TURUN YLIOPISTO
Kauppakorkeakoulu

# Tekoälyn hyödyntäminen haitallisten sisältöjen hallinnassa sosiaalisessa mediassa

Tietojärjestelmätieteen
kandidaatintutkielma

Laatija:
Jaakko Ylinen

Ohjaaja:
FT Kai Kimppa

5.5.2025
Turku

Kanditutkielma

**Oppiaine**: Tietojärjestelmätiede
**Tekijä**: Jaakko Ylinen
**Otsikko**: Tekoälyn hyödyntäminen haitallisten sisältöjen hallinnassa sosiaalisessa mediassa
**Ohjaaja**: FT Kai Kimppa
**Sivumäärä**: 44 sivua
**Päivämäärä**: 5.5.2025

**Tiivistelmä:**

Sosiaalisen median alustojen suosio on kasvanut merkittävästi ja nykyään sosiaalisen median käyttäjiä on miljardeja. Moni alusta on keskittynyt johonkin tiettyyn sisältötyyppiin ja uusia alustoja tulee jatkuvasti lisää. Käyttäjien luomaa sisältöä on valtavasti ja sitä esiintyy tekstinä, kuvina, videoina ja äänitteinä. Sisältömäärien kasvaessa myös erilaisten haitallisten sisältöjen, kuten vihapuheen määrä on kasvanut. Alustojen ja teknologian kehittyessä myös sisällönhallinta on uudistunut. Siinä käytettävät menetelmät ovat kehittyneet ja muuttuneet ajan myötä vastaamaan kasvaneeseen sisältömäärään. Sisällönhallinnan tehtävissä on alettu hyödyntää laajemmin tekoälymenetelmiä, joiden avulla voidaan suorittaa laajamittaista ja nopeaa sisällön analysointia.

Tekoälyn rooli sisällönhallinnassa on erityisen tärkeä, koska sen avulla sisällönhallintaa voidaan skaalata resurssitehokkaasti. Eri alustatyypeillä voidaan hyödyntää eri tekoälymenetelmiä riippuen sisältötyypistä. Tutkielmassa käsiteltävien tekoälymenetelmien, eli koneoppimisen, syväoppimisen ja luonnollisen kielen käsittelyn avulla voidaan käsitellä vaihtelevia sisältötyyppejä. Alustojen sisällönhallinnassa on vaihtelevuutta myös ihmisten ja tekoälyn käytön suhteen. Jotkut alustat voivat käyttää pelkästään ihmisiä sisällönhallintaan, toiset pelkästään tekoälyä tai lähestymistapaa, jossa hyödynnetään molempia. Parhaat lopputulokset sisällönhallinnassa saavutetaan yhdistämällä tekoälyn ja ihmisten parhaat puolet.

Vaikka tekoäly tarjoaa merkittäviä mahdollisuuksia, sen käyttöön liittyy myös teknisiä ja käytännön haasteita. Näitä ovat muun muassa virheelliset luokitukset, kontekstin väärinymmärrys, hallusinointi eli tekaistun tiedon tuottaminen sekä opetusdataan liittyvät haasteet. Tutkielmassa pohditaan myös tekoälyn käyttöön liittyviä eettisiä kysymyksiä, kuten puolueellisuutta ja tekoälyn toiminnan läpinäkymättömyyttä.

Tässä tutkielmassa määritellään mitä haitallinen sisältö on, jaotellaan sosiaalisen median alustatyypit, käsitellään haitallisten sisältöjen hallinnan kannalta keskeisiä tekoälymenetelmiä ja tekoälyn haasteita. Tutkielma on toteutettu kirjallisuuskatsauksena. Suurin osa käytetyistä lähteistä ovat vertaisarvioituja tieteellisiä artikkeleja, jotka käsittelevät tutkielman eri aihealueita.

**Avainsanat:** tekoäly, sisällönhallinta, sosiaalinen media

# SISÄLLYS

## KUVIOT



## TAULUKOT

# 1 Johdanto

Sosiaalisesta mediasta on tullut osa jokapäiväistä elämäämme ja sitä käyttää nykyään yli puolet ihmiskunnasta, sosiaalisen median käyttäjämäärän kasvaessa vuoden 2017 2,73 miljardista vuoden 2024 5,15 miljardiin (Dixon, 2024b). Teknologisen kehityksen avulla sosiaalisen median kanavat ovat tehneet sosiaalisen kanssakäymisen entistä helpommaksi ja ihmiset ovat enemmän yhteydessä toisiinsa (Pittman & Reich, 2016). Sosiaalisen median alustat mahdollistavat henkilökohtaisen viestinnän, sisällön jakamisen ja seuraamisen, tiedonhaun, oppimisen, viihteen, verkostoitumisen, kaupankäynnin, markkinoinnin sekä yhteiskunnalliseen ja poliittiseen vaikuttamiseen osallistumisen.

Sosiaalisen median alustojen käyttäjämäärien kasvua voidaan selittää myös mobiililaitteiden, kuten älypuhelimien ja tablettien kehittymisellä, joka on tehnyt sosiaalisen median käytöstä entistä helpompaa. Mobiililaitteet ovat tyypillisesti jatkuvasti yhteydessä internettiin ja niihin on tarjolla sovelluksia, jotka tekevät sosiaalisen median käyttämisestä vaivatonta. Myös sosiaalisen median alustat ovat kehittyneet ajan myötä. Niissä esiintyvä sisältö on muuttunut pelkästä tekstipohjaisuudesta sisältämään myös kuvia ja videoita, jonka myötä uusia alustoja on erikoistunut eri sisältötyyppeihin (Appel ym., 2020). Eri alustat mahdollistavat erilaisen kanssakäymisen ja sisällön, esimerkiksi YouTube keskittyy videoalustana toimimiseen, Facebook tarjoaa monipuolisen alustan viestintään, ryhmätoimintaan ja sisällön jakamiseen, Instagram keskittyy visuaaliseen sisältöön, kuten kuviin ja videoihin, X (aiemmin Twitter) toimii reaaliaikaisen viestinnän ja uutisten jakamisen välineenä ja TikTok korostaa lyhytvideosisältöä. Sosiaalisen median monipuolinen luonne sekä helppo saavutettavuus ja käytettävyys ovat tehneet sosiaalisesta mediasta osan jokapäiväistä elämää.

Sosiaalisen median alustojen sisällön monimuotoisuus, datan nopeus ja valtava määrä ovat tehneet sisällönhallinnasta haastavan tehtävän (Gillespie, 2020). Viime vuosien aikana sosiaalisen median negatiiviset puolet, kuten käyttäjille haitallisen sisällön jakaminen ovat kasvaneet runsaasti. Haitallisen sisällön tarkoituksena on aiheuttaa vahinkoa tai häiritä sen katsojia. Käyttäjille turvallisen ja miellyttävän käyttökokemuksen tuottaminen vaatii entistä tehokkaampia sisällön hallintakeinoja. Tekoälyä hyödyntämällä sisällönhallinnassa voidaan automatisoida joitain tehtäviä ja käsitellä suuria tietomääriä tehokkaasti. Tekoälyä hyödynnetään tänä päivänä apuvälineenä sisällönhallinnassa sisällön luokitteluun, haitallisen sisällön tunnistamiseen ja moderointiin. (Gongane ym., 2022.)

Tässä tutkielmassa tutustutaan tarkemmin tekoälyn hyödyntämisen mahdollisuuksiin haitallisen sisällön hallinnassa sosiaalisen median alustoilla. Tutkimuskysymykset ovat:

1. Miten tekoälyä hyödynnetään haitallisen sisällön hallinnassa sosiaalisen median alustoilla?
2. Miten tekoäly toimii haitallisen sisällön hallinnassa?
3. Mitä haasteita tekoälyn hyödyntämiseen liittyy?

Tutkielman luvussa 2 käsitellään haitallista sisältöä ja tutustutaan erilaisiin sosiaalisen median alustatyyppeihin ja keskitytään niiden ominaispiirteisiin. Luvussa 3 perehdytään eri tekoälymenetelmiin, niiden toimintaperiaatteisiin ja käsitteisiin. Luvussa käsitellään myös eri tekoälymenetelmien hyödyntämistä haitallisen sisällön hallinnassa. Luvussa 4 käsitellään haasteita, joita tekoäly kohtaa sisällönhallinnassa ja kritiikkiä tekoälyn hyödyntämistä kohtaan. Luvussa 5 kerrataan käsiteltyjä aiheita, vastataan tutkimuskysymyksiin ja esitetään johtopäätöksiä tutkielmasta.

# 2 Haitallinen sisältö sosiaalisen median alustoilla

## 2.1 Haitallinen sisältö

Koska käyttäjien luoman sisällön määrä on kasvanut valtavasti, on myös haitallisen sisällön, kuten vihapuheen määrä yleistynyt ja kasvanut tasaisesti (Schmidt & Wiegand, 2017). Haitallisen sisällön vaikutus yleisesti yksilöön ja yhteisöihin on negatiivista ja se voi johtua monesta syystä ja tekijästä. Haitalliselle sisällölle altistuminen voi heikentää käyttäjien mielenterveyttä lisäten ahdistusta ja masennusta sekä luoda negatiivisen sosiaalisen median käyttökokemuksen. (Louati ym., 2024.) Gongane ym. (2022) määrittelevät haitallisen sisällön sosiaalisen median alustoilla sellaiseksi sisällöksi, jonka julkaisemisella tai jakamisella on tarkoitus aiheuttaa vahinkoa, haittaa tai häiritä henkilöä tai yhteisöä. Sisältö voi olla haitallista, vaikka siitä koettu haitta olisikin pientä. Koetun haitallisuuden aste vaihtelee sisällöstä riippuen. (Jiang ym., 2021.) Tässä tutkielmassa haitallisella sisällöllä viitataan sisältöön, joka kuuluu johonkin seuraavan kappaleen aihealueista.

Jiang ym. (2021) vertasivat, kuinka eri maissa koetaan sisällön haitallisuutta. Tutkimuksessa haitallisen sisällön määriteltiin kuuluvan seuraaviin aihealueisiin, joita osallistujat arvostelivat sisällön haitallisuuden vakavuuden perusteella: joukkovahinko (esim. terrorismi), haavoittuvat ryhmät (esim. lasten hyväksikäyttöä ja pahoinpitelyä sisältävä materiaali), väkivalta (esim. raaka kuvattu väkivalta), alustojen väärinkäyttö ja roskaposti (esim. alustan palveluiden häirintä), seksuaalinen sisältö (esim. seksuaalinen toiminta ja seksuaalisesti eksplisiittinen kieli), säännellyt tuotteet (esim. huumekauppa), itsetuhoisuus (esim. itsemurhan tai itsensä vahingoittamisen kuvaaminen), taloudellinen haitta (esim. petokset, huijaukset ja yksityisyyden loukkaukset) sekä muu haitta (esim. varkaus, vandalismi). Tutkijat tiedostivat, että heidän yhdysvaltalaiskeskeinen perspektiivinsä todennäköisesti vaikutti ainakin osittain aihealueiden valintaan, mutta he kokivat heidän tekemien aluerajauksien olevan hyvä lähtökohta maiden välisten erojen ja yhtäläisyyksien arviointiin sisällön haitallisuuden vakavuutta arvioitaessa. Näiden aihealueiden lisäksi haitalliseksi sisällöksi on tunnistettu kirjallisuudessa myös vihapuhe, väärän tiedon levittämien ja valeuutiset (engl. fake news), rasismi sekä kiusaaminen (ks. esim. Allcott & Gentzkow, 2017; Gongane ym., 2022; Haimson ym., 2021).

Taulukko 1 Haitallisen sisällön aihealueet

| HAITALLISEN SISÄLLÖN AIHEALUE | Esimerkki esiintymisestä sosiaalisessa mediassa: |
|---|---|
| **Joukkovahinko** | Julkaisu, jossa esiintyy terrorismia |
| **Haavoittuvat ryhmät** | Julkaisu, jossa esiintyy lasten hyväksikäyttöä |
| **Väkivalta** | Kuva tai video, jossa esitetään raakaa väkivaltaa, kuten silpomista |
| **Alustojen väärinkäyttö ja roskaposti** | Massaviestien lähettäminen bottien avulla |
| **Seksuaalinen sisältö** | Video, jossa esiintyy pornograafista sisältöä |
| **Säännellyt tuotteet** | Julkaisu, jossa myydään huumausaineita |
| **Itsetuhoisuus** | Video, jossa tehdään itsemurha |
| **Taloudellinen haitta** | Petosmainen viesti, jossa käyttäjiä houkutellaan antamaan tilitietoja |
| **Muu haitta** | Vandalismiin kannustava sisältö |
| **Vihapuhe** | Syrjivä, hyökkäävä kielenkäyttö jotain ihmisryhmää kohtaan |
| **Väärän tiedon levittäminen ja valeuutiset** | Valheellisia väitteitä esim. poliittisesta henkilöstä ilman todisteita |
| **Rasismi** | Kuva, jossa pilkataan tummaihoisia stereotypioiden avulla |
| **Kiusaaminen** | Julkaisu, jossa yksittäistä käyttäjää nimitellään ja nöyryytetään |

Haitallinen sisältö on laaja käsite, joka pitää sisällään monia eri osa-alueita. Sisällön haitalliseksi kokemiseen vaikuttaa monet asiat. Kulttuuriset erot vaikuttavat suuresti siihen, mitä pidetään haitallisena. Grayn ja Prattin (2025) mukaan kulttuurit muokkaavat ihmisten näkemystä siitä, mikä on väärin ja mitä koetaan haitalliseksi. Maiden väliset kulttuurit voivat olla todella erilaisia. Esimerkiksi Lähi-idän kulttuuri ja länsimainen kulttuuri eroavat toisistaan merkittävästi suhtautumisessa seksuaaliseen sisältöön, alastomuuteen ja naisen asemaan. Kulttuurin lisäksi uskonnolla on suuri vaikutus henkilöiden arvoihin ja toimintaan (Costa & Goodwin, 2006). Uskonto voi myös muokata kulttuureita merkittävästi. Nämä asiat vaikuttavat ihmisten ajatuksiin sisällön haitallisuudesta. Jiangin ym. (2021) tutkimuksessa eri kansalaisuuksien väliset erot sisällön haitallisuuden vakavuutta arvioidessa ilmenivät selvästi monien aihealueiden kohdalla. Tutkimuksessa eri kansalaisuuksien edustajilla oli kuitenkin yhtäläinen näkemys kaikista haitallisimmista aiheista, kuten lasten hyväksikäytöstä ja ihmisten tappamisesta, sekä vähiten haitallisista, kuten alustan hyväksikäytöstä ja roskapostista. Ääripäiden välille sijoittuvissa aihealueissa havaittiin eniten eroavaisuuksia kansalaisuuksien kesken. Näitä eroavaisuuksia voidaan selittää aikaisemmin mainituilla aiheilla, kuten kulttuurilla, uskonnolla ja paikallisten lakien vaikutuksella.

