Elsalotta Grönberg

NETTIKIUSAAMISEN AUTOMAATTINEN HAVAINNOINTI SOSIAALISESSA MEDIASSA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2023

TIIVISTELMÄ

Grönberg, Elsalotta
Nettikiusaamisen automaattinen havainnointi sosiaalisessa mediassa
Jyväskylä: Jyväskylän yliopisto, 2023, 40 s.
Tietojärjestelmätiede, kandidaatintutkielma
Ohjaaja(t): Riekkinen, Janne

Tässä tutkielmassa tutkittiin erilaisia keinoja nettikiusaamisen automaattiseen havainnointiin sosiaalisessa mediassa. Sosiaalinen media on kehittynyt 2000-luvun alusta lähtien ja nykyään miljoonat ihmiset maailmanlaajuisesti käyttävät sosiaalisen median eri alustoja. Sosiaalisessa mediassa voi esimerkiksi pitää yhteyttä ystäviin ja perheeseen sekä etsiä töitä tai verkostoitua ammatillisesti. Sosiaalisessa mediassa tapahtuu kuitenkin myös nettikiusaamista. Nettikiusaaminen tarkoittaa kiusaamista, joka tapahtuu digitaalisissa ympäristöissä tai digitaalisia laitteita avuksi käyttäen. Erityisesti lapset, nuoret ja nuoret aikuiset kohtaavat nettikiusaamista sosiaalisessa mediassa. Tämän vuoksi nettikiusaamisen havaitseminen on tärkeää, sillä siten nettikiusaamista pystyttäisiin ehkäisemään. Tämän tutkielman tarkoituksena oli koota yhteen erilaisia keinoja, joilla nettikiusaamista voidaan automaattisesti havaita sosiaalisessa mediassa. Nettikiusaamisen automaattinen havainnointi tarkoittaa tässä yhteydessä sitä, että jokin järjestelmä havaitsee nettikiusaamista automaattisesti ilman ihmisen apua. Tutkielma on toteutettu kirjallisuuskatsauksena ja tutkielmassa on pyritty vastaamaan yhteen tutkimuskysymykseen: ”Millä tavoin nettikiusaamista voidaan automaattisesti havainnoida sosiaalisessa mediassa?” Tutkielmassa on määritelty yleisesti sosiaalista mediaa sekä nettikiusaamista, jonka jälkeen on esitelty erilaisia automaattisen havainnoinnin keinoja. Näitä automaattisen havainnoinnin keinoja vertailtiin toisiinsa ja niistä pyrittiin löytämään kaikista toimivimpia keinoja. Tutkielmassa tultiin siihen tulokseen, että nettikiusaamisen automaattista havainnointia on tutkittu paljon viimeisten vuosien aikana. Erityisesti erilaisia järjestelmiä, jotka pyrkivät havaitsemaan kiusaamisviestejä, on kehitetty. Useissa havainnointijärjestelmissä käytetään apuna muun muassa koneoppimista. Kuitenkin edelleen on iso puute toimivista nettikiusaamisen automaattisen havainnoinnin järjestelmistä, jotka onnistuisivat havaitsemaan nettikiusaamista laajasti ja tarkasti. Tämä on yksi syy, miksi nettikiusaaminen on edelleen ongelma sosiaalisessa mediassa.

Asiasanat: nettikiusaaminen, nettikiusaamisen havainnointi, nettikiusaamisen automaattihavainnointi, nettikiusaaminen sosiaalisessa mediassa

ABSTRACT

Grönberg, Elsalotta
Automatic detection of cyberbullying in social media
Jyväskylä: University of Jyväskylä, 2023, 40 pp.
Information Systems, Bachelor's Thesis
Supervisor(s): Riekkinen, Janne

This thesis examines the diverse ways cyberbullying can be automatically detected in social media. Bullying that happens in digital platforms is called cyberbullying. A lot of the cyberbullying is happening in social media nowadays. A lot of kids, teens and young adults have experienced cyberbullying is social media. Social media started to evolve in the begging of the 21st century and nowadays the most used social media platforms are used by millions of people. Social media is used for example for socializing with friends and family, interacting with companies and for professional networking. There are diverse ways to detect cyberbullying in social media, but cyberbullying is still a big problem. The goal in this thesis is trying to gather up diverse ways to detect cyberbullying automatically. Automatic detection means that some system can detect cyberbullying automatically without the help of a human. Primary research question will be addressed by the study, which is: "What are the ways that cyberbullying can be automatically detected in social media?" The thesis defines social media and cyberbullying and lastly gathers up the diverse ways to detect cyberbullying automatically. The key findings of the thesis are that there is a lot of research about automatic cyberbullying detection and many different systems has been made. Lot of the detection systems use for example machine learning to automatically detect cyberbullying. Still there is serious lack of usable systems and cyberbullying being a big problem in social media.

Keywords: cyberbullying, cyberbullying detection, automatic detection of cyberbullying, cyberbullying in social media

TAULUKOT

KUVIOT

SISÄLLYS

1 JOHDANTO

Sosiaalisen median käyttö on lisääntynyt hurjasti viimeisten vuosien aikana. Niin alustojen kuin käyttäjienkin määrä on kasvanut merkittävästi viimeisen kahden vuosikymmenen aikana (Aichner ym., 2021, s. 215). Sosiaalisen median yleisyydestä kertoo hyvin se, että sosiaalisen median suosituinta alustaa eli Facebookia käyttää maailmanlaajuisesti yli 3 miljardia ihmistä (Dixon, 2023). Sosiaalinen media luo ihmisille paljon erilaisia hyötyjä, koska sillä on paljon erilaisia käyttötarkoituksia. Se vaikuttaa muun muassa yrityksien sekä organisaatioiden menestymiseen (Kumar ym., 2016; Paniagua ym., 2017). Lisäksi se voi vaikuttaa muun muassa erilaisten vähemmistöjen hyvinvointiin (Craig ym., 2021; Gillespie-Smith ym., 2021). Kuitenkin sosiaalinen media tuo mukanaan myös haittoja, kuten trollausta, valeuutisia ja riippuvuutta (Allcott & Gentzkow, 2017; Blackwell ym., 2017; Buckels ym., 2014). Näiden lisäksi sosiaalisessa mediassa tapahtuu myös paljon nettikiusaamista.

Nettikiusaamisen yleinen määritelmä on harmin tai vahingon aiheuttamista, tai kiusaamista, käyttäen digitaalista teknologiaa avuksi (Englander ym., 2017, s. 149). Kiusaamistoimia voi olla esimerkiksi huhujen tai kuvien levittämistä, ulkopuolelle jättämistä netissä, uhkaavia yhteydenottoja sekä seksuaalista häirintää (Rahja ym., 2021, s. 6–7). Sosiaalisen median yleistyessä myös nettikiusaaminen on yleistynyt varsinkin nuorten keskuudessa. Noin puolet niin eurooppalaisista kuin yhdysvaltalaisista alle 18-vuotiaista nuorista ovat kokeneet nettikiusaamista jossain vaiheessa elämäänsä (Lobe ym., 2021; J. W. Patchin, 2021). Nettikiusaaminen on siis erittäin yleinen ja huolestuttava ilmiö sosiaalisessa mediassa, jonka vuoksi sen havaitseminen ja täten torjuminen on erittäin tärkeää.

Nettikiusaamista voidaan pyrkiä automaattisesti havainnoimaan erilaisten järjestelmien avulla. Nettikiusaamisen automaattinen havainnointi voidaan määritellä kiusaamistoimien tunnistamiseksi elektronisissa viestintävälineissä ja erityisesti yksittäisten kiusaamisviestien tunnistaminen on tärkeä osa automaattista havainnointia (Salawu ym., 2020, s. 2). Salawun, Hen ja Lumsdenin (2020) jakavat artikkelissaan nettikiusaamisen automaattisen havainnoinnin neljään eri luokkaan. Nämä luokat ovat jaettu lähestymistapojen mukaisesti ja ne ovat: ohjatun

oppimisen lähestymistapa (eng. *supervised learning approach*), sanastopohjainen lähestymistapa (eng. *lexicon-based approach*), sääntöpohjainen lähestymistapa (eng. *rule-based approach*) ja seka-aloitteellinen lähestymistapa (eng. *mixed-initiative approach*) (Salawu ym., 2020, s. 3). Tässä tutkielmassa perehdytään pelkästään ohjatun oppimisen lähestymistavan tutkimuksiin, koska se todettiin tutkielman rajaussyistä parhaaksi tavaksi esittää mahdollisimman monta ajankohtaista keinoa nettikiusaamisen automaattiseen havainnointiin.

Tutkielmassa pyritään vastaamaan yhteen tutkimuskysymykseen: ”Millä tavoin nettikiusaamista voidaan automaattisesti havainnoida sosiaalisessa mediassa?” Tutkielmassa tukitaan keinoja, joilla voidaan automaattisesti havainnoida nettikiusaamista sosiaalisessa mediassa. Tässä tutkielmassa perehdytään erityisesti tekstipohjaisten kiusaamisviestien automaattiseen havainnointiin. Tutkielmassa perehdytään Salawun, Hen ja Lumsdenin (2021) tekemään nelijakoon nettikiusaamisen automaattisen havainnoinnin lähestymistavoista, mutta tutkitaan tarkemmin ohjatun oppimisen lähestymistavan keinoja automaattisesti havainnoida nettikiusaamista.

Tutkielma toteutetaan kirjallisuuskatsauksena. Tutkielmassa käytetty tutkimusaineisto koostuu pääosin kirjallisuudesta sekä aikakausjulkaisujen artikkeleista. Tutkimusaineistoa on kerätty Google Scholar-tietokantaa ja JYKDOK-kirjaston tietokantaa apua käyttäen. Lähdekirjallisuuden etsimiseen on käytetty muun muassa seuraavia hakusanoja: *social media, cyberbullying, automatic detection of cyberbullying, cyberbullying detection, cyberbullying in social media, bullying in social media.* Tutkielmassa hyödynnettyjen artikkeleiden luotettavuutta on tarkistettu Julkaisufoorumi.fi-sivustolla ja ne täyttävät vähintään luokan 1, joka on Julkaisufoorumi-tasojen perustaso (Julkaisufoorumi, 2023). Kaikki artikkelit ovat siis vertaisarvioituja ja artikkelien viittauskerrat on myös tarkistettu luotettavuuden tunnistamiseksi. Lisäksi sosiaalisen median sekä nettikiusaamisen sisältöluvuissa on käytetty tilastotietoja erilaisia tutkimuksia apuna käyttäen. Tämän lisäksi tietoa on etsitty myös erilaisista blogeista ja muista nettilähteistä. Tässä tutkielmassa viitataan yhteensä 48 lähteeseen, joista 28 ovat tutkimusartikkeleita.

Tutkielma koostuu viidestä luvusta, joista kolme ovat sisältölukuja. Ensimmäinen luku on johdanto. Toisessa luvussa perehdytään sosiaaliseen mediaan. Käsitellään sosiaalisen median määritelmää, historiaa sekä sosiaalisen median hyötyjä ja haittoja. Lisäksi esitellään sosiaalisen median käytetyimmät alustat niin Suomessa kuin maailmanlaajuisestikin. Kolmannessa luvussa perehdytään nettikiusaamiseen. Käsitellään nettikiusaamisen määritelmä, nettikiusaamistapauksia sekä tilastoja nettikiusaamisen yleisyydestä Suomessa ja maailmalla. Neljännessä luvussa perehdytään itse tutkimusaiheeseen eli nettikiusaamisen automaattiseen havainnointiin. Luvussa pyritään vastaamaan tutkimuskysymykseen. Ensin määritellään nettikiusaamisen automaattinen havainnointi, jonka jälkeen perehdytään eri tutkimuksien avulla keinoihin, jolla nettikiusaamista voidaan automaattisesti havainnoida. Tavoitteena on käydä läpi erilaisia keinoja, miten nettikiusaamista voidaan automaattisesti havainnoida sosiaalisessa mediassa. Sisältölukujen jälkeen seuraa yhteenveto. Yhteenvedossa kootaan tutkielman kannalta merkityksellisin sisältö sekä pohditaan sisältöluvuissa käsiteltyjä aiheita. Lisäksi yhteenvedossa esitetään tutkimustuloksia sekä

perehdytään jatkotutkimusaiheisiin. Tutkielman lopussa on lähdeluettelo, josta löytyy tutkielmassa käytetty lähdekirjallisuus.

2 SOSIAALINEN MEDIA

Tässä sisältöluvussa käsitellään sosiaalista mediaa. Jotta voimme ymmärtää mitä sosiaalinen media tarkoittaa on hyvä perehtyä sosiaalisen median historiaan sekä käyttötarkoituksiin. Lisäksi tässä luvussa käydään läpi mitä hyötyjä ja haittoja sosiaalinen media tuo mukanaan. Lopuksi käsitellään vielä tunnetuimpia sosiaalisen median alustoja.

2.1 Sosiaalisen median määritelmä

Sosiaalisen median käyttö on tullut osaksi arkeamme viimeisten vuosien aikana. Niin alustojen kuin käyttäjienkin määrä on kasvanut merkittävästi viimeisen kahden vuosikymmenen aikana (Aichner ym., 2021, s. 215). Ymmärtääksemme sosiaalista mediaa kokonaisuudessaan on tärkeää tarkastella sen historiaa sekä sitä, mihin kaikkeen sosiaalista mediaa käytetään.

Media on kehittynyt hurjasti viimeisien vuosikymmenien aikana ja viimeisen kahden vuosikymmenen aikana sosiaalinen media on tullut perinteisen median ohelle. 1900-luvun loppupuolella media alkoi kehittymään analogisesta digitaaliseksi ja tämän mahdollistavat Internet ja World Wide Web (Goff, 2013, s. 17). Schakman (2013) kertoo, että vaikka Internet alun perin oli tarkoitettu tutkimuskäyttöön, salli se kuitenkin myös ihmisten kommunikoida samojen kiinnostuksen kohteiden omaavien ihmisten kanssa. Kiinnostuksen kohteiden perusteella luotiin yhteisöjä, jotka kommunikoivat internetin kautta. Näiden yhteisöjen syntymisestä alkoi varhainen sosiaalisen median kehittyminen (Schakman, 2013, s. 111). Aichnerin ym. (2021) artikkelissa mainitaan, että sosiaalisen median termiä käytettiin ensimmäisen kerran vuonna 1994. Siitä lähtien sosiaalisen median alustat ja käyttäjät ovat lisääntyneet merkittävästi (Aichner ym., 2021, s. 215). Puolestaan Goff (2013) kertoo, että The Pew Research Center's Internet ja American Life Project ovat seuranneet internetin vaikutusta yhteiskuntaan vuodesta 1999 asti ja vuonna 2005 sosiaalisesta mediasta on ensimmäistä kertaa alettu

raportoida. Lähteiden mukaan sosiaalisen median käyttö ja kehittyminen on siis alkanut suurin piirtein 2000-luvun vaihteessa ja kehittyminen jatkuu edelleen.

