https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment (käytetty 15.7.2022)

Pace University (2022). Fake news: misinformation, disinformation and malinformation. https://libguides.pace.edu/fakenews (käytetty 20.5.2022)

Ryu, S. & Song, T. (2015). Big Data Analysis Framework for Healthcare and Social Sectors in Korea. *Healthcare Informatics Research, 21(1), 3–9.* doi:10.4258/hir.2015.21.1.3

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics, 2(1),* 325–347. doi:10.1146/annurev-linguistics-011415-040518

THL (2021). *Terveydenhuollon menot ja rahoitus 2019.* Helsinki: THL. https://thl.fi/fi/tilastot-ja-data/tilastot-aiheittain/sosiaali-ja-terveydenhuollon-resurssit/terveydenhuollon-menot-ja-rahoitus

Tietovaltuutetun toimisto (2022). Tieteellinen tutkimus ja tietosuoja. https://tietosuoja.fi/tieteellinen-tutkimus (käytetty 10.7.2022)

Tsytsarau, M. & Palpanas, T. (2011). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3), 478–514. DOI: 10.1007/s10618-011-0238-6

Tuomisto, J. (2015). Massadata kansanterveyden edistämisessä. *Duodecim 22(juhlanumero).* http://urn.fi/URN:NBN:fi-fe201601071478 (käyetty 29.11.2021)

Stephens, E. (2020). We've known about pandemic health messaging since 1918. So when it comes to coronavirus, what has Australia learnt? *The Conversation.* https://theconversation.com/weve-known-about-pandemic-health-messaging-since-

1918-so-when-it-comes-to-coronavirus-what-has-australia-learnt-134797 (Käytetty 29.11.2021).

Thomas, R. (2006). *Health Communication*. New York, NY: Springer.

Tilastokeskus (2022a). *Iterointi.* https://www.stat.fi/meta/kas/iterointi.html (käytetty 19.5.2022)

Tilastokeskus (2020b). *Väestön tieto- ja viestintätekniikan käyttö 2020: Liitetaulukko 17. Internetin käyttötarkoitusten yleisyys 2020, %-osuus väestöstä.* Helsinki: Tilastokeskus. https://www.stat.fi/til/sutivi/2020/sutivi_2020_2020-11-10_tau_017_fi.html

Tumasjan, A., Sandner, P., Sprenger T., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, 4(1), 178–185.* https://ojs.aaai.org/index.php/ICWSM/article/view/14009 (käytetty 23.6.2022)

Twitter (2022). *Twitter announces first quarter 2022 results.* https://s22.q4cdn.com/826641620/files/doc_financials/2022/q1/Final-Q1%E2%80%9922-earnings-release.pdf (käytetty 14.7.2022)

WHO (2022). *WHO principles for effective communications.* https://apps.who.int/dco/strategy/principles/en/ (käytetty 18.6.2022)

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority.* Westport, Connecticut: Greenwood Press.

Young, S. (2014). Behavioral insights on big data: using social media for predicting biomedical outcomes. *Trends in Microbiology, 22(11), 601–602*. doi:10.1016/j.tim.2014.08.004

Zipf, G. (1949). Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, Massachusetts: Addison-Wesley Press.