Toisaalta Scheuermanin ym. (2021) tutkimuksessa havaittiin, ettei lakien vaikutus ollut merkittävää sisällön haitallisuutta arvioidessa. Vaikka monien alustojen säännöt pohjautuvat jollain määrin lakeihin, ja osallistujien joukossa oli juridista asiantuntemusta omaavia henkilöitä, ei laillisuuden koettu olevan vaikuttava tekijä sisällön haitallisuutta arvioidessa. Haitan vakavuuteen vaikuttivat enemmän muut tekijät, kuten sisällön haitallinen vaikutus (fyysinen, henkinen, taloudellinen tai sosiaalinen) ja konteksti. Esimerkiksi se, onko sisältö tarkoituksenmukaisesti haitallista, vaikuttaako sisällön haitta yksittäisiin henkilöihin vai isompaan ihmisryhmään, haitan kohteen haavoittuvaisuus (esim. lapset) ja sisällön muoto (teksti, kuva, video, äänite). Myös osallistujien henkilökohtaiset kokemukset olivat yksi vaikuttava tekijä haitallisuuden arvioinnissa. On kuitenkin huomioitava, että tutkimus suoritettiin Yhdysvalloissa, joten lakien vaikutusta ei voi vähätellä muissa maissa, joissa tulokset olisivat voineet olla hyvin erilaisia. Tutkimus osoitti myös sen, että yksilöt kokevat haitan erilailla omien kokemuksien ja näkemyksien myötä. Joku voi kokea väkivaltaisen sisällön haitattomaksi toisen kokiessa saman sisällön erittäin haitalliseksi. Eroavaisuuksista huolimatta haitallisen sisällön olemassaolo tunnustetaan eri puolilla maailmaa (Jiang ym., 2021).

Sosiaalisen median suuri suosio ja valtava sisältömäärä aiheuttavat merkittäviä haasteita haitallisen sisällön hallinnointiin. Suurten tietomäärien lisäksi myös esimerkiksi algoritmien kontekstin ymmärtäminen on osoittautunut vaativaksi tehtäväksi, kun satiiri tai poliittinen keskustelu voidaan virheellisesti tulkita sääntöjen vastaiseksi. Haitallisen sisällön hallintaa hankaloittaa myös haitallisuuden monitulkintainen luonne, jonka takia samaa sisällönhallinnan järjestelmää ei voida käyttää kaikkialla. Automaattisia haitallisen sisällön hallintajärjestelmiä tulee muokata alueelle ja kulttuuriin sopivaksi, jotta niistä saataisiin paras hyöty irti. (Jiang ym., 2021.)

Digitalisaation ja sosiaalisen median yleistymisen myötä kasvava määrä ihmisiä käyttää sosiaalisen median alustoja uutisten lähteenä. Suuret yleisömäärät sekä tiedon nopea leviäminen tekevät sosiaalisen median alustoista tehokkaan työkalun informaation jakamiseen. Se on samalla muuttanut ihmisten käsitystä uutisista. Esimerkiksi X:stä on tullut alusta nopealle journalismille, jossa yksittäinen postaus voidaan nähdä uutisena, erityisesti silloin, kun se on peräisin auktoriteettiasemassa olevalta henkilöltä. Tämä muutos on tehnyt myös valheellisen informaation ja valeuutisten levittämisen helpommaksi. (Tandoc ym., 2018.) Teknologian kehityksen myötä valheellista tietoa ja valeuutisia voidaan esittää vakuuttavasti tekoälyllä tehtyjen syvävääreennösvideoiden (engl. deepfake video) avulla. Syvävääreennösvideoiden mahdollistama haitallinen toiminta on laajaa ja niitä voidaan käyttää esimerkiksi sabotaasiin, hyväksikäyttöön sekä luottamuksen ja yleisen turvallisuuden heikentämiseen. (Chesney & Citron, 2019.) Internetistä ja sosiaalisen median alustoista on tullut tärkeitä välineitä poliittiseen vaikuttamiseen, kampanjointiin

ja kansalaisosallistumiseen. Tämän vuoksi valeuutisten ja väärän tiedon leviämisellä voi olla vakavia seurauksia. Chesneyn ja Citronin (2019) mukaan nämä seuraukset voivat ilmetä esimerkiksi vaalien manipulointina, sosiaalisen luokkajaottelun vahvistumisena, instituutioihin kohdistuvan luottamuksen heikkenemisenä, kansainvälisten suhteiden vahingoittumisena tai taloudellisena haittana.

## 2.2 Sosiaalinen media

Kaplanin ja Haenleinin (2010) sekä Kietzmannin ym. (2011) mukaan sosiaalinen media on joukko verkkopohjaisia teknologioita ja mobiililaitteita hyödyntäviä sovelluksia, jotka luovat interaktiivisia alustoja käyttäjien tuottaman sisällön luomiseen, jakamiseen ja muokkaamiseen. Sosiaalisessa mediassa esiintyy monipuolista sisältöä, joka on johtanut tiettyyn sisältöön erikoistuneisiin alustoihin. Jotkut suuret sosiaalisen median alustat, kuten Facebook, ovat suunnattuja laajalle käyttäjäkunnalle ja tarjoavat monipuolisesti eri sisältötyyppejä. Toiset alustat taas keskittyvät tietynlaisen sisällön jakamiseen, kuten videoihin (esim. YouTube), kuviin (esim. Pinterest) tai ammatilliseen verkostoitumiseen (esim. LinkedIn). (Kietzmann ym., 2011.) Teknologian kehittyessä ja käyttäjämäärien kasvaessa uusien alustojen kysyntä kasvaa ja niitä tulee digitaaliseen ympäristöön jatkuvasti (Kaplan & Haenlein, 2010). Termiä sosiaalinen media on käytetty viimeiset vuosikymmenet kuvaamaan eri verkkopohjaisia alustoja, kuten yhteisöpalveluita (esim. Facebook), mikroblogeja ja blogeja (esim. X), sisällönjako- ja keskustelualustoja (esim. Reddit) sekä video- ja kuvamateriaalin jakoalustoja (esim. YouTube) (Ellison & Vitak, 2015; Lau, 2017; Mangold & Faulds, 2009).

Vaikka uusia alustoja syntyy jatkuvasti, suuret teknologiayritykset päätyvät usein ostamaan niitä tai kopioimaan niiden ominaisuuksia omille alustoilleen. Nykyään suuremmat sosiaalisen median alustat ovat laajentaneet toimintojaan kattamaan ominaisuuksia, joita niillä alun perin ei ole ollut. Esimerkiksi lyhytvideoihin erikoistuneen TikTokin suuren menestyksen myötä monet suuret alustat, kuten Facebook, Instagram ja YouTube lisäsivät omille alustoilleen lyhytvideotoiminnon. Tällainen muutos alustojen toiminnassa tekee niiden jaottelusta tiettyihin kategoriatyyppeihin haastavaa. Seuraavissa alaluvuissa käsitellään kolmea eri sosiaalisen median alustatyyppiä: yhteisöpalveluita, sisällönjako- ja keskustelualustoja sekä videoalustoja tarkemmin.

### 2.2.1 Yhteisöpalvelut

Yhteisöpalveluilla (engl. social networking sites) tarkoitetaan verkkopohjaisia palveluita, jotka mahdollistavat julkisen tai puolijulkisen tilin luomisen, muiden käyttäjien profiilien ja yhteyksien

katsomisen ja yhteyden luomisen muihin käyttäjiin esimerkiksi kaveripyynnöllä tai seuraamisella (Boyd & Ellison, 2007). Ellison ja Vitak (2015) kertovat Boydin ja Ellisonin (2013) aiemman pohjalta päivittämän määritelmän mukaan yhteisöpalveluilla olevan kolme keskeistä piirrettä:

1. Yksilöitävissä olevat profiilit, jotka koostuvat käyttäjän tuottamasta ja muiden käyttäjien tarjoamasta sisällöstä ja järjestelmätason tiedoista.
2. Mahdollisuus julkisesti ilmaista yhteyksiään, joita muut voivat tarkastella ja käydä läpi
3. Mahdollisuus kuluttaa, tuottaa ja olla vuorovaikutuksessa yhteydessä olevien käyttäjien luoman sisällön kanssa.

Yhteisöpalveluissa jokainen profiili on oma uniikki sivunsa, johon käyttäjä voi julkaista haluamaansa sisältöä. Monissa yhteisöpalveluissa on mahdollista luoda oma profiili, julkaista sisältöä, lisätä ystäviä, kommentoida, lähettää yksityisviestejä sekä jakaa kuvia ja videoita (Kaplan & Haenlein, 2010). Monissa suurissa yhteisöpalveluissa keskitytään verkostoitumisen ja uusien ihmisten tapaamisen sijaan kanssakäymiseen omien tuttujen kanssa (Boyd & Ellison, 2007).

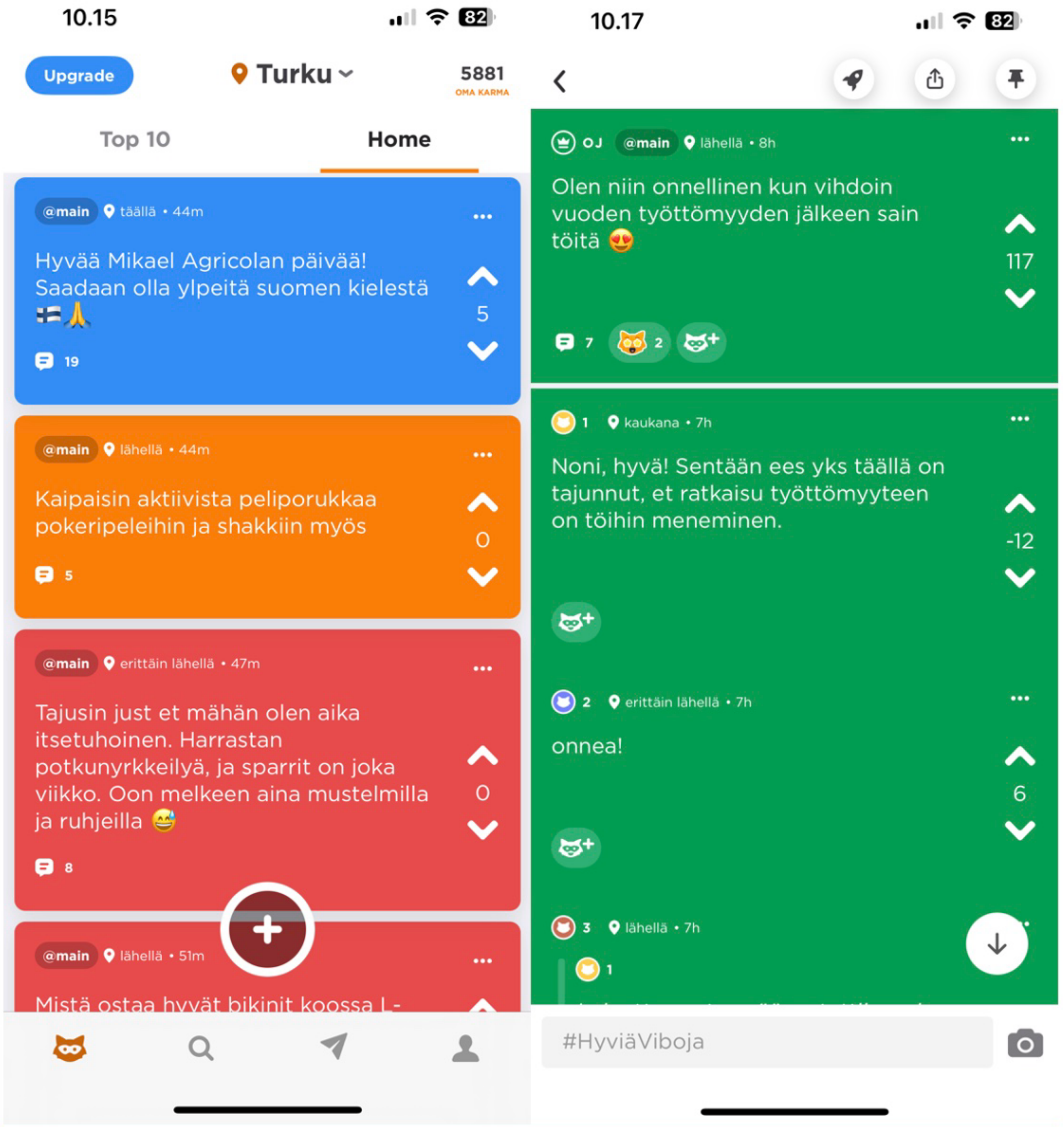
Ellison ja Vitak (2015) kertovat Facebookin olevan hallitseva alusta vuorovaikutukseen ja suhteiden ylläpitämiseen ja monet muut alustat kasvattavat suosiotaan toimimalla erilailla. Nykyään suuret yhteisöpalvelut ovat enemmän monipuolisia alustoja, jotka tarjoavat käyttäjilleen enemmän mahdollisuuksia perinteisen vuorovaikutuksen lisäksi (Nylén ym., 2024). Esimerkiksi Facebook tarjoaa käyttäjilleen mahdollisuuden mm. deittailuun, kaupankäyntiin, yksityiseen viestimiseen, pelaamiseen sekä lyhytvideoiden katsomiseen ja julkaisemiseen.

### 2.2.2 Keskustelualustat

Keskustelualustoilla ja foorumeilla on yleisiä piirteitä ja eroavaisuuksia, jotka tulevat esiin tarkastelemalla muutamia esimerkkialustoja. Reddit on yksi maailman suosituimmista verkkosivustoista, ja sillä on yli 50 miljoonaa päivittäistä käyttäjää (Potter, 2021). Redditiä voi käyttää suoraan verkkoselaimessa tai sovelluksessa. Alusta koostuu yli sadastatuhannesta alafoorumista, joilla käsitellään eri aiheita ja teemoja (Britt ym., 2023). Käyttäjät voivat julkaista sisältöä, kuten tekstiä, kuvia, videoita tai linkkejä ulkoisille sivuille ja muut käyttäjät voivat kommentoida sekä äänestää niitä ylös- tai alaspäin. Äänestämällä käyttäjät voivat vaikuttaa sisällön näkyvyyteen. (Davis & Graham, 2021). Redditin profiileilla on anonyymi luonne, koska käyttäjänimi voidaan valita itse, eikä henkilökohtaisia tietoja tarvitse jakaa.