Sosiaalista median kehittyessä sen määritelmä on myös muuttunut ajan kuluessa. Sosiaalisesta mediasta on tehty paljon tutkimusta ja täten myös erilaisia määritelmiä on paljon. Wellman ym. (1996) käyttää sosiaalisesta mediasta termiä sosiaalinen verkosto. Nämä tietokonetuetut sosiaaliset verkostot syntyvät, kun tietokoneverkot yhdistävät ihmisiä ja koneita (Wellman ym., 1996, s. 213). Kuusi vuotta myöhemmin Ridings, Gefen ja Arnize (2002) kertovat sosiaalisen median tarkoittavan virtuaaliyhteisöjen tapaamispaikkaa. Virtuaaliyhteisöillä on samoja kiinnostuksen kohteita sekä toimintatapoja, ja he kommunikoivat säännöllisesti internetin kautta yhteisen sijainnin tai mekanismin kautta. Näitä sijainteja tai mekanismeja voivat olla esimerkiksi chat-huoneet, ilmoitustaulut tai sähköpostilistapalvelimet (Ridings ym., 2002, s. 273). Näissä lähteissä on huomattavissa samankaltaisuuksia. Voidaan päätellä, että 2000-luvun vaihteessa sosiaalisen median ajatus perustui siihen, että vain tietyt ihmisryhmät kommunikoivat toistensa kanssa internetiä avuksi käyttäen. Puolestaan vuonna 2010 Kaplan ja Haenlein (2010) kertovat sosiaalisen median olevan ryhmä internetpohjaisia sovelluksia, jotka mahdollistavat näiden sovellusten käyttäjien tuottaman sisällön luomisen ja vaihtamisen (Kaplan & Haenlein, 2010, s. 61). On nähtävissä, että 2000-luvusta 2010-lukuun sosiaalinen media kehittyi paljon, eikä kyse ollut enää pelkästään erilaisten yhteisöjen tai ryhmien kommunikoinnista. Vuonna 2018 Kapoor ym. (2018) kuvaavat sosiaalisen median koostuvan erilaisista käyttäjälähtöisistä alustoista, joiden tarkoituksena on levittää käyttäjän luomaa sisältöä, luoda vuoropuhelua käyttäjien välille ja viestiä laajemmalle yleisölle. Kirjoittajien mukaan se on pohjimmiltaan ihmisen ja ihmisten luoma digitaalinen tila, jossa edistetään vuorovaikutusta ja verkostoitumista eri tasoilla. Näitä tasoja ovat esimerkiksi ammatillinen, liiketoiminnallinen, markkinoinnillinen ja poliittinen (Kapoor ym., 2018, s. 536). Kuten näistä neljästä määritelmästä on huomattavissa, sosiaalisen median merkitys on laajentunut huomattavasti vuosien aikana. Ridingsin, Gefenin ja Arnizen (2002) määritelmässä puhuttiin vain virtuaaliyhteisöistä ja ihmisryhmistä, jotka kommunikoivat hyödyntäen verkkoa, kun taas Kapoor ym. (2018) kuvaavat kuusitoista vuotta myöhemmin sosiaalista mediaa digitaaliseksi tilaksi, joka yleisesti edistää vuorovaikutusta ja verkostoitumista. On siis lähteiden mukaan nähtävissä, että sosiaalinen media on nykyään tullut osaksi ihmisten kommunikointia ja vuorovaikutusta, kun vuosituhannen alussa vain tietyt yhteisöt ja ihmisryhmät käyttivät internetiä kommunikoinnin apuna.

2.2 Sosiaalisen median hyödyt ja haitat

Sosiaalinen media on kehittynyt viimeisten vuosikymmenien aikana monipuolisemmaksi. Kuten aikaisemmin mainituista lähteistä käy ilmi, sosiaalinen media on tärkeä vuorovaikutuksen, kommunikoinnin, verkostoitumisen, sisällön luomisen väline, joka toimii erilaisissa internetpohjaisissa alustoissa. Kapoor ym. (2018) mainitsi artikkelissaan sosiaalisen median edistävän vuorovaikutusta

ja verkostoitumista eri tasoilla. Näitä tasoja ovat esimerkiksi henkilökohtainen, ammatillinen, liiketoiminnallinen, markkinoinnillinen, poliittinen ja yhteiskunnallinen (Kapoor ym., 2018, s. 536). Koska näitä eri tasoja sosiaalisen median vuorovaikutuksessa on paljon, myös käyttötarkoituksia on monia. Aichenrin ym. (2021) artikkelissa esitellään viisi eri käyttötapaa sosiaaliselle medialle, jotka ovat yhteydenpito kavereiden ja perheen kanssa, romanttinen käyttö, kommunikointi yritysten ja brändien kanssa, töiden etsiminen ja ammatillinen verkostoituminen sekä liiketoiminta. Sosiaalisen median avulla voi siis pitää yhteyttä kavereihin ja perheeseen, mutta myös löytää uusia ystäviä tai romanttisia suhteita sekä verkostoitua ammatillisesti ja löytää jopa työpaikkoja. Lisäksi sosiaalisessa mediassa voi lähestyä pienellä kynnyksellä erilaisia yrityksiä sekä tehdä myös mahdollisesti omaa liiketoimintaa näkyvämmäksi (Aichner ym., 2021, s. 216–217).

Sosiaalisen median monipuolisten käyttötarkoituksien ansiosta, on se mullistanut yhteydenpidon niin ihmisten kuin yritysten välillä ja täten siitä hyötyy laajalti eri tahot kuten yksittäiset ihmiset, yritykset, kaupungit ja jopa valtiot. Kuitenkin tuo se mukanaan myös haittoja, riskejä ja vaaroja. Seuraavaksi tutustutaan sosiaalisen median hyötyihin ja haittoihin, ja täten myös sosiaalisen median käyttötarkoituksiin.

2.2.1 Sosiaalisen median hyödyt

Sosiaalisen median ehdoton hyöty on yleisesti se, että se helpottaa kommunikointia ihmisten välillä. Lisäksi se helpottaa yritysten kommunikointia asiakkaiden, työntekijöiden ja muiden yritysten kesken (Aichner ym., 2021, s. 216). Seuraavaksi perehdytään muutamaan eri tutkimukseen liittyen sosiaalisen median hyötyihin.

Kumarin ym. (2013) tutkimusartikkelissa tutkitaan miten firmojen sosiaalisen median käyttö vaikuttaa asiakkaiden ostokäyttäytymiseen. Tarkemmin tutkitaan miten FGC (eng. *firm-generated content*, firmojen markkinointiviestinä sosiaalisen median alustoissa) vaikuttaa asiakkaan rahankäyttöön, asiakkaiden ostotottumuksiin ja lisäksi asiakkaasta saatuun tuottoon. Tutkimuksen tulokset osoittavat, että FGC:llä on positiivinen vaikutus asiakkaan käytökseen. Mitä paremmin asiakas ymmärtää teknologiaa ja on kokenut sekä taipuvainen sosiaalisen median käyttämiseen sitä suurempi vaikutus FGC:llä on (Kumar ym., 2016).

Paniagua, Korzynski ja Mas-Tur (2017) kertovat tutkimusartikkelissaan, että vuonna 2016 Espanjassa alettiin sosiaalisen median alustojen Facebookin, Twitterin ja change.org:in kautta kampanjoimaan, jotta Tesla Motors niminen sähköautojen valmistamisen yritys perustaisi tuotantolaitoksen Espanjaan. Mikäli Tesla Motors olisi näin toiminut, tämä olisi ollut ulkomainen suora sijoitus (FDI). Artikkelissa tutkitaan ulkomaisten suorien sijoituksien ja sosiaalisten verkostojen välistä suhdetta. Tutkimuksen tulokset kertovat, että yksinkertaistettuna sosiaalisilla verkostoilla on vaikutusta ulkomaisiin suoriin sijoituksiin, erityisesti jos on kyse isoista monikansallisista yrityksistä (Paniagua ym., 2017, s. 314, 323–324).

Craig ym. (2021) ovat tutkineet artikkelissaan miten sosiaalinen media voi vaikuttaa seksuaali- ja sukupuolivähemmistöjen nuorten hyvinvointiin.

Artikkelissa kerrotaan, että sosiaalinen media tarjoaa mahdollisuuksia lesboille, homoille, trans-, queer- ja muun seksuaalisuuden omaaville sekä sukupuolivähemmistö- (lyh. LGBTQ+) nuorille parantaa hyvinvointiaan tutkimalla omaa identiteettiään, käyttämällä sosiaalisen median resursseja ja olemalla yhteydessä ikätovereihin. Tutkimuksessa tehtiin verkkokysely, johon vastasi yli kuusituhatta 14-29-vuotiasta LGBTQ+ -nuorta. Tutkimuksen tuloksista selvisi, että sosiaalinen media tarjoaa näille vähemmistönuorille emotionaalista tukea ja auttaa heitä kehittämään identiteettiään, löytämään tärkeitä tietoja sekä viihdyttämään heitä (Craig ym., 2021).

Gillespie-Smith ym. (2021) tutkivat artikkelissaan autististen nuorten sosiaalisen median käyttöä. Artikkelissa kerrotaan, että aiemmissa tutkimuksissa on osoitettu, että sosiaalisen median käyttö voi auttaa helpottamaan sosiaalista toimintaa. Kuitenkin käsitykset sosiaalisen median käytön riskeistä ja hyödyistä on vielä epäselviä. Tutkijat haastattelivat autistisia nuoria sekä autististen nuorien vanhempia. Tutkimuksista selvisi, että sosiaalinen mediassa tapahtuva vuorovaikutus on erityisen arvokasta autistisille nuorille, koska se tarjoaa heille helpomman sosiaalisen vuorovaikutuksen kuin ”tosielämässä”. Tuloksista huomattiin myös, että autistiset nuoret ovat tietoisia riskeistä verkossa ja he pohtivat tapoja, joilla välttää näitä riskejä (Gillespie-Smith ym., 2021, s. 1).

Kuten aikaisemmin mainittu sosiaalisesta mediasta on paljon hyötyä sen monipuolisten käyttötarkoitusten vuoksi. Kumarin ym. (2013) ja Paniaguan, Korzynskin ja Mas-Turin (2017) tutkimusartikkeleissa tuotiin hyvin ilmi, kuinka monipuolisesti sosiaalinen media voi vaikuttaa yrityksiin ja niiden menestymiseen. Sosiaalinen media siis luo täysin uuden alustan yrityksille markkinoida ja menestyä. Puolestaan Gillespie-Smithin ym. (2021) ja Craigin ym. (2021) tutkimuksissa tuotiin hyvin ilmi, miten sosiaalinen media voi vaikuttaa kaikkien ihmisten lomassa myös erilaisten vähemmistöjen hyvinvointiin ja helpottaa heidän elämäänsä. Ihmisten voi olla vaikea tutustua uusiin ihmisiin kasvokkaisessa vuorovaikutuksessa monista eri syistä, jolloin sosiaalinen media luo uuden tavan tutustua ihmisiin. Sosiaalinen media vaikuttaa positiivisesti siis monien eri ihmisten, ihmisryhmien, yritysten ja yhteiskuntien elämään.

2.2.2 Sosiaalisen median haitat

Kuten mainittu sosiaalista mediaa voidaan käyttää hyvin monipuolisesti. Tämän vuoksi sosiaalisen median käyttöön liittyy myös paljon riskejä ja haittoja. Sosiaalisen median yleistyessä osaksi ihmisten arkea, myös sen haittoja alettiin tutkia. Yksi tunnetuimmista sosiaalisen median haitoista on nettikiusaaminen, mutta siihen perehdytään tässä tutkielmassa myöhemmin. Seuraavaksi esitellään muutama tutkimus, joissa on tutkittu sosiaalisen median haittoja.

Blackwell ym. (2017) kertovat artikkelissaan sosiaalisen median käytön olevan nykyään niin yleistä, että se voi joillekin käyttäjille aiheuttaa jopa sosiaalisen median riippuvuuden. Tutkimuksessa tutkittiin mitkä tekijät vaikuttavat eniten sosiaalisen median addiktioon. Tutkijat valitsivat neljä eri tekijää eli ulospäinsuuntautuneisuuden, neuroottisuuden, kiintymyssuhteiden ja FOMO:n (”*fear of missing out*” eli pelko siitä, että jää jostain paitsi) ja tutkivat miten nämä

vaikuttavat sosiaalisen median addiktion syntymiseen. Tutkimuksissa selvisi, että pelkästään FOMO vaikuttaa sosiaalisen median addiktion syntymiseen (Blackwell ym., 2017, s. 69).

Buckels, Trapnell ja Paulhus (2014) tutkivat artikkelissaan trollausta. Artikkelissa viitataan Hyden (1998) artikkeliin, jonka mukaan verkkotrollaaminen tarkoittaa petollista, tuhoisaa tai häiritsevää käytöstä internetin sosiaalisessa ympäristössä ilman näkyvää syytä (Hyde 1998, viitattu lähteessä Buckels ym., 2014, s. 97). Tutkimuksessa vertailtiin kahta verkkotutkimusta ja tutkittiin yhteyttä trollauksen ja persoonallisuuden ”Dark Tetradin” välillä. ”Dark Tetrad” tarkoittaa neljää persoonallisuuden piirrettä, jotka ovat machiavellismi (häikäilemätön valtapolitiikka), narsismi, psykopaattisuus ja sadismi. Tuloksista selvisi, että machiavellismi, sadismi ja psykopatia korreloivat positiivisesti trollauksen kanssa. Sadismin ja trollauksen yhteys oli suurinta ja tutkijat päätyivät tulokseen, että verkkotrollaus onkin jokapäiväisen sadismin ilmentymä internetissä (Buckels ym., 2014, s. 97, 100–101).

Allcotin ja Gentzkowin (2017) artikkelissa käydään läpi sosiaalisessa mediassa levinneiden valeuutisten vaikutusta USA:n vuoden 2016 presidentinvaaleihin. Artikkelissa kerrotaan, että sosiaalinen media eroaa perinteisestä mediasta siten, että sisältöä voidaan välittää toisille käyttäjille ilman kolmannen osapuolen tarkistusta tai toimituksen päätöstä. Artikkelin tutkimuksessa käydään laajalti läpi valeuutisten vaikutusta USA:n vuoden 2016 vaaleihin. Tutkimusten tulosten mukaan keskiarvollisesti jokainen USA:n aikuinen kansalainen luki ja painoi mieleensä yhden tai useamman valeuutisen vaalikauden aikana. Tuloksista ei voida kuitenkaan suoraan sanoa vaikuttivatko valeuutiset vaalien tuloksiin ja kuinka paljon (Allcott & Gentzkow, 2017, s. 211, 232–233).

Kuten aikaisemmin mainittu sosiaalisesta mediasta on tullut olennainen osa markkinointiviestintää. Näin todetaan myös Pfefferin, Zorbachin ja Carleyn (2014) artikkelissa. Vaikka sosiaalisesta mediasta on paljon hyötyä markkinointiviestinnässä, markkinoijia häiritsevät ”online firestorm” -nimiset ongelmat sosiaalisessa mediassa. ”Online firestorm” tarkoittaa, sitä kun sosiaalisen median käyttäjät reagoivat mihin tahansa kyseenalaiseen väittämään tai käytökseen. Käyttäjät voivat jopa muutamassa tunnissa aiheuttaa paljon paheksuntaa tiettyyn henkilöön tai yritykseen. Tästä ilmiöstä kärsivät niin poliitikot, yritykset kuin julkisuuden henkilötkin ja lisäksi tämä on myös haaste aiemmin mainitulle markkinointiviestinnälle (Pfeffer ym., 2014, s. 117).

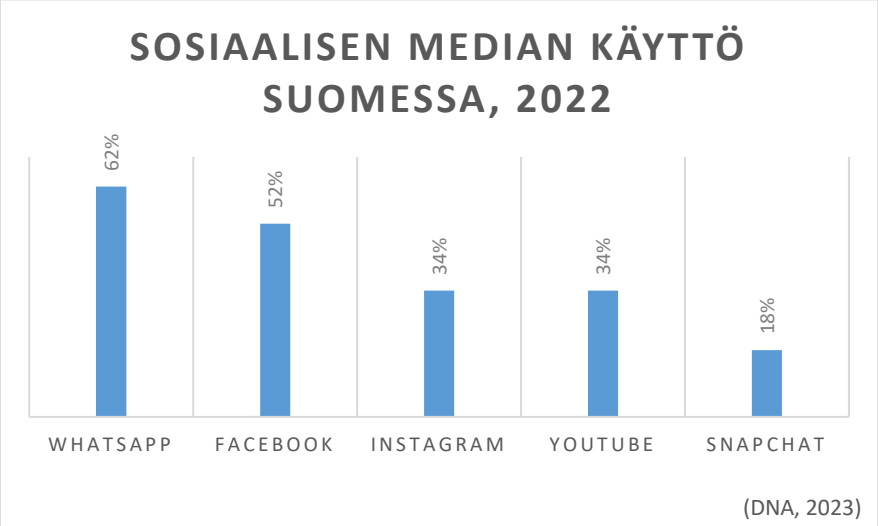
Kuten aiemmin mainittu sosiaalisessa mediassa käyttäjät jakavat itse tuottamaansa sisältöä eri alustoille ja tämä antaa monien hyvin asioiden lomassa mahdollisuuden myös muun muassa trollaamiselle (Buckels ym., 2014), valeuutisille (Allcott & Gentzkow, 2017) sekä ylipäänsä negatiiviselle puheelle, joka voidaan myös kohdistaa yhteen tiettyyn kohteeseen (Pfeffer ym., 2014). Näiden haittojen lomassa sosiaalinen media on addiktoivaa (Blackwell ym., 2017) ja täten altistaa ihmisiä sosiaalisen median haitoille. Näistä haitoista voi olla paljon harmia niin ihmisille kuin myös yrityksille ja brändeille.

2.3 Sosiaalisen median alustat

Kuten mainittu, sosiaalista mediaa käytetään laajalti eri käyttötarkoituksiin. Tämä tietenkin tarkoittaa, että myös alustoja on monia ja erilaisia. Seuraavaksi onkin hyvä tutkia tilastojen avulla, miten sosiaalista mediaa käytetään Suomessa sekä maailmalla ja perehtyä myös hieman suosituimpiin sosiaalisen median alustoihin.

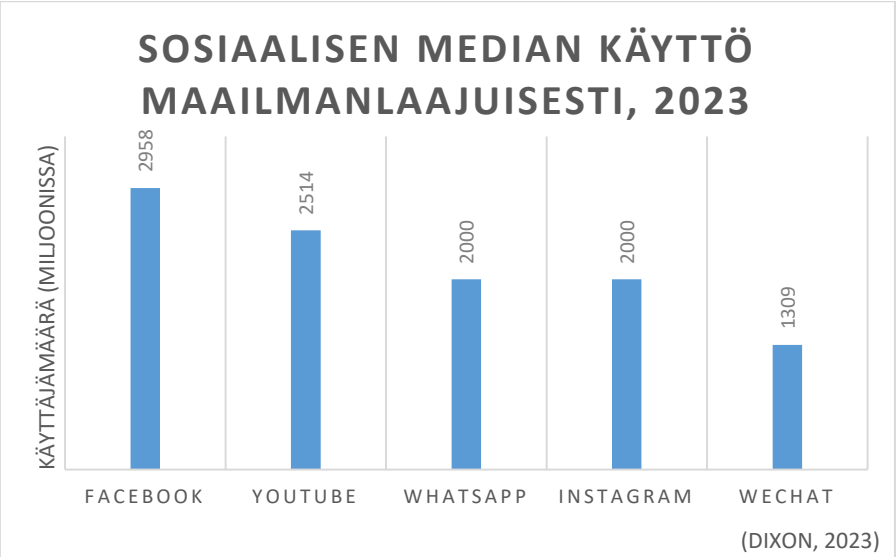
DNA (2023) on sivuillaan tehnyt tutkimusta suomalaisten digitaalisista elämäntavoista. Digitaaliset elämäntavat- tutkimuksessa on myös tutkittu suomalaisten sosiaalisen median käyttöä. Tutkimuksen kohderyhmä on 16–74-vuotiaat suomalaiset ja kokonaisuudessaan tutkimukseen vastasi tuhat ihmisitä. Tiedot kerättiin 4. - 12.5.2022 välisenä aikana (*Digitaaliset elämäntavat -tutkimus 2022*, 2023). Taulukossa (taulukko 1) on esitetty suomalaisten viisi suosituinta sosiaalisen median alustaa ja kerrottu kuinka monta prosenttia tuhannesta vastaajasta käyttää kyseisiä alustoja vähintään 1–2 kertaa päivässä.

TAULUKKO 1 Sosiaalisen median käyttö Suomessa, 2022



Dixon (2023) on julkaissut Statista-nimisellä verkkosivulla tilaston suosituimmista sosiaalisen median alustoista maailmanlaajuisesti. Tilastossa on esitetty suosituimpia alustoja tammikuulta 2023 näiden aktiivisten käyttäjämäärien mukaisesti (Dixon, 2023). Taulukossa (taulukko 2) on esitetty viisi suosituinta sosiaalisen median alustaa maailmanlaajuisesti käyttäjämäärien mukaisesti.

TAULUKKO 2 Sosiaalisen median käyttö maailmanlaajuisesti, 2023



Kuten molemmista taulukoista (taulukko 1, taulukko 2) on huomattavissa neljä suosituinta sosiaalisen median alustaa niin Suomessa kuin maailmanlaajuisestikin ovat Facebook, Instagram ja WhatsApp ja YouTube. Eroavaisuuksia suomalaisten sosiaalisen median käytössä kuitenkin löytyy verrattaessa sitä maailmanlaajuiseen käyttöön. Yksi eroavaisuus käytössä on se, että Suomessa WhatsAppia käytetään huomattavasti eniten, kun taas maailmanlaajuisesti se on vasta kolmanneksi suosituin alusta. Huomattavaa on myös se, että YouTube on maailmanlaajuisesti toisiksi suosituin alusta, mutta suomalaisten keskuudessa vasta neljänneksi suosituin. WeChat on puolestaan maailmanlaajuisesti viidenneksi suosituin alusta, mutta DNA:n (2023) teettämässä tutkimuksessa kyseistä alustaa ei edes mainita, joten Suomessa sen käytön täytyy olla hyvin vähäistä, jos käyttöä ylipäänsä on. Perehdytään seuraavaksi hieman suosituimpiin sosiaalisen median alustoihin.

Metan nettisivuilla (ei pvm.) kerrotaan Metan olevan yksi suurimmista sosiaalisen median alan yrityksistä. Facebook perustettiin vuonna 2004 ja se muutti tavan, jolla ihmiset pitävät yhteyttä toisiinsa. Vuonna 2012 Facebook yrityksenä (Facebook, Inc.) osti Instagramin ja vuonna 2014 he ostivat myös WhatsAppin. Vuonna 2021 Facebook yrityksenä muutti nimekseen Meta (Meta Platforms Inc.), sillä yrityksellä on monia erilaisia tuotteita sekä palveluita ja täten Facebook yrityksen nimenä oli harhaanjohtava (*Meta*, ei pvm.).

Facebook siis perustettiin vuonna 2004. Goff (2013) kertoo, että Facebook oli aluksi käytössä vain Harvardin yliopiston opiskelijoille, jotka pystyivät luomaan ja jakamaan omia verkkoprofiilejaan sivustolla. Vuosien varrella Facebook laajeni muihin yliopistoihin sekä lukioihin ja vuoden 2006 syksyllä kaikkien yli 13-vuotiaiden käyttöön (Goff, 2013, s. 35–36). Facebookista on kasvanut yksi maailman suurimmista sosiaalisen median sivustoista (Goff, 2013, s. 35) ja nykyään Facebookilla on melkein 3 miljardia aktiivista käyttäjää (Dixon, 2023). Facebook kuvaa nettisivuillaan alustansa auttavan ihmisiä pitämään yhteyttä kavereihinsa

ja perheisiinsä sekä löytämään uusia ystäviä sekä yhteisöjä, joilla on samat mielenkiinnon kohteet (*Facebook | Meta*, ei pvm.).

Goff (2013) kertoo, että YouTube on videoiden jakamiseen tarkoitettu sivusto, joka perustettiin vuonna 2005. YouTube havainnollistaa hyvin sosiaalisen median alustoille tyypillisiä ratkaisuja. Se on suunniteltu tukemaan mainostajia, tarjoaa käyttäjien luomaa sisältöä ja sen käyttämiseen ei vaadita monimutkaista teknologiaosaamista. Google osti YouTuben vuonna 2006 (Goff, 2013, s. 24–25). Nykyään YouTubella on 2,5 miljardia aktiivista käyttäjää ja täten se on maailmanlaajuisesti toisiksi suosituin sosiaalisen median alusta (Dixon, 2023).

Instagram julkaistiin syksyllä vuonna 2010 (”Instagram Launches”, 2010) ja vuonna 2012 silloinen Facebook osti alustan (Systrom, 2012). Nykyään Instagramilla on 2 miljardia aktiivista käyttäjää ja tämä määrä tekee siitä maailmanlaajuisesti kolmanneksi suosituimman sosiaalisen median alustan yhdessä WhatsAppin kanssa (Dixon, 2023). Instagramissa pystyy jakamaan kuvia sekä videoita ja juttelemaan muiden käyttäjien kanssa (*About Instagram*, ei pvm.).

WhatsApp julkaistiin alkukeväästä vuonna 2009 ja vuonna 2014 silloinen Facebook osti alustan (”Thank You for 10 Years”, 2019). Nykyään WhatsAppilla on 2 miljardia aktiivista käyttäjää ja tämä määrä tekee sitä maailmanlaajuisesti kolmanneksi suosituimman sosiaalisen median alustan yhdessä Instagramin kanssa (Dixon, 2023). WhatsApp on saatavilla kaikkialla maailmassa ja sen avulla ihmiset voivat lähettää toisilleen viestejä sekä soittaa puheluita maksuttomasti (*WhatsApp*, ei pvm.).

3 NETTIKIUSAAMINEN

Tässä sisältöluvussa perehdytään nettikiusaamisen. Nettikiusaamisella tarkoitetaan kiusaamista, jonka apuna käytetään teknologisia laitteita. Nettikiusaaminen on kuitenkin kehittynyt teknologian kehittyessä, jonka vuoksi siitä on monia erilaisia määritelmiä. Sisältöluvussa käsitellään nettikiusaamisen määritelmää sekä annetaan esimerkkejä nettikiusaamistapauksista. Lisäksi käydään hieman läpi nettikiusaamistilastoja niin Suomesta kuin maailmalta.

3.1 Nettikiusaamisen määritelmä

Nettikiusaamisesta on monia eri määritelmiä, eikä sen määritelmästä ole päästy vielä yhteisymmärrykseen (Englander ym., 2017, s. 149). Yleensä nettikiusaaminen määritellään kuitenkin harmin tai vahingon aiheuttamiseksi, tai kiusaamiseksi, käyttäen digitaalista teknologiaa avuksi (Englander ym., 2017, s. 149).

Nettikiusaamisesta on puhuttu ensimmäisen kerran vuonna 1999 ja sen jälkeen määritelmää on alettu luomaan. Vuonna 2006 Patchin ja Hinduja (2006) määrittelivät nettikiusaamisen olevan tahallista ja toistuvaa vahingonaiheuttamista käyttämällä tietokonetta, matkapuhelinta tai muuta elektronista laitetta (J. Patchin & Hinduja, 2006, s. 152). Puolestaan vuonna 2014 Kowalski ym. (2014) määrittelivät nettikiusaamisen olevan elektronisten kommunikaatioteknologioiden käyttämistä kiusaamistarkoitukseen (Kowalski ym., 2014, s. 1074). Määritelmä ei siis vajaassa kymmenessä vuodessa ole muuttunut paljoa, vaan määritelmät ovat hyvin samankaltaisia. Vuonna 2021 Mannerheimin lastensuojeluliitosta Rahjan ym. (2021) teettämässä kyselyraportissa kerrotaan kyselyyn vastanneiden 12–17-vuotiaiden nuorien kokemuksia nettikiusaamisesta. Raportissa nettikiusaamisen kuvataan olevan moninainen ilmiö, jota tapahtuu verkko- ja sosiaalisen median -alustoilla sekä nettipeliyhteisöissä (Rahja ym., 2021, s. 6). Verrattuna Patchin ja Hindujan (2006) ja Kowalskin ym. (2014) määritelmiin Rahja

(2021) kuvaa nettikiusaamista hyvin yleismaailmallisesti. Huomattavaa kuitenkin on, että nettikiusaaminen määritellään tapahtuvan eri alustoilla kuin aikaisemmissa määritelmissä sen todettiin tapahtuvan erilaisia elektronisia laitteita avuksi käyttäen. Toki verkkoalustoille pääsy vaatii jonkun elektronisen laitteen, mutta voidaan päätellä näiden laitteiden olevan niin yleisiä, ettei niistä tarvitse erikseen mainita. On myös huomattavaa, että elektronisten laitteiden kuten matkapuhelimien välityksellä tapahtunut kiusaaminen on aikaisemmin voinut tapahtua esimerkiksi tekstiviestein, kun nykyään se tapahtuu Rahjan ym. (2021) mukaan juuri eri alustoilla, kuten sosiaalisessa mediassa.