Monien keskustelualustojen ja foorumien toimintatapa eroaa hieman Redditistä. Internetissä on useita eri foorumeita, joissa pääsee osallistumaan täysin anonyymisti ilman profiilia. Esimerkiksi 4chan on foorumisivusto, jossa käyttäjät voivat luoda keskustelulautoja, joihin voi ladata kuvia ja kirjoittaa kommentteja. Jokaisen lauta aloitetaan julkaisulla, jossa on kuva ja mahdollisesti tekstiä. Näihin lautoihin voi vastata kuvilla ja kommenteilla. Toisin kuin Redditissä, jossa julkaistu sisältö voi säilyä pitkiä aikoja ja jonka näkyvyyteen vaikutetaan äänestämällä sekä foorumikohtaisten moderaattoreiden avulla, 4chanissa ei ole käyttäjäprofiileja eikä pysyviä tietoarkistoja. Viestit katoavat ajan myötä, mikä tekee sisällöstä ajankohtaista, mutta vaikeasti jäljitettävää. (Nissenbaum & Shifman, 2017.) Alustalla on säännöt, jotka ohjaavat käyttäjiä jossain määrin, mutta ne sallivat joitain haitalliseksi määriteltyjä sisältöjä julkaistavan. 4chanin sivustolta puuttuva tehokas moderointi ja käyttäjien anonyymiys johtaa haitallisen sisällön jatkuvaan julkaisuun sivustolla (Thorleifsson, 2022).

Anonyymit keskustelufoorumit ovat suosittuja myös Pohjoismaissa ja Suomessa, missä esimerkiksi Jodel-sovellus on kasvattanut suosiotaan. Jodel ei vaadi profiilia osallistumiseen, ja sen käyttö on täysin anonyymiä. Redditin tapaan käyttäjät voivat äänestää julkaisuja ja kommentteja ylös- tai alaspäin, mikä vaikuttaa sisällön näkyvyyteen ja hallintaan. Jodelilla ja monilla muilla keskustelualustoilla on omat yhteisön säännöt, jotka määrittelevät sallitun sisällön (ks. esim. *Yhteisön Säännöt | Jodel Support Hub*, ei pvm.).




Kuva 1 Jodel-sovellus

Keskustelualustojen ja foorumeiden anonyymi luonne voi tarjota tilan avoimelle keskustelulle, mutta se voi myös kannustaa käyttäjiä provosoivaan käytökseen ja vastuuttomuuteen, jotka esiintyvät esimerkiksi asiattomana viestintänä. Alustojen väliset erot sisällön hallinnoimisessa ja säännöissä mahdollistavat käyttäjille vapauden erilaisten sisältöjen julkaisemiselle. Esimerkiksi pornograafisen sisällön tai raa'an väkivallan julkaiseminen on kiellettyä Jodelissa, Redditin ja 4chanin salliessa sen (*Reddit Rules*, ei pvm.; *Rules - 4chan*, ei pvm.; *Yhteisön Säännöt | Jodel Support Hub*, ei pvm.). Alustojen lisäksi myös käyttäjien välillä on eroja. Alustoja, kuten 4chania saattaa käyttää enemmän henkilöt, jotka tiedostavat siellä julkaistavan sisällön luonteen. He eivät välttämättä pidä sitä haitallisena ja ovat tehneet tietoisen valinnan käyttää alustaa.

### 2.2.3 Video- ja kuva-alustat

Käyttäjien luoma sisältö (engl. user generated content, UGC) on mullistanut ihmisten viihdekulutuksen. Nykyään miljoonat ihmiset tuottavat ja miljardit kuluttavat videoita. Videosovelluksia ladattiin vuonna 2023 yli kolme miljardia kertaa (Ceci, 2024). Jatkuva uuden sisällön virta, yhdistettynä internetin ja sovellusten helppokäyttöisyyteen sekä kehittyneisiin algoritmeihin tekee käyttö- ja katselukokemuksesta miellyttävän ja henkilökohtaisen.

Internetissä tapahtui räjähdysmäinen kasvu videoiden jakamisessa, kun ensimmäiset videonjakoalustat julkaistiin (Cheng ym., 2008). YouTube on ollut julkaisustaan vuonna 2005 lähtien alan suurin toimija, ja nykyään sillä on yli 2,5 miljardia päivittäistä käyttäjää (Dixon, 2024a). YouTube toimii videonjakoalustana, jossa käyttäjät voivat julkaista ja katsoa erimittaisia videoita sekä tilata kanavia, joiden sisällöstä he pitävät, pysyäkseen paremmin ajan tasalla uusista julkaisuista. Videoihin voi kommentoida ja antaa peukun ylös tai alaspäin ilmaistakseen mielipiteensä. Alustan toimintatapa on vuosien varrella muuttunut siten, että ylöspäin annettujen peukkujen määrä näkyy yhä (ellei videon lataaja erikseen estä sitä), mutta alaspäin annettujen määrä ei enää ole julkisesti nähtävissä ilman ulkopuolista sovellusta, kuten ”Return YouTube Dislike”-selaimen laajennusta.

Käyttäjien katselupreferenssit ovat muuttuneet vuosien saatossa. Viime vuosina lyhytvideot ovat kasvattaneet suosiotaan, erityisesti TikTokin julkaisun ja nopean kasvun myötä. TikTokissa käyttäjät voivat luoda ja jakaa lyhyitä noin 3–15 sekunnin (nykyään pidempiäkin) mittaisia videoita esimerkiksi tanssimisesta, komediasta tai muusta sisällöstä. Videoiden taustalle voi lisätä haluamansa ääniraidan. TikTokin suosion taustalla on muun muassa älypuhelimien yleistyminen, joka tekee sovelluksen käytöstä vaivatonta. Lisäksi sen kehittynyt algoritmi tarjoaa käyttäjille henkilökohtaista ja jatkuvasti ajan tasalla olevaa sisältöä, mikä on tehnyt sovelluksesta entistä suositumman. (Fan & Hemans, 2022.) Käyttäjät voivat luoda, tykätä, jakaa, kommentoida, tallentaa ja ilmiantaa julkaisuja. TikTokin menestys on vaikuttanut myös muihin alustoihin, jotka ovat alkaneet tarjota samanlaista mahdollisuutta lyhytvideoiden julkaisemiseen ja katseluun.

Videoalustojen tapaan myös kuvien jakamiseen on olemassa omia, siihen erikoistuneita alustoja. Mobiililaitteiden yleistyminen ja sosiaalisen median kasvu ovat tehneet kuvien jakamisesta merkittävän ilmiön, mikä näkyy esimerkiksi Facebookin, Instagramin ja X:n käyttäjien lataamien satojen miljardien kuvien määrässä. Vaikka monet alustat eivät alun perin keskittyneet kuvien jakamiseen, siitä on tullut yksi yleisimmistä toiminnoista. Nykyään lähes kaikki sosiaalisen median

alustat tukevat kuvajulkaisuja, mutta jotkin alustat, kuten Pinterest ja Flickr, ovat erikoistuneet nimenomaan kuvien jakamiseen. (Oeldorf-Hirsch & Sundar, 2016.)

# 3 Tekoälymenetelmien käyttö sisällönhallinnassa

Chartrand ym. (2017) määrittelevät tekoälyn (engl. artificial intelligence, AI) tietojenkäsittelytieteen osa-alueeksi, joka keskittyy luomaan ja kehittämään järjestelmiä, joiden avulla voidaan suorittaa ihmisen älykkyyttä vaativia tehtäviä. Heidän mukaansa tekoäly on enemmänkin kattotermi, joka kattaa useita erilaisia alalajeja ja tekniikoita. Samankaltaisesti tekoäly voidaan nähdä kattoterminä, joka kattaa tieteenalat ja teknologiat, joissa koneita hyödynnetään ihmisen älykkyyden jäljittelemiseen, laajentamiseen tai parantamiseen. Se tarkoittaa koneen kykyä oppia kokemuksesta, mukautua uusiin syötteisiin ja suorittaa ihmismäisiä tehtäviä. (Prakash & Das, 2021.)

Tekoälyllä on merkittävä rooli digitaalisen sisällön hallinnassa, sillä sen avulla voidaan automatisoida monia prosesseja, kuten moderointia, suodattamista ja analysointia (Gorwa ym., 2020). Tämä tutkielma keskittyy sisällönhallinnan näkökulmasta keskeisiin tekoälyn alalajeihin: koneoppimiseen (engl. machine learning) ja sen osa-alueisiin, syväoppimiseen (engl. deep learning) ja luonnollisen kielen käsittelyyn (engl. natural language processing, NLP). Gorwan ym. (2020) mukaan koneoppimisen kehitys on mahdollistanut useiden sisällönhallinnan tehtävien siirtämisen tekoälyn hoidettavaksi ja automaattiset järjestelmät vihapuheen, henkilökohtaisten hyökkäysten ja muiden toksisten ilmiöiden tunnistamiseen ovat kehittyneet luonnollisen kielen käsittelyn edistysaskelten myötä.

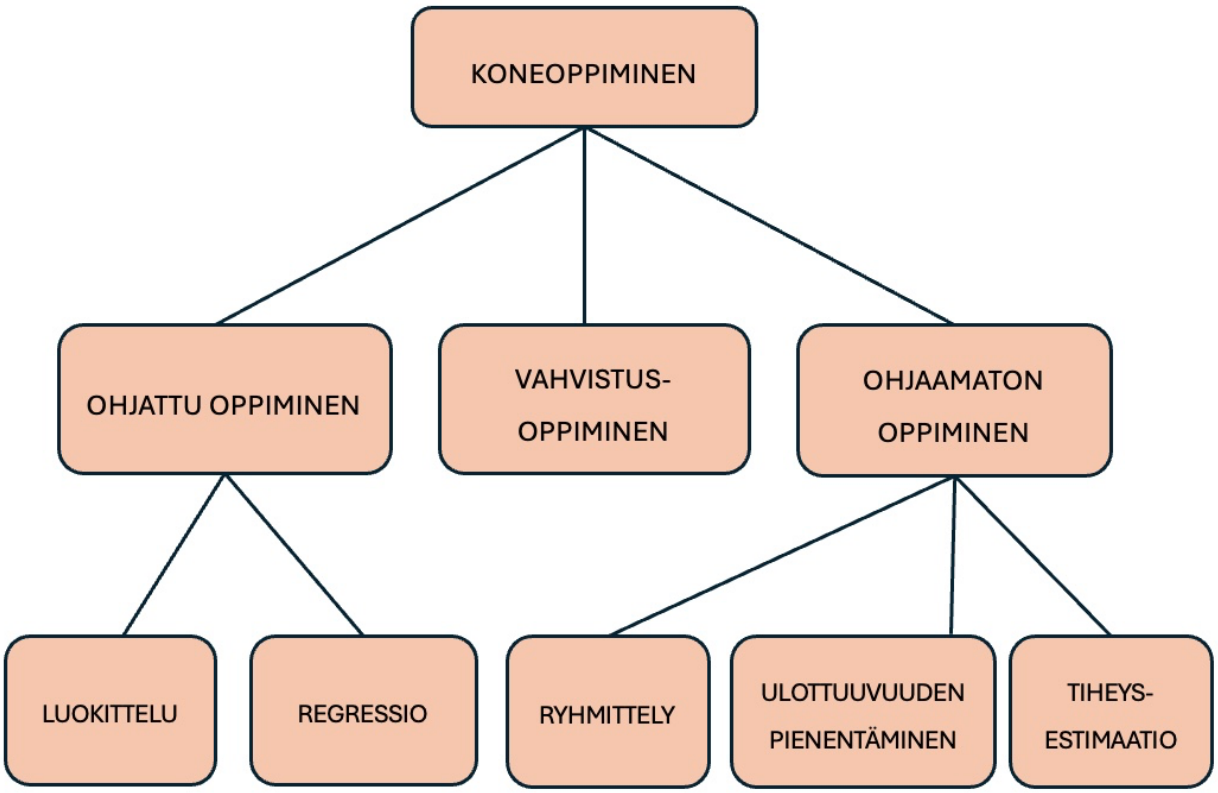
## 3.1 Koneoppiminen

Koneoppiminen on tekoälyn osa-alue, joka tutkii algoritmeja, jotka kykenevät oppimaan itsenäisesti suoraan syötedata (Bertolini ym., 2021). Koneoppimiselle esitetään useita määritelmiä kirjallisuudessa. Murphy (2012) määrittelee koneoppimisen joukoksi metodeja, jotka voivat automaattisesti havaita toistuvia malleja syötetystä datasta ja käyttää näitä havaintoja tulevan datan ennustamiseen tai päätöksentekoon epävarmuuden alla. Bertolinin ym. (2021) mukaan Murphyn (2012) määritelmä, selkeydestään huolimatta, painottaa liikaa mallien tunnistamista ja päätöksentekoa. Määritelmä ei kata kokonaan laajaa koneoppimisen metodien ja lähestymistapojen joukkoa. He määrittelevät koneoppimisen hieman yleisemmin joukoksi algoritmeja ja metodeja, jotka pystyvät tuottamaan tietoa datasta ja parantamaan suorituskykyään oppimalla kokemuksesta.

Tekoälyn tavoin koneoppiminen ei ole uusi teknologia. Koneoppiminen on ollut olemassa ainakin 1970-luvulta asti. Ajan saatossa koneiden laskentateho on kasvanut, joka on mahdollistanut koneoppimisen käytön entistä haastavampien ja monimutkaisempien ongelmien parissa. (Louridas & Ebert, 2016.) Etenkin viimeisten vuosikymmenien aikana koneoppiminen on ottanut suuria

kehitysaskeleita eteenpäin, jotka ovat näkyneet esimerkiksi itseohjautuvien autojen algoritmeissa (Bertolini ym., 2021).

Koneoppiminen voidaan Murphyn (2012) mukaan jaotella yleisesti kolmeen osa-alueeseen: ohjattuun oppimiseen (engl. supervised learning), ohjaamattomaan oppimiseen (engl. unsupervised learning) ja vahvistusoppimiseen (engl. reinforcement learning).