On olemassa myös monia eri tapoja kiusata netissä. Willardin (2006) kirjassa kuvataan eri tapoja nettikiusaamiseen. Näitä tapoja ovat ulkopuolelle jättäminen, tekeytyminen toiseksi henkilöksi, haukkuminen, henkilökohtaisten tietojen levittäminen, loukkaaminen ja uhkaaminen sekä ahdistelu ja häirintä (Willard 2006, viitattu lähteestä Horowitz & Bollinger, 2014, s. 9). Mannerheimin lastensuojeluliitosta Rahjan ym. (2021) teettämässä kyselyraportissa kyselyyn vastanneiden mukaan nettikiusaaminen voi olla esimerkiksi ilkeitä kommentteja, huhujen tai kuvien levittämistä, ulkopuolelle jättämistä netissä, uhkaavia yhteydenottoja sekä seksuaalista häirintää. (Rahja ym., 2021, s. 6–7). Lähteistä on huomattavissa, että nettikiusaamisen tavat eivät ole muuttuneet juurikaan vuosien aikana.

Nettikiusaaminen voidaan siis lähteiden perusteella määritellä kiusaamiseksi, jossa käytetään teknologisia laitteita avuksi. Vaikka (Englander ym., 2017, s. 149) väitti artikkelissaan, että nettikiusaamisen määritelmiä on monia, eikä niistä ole päästy yhteisymmärrykseen, on kuitenkin eri määritelmät hyvin samankaltaisia. Nykyään nettikiusaaminen tapahtuu enimmäkseen erilaisilla verkkoalustoilla sekä sosiaalisessa mediassa, mikä voidaan nähdä muutoksena entiseen nettikiusaamiseen. Tässä tutkielmassa perehdytäänkin tarkemmin nettikiusaamiseen juuri sosiaalisessa mediassa. Tavat nettikiusaamiseen eivät kuitenkaan ole muuttuneet vuosien varrella vaan Willardin (2006) kirjassa kertomia nettikiusaamisen tapoja on mainittu myös Rahjan ym. (2021) kyselyraportissa.

3.2 Nettikiusaamistapauksia

Nettikiusaamista on monenlaista ja sillä on vakavia seurauksia uhreille. Jotta voimme ymmärtää käytännössä, mitä nettikiusaaminen on ja mitä siitä aiheutuu sen uhreille, voidaan käydä läpi tosielämän nettikiusaamistapauksia. Horowitzin ja Bollingerin (2014) kirjassa käsitellään erilaisia esimerkkitapauksia nettikiusaamisesta. Kirjassa kerrotaan muun muassa Nafeesa Onquen, Janey Rodemeyerin ja Amanda Toddin tarinat (Horowitz & Bollinger, 2014, s. 26–29) ja seuraavaksi käsitellään niitä.

Horowitz ja Bollinger (2014) kertovat Nafeesan tarinan. Nafeesa Onque oli teini-ikäinen cheerleader, joka käytti Facebookia ja muita sosiaalisen median

alustoja tavallisten nuorten tavoin. Kuitenkin joku päätti luoda Nafeesasta valeprofiilin niin Facebookiin kuin muihinkin sosiaalisen median palveluihin. Kiusaaja esitti olevansa Nafeesa ja pyysi Facebookissa kaverikseen Nafeesan kaupungissa asuvia nuoria sekä Nafeesan sukulaisia. Kun nämä ihmiset hyväksyivät pyynnöt, kiusaaja alkoi lähetellä heille uhkaavia viestejä, seksuaalista sisältöä sekä muita hävyttömyyksiä. Nafeesan ystävät ja tutut alkoivat kysellä Nafeesalta miksi hän lähetteli viestejä. Nafeesa oli peloissaan, eikä tiennyt keneen luottaa, joten hän alkoi vetäytyä sosiaalisista tilanteista. Lopulta kiusaaja alkoi tekemään vielä häiriintyneempiä julkaisuja valetileilleen ja julkaisi seksuaalista sisältöä tilille väittäen, että videoissa ja kuvissa esiintyi oikea Nafeesa. Lopulta useiden kuukausien jälkeen, poliisi onnistui löytämään kiusaajan. Kiusaaja oli viisitoistavuotias tyttö, joka asui samassa rakennuksessa Nafeesan ja tämän perheen kanssa. Kiusaaja pidätettiin ja häntä vastaan nostettiin syyte laittomasta tekeytymisestä toiseksi ihmiseksi (Horowitz & Bollinger, 2014, s. 26–27).

Horowitz ja Bollinger (2014) kertovat Jamey Rodemeyerista, joka oli neljätoistavuotias nuori, jota kiusattiin koulussa hänen biseksuaalisuutensa vuoksi. Jamey piti blogia, jossa kertoi kokemuksistaan. Blogi keräsi paljon nimettömiä kommentteja, joissa Jameyta kiusattiin ja yllytettiin itsemurhaan. Tästä huolimatta Jamey julkaisi inspiroivia ja itsemurhavastaisia julkaisuja niin YouTubeen kuin muihinkin sosiaalisen median alustoihin. Vuosien kiusaamisen jälkeen Jamey kuitenkin päätyi tekemään itsemurhan (Horowitz & Bollinger, 2014, s. 28).

Horowitz ja Bollinger (2014) käsittelevät kirjassaan myös tunnettua nettikiusaamistapausta eli Amada Toddin tarinaa. Amandan kiusaaminen alkoi, kun hän seitsemännellä luokalla alkoi käyttää videochat-palveluita etsiäkseen uusia ystäviä. Eräs henkilö palvelussa suostutteli Amandan ottamaan paitansa pois. Amandalle selvisi myöhemmin, että kuva hänestä ilman paitaa levisi internetissä. Tästä aiheutunut kiusaaminen koulussa aiheutti Amandalle ahdistusta, masennusta ja paniikkihäiriön. Myöhemmin Amanda puolestaan alkoi käyttää alkoholia ja huumeita helpottaakseen oloansa. Kiusaaminen jatkui ja joku loi myös Facebook profiilin, jonka profiilikuvana oli kyseinen paidaton kuva Amandasta. Amandan vaihdettua koulua henkilö alkoi ottaa profiilin kautta yhteyttä uuden koulun oppilaisiin. Kiusaamisen jatkuessa Amanda yritti myös itsemurhaa, mutta hänet pystyttiin pelastamaan. Kun Amanda palasi kotiin sairaalasta itsemurhayrityksen jälkeen, hän huomasi Facebook-sivullaan loukkaavia viestejä epäonnistuneesta itsemurhayrityksestä. Kiusaamisen seurauksena Amanda alkoi myös lääkitä itseään alkoholin ja muiden päihteiden avulla. Amanda perheineen muutti aloittaakseen puhtaalta pöydältä, mutta Amanda ei päässyt eroon menneisyyttään. Vielä kuuden kuukauden jälkeenkin loukkaavia viestejä lähetteltiin Amandalle eri sosiaalisen median alustoissa. Amanda yritti uudestaan itsemurhaa lääkkeiden yliannostuksella, mutta hänet pystyttiin pelastamaan. Vuonna 2012 Amanda Todd julkaisi YouTubeen videon, jossa itse kirjoittamiensa lappujen avulla kertoi omista kiusaamiskokemuksistaan. Noin kuukausi videon

julkaisun jälkeen, Amanda teki itsemurhan ja tällä kertaa häntä ei ehditty pelastamaan (Horowitz & Bollinger, 2014, s. 29).

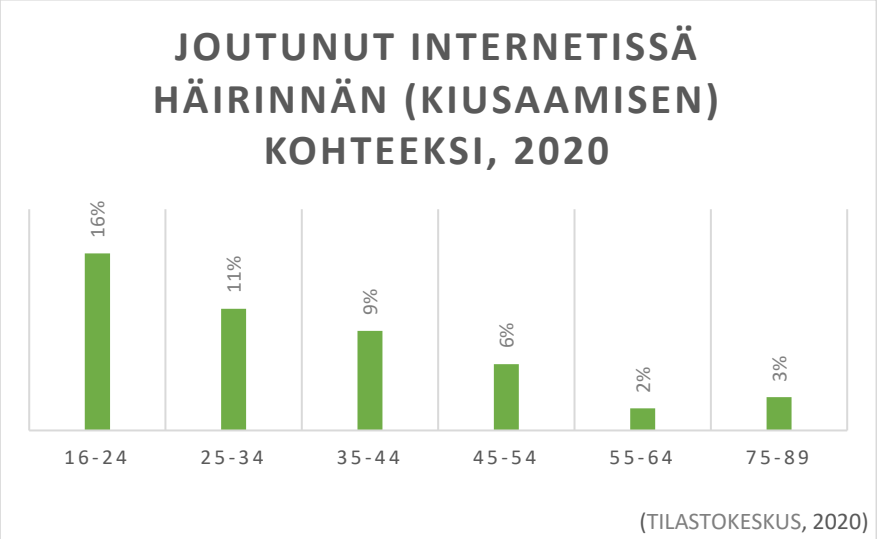
Kuten aikaisemmin mainituista esimerkkitapauksista voidaan päätellä, nettikiusaamisen seuraukset uhreilleen ovat hyvin vakavia. Kiusaaminen voi aiheuttaa sosiaalisten tilanteiden pelkoa, masennusta, ahdistusta, päihteiden väärinkäyttöä ja jopa itsemurhan. Koska nettikiusaamisesta koituu uhreille vakavia seurauksia, pitäisi myös nettikiusaajille koitua kiusaamisesta seurauksia. Horowitzin ja Bollingerin (2014) kirjan esimerkkitapauksissa vain Nafeesan kiusaajaa syytettiin rikoksesta. Kirjassa mainitaan, että suurin osa nettikiusaamistapauksista eivät etene oikeuteen asti (Horowitz & Bollinger, 2014, s. 25). Rikosuhripäivystyksen blogissa (ei pvm.) kerrotaan, että kiusaamisen rikosnimikkeitä voi olla esimerkiksi kunnianloukkaus, yksityiselämää loukkaavan tiedon levittäminen, laiton uhkaus tai identiteettivarkaus. Finlex-sivustolla (ei pvm.) on nähtävillä ajantasainen lainsäädäntö rikoslaista, jonka mukaan kunnianloukkauksesta voidaan tuomita sakkoihin ja törkeästä kunnianloukkauksesta sakkoihin tai enintään kahdeksi vuodeksi vankilaan. Sama rangaistus pätee myös yksityiselämää loukkaavan tiedon levittämiseen sekä laittomaan uhkaukseen. Identiteettivarkaudesta puolestaan voidaan tuomita sakkoihin. (*FINLEX ® - Ajantasainen lainsäädäntö*, ei pvm.) Nettikiusaamisella saattaa siis olla vakavia seurauksia myös kiusaajille, mutta vakavimpia seuraukset ovat kuitenkin uhreille.

3.3 Nettikiusaamistilastoja

Seuraavaksi käsitellään nettikiusaamiseen liittyviä tilastoja. Tilastojen avulla pyritään selvittämään, kuinka yleistä nettikiusaaminen on Suomessa ja maailmalla. Tilastoja on Suomesta, Euroopasta ja Yhdysvalloista. Nettikiusaaminen on huomattavasti tutkitumpaa nuorten ja lasten keskuudessa, joten suurin osa tilastoista käsittelee nuorten kokemaa nettikiusaamista.

Suomessa Tilastokeskus (2020) on julkaissut tilaston siitä, kuinka suuri osuus väestöstä on joutunut joskus internetissä häirinnän eli nettikiusaamisen kohteeksi (*Suomen virallinen tilasto (SVT): Väestön tieto- ja viestitätekniikan käyttö*, 2020). Taulukossa (taulukko 3) esitetään Tilastokeskuksen (2020) julkaisema tilasto.

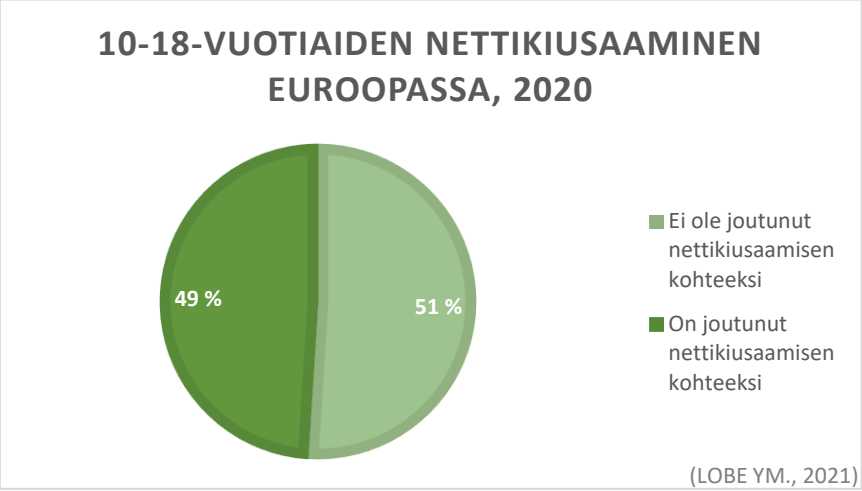
TAULUKKO 3 Suomessa internetissä tapahtuva häirintä (kiusaaminen), 2020



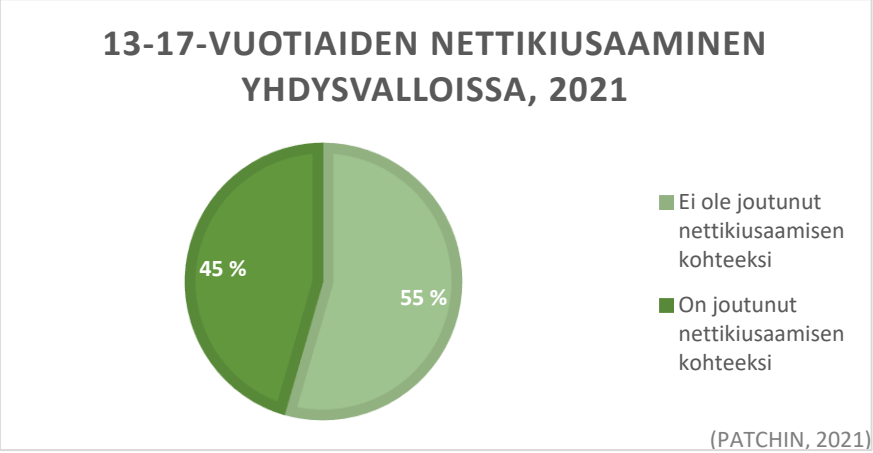
Kuten taulukosta (taulukko 3) on huomattavissa, nettikiusaaminen on Suomessa huomattavasti yleisempää nuorilla ja nuorilla aikuisilla. Mannerheimin Lastensuojeluliitosta Rahja ym. (2021) on teettänyt vuonna 2020 kattavan kyselyn nuorten nettikiusaamiseen liittyen. Kyselyyn vastasi 1123 12–17-vuotiasta nuorta, joista puolet (50 %) kertoi joutuneensa itse nettikiusaamisen kohteeksi vähintään kerran. Yli 80 %:a vastaajista kertoi, että on nähnyt verkossa muihin ihmisiin kohdistuvaa kiusaamista (Rahja ym., 2021, s. 5). Tilastoista on siis havaittavissa, että nettikiusaamista tapahtuu eniten nuorten keskuudessa.

Myös Euroopassa ja Yhdysvalloissa on tehty tutkimusta nettikiusaamisen yleisyydestä. Tutkimusta on kuitenkin tehty juuri nuorten kokemasta nettikiusaamista ja siksi tilastoja koko väestön nettikiusaamisesta on vaikea löytää. Siksi tutustummekin seuraavaksi eurooppalaisten 10–18-vuotiaden nuorien sekä 13–17- vuotiaiden yhdysvaltalaisten nuorten nettikiusaamiseen. Euroopan Komissiosta Lobe ym. (2021) on teettänyt tutkimuksen, jossa on kysytty yhdentoista Euroopan maan 10–18-vuotialta nuorilta heidän kokemuksiaan nettikiusaamisesta (Lobe ym., 2021, s. 23). Yhdysvalloissa puolestaan Patchin (2021) on tutkinut 13-17-vuotiaiden yhdysvaltalaisnuorien nettikiusaamista (J. W. Patchin, 2021). Taulukossa 4 esitetään eurooppalaisten nuorten kokemaa nettikiusaamista ja taulukossa 5 esitetään yhdysvaltalaisten nuorien kokemaa nettikiusaamista.

TAULUKKO 4 10–18-vuotiaiden nettikiusaaminen Euroopassa, 2020



TAULUKKO 5 13–17-vuotiaiden nettikiusaaminen Yhdysvalloissa, 2021



Kuten tilastoista (taulukko 4 ja taulukko 5) on huomattavissa niin eurooppalaisista kuin yhdysvaltalaisista nuorista melkein puolet ovat kokeneet nettikiusaamista jossain vaiheessa elämäänsä. Myös Mannerheimin lastensuojeluliiton (2021) teettämän kyselyn mukaan noin puolet suomalaisista nuorista olivat kokeneet nettikiusaamista (Rahja ym., 2021, s. 5). Tilastojen ja tutkimuksien mukaan on siis havaittavissa, että nettikiusaaminen on huomattavasti yleisempää nuorten keskuudessa niin Suomessa kuin globaalistikin. Tilastoista on huomattavissa, että nettikiusaaminen varsinkin nuorten keskuudessa, on liian yleistä. Jos melkein puolet nuorista ovat jossain vaiheessa elämäänsä kokeneet nettikiusaamista, voidaan päätellä, että nettikiusaaminen on liian helppoa. Sen perusteella mitä seurauksia nettikiusaamisesta on sen kokemille uhreille, voidaan päätellä nettikiusaamisen olevan vakava ongelma tämän päivän yhteiskunnassa.

4 NETTIKIUSAAMISEN AUTOMAATTINEN HAVAINNOINTI SOSIAALISESSA MEDIASSA

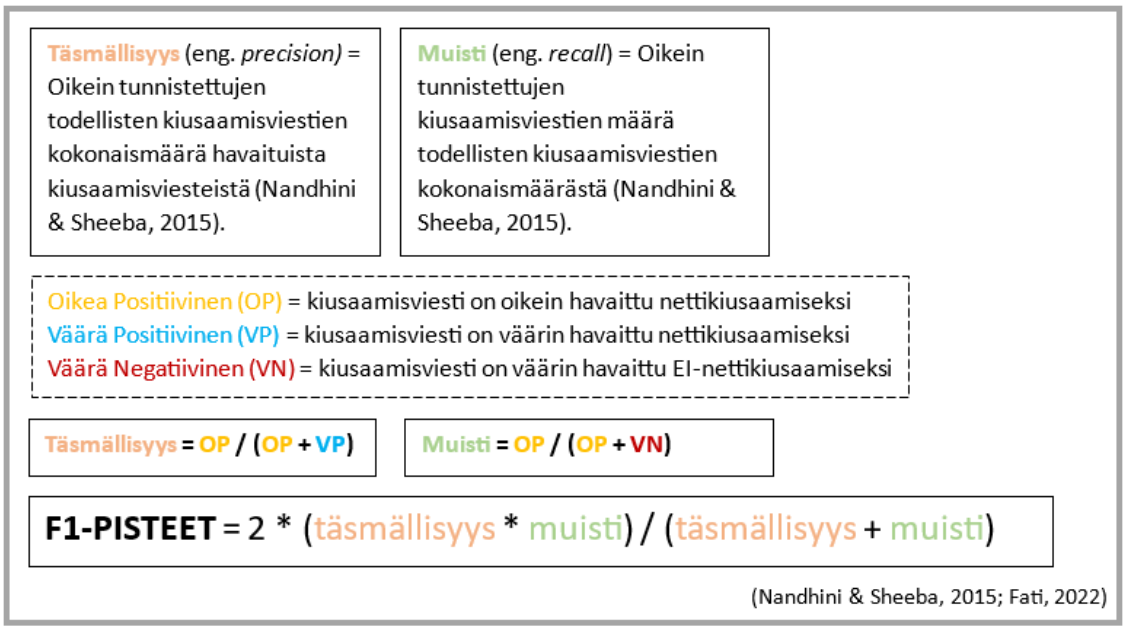
Nettikiusaaminen sosiaalisessa mediassa on suuri ongelma tämän päivän yhteiskunnassa. Keinoja nettikiusaamisen havainnointiin sekä estämiseen on yritetty luoda sekä parantaa sosiaalisen median sekä teknologian kehittyessä. Tunnetuin tapa yrittää havainnoida ja estää nettikiusaaminen on raportoiminen. Kaikissa käytetyimmissä sosiaalisen median alustoissa, kuten Facebookissa, Instagramissa, YouTubessa ja WhatsAppissa pystyy raportoimaan loukkaavasta tai häiritsevästä sisällöstä alustan ylläpitäjille (*Harassment & cyberbullying policies - YouTube Help*, ei pvm.; *Promoting Safety and Expression | Meta*, ei pvm.). Kuitenkin koko ajan pyritään kehittämään lisää keinoja nettikiusaamisen havainnointiin. Esimerkiksi Meta (ei pvm.) kertoo nettisivuillaan heidän havaitsevan suurimman osan sisällöstä, jota he joutuvat poistamaan, ennen kuin kukaan kerkeää raportoimaan kyseisestä sisällöstä. Lisäksi kerrotaan, että Instagramissa käyttäjää varoitetaan, jos hän yrittää kommentoida jotain joka saattaa olla loukkaavaa (*Promoting Safety and Expression | Meta*, ei pvm.).

Tutkimus nettikiusaamisen automaattisesta havainnoinnista on hyvin ajankohtaista ja uusia keinoja havainnointiin pyritään luomaan koko ajan. Salawu ym. (2020) käsittelevät artikkelissaan erilaisia keinoja havainnoida nettikiusaamista automaattisesti. Artikkelissa määritellään nettikiusaamisen havaitseminen kiusaamistoimien (esimerkiksi tekeytyminen toiseksi henkilöksi, henkilökohtaisten tietojen levittäminen, ahdistelu tai häirintä) tunnistamiseksi elektronisissa viestintävälineissä. Tämä havaitseminen sisältää neljä keskeistä tehtävää: yksittäisten kiusaamisviestien tunnistaminen, kiusaamistapahtuman vakavuuden laskeminen, asianomaisten henkilöiden roolien yksilöinti ja nettikiusaamistapahtuman jälkeen tapahtuvien tilanteiden luokittelu (esimerkiksi uhrin tunteiden havaitseminen kiusaamisviestin vastaanottamisen jälkeen). Näistä neljästä keskeisestä tehtävästä yksittäisten kiusaamisviestin tunnistaminen on tärkein (Salawu ym., 2020, s. 2). Nettikiusaamisen automaattisessa havainnoinnissa tärkeintä siis on, että kiusaamisviesti tai -julkaisu pystytään havaitsemaan ja täten se voidaan esimerkiksi poistaa. Nettikiusaamisen automaattisen havainnoinnin tutkimuksessa on siis tarkoituksena luoda keino, jolla nettikiusaamista voidaan havainnoida

automaattisesti erilaisia teknologioita hyödyntäen. Havainnointiin ei siis tarvita ihmistä vaan tietty teknologia hoitaa havainnoinnin automaattisesti. Useissa lähteissä puhutaan pelkästään nettikiusaamisen havainnoinnista ja sana ”automaattinen” on jätetty pois. Kuitenkin kaikki tässä tutkielmassa esiteltävät tutkimukset tutkivat juuri nettikiusaamisen automaattista havainnointia.

Salawu, He ja Lumsden (2020) jakavat artikkelissaan keinoja nettikiusaamisen automaattiseen havainnointiin neljään eri luokkaan niiden lähestymistapojen mukaisesti. Nämä neljä eri luokkaa ovat vapaasti suomennettuna: ohjatun oppimisen lähestymistapa (eng. *supervised learning approach*), sanastopohjainen lähestymistapa (eng. *lexicon-based approach*), sääntöpohjainen lähestymistapa (eng. *rule-based approach*) ja seka-aloitteellinen lähestymistapa (eng. *mixed-initiative approach*). Ohjatun oppimisen lähestymistavassa käytetään tyypillisesti esimerkiksi erilaisia koneoppimisluokittelijoita, kehittämään ennustavia malleja nettikiusaamisen havaitsemiseksi. Sanastopohjainen lähestymistapa hyödyntää sanalistoja ja käyttää näiden sanojen merkittävyyttä havaitakseen nettikiusaamista tekstistä. Sääntöpohjainen lähestymistapa yhdistää tekstiä ennalta määriteltyihin sääntöihin havaitakseen tekstistä nettikiusaamista. Seka-aloitteellinen lähestymistapa yhdistää ihmislähtöisen päättelyn yhteen tai useampaan yllä mainituista lähestymistavoista (Salawu ym., 2020, s. 3). Näistä neljästä lähestymistavasta eniten tutkittu on ohjattu oppiminen. Muista lähestymistavoista on vaikea löytää ajankohtaista ja luotettavaa tutkimusta. Jotta tässä tutkielmassa voidaan perehtyä mahdollisimman moniin ajankohtaisiin ja luotettaviin tutkimuksiin, keskitytään tässä tutkielmassa ainoastaan ohjatun oppimisen lähestymistavan tutkimuksiin.

Seuraavaksi perehdytään tutkimuksiin ohjatun oppimisen lähestymistavasta. Esimerkkitutkimuksissa esitellään aina sen tekijöiden nettikiusaamisen automaattiseen havainnointiin luoma havaitsemisjärjestelmä tai vastaava ehdotus. Näiden järjestelmien teknologiseen toteutukseen ei kuitenkaan perehdytä kovinkaan syvällisesti vaan järjestelmästä kerrotaan yleisesti. Mikäli tutkimuksessa on tehty testejä, joilla on testattu järjestelmän toimivuutta, myös näiden testien tulokset tuodaan ilmi. Testeissä järjestelmän toimivuutta pyritään testaamaan usein eri parametrien pisteitä laskemalla. Parametreista tärkeimpänä pidetään usein F1-pisteitä (eng. *F1-score, F-1 measure, F-measure*). F1-pisteet tarkoittavat täsmällisyys- ja muistipisteiden harmonista keskiarvoa. Kuviossa (kuvio 1) esitetään täsmällisyyspisteiden, muistipisteiden ja F1-pisteiden kaavat.



KUVIO 1: F1-pisteiden laskukaava

F1-pisteet voidaan esittää niin pelkkinä pisteinä kuin prosentteina ja tässä tutkielmassa ne esitetään prosentteina riippumatta tutkimusartikkelin esitystavasta. Tutkielmassa vertaillaan eri havaitsemisjärjestelmien toimivuutta F1-pisteitä vertailemalla. Koska tutkielman aiheena on nettikiusaaminen sosiaalisessa mediassa, on tutkimukset valittu siten, että suurimmassa osassa testauksissa käytetään tietoaineistona juuri sosiaalisesta mediasta kerättyjä tekstipohjaisia kommentteja ja julkaisuja.

4.1 Ohjatun oppimisen lähestymistapa nettikiusaamisen automaattiseen havainnointiin

Seuraavaksi käsitellään eri ohjatun oppimisen lähestymistavan havaitsemisjärjestelmiä. Kuten mainittu, ohjatun oppimisen lähestymistavassa käytetään apuna erilaisia luokittelijoita. Oman päätelmäni mukaan juuri ohjatun oppimisen lähestymistavan havaitsemisjärjestelmät ovat kaikista kehittyneimpiä ja niissä käytetään apuna moderneja keinoja. Seuraavaksi esitellään yksitoista eri tutkimusartikkelia, joissa on luotu havaitsemisjärjestelmä ohjatun oppimisen lähestymistavan mukaisesti. Välttääksemme vanhentunutta ja täten epäpätevää tutkimustietoa, käytetään tutkimuksia, jotka on tehty 2010- ja 2020-luvulla. Tutkimukset esitellään aikajärjestyksessä alkaen vuodesta 2015 ja päättyen vuoteen 2022. Näin päästään tutkimaan myös havaitsemiskeinojen kehitystä vajaan kymmenen vuoden ajalta.