Kuva 2 Koneoppimisen osa-alueet

Ohjattu oppiminen on käytetyin koneoppimisen muoto (Murphy, 2012). Se sisältää monia algoritmeja, joista yleisimpien joukkoon kuuluvat muun muassa keinotekoiset neuroverkot, päätöspuut, tukivektorikone ja looginen regressio. Vaikka algoritmien välillä on eroavaisuuksia implementoinnissa ja toiminnassa, niillä kaikilla on tavoitteena oppia approksimaatiofunktio. Approksimaatiofunktio oppii syötteiden (x) ja haluttujen tulosten (y) välisen yhteyden merkityn opetusdatan perusteella. Oppimisdata sisältää syötteet sekä niiden oikeat vastaukset. Tämä oppimisdata tulee siis merkitä (engl. label) jonkun toimesta ja sitä tulee olla runsaasti, jotta malli voi oppia tarkasti. Toisin sanoen, ohjatun oppimisen algoritmit oppivat tekemään ennusteita ja päätöksiä merkatun opetusdatan avulla, ja opitun approksimaatiofunktion avulla algoritmi voi soveltaa oppimaansa ennalta tuntemattomiin tapauksiin. (Bertolini ym., 2021; Chen ym., 2023.) Louridas ja Ebert (2016) vertasivat ohjattua oppimista tilanteeseen, jossa opiskelijalle annetaan

joukko tehtäviä ja niiden ratkaisut ja hänen tehtävänään on ymmärtää niiden avulla, miten ratkaista tulevaisuudessa vastaavanlaisia ongelmia.

Ohjattua oppimista käytetään erityisesti kahden tyyppisiin ongelmiin: luokitteluun ja regressioon. Näiden pääeroavaisuutena on tulosten (engl. output) datan tyyppi. Regressiossa ennustetaan jatkuvia numeerisia arvoja ja luokittelussa pyritään selvittää mihin ennalta määriteltyyn kategoriaan syöte kuuluu. Luokittelua käytetään esimerkiksi eri tiedostojen luokittelemiseen, kuten sähköpostiviesteistä roskapostin suodattamiseen. Regression avulla taas voidaan esimerkiksi ennustaa tulevaa osakkeen hintaa huomioiden markkinoiden olosuhteet ja muut tiedot. (Murphy, 2012.)

Gorwan ym. (2020) kertovat nykyisten modernien sisällönluokittelualgoritmien yleensä käyttävän koneoppimista ja luokittelua toiminnassaan. Nämä modernimmat algoritmit ovat korvanneet aiemmin käytettyjä manuaalisesti koodattuja luokitteluohjelmia, joilla oli rajoituksia, kuten tehokkaan ja ajan tasalla olevan ”mustan listan” ylläpitäminen jatkuvasti muuttuvassa ja kehittyvässä ympäristössä. Khandayn ym. (2022) tutkimuksessa tarkasteltiin eri ohjatun oppimisen menetelmiä vihapuheen tunnistamiseen X:ssä Covid-19 pandemian aikana. Vihapuheen tunnistamiseen käytettiin eri luokittelutekniikoita, joilla postaukset jaettiin sen mukaan, sisälsivätkö ne vihapuhetta vai eivät. Ohjatun oppimisen menetelmät suoriutuivat luokittelusta hyvin tuloksin. Etenkin päätöspuiden avulla saavutettiin 98 % tarkkuus ja 97 % herkkyys, eli kyky löytää oikeat tapaukset tietyssä luokassa. Tutkimus osoitti koneoppimisen hyödyntämisen tehokkuuden vihapuheen havaitsemisessa, mutta se myös korosti ohjatun oppimisen suurinta haastetta, eli tarvetta merkitylle opetusdatalle sekä manuaaliselle piirteiden valinnalle.

Ohjaamaton oppiminen eroaa ohjatusta siten, että tietojoukot sisältävät merkkaamatonta dataa. Sen tavoitteena ei ole ennusteen tekeminen, vaan saada algoritmit tunnistamaan ja erottamaan toistuvia malleja tietojoukoista. Koska dataa ei tarvitse merkata kenenkään toimesta, se tekee ohjaamattomasta oppimisesta laajemmin hyödynnettävän koneoppimisen vaihtoehdon, tarjoten ohjattuun oppimiseen verrattuna joustavampia, halvempia ja automatisoituja koneoppimisalgoritmeja. (Bertolini ym., 2021; Chen ym., 2023.)

Murphyn (2012) mukaan ohjaamaton oppiminen voidaan jakaa kolmeen pääosa-alueeseen: ryhmittelyyn, tiheysestimaatioon ja ulottuvuuden pienentämiseen. Ryhmittelyalgoritmien (engl. clustering) tehtävänä on jakaa kohteet ryhmiin niin, että saman ryhmän kohteet ovat mahdollisimman samankaltaisia keskenään ja eroavat muista ryhmistä. Ryhmittelyä käytetään esimerkiksi markkinoinnissa, kun halutaan löytää joukko kuluttajia, joilla on samanlaista

ostokäyttäytymistä. Tiheysestimaatio (engl. density estimation) on joukko tekniikoita, joiden avulla voidaan paljastaa ja havaita hyödyllisiä ominaisuuksia aineistosta, kuten vinoumia tai monihuippuisuutta. (Bertolini ym., 2021.) Ulottuvuuden pienentämisen (engl. dimensionality reduction) algoritmit käsittelevät moniulotteista dataa ja projisoivat sitä, eli ns. esittävät sitä pienemmässä ulottuvuudessa. Sen tavoitteena on yksinkertaistaa tietoa, jonka avulla olennaisten piirteiden havaitseminen datasta helpottuu. (Louridas & Ebert, 2016.)

Ulottuvuuden pienentämisellä on merkittävä rooli vihapuheen tunnistamiseen käytetyissä koneoppimismalleissa. Luokittelu- ja ryhmittelyalgoritmeja käytetään paljon vihapuheen tunnistamiseen ja ne toimivat parhaiten, kun niille syötettyä dataa on siistitty ulottuvuuden pienentämisen avulla. Siistitystä datasta on poistettu häiriöitä ja karsittu ominaisuuksia, jotka eivät tuo algoritmille lisäarvoa. Ulottuvuuden pienentäminen on tärkeää nykypäivänä, koska dataa on valtavasti ja se sisältää runsaasti epäoleellista tietoa koneoppimismallille, mikä tekee merkityksellisten trendien havaitsemisen haastavaksi. (Mullah & Zainon, 2021.)

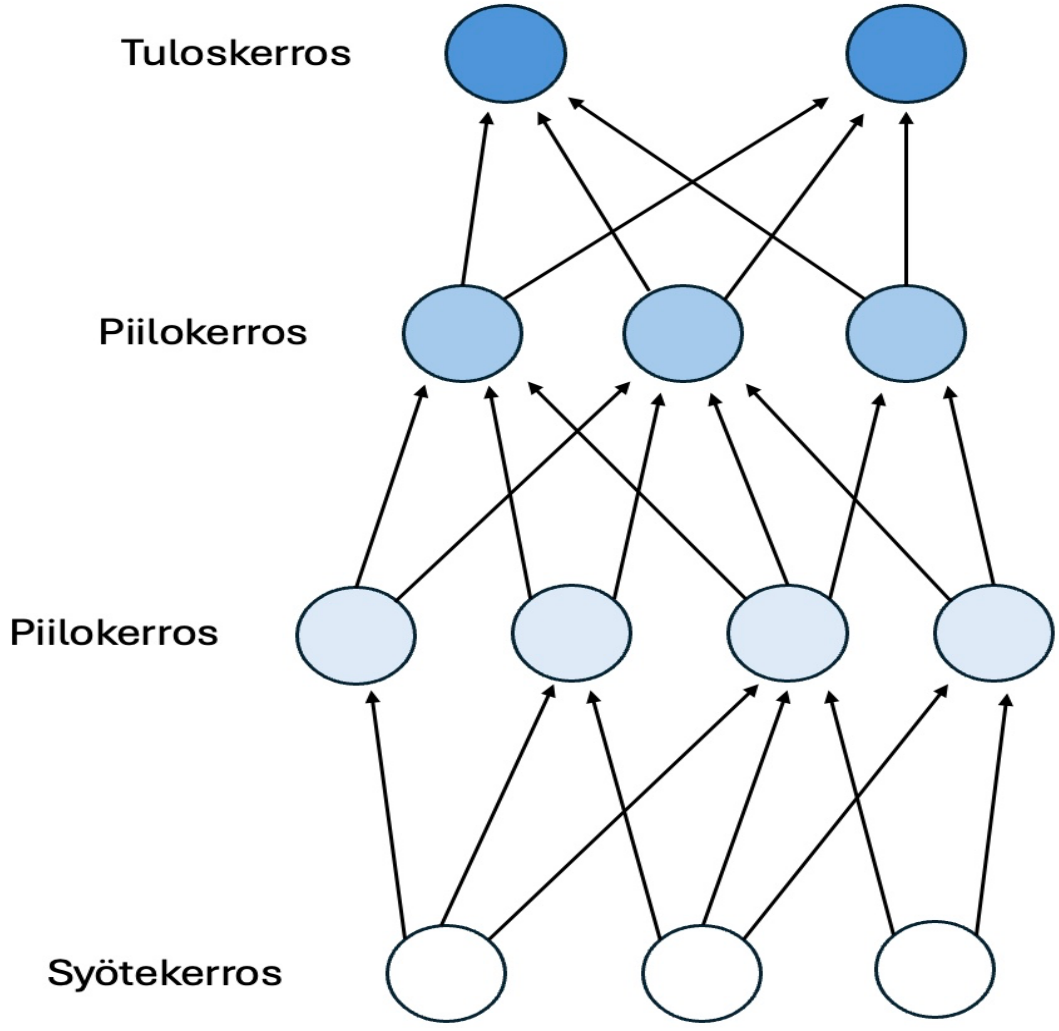
Muihin koneoppimismenetelmiin verrattuna vahvistusoppiminen toimii eri periaatteella, sillä sen algoritmit perustuvat oppimiseen vuorovaikutuksesta ympäristön kanssa, eikä valmiin opetusaineiston avulla. Vahvistusoppimisalgoritmien oppimisprosessit muistuttavat ihmisen tapaa oppia, sillä niiden tavoitteena on oppia itsenäisesti tekemään päätöksiä kokeilemalla ja saamalla palautetta onnistumisesta. Vahvistusoppimisen ytimessä on agentti, joka toimii ympäristössään ennalta määriteltyjen toimintojen avulla. Jokainen agentin tekemä toiminto muuttaa järjestelmän tilaa ja muutoksesta annetaan agentille palkkiosignaali. Agentin tavoitteena on maksimoida saamansa kokonaispalkkio, jonka seurauksena agentti oppii tekemään optimaalisia päätöksiä eri tilanteissa. Kaikkien vahvistusoppimismenetelmien lopullisena tavoitteena on kehittää agentteja, jotka pystyvät tekemään hyviä päätöksiä ympäristönsä ja kokemuksensa perusteella. (Bertolini ym., 2021.)

## 3.2 Syväoppiminen ja neuroverkot

Perinteisten koneoppimismenetelmien rajoitteena on ollut niiden kyvyttömyys raakadatan suoraan prosessointiin. Koneoppimisjärjestelmien rakentaminen edellytti manuaalista piirteiden valintaa sekä datamuuntajan luomista, jonka avulla raakadata saatiin oppimisjärjestelmälle sopivaan muotoon. Esitysmuotojen oppimismenetelmät (engl. representation learning methods) kykenevät automaattisesti, ilman etukäteen tehtyjä manuaalisia määrittelyjä, havaitsemaan raa'asta datasta piirteitä, joita tunnistamiseen tai luokitteluun tarvitaan. Syväoppimisen menetelmät ovat monikerroksisia esitysmuotojen oppimismenetelmiä, joista lähes kaikkien toiminta perustuu

neuroverkkoihin. (Lecun ym., 2015.) Tästä syystä keinotekoiset neuroverkot (engl. artificial neural networks, ANNs) ovat oleellinen osa tekoälykeskustelua, erityisesti syväoppimista.

Macukow (2016) määrittelee neuroverkot tiedon ja signaalien prosessointijärjestelmäksi, joka käsittelee ihmisten aivojen tavalla tietoa. Neuroverkot, keinotekoiset tai biologiset, koostuvat suuresta määrästä yksinkertaisia laskentayksiköitä (hermosoluja/neuroneja), jotka käsittelevät tietoa ja signaaleja. Nämä yksiköt ovat suoraan yhteydessä toisiinsa ja toimivat rinnakkaisesti. Tämä mahdollistaa neuroverkoille monimutkaisten laskennallisten tehtävien tehokkaan ratkaisemisen. (Macukow, 2016.) Monikerroksinen neuroverkko koostuu syötekerroksesta, piilokerroksista ja tuloskerroksesta.




Kuva 3 Monikerroksinen neuroverkko

Kuvan 3 neuroverkko on pelkistetty versio vain muutamilla yksiköillä/neuroneilla kerroksittain. Lecunin ym. (2015) mukaan esimerkiksi luonnollisen kielen prosessointiin käytetyissä neuroverkoissa voi olla kerroksittain jopa satoja tuhansia laskentayksiköitä.

Monikerroksiselle neuroverkolle syötetään dataa, jonka jokainen kerros käsittelee laskemalla painotetun summan edellisen kerroksen syötteistä, jonka se kuljettaa epälineaarisen funktion läpi. Näin jokaisessa kerroksessa voidaan poimia syötteestä olennaisia piirteitä, ja askel askeleelta data muuttuu koneoppimismallille helpommin tulkittavaan muotoon. Neuroverkko voi oppia datasta monipuolista tietoa, kun se on kulkenut riittävän monen prosessointikerroksen läpi. Luokittelutehtävissä korkeammat prosessointikerrokset vahvistavat syötteessä olevia luokittelun kannalta tärkeitä piirteitä ja suodattavat pois epäolennaista tietoa. Syväoppimisessa ydinasia on, että prosessi tapahtuu automaattisesti ilman, että ihmisen täytyy erikseen määritellä, mitä piirteitä mallin tulisi etsiä. (Lecun ym., 2015.)

Lecun ym. (2015) sekä Saleem ym. (2022) painottavat syvien neuroverkkojen, erityisesti konvoluutioneuroverkkojen (engl. convolutional neural networks, CNN) erinomaista onnistumista kuvien luokittelussa, segmentoinnissa ja tunnistamisessa. Heidän mukaansa CNN:t ovat paras vaihtoehto kuvatiedon oppimiseen ja niitä hyödynnetään enemmistöön ongelmista, jotka liittyvät kuvien käsittelyyn ja tietokonenäköön (engl. computer vision). CNN:t ovat neuroverkkoja, jotka sisältävät konvoluutiokerroksia. Konvoluutiokerroksessa on suodattimia, joiden avulla kuvista tunnistetaan piirteitä, kuten reunoja, eri värisiä pisteitä ja kuvioita. Konvoluutiokerroksien jälkeen neuroverkossa on yhdistelykerroksia (engl. pooling), jotka yhdistävät semanttisesti samankaltaisia piirteitä ja pienentävät kuvan kokoa samalla säilyttäen oleelliset tiedot. Yhdistelykerrokset lisäävät mallin kykyä käsitellä piirteitä niiden tarkasta sijainnista riippumatta. Lopulta verkossa on täysin yhdistetty kerros, joka yhdistää opitut piirteet ja suorittaa lopullisen luokittelun. CNN:ien tehokkuus perustuu niiden hierarkkiseen tapaan muodostaa esityksiä datasta. Esimerkiksi kuvantunnistuksessa CNN oppii alemmilla tasoilla yksinkertaisia piirteitä, kuten reunoja ja rakenteita, kun taas ylemmät tasot yhdistelevät näitä monimutkaisemmiksi muodoiksi, kuten ympyröiksi tai suorakulmioiksi. Lopulta verkko tunnistaa kuvasta esimerkiksi polkupyörän tai ihmisen kasvot (Lecun ym., 2015; Saleem ym., 2022.).