Vuonna 2015 Galán-GarcÌa ym. (2015) tutkivat, miten vale- tai trolliprofiileja voidaan havaita Twitter nimisestä sosiaalisen median alustasta. Artikkelin kirjoittajien luoma havaitsemisjärjestelmä perustuu siihen ajatukseen, että

jokaista vale-/trolliprofiilia seuraa kyseisen profiilin luojan oikea henkilökohtainen käyttäjä. Havaitsemisjärjestelmä luokittelee profiilit julkaisujen kirjoittamistyylien mukaisesti eri koneoppimisluokittelijoita avuksi käyttäen. Tutkimuksessa tutkittiin mitkä koneoppimisluokittelijat toivat parhaan tuloksen ja näitä olivat SMO-PolyKernel, J48, SMO-NormalizedPolyKernel ja RandomForest, joista SMO-PolyKernel tuotti parhaan tuloksen täsmällisyyden ollessa 68.47. Tätä havaitsemisjärjestelmää on testattu myös oikeassa tosielämän tilanteessa, jossa espanjalaisessa koulussa oppilaita nettikiusattiin Twitterissä. Twitteriin oli luotu profiili, joka levitti erilaisia juoruja koulun oppilaista. Tätä havaitsemisjärjestelmää avuksi käyttäen pystyttiin selvittämään kuka tai ketkä kyseistä profiilia ylläpiti. Onnistuneesta tosielämän testauksesta huolimatta kirjoittavat tuovat esiin havaitsemisjärjestelmän haasteen olevan se, että se perustuu ajatukselle, jonka mukaan oikea ihminen seuraa itse luomaansa vale-/trolliprofiilia. Mikäli näin ei kuitenkaan ole, tällöin havaitsemisjärjestelmä ei pysty löytämään profiilin ylläpitäjää (Galán-GarcÌa ym., 2015). Tämä tutkimus ja siinä luotu havaitsemisjärjestelmä on yksi varhaisimpia havaitsemisjärjestelmiä, joissa käytetään apuna erilaisia koneoppimisluokittelijoita. Tässä tutkimuksessa ei niinkään pyritä havaitsemaan kiusaamisviestejä, vaan pyritään löytämään kiusaaja ja täten pysäyttämään nettikiusaaminen. Tämän vuoksi vertaaminen tuleviin havaitsemisjärjestelmiin on myös hankalaa. On kuitenkin mainittavaa, että tämä havaitsemisjärjestelmä on tuottanut tulosta tosielämän testauksessa ja se kertoo hyvin havaitsemisjärjestelmän pätevyydestä. Tämä havaitsemisjärjestelmän toimintaa voisi laajentaa toimimaan myös muissa sosiaalisen median palveluissa kuin vain Twitterissä.

Myös vuonna 2015 Nandhini ja Sheeba (2015) tekivät hieman erilaista tutkimusta liittyen nettikiusaamisen automaattiseen havainnointiin. Heidän havaitsemisjärjestelmänsä tunnistaa nettikiusaamisen termien esiintymisen. Lisäksi järjestelmä pystyy luokittelemaan nettikiusaamistoimintaa, kuten häirintää ja rasismia, sosiaalisessa mediassa. Tässä havaitsemisjärjestelmässä on useita vaiheita ja se etenee seuraavasti: ensin tietoaineisto eli teksti esikäsitellään ja siitä erotellaan erilaisia ominaisuuksia, kuten substantiivit, adjektiivit ja pronominit. Tämän jälkeen nämä erotellut ominaisuudet käsitellään FuzGen-oppimisalgoritmin sekä Naïve Bayes koneoppimisluokittelijan avulla. Näin saadaan tietoaineiston tekstistä tunnistettua sekä luokiteltua nettikiusaamiseen liittyviä sanoja. Havaitsemisjärjestelmää testattiin käyttämällä tietoaineistoa MySpace sekä Fromspring.me nimisistä sosiaalisen median alustoista. Testeissä mitattiin havaitsemisjärjestelmän F1-pisteitä, joka kertoo hyvin järjestelmän virheettömyydestä. MySpace-testissä F1-pisteiksi saatiin 91 %. Fromspring.me-testissä puolestaan F1-pisteet olivat 92 % (Nandhini & Sheeba, 2015). Tulokset tässä tutkimuksessa olivat hyvin vakuuttavia korkeiden F1-pisteiden vuoksi. Tässä tutkimuksessa luotua havaitsemisjärjestelmää voidaan siis pitää hyvin luotettavana.

Muutama vuosi myöhemmin, vuonna 2017, Zhao ja Mao (2017) tutkivat uusia keinoja havaita nettikiusaamista sosiaalisesta mediasta. He kehittivät smSDA:n, joka on nettikiusaamisen havaitsemiseen erikoistunut oppimismalli. Tätä oppimismallia testattiin Twitter ja MySpace sosiaalisen median alustoissa ja

sen toimintaa verrattiin seitsemään muuhun nettikiusaamisen havainnointiin tarkoitettuun järjestelmään. Niin Twitter-testissä kuin MySpace-testissä smSDA pärjäsi parhaiten muihin järjestelmiin verrattuna ja F1-pisteiksi se sai Twitter-testissä 71,9 % ja MySpace-testissä 77,6 % (Zhao & Mao, 2017). Tämä oppimismalliksi kutsuttu havaitsemisjärjestelmä on hyvin monimutkainen, mutta se kuitenkin sai lupaavia tuloksia ja tuotti parhaimman tuloksen verrattaessa muihin järjestelmiin. Kuitenkin Nandhinin ja Sheeban (2015) havaitsemisjärjestelmä sai huomattavasti paremmat F1-pisteet omassa testissään ja onkin mielenkiintoista, ettei heidän luomaa havainnointijärjestelmää ole otettu mukaan Zhaon ja Maon (2017) testiin, jossa he testasivat muita havaitsemisjärjestelmiä ja vertasivat niitä heidän järjestelmäänsä.

Vuonna 2018 Van Hee ym. (2018) huomasivat, että nettikiusaamistutkimus on keskittynyt enimmäkseen nettikiusaamisen ”hyökkäyksien” havaitsemiseen. Tässä tutkimuksessa nettikiusaamisen automaattista havainnointia lähdettiin tutkimaan erilaisesta näkökulmasta. Tutkimuksessa haluttiin luoda havaitsemisjärjestelmä, joka pystyisi havaitsemaan kiusaamista ennen kuin sitä pääsisi tapahtumaan. Tämän vuoksi luotiin havaitsemisjärjestelmä, jonka ideana oli havaita automaattisesti nettikiusaamisen signaalit sosiaalisessa mediassa. Tutkimuksessa tehtiin myös joukko kokeita, joilla tutkittiin havaitsemisjärjestelmän toteutettavuutta sosiaalisessa mediassa. Lisäksi tutkittiin, mitkä tietolähteet, kuten subjektiiviset sanastot ja termilistat, edistävät havaitsemista parhaiten. Kokeita tehtiin englannin sekä hollannin kielisistä aineistoista. Kokeiden jälkeen parhaat F1-tulokset olivat englannin kielisestä aineistosta 64,32 % ja hollannin kielisestä 58,72 % (Van Hee ym., 2018). Tämän tutkimuksen tulokset eivät ensilukemalta vaikuta kovinkaan vakuuttavilta verrattaen esimerkiksi Nandhinin ja Sheeban (2015) tutkimukseen. Tämän tutkimuksen tuloksia ei kuitenkaan voi suoraan verrata aikaisemmin esitettyihin tutkimuksiin, sillä ne perustuivat juuri nettikiusaamisen ”hyökkäysten” havainnointiin. Tämä tutkimus puolestaan pyrki luomaan havaitsemisjärjestelmän, joka havaitsee kiusaamisen jo ennen ”hyökkäystä”. Tulokset ovat vakuuttavia, vaikkakin F1-pisteet ovat vielä hieman alhaisia. Tämä uusi näkökulma kiusaamisen havainnointiin sosiaalisessa mediassa voisi toimiessaan olla hyvinkin hyödyllinen, sillä järjestelmän avulla nettikiusaamista pystyttäisiin hyvin ennakoimaan ja täten mahdollisesti pysäyttämään.

Vuonna 2020 Muneer ja Fati (2020) testasivat seitsemää eri koneoppimisluokittelijaa löytääkseen parhaan luokittelijan heidän järjestelmäänsä nettikiusaamisen havainnointia varten. Nämä seitsemän luokittelijaa olivat Logistic Regression (LR), Light Gradient Boosting Machine (LGBM), Stochastic Gradient Descent (SGD), Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB) ja Support Vector Machine (SVM). Näistä luokittelijoista Random Forestia käytettiin jo Galán-GarcÌa ym. (2015) tutkimuksessa ja myös Naive Bayesia on käytetty Nandhinin ja Sheeban (2015) tutkimuksessa. Tässä tutkimuksessa käytettiin tietoaineistona Twitterin julkaisuja. Kuten Nandhinin ja Sheeban (2015) havaitsemisjärjestelmässä tietoaineiston prosessointi aloitettiin tekstin esikäsittelyllä ja siitä eroteltiin erilaisia ominaisuuksia. Tämän jälkeen aineistoa käsiteltiin eri

koneoppimisluokittelijoiden avulla. Luokittelijoista laskettiin F1-pisteet ja seitsemästä luokittelijasta paras luokittelija oli Logistic Regression, joka saavutti 92,80 % F1-pisteet. Lisäksi kokeiden aikana havaittiin, että Logistic Regression toimii paremmin aineiston koon kasvaessa (Muneer & Fati, 2020). Aikaisempiin tutkimuksiin verrattuna Muneerin ja Fatin (2020) tutkimuksessa saatiin ehdottomasti parhaita F1-pisteitä. Vaikka tutkimus olikin vertailua erilaisten koneoppimisluokittelijoiden välillä, oli tutkimuksessa kuitenkin luotu myös uusia keinoja havaitsemisjärjestelmän parantamiseksi. Muun muassa tutkimuksen tekijät olivat käyttäneet uutta teknologiaa tietoaineiston ominaisuuksien erotteluvaiheeseen. Tähän mennessä esitetyistä samantyylisistä havaitsemisjärjestelmistä, tässä järjestelmässä on eniten potentiaalia nettikiusaamisen automaattiseen havaitsemiseen, kun käytetään apuna Logistic Regression koneoppimisluokittelijaa.

Myös vuonna 2020 AlHarbi ym. (2020) tutkivat erilaisia koneoppimisluokittelijoita, mutta englannin kielen sijaan he tutkivat niiden toimivuutta arabian kielessä. Tutkimuksessa tutkittiin seuraavien luokittelijoiden toimivuutta: Ridge Regression (RR), Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), K-nearest neighbors (KNN), Deep neural network (DNN) ja C5.0. Tutkimuksen tekijät testasivat näitä luokittelijoita semanttisessa analyysissä arabian kielisessä tekstissä ja tutkimus toteutettiin analyysitutkimuksena. Tutkimuksen tekijät päätyivät Ridge Regression ja Logistic Regression luokittelijoihin, sillä ne saavuttivat korkeimman täsmällisyyden nettikiusaamisen automaattisessa havainnoinnissa arabian kielestä (AlHarbi ym., 2020). Tämä tutkimus poikkeaa paljolti muista tutkimuksista, sillä tässä tutkimuksessa tutkimuksen tekijät eivät vielä luoneet havaitsemisjärjestelmää vaan tutkivat mikä luokittelijoista olisi paras havaitsemisjärjestelmään. Jo Muneerin ja Fatin (2020) tutkimuksesta tuttu Logistic Regression tuotti hyvän tuloksen myös tässä tutkimuksessa. Tämä tutkimus oli aikaisempiin tutkimuksiin verrattuna hieman vajavainen siksi, että se oli toteutettu vain analyysinä. Lisäksi luokittelijoiden täsmällisyyspisteitä oli ainoastaan verrattu, eikä ollut otettu huomioon esimerkiksi F1-pisteitä.

Vuoden 2020 loppupuolella Talpur ja O'Sullivan (2020) tekivät tutkimusta nettikiusaamisen automaattisesta havainnoinnista ja loivat uuden havaitsemisjärjestelmän. Jo Nandhinin ja Sheeban (2015) sekä Muneerin ja Fatin (2020) tutkimuksista tuttujen havaitsemisjärjestelmien mukaisesti tämäkin järjestelmä aloittaa toiminnan tietoaineiston käsittelyllä. Tietoaineistona tässä tutkimuksessa toimii Twitterin tekstijulkaisut. Aluksi tietoaineiston tekstiä esikäsitellään ja tämän jälkeen tekstin ominaisuuksia, kuten substantiiveja, verbejä ja adjektiiveja, eritellään. Näiden tuttujen vaiheiden jälkeen tietoaineistosta eriteltyjä sanoja luokitellaan sen perusteella ovatko ne kiusaamiseen liittyviä sanoja vai ei. Lisäksi tietoaineistoa eritellään ja luokitellaan vielä muidenkin piirteiden mukaisesti. Tämän jälkeen tutkimuksen tekijät vertasivat viiden eri luokittelijan toimivuutta tässä havaitsemisjärjestelmässä. Nämä luokittelijat olivat Naïve Bayes, K-nearest neighbors (KNN), Decision Tree, Random Forest, ja Support Vector Machine. Näistä luokittelijoista kaikkia muita paitsi Decision Treetä on kokeiltu aikaisemmissa tutkimuksissa. Näistä viidestä luokittelijasta parhaaksi osoittautui Random Forest, jonka F1-pisteet olivat 89,8 % (Talpur & O'Sullivan, 2020). On

yllättävää huomata, että tässä havaitsemisjärjestelmässä Random Forest toimi parhaiten, kun aikaisemmissa tutkimuksissa Random Forestin toimivuus on ollut hyvin alhaisella tasolla muihin luokittelijoihin verrattuna. Kuitenkin on huomattavaa, että tässä tutkimuksessa havaitsemisjärjestelmää oli paranneltu, esimerkiksi tietoaineiston käsittelyä ja prosessointia oli muutettu. Kuitenkaan tämäkään havainnointijärjestelmä ei saavuta parhaita mahdollisia F1-pisteitä verrattuna aikaisempiin samankaltaisiin tutkimuksiin. Mielenkiintoista on myös, että Muneerin ja Fatin (2020) tutkimuksessa Random Forest-luokittelijan F1-pisteet olivat 92,8 % eli huomattavasti suuremmat kuin tässä tutkimuksessa.

Vuonna 2021 Bharti ym. (2021) loivat oman havaitsemisjärjestelmänsä. Heidän havaitsemisjärjestelmänsä toiminta alkaa myös tietoaineiston prosessoinnilla. Myös tässä tutkimuksessa tietoaineistona käytettiin Twitterin tekstijulkaisuja. Tutkijat vertailivat jo tutuksi tulleita koneoppimisluokittelijoita keskenään, kuten Naive Bayesia, Random Forestia ja Logistic Regressionia. Näistä tutuista luokittelijoista parhaaksi todettiin Logistic Regression, joka sai 94,09 % F1-pisteet. Kuitenkin tässä tutkimuksessa testattiin myös syväoppimismallia nettikiusaamisen automaattisen havaitsemisen avuksi. Näitä malleja olivat esimerkiksi One hot encoding, GloVe 6, GloVe 42 ja GloVe 840. Näistä parhaaksi arvioitiin GloVe 840, joka saavutti 94,20 %:n F1-pisteet. Täten GloVe 840 todettiin parhaaksi malliksi tähän havaitsemisjärjestelmään (Bharti ym., 2021). Tässä tutkimuksessa otettiin uusi lähestymistapa nettikiusaamisen automaattiseen havainnointiin ja tutkittiin koneoppimisluokittelijoiden lisäksi myös syväoppimismalleja. Tähän havaitsemisjärjestelmään sopikin parhaiten juuri syväoppimismalli ja saatiin suurimmat F1-pisteet verrattuna aikaisempiin tutkimuksiin ja havaitsemisjärjestelmiin. Kuitenkin on huomattavaa, että myös havaitsemisjärjestelmä toimii luokittelijasta tai mallista huolimatta paremmin, sillä se saavutti myös hyvin korkeat F1-pisteet Logistic Regressionia avuksi käyttäen.

Vuonna 2022 Trandabăţ, Gifu ja Adrian (2022) halusivat kehitellä kieliaineiston eli korpuksen romanian kieliselle loukkaavalle sisällölle. Lisäksi he halusivat testata useita koneoppimisluokittelijoita löytääkseen parhaan lähestymistavan romanian kielisen nettikiusaamisen havainnointiin. Artikkelin tekijät loivat kieliaineiston jo olemassa olevan kieliaineiston pohjalta ja lisäsivät siihen aineistoa YouTuben, Facebookin ja Twitch-striimauspalvelun kommenteista. Koneoppimisluokittelijoita lähdettiin testaamaan tämän aineiston pohjalta. Tuttuun tapaan aineisto ensin esikäsiteltiin, jonka jälkeen tekstiominaisuuksia eroteltiin. Artikkelin tekijät testasivat jo tutuksi tulleita luokittelijoita, kuten Naive Bayes, AdaBoost, DecisionTree, SVM, Logistic Regression ja Random Forest luokittelijoita. Kirjoittajat testasivat myös Passive Aggressive -luokittelijaa, jota ei ole testattu aikaisemmin esitellyissä tutkimuksissa. Tulokset osoittavat, että SVM (*support vector machine*), Naive Bayes ja Passive Aggressive -luokittelijoiden yhdistelmä voi auttaa tunnistamaan ja luokittelemaan loukkaavaa sisältöä sosiaalisessa mediassa romanian kielellä. Tälle luokittelijayhdistelmälle saatiin F1-pisteiksi 92,09 % (Trandabăţ ym., 2022). Tässä tutkimuksessa on hieman erilainen tulos, sillä aikaisemmissa koneoppimisluokittelutestauksissa ei ole käytetty

luokittelijoiden yhdistelmiä. Luokittelijoiden yhdistelmä saikin hyvät ja lupaavat F1-pisteet, muttei siltikään päihitä esimerkiksi Bharti ym. (2020) syväoppimismallia apunaan käyttävää järjestelmää tai Muneerin ja Fatin (2020) havaitsemisjärjestelmää. On kuitenkin muistettava, että tässä tutkimuksessa kyse oli romanian kielisestä nettikiusaamisen havainnoinnista, jota ei voi täysin verrata englannin kielisen nettikiusaamisen havainnoinnin kanssa.

Vuonna 2022 Mangaonkar ym. (2022) lähtivät parantelemaan aikaisemmin luomaansa havaitsemisjärjestelmää. Tekijät kuvaavat havaitsemisjärjestelmäänsä yhteiskaavaksi nettikiusaamisen havaitsemiseen. Järjestelmä käyttää erilaisia hajautettuja yhteismalleja ja sen kattavaa arviointia. Tätä järjestelmää on testattu Twitterin tekstijulkaisujen avulla. Erilaiset kokeet osoittavat, että yhteismallit toimivat paremmin kuin itsenäinen lähestymistapa ja ne toimivat hyvin myös epäonnistumisten sattuessa. Tutkimuksessa vahvistettiin myös, että pilvipohjainen havaitseminen päihittää muut kaikissa kokeiluissa (Mangaonkar ym., 2022). Tässä tutkimuksessa testattu havaitsemisjärjestelmä on täysin uudenlainen ja hyvin monimutkainen verrattuna aikaisempiin järjestelmiin. Tätä järjestelmää on vaikea alkaa vertailemaan muiden järjestelmien kanssa, mutta Mangaonkarin ym. (2022) mukaan järjestelmä kuitenkin päihittää monet muut havaitsemisjärjestelmät. Kokeissa ei laskettu F1-pisteitä, joten senkään puolesta ei voida lähetä vertaamaan tätä järjestelmää muihin.

Viimeisenä esitellään vuonna 2022 tehty tutkimus. Fati (2022) esittelee artikkelissaan havaitsemisjärjestelmän, jolla voidaan havaita nettikiusaamista Twitteristä. Tuttuun tapaan tekstiaineisto esikäsitellään sekä tekstin ominaisuudet eritellään. Tämän jälkeen järjestelmä luokittelee aineiston sekä asiantuntijoilta että potentiaalisilta uhreilta poimittujen avainsanojen perusteella käyttämällä syväoppimismallia, joka perustuu tunneanalyysiin (eng. *sentiment analysis*). Tämän järjestelmän testauksessa F1-psiteiksi laskettiin 89 % (Fati, 2022). Myös tämä tutkimus seuraa monista tutkimuksista tuttua kaavaa, kun havaitsemisjärjestelmä prosessoi tekstiä ensin erilaisin keinoin. Tämän jälkeen kuitenkin käytetään tunneanalyysiä, jota ei ole käytetty aikaisemmin esitellyissä tutkimuksissa. Tutkimuksessa on käytössä uusi mielenkiintoinen näkökulma nettikiusaamisen havaitsemiseen, kun käytetään apuna asiantuntijoilta ja potentiaalisilta uhreilta poimittuja avainsanoja syväoppimismallissa. Tulokset ovat vakuuttavia, mutta eivät kuitenkaan pääse parhaimmistoon tässä tutkielmassa esiteltyihin muihin tutkimuksiin verrattaessa.

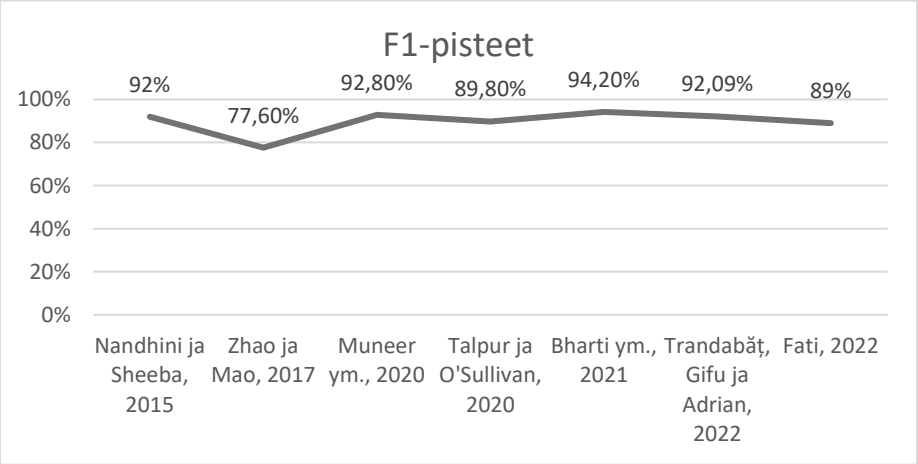
4.2 Havaitsemisjärjestelmien vertailu

Tutkimuksia ja erilaisia havaitsemisjärjestelmiä toisiinsa verrattaessa on huomattavaa, että viimeisen vajaan kymmenen vuoden aikana on pyritty tutkimaan ja löytämään monia erilaisia keinoja nettikiusaamisen automaattiseen

havainnointiin erityisesti koneoppimista ja syväoppimista avuksi käyttäen. On huomattavaa, että useissa tutkimuksissa tärkeimpänä oli löytää tietty koneoppimisluokittelija tai ryhmä luokittelijoita, jotka pystyivät havaitsemaan nettikiusaamista mahdollisimman hyvin. Useissa havaitsemisjärjestelmissä oli yleistetysti samanlainen alkuvaihe, jossa tietoaineistoa prosessoitiin. Kuitenkin jokaisessa tutkimuksessa tietoaineistoa prosessoitiin hieman eri keinoin ja erilaisia teknologioita hyödyntäen. Puolestaan toisissa tutkimuksissa lähdettiin tekemään täysin erilaista havaitsemisjärjestelmää uudesta näkökulmasta ja yhdessä tutkimuksessa huomattiin myös ongelma nettikiusaamisen varhaishavainnoinnissa.

Vaikka kehitystä viimeisen vajaan kymmenen vuoden aikana nettikiusaamisen automaattisen havainnoinnin tutkimuksessa on tapahtunut, on kuitenkin mielenkiintoista huomata, että F1-pisteet eivät ole erityisesti kehittyneet suuntaan tai toiseen vuosien aikana. Taulukossa (taulukko 6) on esitelty F1-pisteiden kehitys niiden havaitsemisjärjestelmien osalta, joista oli kokeellisesti F1-pisteet laskettu ja joiden pääsääntöisenä tarkoituksena on pyrkiä havaitsemaan kiusaamisviestejä.

TAULUKKO 6 F1-pisteiden kehitys



Jos lähdetään vertailemaan havaitsemisjärjestelmiä niiden saavuttamia F1-pisteitä vertaamalla, niin Bharti ym. (2021) tekemä havaitsemisjärjestelmä saavutti parhaat pisteet. Tästä voitaisiin siis päätellä, että syväoppimismalleja hyödyntämällä voidaan kehittää parempaa ja tarkempaa nettikiusaamisen automaattista havainnointia. Kuitenkin on tärkeää huomata, että samoilla luokittelijoilla tai malleilla voidaan saada täysin erilaisia tuloksia, jos havaitsemisjärjestelmä on muuten erilainen. Esimerkiksi Talpur ja O'Sullivan (2020) kehittivät havaitsemisjärjestelmän, jonka parhaaksi koneoppimisluokittelijaksi todettiin Random Forest, joka saavutti 89.9 % F1-pisteet. Muneerin ja Fatin (2020) havaitsemisjärjestelmän tutkimuksessa Random Forest-luokittelijan F1-pisteet oli 92.8 %, eikä tätä edes todettu parhaimmaksi luokittelijaksi. Tärkeintä on siis kehittää havaitsemisjärjestelmää monipuolisesti ja kokonaisvaltaisesti aina tietoaineiston

alkuprosessoinnista lähtien, eikä keskittyä pelkästään luokittelijoiden tai mallien vertailuun.

5 YHTEENVETO

Sosiaalisen median käyttö on lisääntynyt hurjasti viimeisten vuosien aikana. Alustojen ja käyttäjien määrä on kasvanut merkittävästi 2000-luvun alusta alkaen. Sosiaalisen median yleisyydestä kertoo hyvin se, että sosiaalisen median suosituimpia alustoja käyttää miljardit ihmiset ympäri maailmaa. Sosiaalisella medialla on paljon erilaisia käyttötarkoituksia, kuten yhteydenpitäminen ystäviin ja sukulaisiin sekä ammatillinen verkostoituminen. Näiden monipuolisten käyttötarkoituksien ansiosta sosiaalinen media luo käyttäjilleen monenlaisia hyötyjä. Kuitenkin hyötyjen lisäksi sosiaalinen media tuo mukanaan myös haittoja, joista yksi suurimmista on nettikiusaaminen. Nettikiusaaminen tarkoittaa kiusaamista, joka tapahtuu digitaalista teknologiaa avuksi käyttäen. Nettikiusaaminen on laaja ongelma ja sitä kohtaavat sosiaalisessa mediassa erityisesti lapset ja nuoret.

Tämän tutkielman tavoitteena oli kirjallisuuskatsauksena selvittää, millaisin erilaisin keinoin nettikiusaamista voidaan automaattisesti havainnoida sosiaalisessa mediassa. Tutkielmassa perehdyttiin ensin sosiaalisen median määritelmään sekä käytiin läpi sosiaalisen median hyötyjä ja haittoja. Lisäksi esiteltiin käytetyimmät sosiaalisen median alustat. Tämän jälkeen perehdyttiin nettikiusaamisen määritelmään sekä käytiin läpi nettikiusaamistapauksia ja nettikiusaamistilastoja. Termien määrittelyn jälkeen pyrittiin vastaamaan tutkimuskysymykseen: millä tavoin nettikiusaamista voidaan automaattisesti havainnoida sosiaalisessa mediassa? Tutkimuskysymykseen pyrittiin vastaamaan tämän tutkielman neljännessä luvussa, jossa esiteltiin erilaisia tutkimuksia nettikiusaamisen automaattiseen havainnointiin liittyen. Tällä tavoin pyrittiin löytämään erilaisia keinoja nettikiusaamisen automaattiseen havainnointiin sosiaalisessa mediassa.