Saleemin ym. (2022) mukaan konvoluutioneuroverkkoja on viime vuosina käytetty yhä enemmän kuvien tunnistamiseen ja luokitteluun sekä puheentunnistukseen. CNN:iä hyödynnetään myös videoiden sisällön tunnistamisessa ja luokittelussa, sillä videot voidaan käsitellä ruutu ruudulta yksittäisinä kuvina. Muun muassa Perez ym. (2017) sekä Mallman ym. (2020) ovat tutkineet syväoppimisen ja etenkin CNN:ien hyödyntämistä sensitiivisen sisällön, kuten pornografian tunnistamiseen ja suodattamiseen. Perez ym. (2017) huomauttavat, että pornograafinen sisältö on usein kuva tai videomuodossa, joka tekee CNN:istä hyvän työkalun niiden käsittelemiseen ja varmistamiseen, ettei tällainen sisältö päätyisi sille kuulumattomalle kohderyhmälle tai väärään

ympäristöön. Perez ym. (2017) yhdistivät syväoppimiseen staattista- ja liiketietoa, joka mahdollisti tarkemman määrittelyn siitä, mitä videolla tapahtuu. Tämä toimintatapa saavutti parempia tuloksia kuin aiemmin käytetyt tunnistuskeinot. Mallman ym. (2020) tutkivat CNN-pohjaista reaaliaikaista pornograafisen sisällön tunnistus ja sensuroimisohjelmaa, jossa ohjelma sensuroi pelkästään tarpeellisen osan tunnistetusta sisällöstä, kuten sukupuolielimet. Tämä oli heidän ratkaisunsa CNN:ien suorittamia vääriä positiivisia (ohjelma luokittelee normaalin kuvan pornograafiseksi) kohtaan. Väärien positiivisten aiheuttama haitta ei ole niin merkittävä ongelma heidän ohjelmalleen, koska se sensuroi vain osan sisällöstä, verrattuna siihen, että sisältö sensuroitaisiin kokonaan.

Konvoluutioneuroverkot ovat erinomainen tapa videomateriaalin käsittelemiseen, koska ne voivat käsitellä videomateriaalia kuvasarjoina, joista ne arvioivat jokaisen kuvan erikseen. Tämä prosessi vaatii kuitenkin paljon prosessointivoimaa ja valtavia määriä muistia toimiakseen, joka asettaa tarkat vaatimukset koneistolle, jolla ohjelmaa käytetään. Nämä vaatimukset tekevät videomateriaalin automaattisesta käsittelystä kallista. (Mallmann ym., 2020.)

## 3.3 Luonnollisen kielen käsittely

Sosiaalisen median alustojen sisältö on monimuotoista ja sitä esiintyy kuvien ja videoiden lisäksi paljon tekstimuodossa esimerkiksi kommentteina tai julkaisuina. Ihmiselle tämän sisällön tulkitseminen on helppoa, koska osaamme lukea tekstiä ja analysoida sen sisältöä. Luonnollisen kielen käsittelyn (engl. natural language processing, NLP) teknologia mahdollistaa koneille ihmisten puhutun ja kirjoitetun luonnollisen kielen oppimisen, ymmärtämisen ja luomisen (Li ym., 2021a). Li ym. (2021b) määrittelevät luonnollisen kielen käsittelyn tietojenkäsittelytieteen, tekoälyn ja kielitieteen monitieteiseksi alaksi, joka tutkii, miten koneita voidaan käyttää luonnollisen kielen ymmärtämiseen ja käsittelemiseen puheesta tai kirjoitetusta kielestä.

Collobertin ym. (2011) artikkelissa avataan luonnollisen kielen käsittelyn neljää perustoimintoa. Näistä ensimmäinen on puheosamerkintä (engl. part-of-speech tagging, POS), jossa lauseen jokaiselle sanalle annetaan kieliopillinen luokka, kuten substantiivi, verbi tai monikko. Toinen toiminto on ryhmittely (engl. chunking), jossa lauseen sanat jaotellaan niiden syntaktisen roolin perusteella. Sanat ryhmitellään peräkkäisiksi lausekkeiksi, kuten substantiivi- tai verbilausekkeiksi, mikä auttaa hahmottamaan lauseen rakennetta. Kolmas toiminto on nimettyjen yksiköiden tunnistus (engl. named entity recognition, NER), jossa luokitellaan lauseen jäseniä erilaisiin merkityksellisiin kategorioihin, kuten henkilönimiin (PERSON), paikkaan (LOCATION) tai aikaan (DATE). Neljäs perustoiminto, jota he käsittelevät on semanttisten roolien luokittelu (engl. semantic role labeling, SRL). Se keskittyy tunnistamaan lauseen eri osien rooleja suhteessa verbiin. Näitä rooleja ovat

esimerkiksi tekijä, tekemisen kohde, aika ja sijainti. Nämä perustoimet itsessään ovat vain sanojen ja lauseenosien luokittelua. Collobertin ym. (2011) mukaan perinteinen luonnollisen kielen käsittelyn lähestymistapa on lauseesta eriteltyjen ominaisuuksien (kuten neljässä perustoiminnossa tapahtuu) syöttäminen luokittelualgoritmiin, josta painottamalla eri ominaisuuksia saavutetaan haluttu lopputulos algoritmin tehtävälle. He suosittelevat vahvasti (syvien) neuroverkkojen hyödyntämistä luonnollisen kielen käsittelyn kanssa, jotta toimintaa saataisiin automatisoitua ja manuaaliselta prosessoinnilta vältyttäisiin. Syvät neuroverkot soveltuvat heidän mukaansa kaikille luonnollisen kielen käsittelyn tehtäville ja ne poistavat perinteisen lähestymistavan tarpeen tehtäväkohtaiselle ominaisuuksien valinnalle.

Gambäck ja Sikdar (2017) havaitsivat konvoluutioneuroverkkojen tehokkuuden luonnollisen kielen käsittelyn tehtävissä, tässä tapauksessa luokittelussa. Konvoluutioneuroverkot suoriutuivat vihapuheen tunnistuksessa tarkemmin kuin perinteinen ohjatun oppimisen luokittelumenetelmä (logistinen regressio). Neuroverkot tekivät vähemmän virheellisiä tunnistuksia ja tuottivat tarkempia tuloksia. Logistinen regressio sen sijaan tunnisti vihapuhetta laajemmin, mutta siihen sisältyi enemmän virheitä.

Sosiaalisen median sisällön valtavan määrän vuoksi luonnollisen kielen käsittelyn menetelmiin perustuvat automaattiset ohjelmat ovat välttämättömiä haitallisen kielen tunnistamiselle (Plaza-del-Arco ym., 2021). Paulin ja Sahan (2022) mukaan luonnollisen kielen käsittely on osoittanut potentiaalinsa haitallisten sisältöjen, kuten vihapuheen, häirinnän, kiusaamisen ja muun haitallisen kielen tunnistamistehtävissä. Luonnollisen kielen käsittelyä ja sen eri toimintoja hyödynnetään monissa haitallisen sisällön tunnistamisalgoritmeissa.

Muun muassa Cartwrightin ym. (2022) tutkimuksessa he vertasivat eri koneoppimisalgoritmeja ja syviä neuroverkkoja disinformaation ja valeuutisten havaitsemiseen X:ssä tehdyistä postauksista ja Facebook-julkaisuista. Ennen julkaisujen syöttämistä algoritmeille, he käyttivät luonnollisen kielen käsittelyyn erikoistunutta koneoppisohjelmaa sisällön tulkitsemiseen. He saivat luonnollisen kielen käsittelyn avulla jokaisesta tekstistä 126 lisäominaisuutta ja tilastoja, jotka lisättiin dataan ennen tunnistamisalgoritmeille syöttämistä. Lisätilastojen ja ominaisuuksien avulla algoritmeille mahdollistettiin tehokkaampi toiminta ja tarkempi tunnistaminen.

Paul ja Saha (2022) tutkivat Googlen kehittämän syväoppimismallin BERT (Bidirectional Encoder Representations from Transformers) (ks. esim. Devlin ym., 2019) hyödyntämistä kiusaamisen (engl. cyberbullying) tunnistamiseen. BERT mahdollistaa sanan kontekstin ymmärtämisen lauseessa tutkien sanaa sitä ennen ja jälkeen. BERT on etukäteen merkkaamattomalla datalla opetettu malli,

jota hienosäädetään tehtävään sopivaksi merkatun datan avulla. (Paul & Saha, 2022.) Heidän mukaansa BERT on osoittanut erinomaisia tuloksia luonnollisen kielen ymmärtämisessä ja suoriutui hyvin heidän tutkimuksessaan kiusaamisen tunnistamisesta. He aikovat laajentaa sen käyttöä sosiaalisessa mediassa automaattiseen kiusaamisen tunnistamiseen yhdistämällä siihen ulkoista tietoa sekä monimuotoista dataa, kuten kuvia, videoita ja ääntä.

# 4 Haasteita ja kritiikkiä tekoälyn hyödyntämisessä

## 4.1 Ihmisten muuttunut rooli sisällönhallinnassa

Miljardien käyttäjien ja valtavan sisältömäärän takia proaktiivinen sisällön arviointi on lähes mahdoton tehtävä, minkä vuoksi monilla alustoilla sisältö arvioidaan vasta julkaisemisen jälkeen. Julkaisemisen jälkeinen arviointi tapahtuu käytännössä vain pienissä osissa kerrallaan, mikä on hyvin rajallista suhteessa julkaistun sisällön määrään. Alustojen tulisi kyetä havaitsemaan haitallinen sisältö nopeasti, mutta se ei ole tällä tavalla mahdollista. Ennen tekoälyratkaisujen yleistymistä monet alustat alkoivat hyödyntämään käyttäjiään haitallisten sisältöjen tunnistamisessa. Käyttäjät voivat ilmiantaa haitalliseksi kokemansa sisällön ja siten auttaa alustoja tunnistamaan ongelmallista sisältöä nopeammin ja laajemmin koko alustan mittakaavassa. Käyttäjien hyödyntäminen tehostaa haitallisen sisällön tunnistamista, ja samalla heille tarjoutuu mahdollisuus vaikuttaa kohdatessaan haittaa ja väärinkäyttöä. Jotkut alustat pyytävät käyttäjiä tarkentamaan ilmiannon syytä, jotta sisältö voidaan luokitella tarkemmin. Usein tämä tapahtuu valitsemalla alustan tarjoamista vaihtoehdoista, joita voivat olla esimerkiksi lasten pahoinpitely, seksuaalinen sisältö tai väkivalta. Käyttäjien hyödyntämiseen liittyy kuitenkin haasteita, vaikka se on yhä keskeinen menetelmä monilla alustoilla. Käyttäjät ovat yksilöitä, jotka kokevat haitallisuutta eri tavoin monista eri syistä (ks. luku 2), ja heidän arviointeihinsa voivat vaikuttaa esimerkiksi ristiriidat toisten käyttäjien kanssa. Tämän vuoksi käyttäjän kokemus sisällön haitallisuudesta ei välttämättä vastaa alustan linjausta. Ilmiantoja voidaan myös strategisesti väärinkäyttää esimerkiksi poliittisten tai sosiaalisten päämäärien saavuttamiseksi. (Gillespie, 2018, s. 87–96.)

Elkin-Korenin (2020) mukaan pelkästään ihmisten suorittamaan sisällön arviointiin ei voida enää tukeutua käyttäjien luoman sisällön eksponentiaalisen kasvun takia. Tähän on kuitenkin vaihtelevia lähestymistapoja alustojen välillä. Jotkut alustat tukeutuvat täysin ihmismoderaattoreihin, jotka arvioivat käyttäjien ilmiantamia julkaisuja, kun taas toiset alustat hyödyntävät kehittyneempiä menetelmiä, kuten koneoppimista ja luonnollisen kielen käsittelyä haitallisen sisällön tunnistamiseen ja käsittelyyn. (Wang & Kim, 2023.)

Ihmisten suorittamia sisällönhallintamenetelmiä on korvattu ja korvataan yhä enemmän automatisoiduilla ratkaisuilla. Tekoälyä hyödyntävien järjestelmien avulla sisältöä, joka on esimerkiksi laitonta tai alustan sääntöjen vastaista, voidaan tunnistaa, poistaa, estää tai suodattaa jo ennen julkaisua. Automatisoinnin tehokkuus on yksi pääsyy alustojen sisällönhallinnassa tekoälyn hyödyntämiselle. Erityisesti koneoppimisalgoritmeja ja muita kehittyneitä menetelmiä

hyödynnetään monien alustojen hallinnointitehtävissä. Tekoäly on muuttanut sisällönhallinnan lähestymistapaa, sillä sen päätöksentekoprosessit perustuvat suurien datamäärien analysointiin ja niistä havaittuihin yhteyksiin ja ennusteisiin. (Elkin-Koren, 2020.)

Tekoälyn mahdollistama työläiden ja rutiininomaisten tehtävien automatisointi ja lisääntynyt tehokkuus eivät poista ihmistyövoiman tarvetta. Gillespien (2018) mukaan laajanmittainen sisällönhallinta edellyttää edelleen merkittävää määrää ihmistyötä. Ihmisen rooli ja tehtävät voivat vaihdella suuresti. Työtehtävät vaihtelevat alustojen eri sivujen tai moderointitiimien johtajista joukkotyöntekijöihin, itse alustan käyttäjiin, jotka hyödyntävät arvostelu- ja ilmiantotyökaluja, tai ulkopuolisiin organisaatioihin ja yrityksiin, jotka ovat palkattu avustamaan sisällönhallinnassa. Tekoälyn käyttöönotto on kuitenkin muuttanut osittain ihmisten. Vaikka tekoäly kykenee käsittelemään ja luokittelemaan valtavia määriä sisältöä, vaikeammat tapaukset jäävät edelleen ihmisten arvioitavaksi. Ihmisten tehtävänä on ratkaista tilanteita, jotka vaativat kontekstin ymmärtämistä, harkintaa ja empatiaa. Klonick (2018) kertoo esimerkiksi Facebookilla olleen kolmitasoinen ihmismoderaattorijärjestelmä, jossa kolmannen tason moderaattorit arvioivat päivittäin tekoälyn tai käyttäjien ilmiantamaa sisältöä. Toisen tason moderaattorit valvovat kolmannen tason toimintaa ja käsittelevät priorisoitua tai eskaloitua sisältöä. Ensimmäisen tason moderaattorit ovat hänen mukaansa yleensä asianajajia ja muita alustojen sääntöjen laatijoita.

Ihmisen roolin merkitys haitallisen sisällön hallinnassa korostui vuonna 2020, kun monet sosiaalisen median alustat siirtyivät lähes täysin automatisoituun sisällönhallintaan koronaviruspandemian seurauksena. Useat alustat joutuivat myöhemmin pahoittelemaan tekoälyn tekemiä virheitä, sillä tekoäly ei kyennyt arvioimaan kontekstia ihmistiimien tavoin. Ihmiset vastaavat sisällönhallinnan lisäksi myös tekoälyjärjestelmien suunnittelemisesta, testaamisesta, ylläpidosta sekä niiden käytön ja suorituskyvyn arvioimisesta. Automatisoidut tunnistusjärjestelmät ja niiden algoritmit ovat ihmisten kehittämiä, ja niiden tehokas toiminta vaatii laadukasta opetusdataa ja valmisteluja, joista ihmiset ovat vastuussa. (Gillespie, 2018; 2020.)

Ihmisen ja tekoälyn yhteistyö on osoittautunut yhdeksi tehokkaimmista keinoista haitallisen sisällön hallinnassa. Tekoäly pystyy käsittelemään ja luokittelemaan suuria määriä sisältöä nopeasti, kun taas ihmismoderaattorit keskittyvät monimutkaisempiin tapauksiin, jotka vaativat syvällisempää kontekstin ymmärtämistä, harkintaa ja empatiaa. (Molina & Sundar, 2022; Wang & Kim, 2023.)

Myös Gillespie (2020) tukee ihmisen ja tekoälyn yhteistyötä sisällönhallinnassa. Hän esittää kahta lähestymistapaa, joista ensimmäisessä tekoälyä hyödynnetään ihmistiimien tukena niiden korvaamisen sijaan, Molinan ja Sundarin (2022) sekä Wangin ja Kimin (2023) mukaisesti. Toinen

lähestymistapa perustuu erityisen haitallisten sisältöjen, kuten lasten hyväksikäytön ja graafisen väkivallan tunnistamisen optimointiin ja automaattiseen poistamiseen. Näin sisällönhallintaa suorittavat työntekijät eivät joutuisi altistumaan psyykkisesti kuormittavalle materiaalille. Tämä lähestymistapa kuitenkin tulisi vaatimaan käyttäjiltä ymmärtämistä, koska väärien positiivisten määrä tulisi todennäköisesti kasvamaan automatisoinnin myötä, mutta se suojelisi sisällönhallintaa suorittavia ihmisiä erityisen haitallisen sisällön aiheuttamalta rasitukselta.

Vaikka tekoäly on muuttanut sisällönhallinnan lähestymistapaa ja tehostanut sen eri osa-alueita, ihmisten rooli on edelleen keskeinen. Ihmisten tehtävät ovat laajentuneet kattamaan tekoälyyn liittyviä osa-alueita ja ihmisten suorittama moderointityö tekoälyn ohella on edelleen erittäin tärkeää. Tehokkaimpia tapoja haitallisten sisältöjen hallinnoimiseen on hyödyntää tekoälyn ja ihmisen yhteistyötä siten, että molempien vahvuudet tukevat toisiaan.

## 4.2 Tekoälyn teknologisia ja käytännön haasteita

Monilla sosiaalisen median alustoilla käytetään laajasti automaattista sisällönhallintaa tekoälyratkaisujen yleistyttyä, mutta tekoäly kohtaa toiminnassaan rajoitteita ja haasteita. Väärin toimiessa sisällönhallintaohjelmat voivat rajoittaa käyttäjien ilmaisunvapautta sensuroimalla ja suodattamalla hyväksyttävää sisältöä tai aiheuttaa haittaa jättämällä haitallista sisältöä tunnistamatta. (Marsoof ym., 2023.)

Väärin suodatettujen ja sensuroitujen tai sensuroimattomien julkaisujen taustalla on ollut alkukantaisten suodatusohjelmien haasteita. Monet alustojen hyödyntämät sisällönsuodatusohjelmat ovat perustuneen IP-osoitteiden, verkkosivujen osoitteiden tai avainsanojen käyttöön. Marsoofin ym. (2023) mukaan tällainen lähestymistapa voi johtaa hyväksytyn sisällön suodattamiseen ja näin liialliseen sensuuriin. Erityisesti avainsanojen ja tekstin arvioinnin kanssa väärien negatiivisten ja väärien positiivisten riski on todellinen.

Väärin suodattaminen voi johtua monesta syystä. Joillain sanoilla on monia merkityksiä, joita tekoäly ei tunnista ja sanojen käyttäminen tietyssä kontekstissa voi olla joko soveliasta tai sääntöjen vastaista. Esimerkiksi ihmisten sukupuolielimistä puhuminen tietyssä kontekstissa, kuten vihapuheessa voi olla alustojen sääntöjen vastaista, mutta opetuksessa käytettynä se on sallittua. (Gillespie, 2018, s. 99.)

Haitallisen sisällön hallinnassa, Davidsonin ym. (2017) tapauksessa vihapuheen tunnistamisessa, haasteeksi tuotiin esille vihapuheen erottaminen loukkaavasta kielenkäytöstä. Automatisoidut vihanhallinnan tunnistusmenetelmät luokittelivat loukkaavaa kieltä sisältäneitä postaukset

vihapuheeksi, vaikka ne eivät täyttäneet vihapuheen määritelmää. Postaukset sisälsivät termejä, jotka olivat mm. rasistisia ja homofobisia, mutta niitä käytettiin kontekstissa, joka ei täyttänyt vihapuheen määritelmää, esimerkiksi rap-kappaleen sanojen toistamista. Automaattiset menetelmät myös jättivät vihapuhetta sisältäviä postauksia tunnistamatta, koska ne eivät sisältäneet kirosanoja, rasistisia ja homofobisia termejä tai suoranaista vihaa. (Davidson ym., 2017.)

Nykyään käytetään myös kehittyneempiä tekoälymenetelmiä, jotka hyödyntävät muun muassa syväoppimista ja luonnollisen kielen käsittelyä toiminnassaan ja kykenevät tunnistamaan kontekstia tekstissä. Niiden avulla haitallisen sisällön hallinnointia voidaan toteuttaa tehokkaammin ja tarkemmin. (Marsoof ym., 2023.) Esimerkiksi Shilpashreen ja Ashokan (2024) kehittämä ominaisuuspainotteinen syväoppimisrakenne, joka hyödyntää konvoluutioneuroverkkoa ja sanavektorien upottamistekniikkaa (engl. word embedding), kykenee tunnistamaan kontekstiriippuvaista vihapuhetta huomattavasti perinteisiä menetelmiä paremmin. He tuovat tutkimuksessaan kulttuuristen erojen ja kontekstin ymmärtämisen lisäksi toisenkin haasteen, joka tuo haasteita automaattisessa vihapuheen tunnistamisessa. Heidän mukaansa kokeellisissa testauksissa hyvin suoriutuvat, kehittyneet tekoälyratkaisut tulevat kohtaamaan runsaasti haasteita reaaliaikaisessa käytössä. Internetin ja sosiaalisen median käyttäjät keksivät ja luovat koko ajan uusia tyylejä vihapuheen ja loukkaavan kielen ilmaisemiselle. Vihapuheen tunnistaminen vaikeutuu, koska käytetyt tekoälymenetelmät eivät osaa etsiä asioita, joita niille ei ole opetettu. Tekoälyn käyttämää opetusdataa tulee siis jatkuvasti päivittää kattamaan näitä uusia termejä ja tapoja ilmaista vihapuhetta. Marsoof ym. (2023) korostavat myös, että erityisesti luonnollisen kielen käsittelyn työkalut toimivat parhaiten ympäristössä, joka on mahdollisimman samankaltainen mallille syötetyn opetusdatan kanssa.

Jotkut sosiaalisen median alustat ovat ottaneet käyttöönsä NSFW-merkkaukset (engl. not safe for work), joilla voidaan merkitä haitallista tai arkaluontoista sisältöä, jota et haluaisi katsoa esimerkiksi julkisessa tilassa tai työpaikalla. Näiden merkkausten takana voi olla esimerkiksi pornograafista materiaalia tai graafista väkivaltaa. NSFW-merkkausten avulla alustat voivat hallita tällaista sisältöä ilman, että sitä tarvitsee poistaa kokonaan. Tämä mahdollistaa alustoille toiminnan laajempien käyttäjäpreferenssien palvelemiseksi. Halukkaat voivat nähdä merkatun sisällön, kun taas käyttäjät, jotka eivät halua altistua sille voivat välttää sitä. (Gillespie, 2018). NSFW-merkatun sisällön saa suodatettua kokonaan pois monien alustojen asetuksista, jolloin sisältöön ei tule törmäämään ollenkaan. Vaikka suodatus ei ole päällä, sisältöä ei näe suoraan. Nähdäkseen NSFW-merkatun julkaisun, käyttäjän tulee erikseen hyväksyä suostumus sisällön katsomiseen. Tämä on

monella alustalla oletusasetuksena. Nämä merkkaukset ovat käytössä useilla sosiaalisen median alustoilla, kuten X:ssä, Instagramissa ja Redditissä.

Tekoälyn antamat väärät positiiviset ja väärät negatiiviset voivat myös johtua myös sille syötetystä opetusdatasta. Opetusdatan rooli oppimispohjaisen tekoälyn toiminnassa on välttämätöntä, sillä kaikki mitä tekoäly oppii pohjautuu sille syötettyyn opetusdataan (Marsoof ym., 2023). Tästä syystä opetusdatan painoarvo automaattisten sisällönhallinnan ohjelmien ja algoritmien toiminnassa on merkittävä ja se tuo mukanaan erilaisia haasteita. Opetusdatan valinnalla on tärkeä rooli näiden ohjelmien toiminnan tarkkuuden kanssa. Opetusdataa tulee jatkuvasti päivittää ja sen tulisi olla mahdollisimman samankaltaista todellisen toimintaympäristön kanssa, jotta sitä käyttävät automaattiset sisällönhallinnan ohjelmat toimisivat mahdollisimman tehokkaasti. Jos jotain tällaista ohjelmaa haluttaisiin käyttää esimerkiksi espanjankielisellä alueella, sille tulisi syöttää opetusdataa, joka sisältäisi espanjankielisiä sisältöä. Jopa nykyaikaisen teknologian avulla on haasteita luoda tunnistusohjelmia, jotka toimisivat hyvin eri sivuilla, kielillä, kulttuureissa tai aihepiireissä. (Marsoof ym., 2023.)

Toinen opetusdataan liittyvä haaste Marsoofin ym. (2023) mukaan on tekoälylle syötettävän opetusdatan valitsijat. Koska ihmisen on valikoitava tekoälylle syötettävä data, mahdollisuus inhimilliselle erheelle on olemassa. Ihmisen rooli opetusdatan kanssa tuo myös puolueellisuuden ongelman. Opetusdatan valitsijalla voi olla henkilökohtaisia ennakkoluuloja ja asenteita, jotka voivat vaikuttaa tekoälylle syötetyn datan valitsemiseen. Esimerkiksi ohjatussa oppimisessa opetusdata merkataan, jolloin se sisältää syötteet ja niiden oikeat vastaukset. Jos opetusdatan merkkaajalla on henkilökohtaisia ennakkoluuloja esimerkiksi tiettyä ihmisryhmää kohtaan, se voi siirtyä opetusdataan tahallisesti tai tahattomasti. Ihminen ja ihmisen heikkoudet, jotka voivat esiintyä muun muassa tahattomana puolueellisuutena, tunnustetaan kolmanneksi opetusdataan liittyväksi haasteeksi. (Marsoof ym., 2023)

Toinen merkittävä tekoälyn toimintaan liittyvä teknologinen haaste on sen niin sanottu hallusinointi. Se tarkoittaa tilannetta, jossa tekoäly tuottaa virheellistä tai täysin keksittyä tietoa ja esittää sen totuutena ja uskottavalla tavalla. Hallusinointia esiintyy erityisesti suurilla kielimalleilla (engl. large language model, LLM). (Jones, 2025.) Sisällönhallinnassa käytettävä tekoälymenetelmä voi hallusinoidessaan arvioida julkaistun sisällön väärin, mikä johtaa hyväksyttävän sisällön poistamiseen tai haitallisen sisällön huomaamatta jättämiseen. Hallusinointia ei voida täysin estää, mutta sen vähentämiseen on useita keinoja, kuten kielimallien kouluttaminen laajemmalla ja laadukkaammalla opetusdatalla. Lisäksi mallien pidempi koulutusprosessi ja sen parametrien

lisääminen voivat vähentää hallusinoinnin esiintymistä. (Jones, 2025.) Hallusinointi korostaa tarvetta ihmisen roolille sisällönhallinnassa, jotta sen vaikutukset pystyttäisiin minimoimaan ja tekemään päätöksistä luotettavia.

### 4.3 Kritiikkiä tekoälyn hyödyntämistä kohtaan

Tarkoituksenmukainen puolueellisuus ja ennakkoluulot tekoälylle syötetyssä opetusdatassa herättävät eettisiä kysymyksiä tekoälyn toimintaan liittyen. Esimerkiksi kiinalaisten valmistama uusi suuri kielimalli DeepSeek ei kykene kertomaan aiheista, jotka ovat Kiinan hallituksen mielestä arkoja, kuten Tiananmenin aukion tapahtumat vuonna 1989 tai Taiwanin itsenäisyys (Smith, 2025). Tekoäly voi toimia puolueellisesti myös ilman ihmisen tarkoituksenmukaista tai tahatonta toimintaa, sillä puolueellisuutta ja ennakkoluuloja voi nousta myös datasta, jota tekoälylle syötetään (Marsoof ym., 2023; Sipior ym., 2024).

Datasta noussutta puolueellisuutta on havaittu esimerkiksi Googlen konekäännöspalvelussa. Google Translate on tekoälyä hyödyntävä työkalu, joka kääntää tekstiä kielten välillä. Joissakin kielissä, kuten unkarissa ja suomessa, käytetään sukupuolineutraaleja pronomineja. Kun näitä kieliä käännetään kieleen, jossa pronominit ovat sukupuolittuneita (esim. englanti), käännöstyökalu joutuu päättelemään, kumpaa pronominia käyttää. (Prates ym., 2020.)