Tutkielmassa todettiin, että erilaisia keinoja nettikiusaamisen automaattiseen havainnointiin on 2010- ja 2020-luvulla kehitetty paljon. Näissä keinoissa käytetään usein apuna muun muassa koneoppimista ja syväoppimista. Viimeisen vajaan kymmenen vuoden aikana on kehitetty monia erilaisia havaitsemisjärjestelmiä, joiden pääsääntöisenä tarkoituksena on pyrkiä havaitsemaan sosiaalisessa mediassa kiusaamisviestejä. Lisäksi on kehitelty järjestelmiä muun muassa nettikiusaamisen varhaiseen havaitsemiseen sekä nettikiusaamisprofiilien

tekijöiden löytämiseen. Kiusaamisviestien havaitsemiseen keskittyvissä havaitsemisjärjestelmissä käytäntö oli usein samanlainen verrattuna esimerkiksi nettikiusaamisen varhaishavainnoinnin järjestelmiin. Kiusaamisviestien havaitsemisen järjestelmissä käytettiin usein joko koneoppimisluokittelijaa tai syväoppimismallia. Jokaisessa havaitsemisjärjestelmässä oli kuitenkin käytössä erilaisia teknologiota esimerkiksi aineiston prosessoinnissa. Tutkielmassa esitetyistä tutkimusartikkeleista käsiteltiin kuitenkin myös sellaisia tutkimuksia, joissa havaitsemisjärjestelmä oli rakennettu täysin erilaisin keinoin eikä esimerkiksi koneoppimismalleja ollut hyödynnetty samalla tavoin kuin muissa.

Keinoja nettikiusaamisen havaitsemiseen pyrittiin myös vertailemaan keskenään. Tutkielmassa päädyttiin vertailemaan keinoja toisiinsa F1-pisteiden avulla, mitkä oli useassa tutkimuksessa laskettu järjestelmää testattaessa. Monessa tutkimuksessa F1-pisteiden avulla pyrittiin kertomaan järjestelmän toimivuudesta. Tässä tutkielmassa esitettyjen tutkimusten pohjalta huomattiin, että nämä F1-pisteet eivät juurikaan ole nousseet suuntaan tai toiseen viimeisen vajaan kymmenen vuoden aikana. Tässä tutkielmassa esitetyistä tutkimuksien havaitsemisjärjestelmistä parhaat F1-pisteet sai Bhartin ym. (2021) tekemä järjestelmä, jossa käytettiin syväoppimismallia avuksi nettikiusaamisen automaattiseen havainnointiin. Bhartin ym. (2021) tutkimus oli ainut, jossa käytettiin apuna kyseistä syväoppimismallia. Tästä voidaankin päätellä, että tulevaisuudessa olisi hyvä tutustua enemmän syväoppimismallien käyttöön nettikiusaamisen automaattisessa havainnoinnissa.

Huomattavaa kuitenkin keinoja toisiinsa verrattaessa oli myös se, että samat koneoppimisluokittelijat saivat eri havaitsemisjärjestelmien testauksissa hyvin erilaisia F1-pisteitä. Esimerkiksi Talpurin ja O'Sullivanin (2020) tutkimuksessa Random Forest -luokittelija saavutti 89,9 % F1-pisteet ja puolestaan Muneerin ja Fatin (2020) tutkimuksessa sama luokittelija sai 92,8 % F1-pisteet. On siis tärkeää muistaa, että pelkällä koneoppimisluokittelijalla tai syväoppimismallilla ei ratkaista havaitsemisjärjestelmän toimivuutta vaan järjestelmän toimivuuteen vaikuttaa myös muut asiat, kuten esimerkiksi aineiston käsittely ja prosessointi. Ratkaisevaa on, että havaitsemisjärjestelmiä pyritään kehittämään monipuolisesti ja kokonaisvaltaisesti parhaimman mahdollisen järjestelmän luomiseksi. Tässä vaiheessa on myös oleellista muistaa, että tutkimuksia, joissa F1-pisteitä ei ollut laskettu, on vaikeaa vertailla muihin. Kuitenkin esimerkiksi Mangaonkarin ym. (2022) tutkimuksessa kerrotaan heidän pilvipohjaisen havaitsemisjärjestelmänsä päihittävän monet muut järjestelmät.

Tämän tutkielman tulosten pohjalta on huomattavissa, että monenlaisia nettikiusaamisen automaattisen havainnoinnin järjestelmiä on viimeisten vuosien aikana kehitetty. Tutkielman perusteella huolestuttavaa onkin, ettei järjestelmien kehityksestä huolimatta nettikiusaamista vieläkään pystytä havaitsemaan sen vaatimalla tarkkuudella. Tulevaisuudessa havaitsemisjärjestelmien kehitystä pitää ehdottomasti jatkaa. Erityisesti syväoppimismalleja apunaan käyttäviä järjestelmiä tai pilvipohjaisia havaitsemisjärjestelmiä pitäisi tulevaisuudessa tutkia sekä kehittää enemmän. On myös tärkeää muistaa, että tässä tutkielmassa perehdyttiin pelkästään ohjatun oppimisen lähestymistavan tutkimuksiin.

Tulevaisuudessa voisi olla tärkeää kehittää myös muita lähestymistapoja ja niiden toimivuutta nettikiusaamisen automaattisessa havainnoinnissa. Nettikiusaamisen ehkäiseminen pitäisi olla sosiaalisen median alustoilla yksi tärkeimmistä asioista. Kaikkien, kuten tutkijoiden, sosiaalisen median alustojen tekijöiden sekä omistajien, pitää myös jatkossa kiinnittää erityistä huomiota nettikiusaamisen havaitsemisen ongelmiin ja pyrkiä tekemään sosiaalisesta mediasta kaikille turvallinen paikka.

LÄHTEET

About Instagram | Capture, Create & Share What You Love. (ei pvm.). Noudettu 23. helmikuuta 2023, osoitteesta https://about.instagram.com/

Aichner, T., Grünfelder, M., Maurer, O. & Jegeni, D. (2021). Twenty-Five Years of Social Media: A Review of Social Media Applications and Definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking*, *24*(4), 215–222. https://doi.org/10.1089/cyber.2020.0134

AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M. & Ibrahim, D. M. (2020). Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media. *Journal of Information Technology Management*, *Online First*. https://doi.org/10.22059/jitm.2020.75796

Allcott, H. & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *The Journal of Economic Perspectives*, *31*(2), 211–235. https://doi.org/10.1257/jep.31.2.211

Bharti, S., Yadav, A. K., Kumar, M. & Yadav, D. (2021). Cyberbullying detection from tweets using deep learning. *Kybernetes*, *51*(9), 2695–2711. https://doi.org/10.1108/K-01-2021-0061

Blackwell, D., Leaman, C., Tramposch, R., Osborne, C. & Liss, M. (2017). Extraversion, neuroticism, attachment style and fear of missing out as predictors of social media use and addiction. *Personality and Individual Differences*, *116*, 69–72. https://doi.org/10.1016/j.paid.2017.04.039

Buckels, E. E., Trapnell, P. D. & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, *67*, 97–102. https://doi.org/10.1016/j.paid.2014.01.016

Craig, S., Eaton, A., McInroy, L., Leung, V. & Kirshnan, S. (2021). Can Social Media Participation Enhance LGBTQ+ Youth Well-Being? Development of the Social Media Benefits Scale. *Social Meda + Society*, *2021*(7). https://doi.org/10.1177/2056305121988931

Digitaaliset elämäntavat -tutkimus 2022. (25.1.2023). DNA. https://corporate.dna.fi/fi/tutkimukset-digitaaliset-elamantavat-22

Dixon, S. (14.2.2023). *Biggest social media platforms 2022*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Englander, E., Donnerstein, E., Kowalski, R., Lin, C. A. & Parti, K. (2017). Defining Cyberbullying. *Pediatrics*, *140*(Supplement_2), S148–S151. https://doi.org/10.1542/peds.2016-1758U

Facebook | Meta. (ei pvm.). Noudettu 23. helmikuuta 2023, osoitteesta https://about.meta.com/technologies/facebook-app/

Fati, S. M. (2022). Detecting Cyberbullying across Social Media Platforms in Saudi Arabia Using Sentiment Analysis: A Case Study. *The Computer Journal*, *65*(7), 1787–1794. https://doi.org/10.1093/comjnl/bxab019

FINLEX ® - Ajantasainen lainsäädäntö: Rikoslaki 39/1889. (ei pvm.). Oikeusministeriö, Edita Publishing Oy. Noudettu 8. maaliskuuta 2023, osoitteesta https://www.finlex.fi/fi/laki/ajantasa/1889/18890039001?search%5Btype%5D=pika&search%5Bpika%5D=V%C3%A4kivaltakuvauksen%20levitt%C3%A4minen

Galán-GarcÌa, P., Puerta, J. G. D. L., Gómez, C. L., Santos, I. & Bringas, P. G. (2015). Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. *Logic journal of the IGPL*. https://doi.org/10.1093/jigpal/jzv048

Gillespie-Smith, K., Hendry, G., Anduuru, N., Laird, T. & Ballantyne, C. (2021). Using social media to be 'social': Perceptions of social media benefits and risk by autistic young people, and parents. *Research in Developmental Disabilities*, *118*, 104081. https://doi.org/10.1016/j.ridd.2021.104081

Goff, D. (2013). A History Of The Social Median Industries. Teoksessa *The Social Media Industires*. Routledge.

Harassment & cyberbullying policies - YouTube Help. (ei pvm.). Noudettu 9. maaliskuuta 2023, osoitteesta https://support.google.com/youtube/answer/2802268?hl=en

Horowitz, M. & Bollinger, D. M. (2014). *Cyberbullying in social media within educational institutions: Featuring student, employee, and parent information*. Rowman & Littlefield.

Instagram Launches. (6.10.2010). *Instagram Blog*. https://about.instagram.com/blog/announcements/instagram-launches

Julkaisufoorumi. (ei pvm.). Noudettu 6. toukokuuta 2023, osoitteesta https://www.tsv.fi/julkaisufoorumi/haku.php

Kaplan, A. M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kapoor, K. K., Tamilmani, K., Rana, N. P., Patil, P., Dwivedi, Y. K. & Nerur, S. (2018). Advances in Social Media Research: Past, Present and Future. *Information Systems Frontiers*, *20*(3), 531–558. https://doi.org/10.1007/s10796-017-9810-y

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N. & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137. https://doi.org/10.1037/a0035618

Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R. & Kannan, P. K. (2016). From Social to Sale: The Effects of Firm-Generated Content in Social Media on Customer Behavior. *Jounarl of Marketing*, *Volume 80*(1), 1–123. https://doi.org/10.1509/jm.14.0249

Lobe, B., Velicu, A., Staksrud, E., Chaudron, S. & Gioia, R. (2021). *How children (10-18) experienced online risks during the Covid-19 lockdown - Spring 2020* [Kyselyraportti].

Mangaonkar, A., Pawar, R., Chowdhury, N. S. & Raje, R. R. (2022). Enhancing collaborative detection of cyberbullying behavior in Twitter data. *Cluster Computing*, *25*(2), 1263–1277. https://doi.org/10.1007/s10586-021-03483-1

Meta. (ei pvm.). Meta | Social Metaverse Company. Noudettu 23. helmikuuta 2023, osoitteesta https://about.meta.com/

Muneer, A. & Fati, S. M. (2020). A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, *12*(11), 187. https://doi.org/10.3390/fi12110187

Nandhini, B. S. & Sheeba, J. I. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia Computer Science*, *45*, 485–492. https://doi.org/10.1016/j.procs.2015.03.085

Paniagua, J., Korzynski, P. & Mas-Tur, A. (2017). Crossing borders with social media: Online social networks and FDI. *European Management Journal*, *35*(3), 314–326. https://doi.org/10.1016/j.emj.2016.09.002

Patchin, J. & Hinduja, S. (2006). Bullies Move Beyond the Schoolyard. *Youth Violence and Juvenile Justice*, *4*(2), 123–216. https://doi.org/10.1177/1541204006286288

Patchin, J. W. (1.6.2021). 2021 Cyberbullying Data. *Cyberbullying Research Center*. https://cyberbullying.org/2021-cyberbullying-data

Pfeffer, J., Zorbach, T. & Carley, K. M. (2014). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, *20*(1–2), 117–128. https://doi.org/10.1080/13527266.2013.797778

Promoting Safety and Expression | Meta. (ei pvm.). Noudettu 30. maaliskuuta 2023, osoitteesta https://about.meta.com/actions/promoting-safety-and-expression/

Rahja, R., Aalto, P., Helenius, J., Järvi, N. & Saari, S. (2021). *Nuoret ja nettikiusaaminen* [Kyselyraportti]. https://cdn.mll.fi/prod/2021/04/21124054/nuoret-ja-nettikiusaaminen-kyselyraportti-mll-2021.pdf

Ridings, C. M., Gefen, D. & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *The Journal of Strategic Information Systems*, *11*(3–4), 271–295. https://doi.org/10.1016/S0963-8687(02)00021-5

Salawu, S., He, Y. & Lumsden, J. (2020). Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing*, *11*(1), 3–24. https://doi.org/10.1109/TAFFC.2017.2761757

Schakman, D. (2013). Social Media Content. Teoksessa *The Social Media industries*. Routledge.

Suomen virallinen tilasto (SVT): Väestön tieto- ja viestitätekniikan käyttö. (2020). Tilastokeskus. https://stat.fi/til/sutivi/2020/sutivi_2020_2020-11-10_tau_029_fi.html

Systrom, K. (9.4.2012). Instagram Joins Facebook. *Instagram Blog*. https://about.instagram.com/blog/announcements/instagram-joins-facebook

Talpur, B. A. & O'Sullivan, D. (2020). Cyberbullying severity detection: A machine learning approach. *PLOS ONE*, *15*(10), e0240924. https://doi.org/10.1371/journal.pone.0240924

Thank You for 10 Years. (25.2.2019). *WhatsAppin blogi*. https://blog.whatsapp.com/thank-you-for-10-years

Trandabăț, D., Gifu, D. & Adrian, P. (2022). Detecting Offensive Language in Romanian Social Media. *Procedia Computer Science*, *207*, 2883–2890. https://doi.org/10.1016/j.procs.2022.09.346

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., De Pauw, G., Daelemans, W. & Hoste, V. (2018). Automatic Detection of Cyberbullying in Social Media Text. *PLOS ONE*, *13*(10), e0203794. https://doi.org/10.1371/journal.pone.0203794

Wellman, B., Salaff, J., Dimitrova, D., Garton, L., Gulia, M. & Haythornthwaite, C. (1996). Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. *Annual Review of Sociology*, *22*, 213–238. https://www.jstor.org/stable/2083430

WhatsApp. (ei pvm.). WhatsApp.com. Noudettu 23. helmikuuta 2023, osoitteesta https://www.whatsapp.com/

Zhao, R. & Mao, K. (2017). Cyberbullying Detection Based on Semantic-Enhanced Marginalized Denoising Auto-Encoder. *IEEE Transactions on Affective Computing*, *8*(3), 328–339. https://doi.org/10.1109/TAFFC.2016.2531682