Pratesin ym. (2020) tutkimuksessa havaittiin, että Google Translate tekee käännöksissä sukupuoleen liittyviä oletuksia, jotka heijastavat stereotypioita. Pronominien jakautuminen ei ollut tasapuolista, vaan maskuliinisia oletuksia esiintyi yleisesti ja ne korostuivat erityisesti aloilla, joihin liitetään yleisesti sukupuolistereotypioita, kuten matemaattis-teknisillä aloilla. Tutkimuksessa havaittiin myös, että sukupuolioletukset eivät rajoitu työtehtäviin. Myös henkilöä kuvaavilla adjektiiveilla oli vaikutusta siihen, käytettiinkö käännöksessä mies- vai naispronominia. Esimerkiksi sanat "ujo" ja "haluttava" johtivat useammin naispronominien käyttöön, kun taas sanat "syyllinen" ja "julma" yhdistettiin useammin miespronomineihin. Tutkijat arvelevat, että nämä vinoumat johtuvat suurelta osin opetusdatasta, jota Google Translate käyttää. Käännöstyökalun taustalla oleva tekoäly oppii sadoista miljoonista verkosta kerätyistä tekstiaineistoista, jotka heijastavat ihmisten tuottamaa kieltä, ja samalla siinä esiintyviä sukupuolistereotypioita. (Prates ym., 2020.)

Leen ja Chan (2024) mukaan läpinäkyvyyden ja luottamuksen välistä positiivista suhdetta on tutkittu monessa eri kontekstissa ja on huomattu, että päätöksentekoon liittyvän informaation jakaminen on kasvattanut luottamusta siitä, että tehty päätös on ollut oikea. Heidän mukaansa läpinäkyvyydellä on merkittävä rooli tekoälyn toimintaan luottamisessa. Monet sisällön

hallinnoimisessa käytettävät läpinäkymättömät tekoälyjärjestelmät herättävät kysymyksiä toiminnan läpinäkyvyyteen liittyen (Gorwa ym., 2020). Gorwan ym. (2020) mukaan automatisoitujen sisällönhallintajärjestelmien tekemien päätösten perusteiden selvittäminen on haastavaa, erityisesti kun ilmiantojen tai poistojen tarkkoja kriteerejä ei tunneta. Jos käyttäjät eivät ymmärrä, eikä heille kerrota miksi tekoälyn tekemä päätökseen päädyttiin, heidän luottamuksensa tekoälyn luotettavaan toimintaan voi kärsiä. Beduén ja Fritzschen (2022) mukaan käyttäjien luottamusta tekoälyä kohtaan voidaan parantaa tekemällä toiminnasta läpinäkyvämpää. Se voisi tapahtua antamalla lisätietoa ja selityksiä algoritmien tekemistä päätöksistä ja tarjoamalla visualisointeja niiden tueksi. Myers Westin (2018) mukaan pienillä muutoksilla sisällönhallintajärjestelmissä, kuten tarjoamalla ietoa sisällönhallinnan prosesseista ja selittämällä, miten käyttäjä on rikkonut alustan sääntöjä, voitaisiin kannustaa käyttäjiä itse valvomaan omaa toimintaansa, mikä voisi vähentää moderoitavan sisällön määrää.

Tekoälyn ja algoritmien läpinäkyvyyden lisääminen julkiselle yleisölle kasvattaa myös sosiaalisen median alustojen vastuuta, kun niiden toimia voidaan tarkkailla. Jos sosiaalisen median alustat toimivat ”salaisesti”, niiden luotettavuus heikkenee, eikä toiminnan eettisyydestä ole takuita. Esimerkiksi vuonna 2021 Facebookin entinen työntekijä Frances Haugen vuoti tuhansia sisäisiä dokumentteja, jotka tunnetaan nimellä Facebook Papers tai Facebook Files. Dokumentit paljastivat, että Facebook muutti algoritmejaan vuosina 2017–2018, siten että ne lisäsivät äärimmäisen, kuten haitallisen ja provosoivan sisällön näkyvyyttä ajan mittaan, jotta alustalla tapahtuvaa vuorovaikutusta saataisiin lisättyä. Alustan algoritmit toimivat läpinäkymättömästi, eikä niiden vaikutuksia olisi todennäköisesti paljastettu ilman sisäpiiriläisen todistusta. Haitallinen sisältö pääsi leviämään alustalla, koska algoritmit suosivat sisältöä, joka herätti tunteita, kuten vihaa ja pelkoa. Esimerkiksi Myanmarissa hallinnon väkivaltaiset kampanjat rohingya-vähemmistöä vastaan tapahtuivat suurelta osin Facebookissa. Alusta mahdollisti äärisisällön ja misinformaation leviämisen ilman merkittäviä rajoituksia tai asiaan puuttumista. Facebookin omien tutkimusten mukaan yhtiö tiesi ongelmista, mutta ei tehnyt riittäviä toimia niiden ratkaisemiseksi. (Olesen, 2025.)

Tekoälyn toiminnan läpinäkyvyyteen liittyy läheisesti vastuu. Herää kysymys siitä, kuka kantaa vastuun tekoälymenetelmien toimiessa väärin sisällönhallinnassa. Marsoof ym. (2023) ehdottavat tähän ratkaisuksi alustoille sitä, että organisaatiot antaisivat kontaktihenkilön, johon ottaa yhteyttä, kun käyttäjät kokevat tulleensa väärin kohdelluiksi automaattisten sisällönhallinnan järjestelmien toimesta. Yhteyshenkilö voisi olla ns. tekoälyvastaava, jolla tulisi olla riittävä osaaminen ja valtuudet käsitellä tapauksia asianmukaisesti ja tarjota lisätietoa päätöksen taustoista ja mahdollistaa

oikaisuprosessi, jos siihen on syytä. Tämä käytäntö lisäisi alustojen tilivelvollisuutta tekoälyn toiminnasta ja antaisi käyttäjille kanavan tulla kuulluksi ja vaikuttaa, jos he ovat kokeneet tulleensa epäoikeudenmukaisesti kohdelluiksi. (Marsoof ym., 2023.)

Käytännössä tällaisen järjestelmän toteuttaminen olisi kuitenkin haasteellista sosiaalisen median alustoilla, joilla on miljoonia käyttäjiä. Päivittäisten yhteydenottojen määrä voisi nousta tuhansiin, mikä tekisi yksittäisten kontaktihenkilöiden käytön mahdottomaksi. Tapauksien läpikäynti ja viesteihin yhteydenottoihin vastaaminen vievät aikaa, joten tämän toteuttaminen vaatisi erillisiä kontaktitiimejä, jotka käsittelisivät tilanteita tehokkaasi. Tämän prosessin automatisointi ei olisi suositeltavaa, koska tekoälyn tekemien virheiden käsitteleminen toisella tekoälyllä voisi heikentää käyttäjien luottamusta entisestään ja tapausten ratkaiseminen voi vaatia ihmisen ymmärrystä ja tilannetajua. Kontaktihenkilöiden tai tiimien rooli kuvastaa kuitenkin hyvin uudenlaista roolia, joka ihmisellä voisi olla tekoälyn rinnalla sosiaalisen median alustojen sisällönhallinnassa.

# 5 Yhteenveto ja johtopäätökset

Tässä tutkielmassa tutkittiin, kuinka eri tekoälymenetelmiä voidaan hyödyntää sosiaalisen median alustojen haitallisten sisältöjen hallinnassa. Teknologian kehittyminen on mahdollistanut ihmisille tehokkaamman tavan kommunikoida toistensa kanssa. Sosiaalisen median alustojen tarjoama vaivaton kanssakäyminen ja itsensä ilmaiseminen on kasvattanut niiden suosiota räjähdysmäisesti. Ajan myötä alustoista tuli entistä helppokäyttöisempiä, kun niiden käyttöliittymät kehittyivät ja mobiililaitteiden käytön lisääntyminen mahdollisti niiden käyttämisen melkein kaikkialla. Uusia alustoja tulee jatkuvasti ja monet alustat ovat erikoistuneet johonkin sisältötyyppiin, kuten videoihin tai kuviin.

Tutkielman toisessa luvussa käsiteltiin haitallista sisältöä ja erilaisia sosiaalisen median alustatyyppejä sekä niiden piirteitä. Sosiaalinen media voidaan määritellä joukoksi sovelluksia, jotka hyödyntävät verkkopohjaista teknologiaa ja mobiililaitteita ja mahdollistavat käyttäjille sisällön luomisen, jakamisen ja muokkaamisen. Sosiaalisen median alustatyyppejä on monia, mutta tässä tutkielmassa ne jaoteltiin yhteisöpalveluihin, keskustelualustoihin sekä video- ja kuva-alustoihin. Sosiaalisen median alustoilla on miljardeja käyttäjiä ja niillä julkaistaan koko ajan valtavia määriä uutta sisältöä. Käyttäjien luoman sisällön määrän kasvun myötä myös haitallisen sisällön määrä on kasvanut. Haitallisella sisällöllä katsotaan yleisesti olevan negatiivinen vaikutus yksilöihin ja yhteisöihin, ja se voi johtua monesta eri syystä. Haitalliselle sisällölle altistuminen voi aiheuttaa käyttäjille ahdistusta, masennusta, mielenterveyden heikentymistä sekä luoda negatiivisen sosiaalisen median käyttökokemuksen. Haitallista sisältöä esiintyy monessa muodossa, kuten tekstinä, kuvina, videoina tai äänitteinä. Tässä tutkielmassa sisältö määritellään haitalliseksi, jos se kuuluu luvussa 2.1 esitettyihin aihealueisiin. Yleistä haitallisen sisällön määritelmää on haastavaa tehdä, koska haitallisuus on subjektiivista. Ihmisten näkemyksiin haitallisuudesta vaikuttaa erot kulttuureissa, uskonnossa, lainsäädännössä, sekä henkilökohtaisissa taustoissa ja näkemyksissä.

Luvussa kolme käsitellään tekoälyn alalajeja, jotka ovat sisällönhallinnan näkökulmasta keskeisiä. Näitä ovat koneoppiminen ja sen alalajit, syväoppiminen ja neuroverkot sekä luonnollisen kielen käsittely. Tässä luvussa avataan jokaisen alalajin toimintatapaa ja käsitellään kuinka niitä on hyödynnetty haitallisten sisältöjen hallinnassa. Tässä luvussa vastataan ensimmäiseen tutkimuskysymykseen: ”*Miten tekoälyä hyödynnetään haitallisten sisältöjen hallinnassa sosiaalisen median alustoilla?*”

Tekoälyä hyödynnetään haitallisten sisältöjen hallinnassa sosiaalisen median alustoilla ensisijaisesti sisällön automaattisessa tunnistamisessa, luokittelussa ja poistamisessa. Automaattinen sisällönhallinta hyödyntää muun muassa koneoppimista, luonnollisen kielen käsittelyä ja syväoppimista tunnistaakseen esimerkiksi vihapuhetta, väkivaltaista kuvamateriaalia tai väärää tietoa. Tekoäly mahdollistaa rutiininomaisten tehtävien, kuten roskapostin tai selkeästi sääntöjä rikkovan sisällön suodatuksen automatisoinnin. Tämä vapauttaa ihmismoderaattoreille aikaa vaikeampien ja tulkintaa vaativien tapausten käsittelyyn. Tässä luvussa vastataan myös toiseen tutkimuskysymykseen: ”*Miten tekoäly toimii haitallisen sisällön hallinnassa?”*

Tekoäly toimii haitallisen sisällön hallinnassa erityisesti analysoimalla, luokittelemalla ja suodattamalla suuria määriä käyttäjien tuottamaa sisältöä. Tekoälyjärjestelmät hyödyntävät koneoppimista ja syväoppimista oppiakseen havaitsemaan haitallisen sisällön piirteitä suurista datajoukoista. Eri tekoälymenetelmien avulla voidaan tunnistaa haitallista sisältöä, kuten vihapuhetta, häirintää tai disinformaatiota huomattavasti nopeammin ja laajamittaisemmin, kuin ihmisten suorittamalla valvonnalla olisi mahdollista.

Monet automaattiset järjestelmät käyttävät haitallisen sisällön tunnistamiseen luokittelualgoritmeja. Näille algoritmeille syötetään esimerkkidataa, jonka avulla ne oppivat tunnistamaan haitallisen sisällön ominaispiirteitä uusista syötteistä. Syväoppimismenetelmät ovat edistyneempiä koneoppimisen muotoja, ja ne hyödyntävät neuroverkkorakenteita, joiden avulla voidaan käsitellä esimerkiksi raakadataa ja tunnistaa siitä merkityksellisiä piirteitä. Konvoluutioneuroverkot ovat erityisen tehokkaita visuaalisen datan, kuten kuvien ja videoiden käsittelyssä. Niitä käytetään paljon kuvasisällön automaattisessa valvonnassa. Luonnollisen kielen käsittelyn menetelmät mahdollistavat tekstisisällön käsittelyn ja lisätiedon tarjoamisen algoritmeille. Syväoppimista ja luonnollisen kielen käsittelyä yhdistämällä voidaan onnistua tulkitsemaan kontekstia tekstisisällöistä. Tekoälyratkaisut voivat vaihdella yksinkertaisista sääntöpohjaisista malleista monimutkaisiin järjestelmiin, jotka yhdistävät useita eri menetelmiä tehokkaaksi kokonaisuudeksi sisällönhallinnan tueksi.

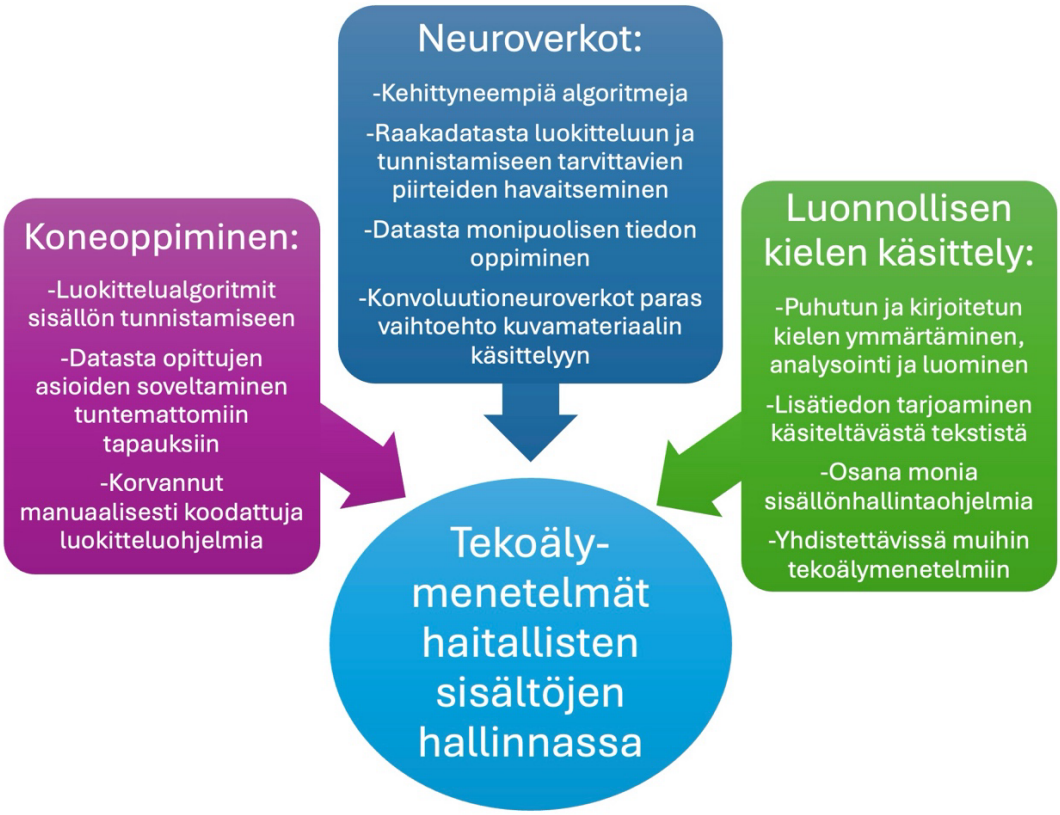
Neljännessä luvussa käsiteltiin ihmisen muuttunutta roolia sisällönhallinnassa sekä tekoälyn kohtaamia haasteita ja kritiikkiä tekoälyä kohtaan. Tekoälyn automatisoidessa sisällönhallinnan tehtäviä, ihmisen rooli on muuttunut. Ihmisen rutiininomaiset tehtävät ovat vähentyneet, mutta ihmismoderaattorit käsittelevät edelleen käyttäjien ja tekoälyn ilmiantamaa sisältöä. Ihmisen ja tekoälyn yhteistyö on todettu tehokkaaksi tavaksi toimia. Tekoäly pystyy käsittelemään ja luokittelemaan suuria määriä sisältöä nopeasti, kun taas ihmismoderaattorit keskittyvät

monimutkaisempiin tapauksiin, jotka vaativat syvällisempää kontekstin ymmärtämistä, harkintaa ja empatiaa. Ihmisen rooli kattaa myös kaiken tekoälyn luomiseen, käyttämiseen, ylläpitoon ja toiminnan arvioimiseen liittyvät tehtävät. Tässä luvussa vastataan myös kolmanteen tutkimuskysymykseen: ” *Mitä haasteita tekoälyn hyödyntämiseen liittyy*?”

Tekoälyn hyödyntämiseen sisällönhallinnassa liittyy sekä teknisiä että eettisiä haasteita. Yksi keskeisimmistä ongelmista on tekoälyn tekemät virheelliset luokitukset. Mikäli järjestelmä tulkitsee sopivan sisällön virheellisesti haitalliseksi, se voi rajoittaa käyttäjien sananvapautta ja aiheuttaa turhautumista sekä epäluottamusta järjestelmää kohtaan. Toisaalta, jos haitallinen sisältö jää järjestelmältä huomaamatta, käyttäjät voivat altistua sisällölle, joka voi aiheuttaa monipuolista haittaa. Tekoälyn toiminta perustuu pitkälti sille syötettyyn opetusdataan, mikä muodostaa merkittävän haasteen. Opetusdatan laatu, monimuotoisuus ja puolueettomuus vaikuttavat suoraan algoritmien toimintaan. Koska dataa valitaan ja käsitellään ihmisten toimesta, siihen voi sisältyä tietoisia tai tiedostamattomia vinoumia ja ennakkoluuloja. Nämä heijastuvat suoraan tekoälyn päätöksentekoon ja voivat johtaa esimerkiksi tiettyjen ihmisryhmien epäreiluun kohteluun. Lisäksi monien tekoälymenetelmien toiminta on läpinäkymätöntä, mikä vaikeuttaa luottamuksen rakentamista käyttäjien suuntaan.

Tämä tutkielma tarjoaa laajempaa kuvaa sosiaalisen median alustoilla esiintyvän haitallisen sisällön hallintaan käytettävistä tekoälymenetelmistä. Tutkielma käsittelee myös hieman muita sisällönhallinnan menetelmiä, kuten käyttäjien hyödyntämistä. Tutkielma voi olla hyödyllinen arvioitaessa sisällönhallintaan käytettäviä tekoälyratkaisuja sekä niiden hyviä ja huonoja puolia. Kuvan 4 viitekehys pyrkii havainnollistamaan tekoälymenetelmien hyödyntämistä haitallisten

sisältöjen hallintaan sosiaalisen median alustoilla.




Kuva 4 Tekoälymenetelmät sisällönhallissa

Tutkielman aihetta voitaisiin jatkossa syventää keskittymällä joko tiettyyn tekoälymenetelmään ja sen hyödyntämiseen haitallisten sisältöjen hallinnassa, tai tarkastelemalla jonkin yksittäisen haitallisen sisältötyypin hallintaa tekoälymenetelmien avulla.

## Lähteet


Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing Science*, *48*(1), 79–95. https://doi.org/10.1007/s11747-019-00695-1

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, *175*, 114820-. https://doi.org/10.1016/j.eswa.2021.114820

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x

Britt, R. K., Franco, C. L., & Jones, N. (2023). Trends and challenges within Reddit and health communication research: A systematic review. *Communication and the Public*, *8*(4), 402–417. https://doi.org/10.1177/20570473231209075

Cartwright, B., Frank, R., Weir, G., & Padda, K. (2022). Detecting and responding to hostile disinformation activities on social media using machine learning and deep neural networks. *Neural Computing & Applications*, *34*(18), 15141–15163. https://doi.org/10.1007/s00521-022-07296-0

Ceci, L. (2024). *Topic: Mobile video worldwide*. Statista. https://www.statista.com/topics/8248/mobile-video-worldwide/

Chartrand, G., Cheng, P. M., Vorontsov, E., Drozdzal, M., Turcotte, S., Pal, C. J., Kadoury, S., & Tang, A. (2017). Deep Learning: A Primer for Radiologists. *RadioGraphics*, *37*(7), 2113–2131. https://doi.org/10.1148/rg.2017170077

Chen, T., Sampath, V., May, M. C., Shan, S., Jorg, O. J., Aguilar Martín, J. J., Stamer, F., Fantoni, G., Tosello, G., & Calaon, M. (2023). Machine Learning in Manufacturing towards Industry 4.0: From 'For Now' to 'Four-Know'. *Applied Sciences*, *13*(3), 1903. https://doi.org/10.3390/app13031903

Cheng, X., Dale, C., & Liu, J. (2008). *Statistics and social network of YouTube videos*. 229–238. https://doi.org/10.1109/IWQOS.2008.32

Chesney, B., & Citron, D. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753–1820. https://heinonline.org/HOL/P?h=hein.journals/calr107&i=1812

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, *12*(null), 2493–2537.

Costa, P., & Goodwin, R. (2006). The role of religion in human values: A case study. *Journal of Beliefs and Values*, *27*(3), 341–346. https://doi.org/10.1080/13617670601001215

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), Article 1. https://doi.org/10.1609/icwsm.v11i1.14955

Davis, J. L., & Graham, T. (2021). Emotional consequences and attention rewards: The social effects of ratings on Reddit. *Information, Communication & Society*, *24*(5), 649–666. https://doi.org/10.1080/1369118X.2021.1874476

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *1*, 4171–4186.

Dixon, S. (2024a). *Biggest social media platforms by users 2024*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Dixon, S. (2024b). *Number of worldwide social network users 2028*. Statista. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

Elkin-Koren, N. (2020). Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720932296

Ellison, N. B., & Vitak, J. (2015). Social network site affordances and their relationship to social capital processes. Teoksessa *The Handbook of the Psychology of Communication Technology* (ss. 203–227). https://doi.org/10.1002/9781118426456.ch9

Fan, M., & Hemans, M. (2022). TikTok: How a Chinese Video Clip App Became a Popular and Successful Global Brand. Teoksessa *Casebook of Chinese Business Management* (ss. 33–45). Springer. https://doi.org/10.1007/978-981-16-8074-8_4

Gambäck, B., & Sikdar, U. K. (2017). Using Convolutional Neural Networks to Classify Hate-Speech. Teoksessa Z. Waseem, W. H. K. Chung, D. Hovy, & J. Tetreault (Toim.), *Proceedings of the First Workshop on Abusive Language Online* (ss. 85–90). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3013

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Teoksessa *Custodians of the Internet: Platforms,*

*Content Moderation, and the Hidden Decisions That Shape Social Media* (s. 288). https://doi.org/10.12987/9780300235029

Gillespie, T. (2020). *Content moderation, AI, and the question of scale*. https://doi.org/10.1177/2053951720943234

Gongane, V. U., Munot, M. V., & Anuse, A. D. (2022). Detection and moderation of detrimental content on social media platforms: Current status and future directions. *Social Network Analysis and Mining*, *12*(1), 129. https://doi.org/10.1007/s13278-022-00951-3

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society*, *7*(1). https://doi.org/10.1177/2053951719897945

Gray, K., & Pratt, S. (2025). Morality in Our Mind and Across Cultures and Politics. *Annual Review of Psychology*, *76*(1), 663–691. https://doi.org/10.1146/annurev-psych-020924-124236

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2). https://doi.org/10.1145/3479610

Jiang, J. A., Scheuerman, M. K., Fiesler, C., & Brubaker, J. R. (2021). Understanding international perceptions of the severity of harmful content online. *PLoS ONE*, *16*(8 August). https://doi.org/10.1371/journal.pone.0256762

Jones, N. (2025). *AI hallucinations can't be stopped—But these techniques can limit their damage—Turku University*. https://utuvolter.fi/discovery/fulldisplay?docid=cdi_proquest_miscellaneous_3158095737&context=PC&vid=358FIN_UTUR:VU1&lang=fi&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,AI%20hallucination&offset=0

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

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*, *2*(2), 100120–100120. https://doi.org/10.1016/j.jjimei.2022.100120

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review*, *131*(6), 1599-.

Lau, W. W. F. (2017). Effects of social media usage and social media multitasking on the academic performance of university students. *Computers in Human Behavior*, *68*, 286–291. https://doi.org/10.1016/j.chb.2016.11.043

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, C., & Cha, K. (2024). Toward the Dynamic Relationship Between AI Transparency and Trust in AI: A Case Study on ChatGPT. *International Journal of Human–Computer Interaction*, 1–18. https://doi.org/10.1080/10447318.2024.2405266

Li, Y., Thomas, M. A., & Liu, D. (2021). From semantics to pragmatics: Where IS can lead in Natural Language Processing (NLP) research. *European Journal of Information Systems*, *30*(5), 569–590. https://doi.org/10.1080/0960085X.2020.1816145

Li, Y., Wang, T., Chen, S., & Zhang, X. (2021). Application and Development of Natural Language Processing Service in Intelligent Customer Service System. Teoksessa *3D Imaging Technologies—Multidimensional Signal Processing and Deep Learning* (Vsk. 236, ss. 157–162). Springer. https://doi.org/10.1007/978-981-16-3180-1_20

Louati, A., Louati, H., Albanyan, A., Lahyani, R., Kariri, E., & Alabduljabbar, A. (2024). Harnessing Machine Learning to Unveil Emotional Responses to Hateful Content on Social Media. *Computers (Basel)*, *13*(5), 114-. https://doi.org/10.3390/computers13050114

Louridas, P., & Ebert, C. (2016). Machine Learning. *IEEE Software*, *33*(5), 110–115. https://doi.org/10.1109/MS.2016.114

Macukow, B. (2016). Neural Networks – State of Art, Brief History, Basic Models and Architecture. Teoksessa *Computer Information Systems and Industrial Management* (Vsk. 9842, ss. 3–14). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-45378-1_1

Mallmann, J., Santin, A. O., Viegas, E. K., dos Santos, R. R., & Geremias, J. (2020). PPCensor: Architecture for real-time pornography detection in video streaming. *Future Generation Computer Systems*, *112*, 945–955. https://doi.org/10.1016/j.future.2020.06.017

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons*, *52*(4), 357–365. https://doi.org/10.1016/j.bushor.2009.03.002

Marsoof, A., Luco, A., Tan, H., & Joty, S. (2023). Content-filtering AI systems–limitations, challenges and regulatory approaches. *Information & Communications Technology Law*, *32*(1), 64–101. https://doi.org/10.1080/13600834.2022.2078395

Molina, M. D., & Sundar, S. S. (2022). When AI moderates online content: Effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, *27*(4). https://doi.org/10.1093/jcmc/zmac010

Mullah, N. S., & Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review. *IEEE Access*, *9*, 88364–88376. https://doi.org/10.1109/ACCESS.2021.3089515

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Myers West, S. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366–4383. https://doi.org/10.1177/1461444818773059

Nissenbaum, A., & Shifman, L. (2017). Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society*, *19*(4), 483–501. https://doi.org/10.1177/1461444815609313

Nylén, D., Arvidsson, V., Carroll, N., & Holmström, J. (2024). Tracing the evolutionary leaps and boundaries of digital platforms: A case study of Facebook. *Innovation*, 1–20. https://doi.org/10.1080/14479338.2024.2363263

Oeldorf-Hirsch, A., & Sundar, S. S. (2016). Social and Technological Motivations for Online Photo Sharing. *Journal of Broadcasting & Electronic Media*, *60*(4), 624–642. https://doi.org/10.1080/08838151.2016.1234478

Olesen, T. (2025). Big Tech whistleblowing: Frances Haugen and the Facebook Files. *Organization (London, England)*. https://doi.org/10.1177/13505084251321785

Paul, S., & Saha, S. (2022). CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification. *Multimedia Systems*, *28*(6), 1897–1904. https://doi.org/10.1007/s00530-020-00710-4

Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., & Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. *Neurocomputing (Amsterdam)*, *230*, 279–293. https://doi.org/10.1016/j.neucom.2016.12.017

Pittman, M., & Reich, B. (2016). Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*, *62*, 155–167. https://doi.org/10.1016/j.chb.2016.03.084

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120-. https://doi.org/10.1016/j.eswa.2020.114120

Potter, M. (2021). Bad actors never sleep: Content manipulation on Reddit. *Continuum*, *35*(5), 706–718. https://doi.org/10.1080/10304312.2021.1983254

Prakash, A. V., & Das, S. (2021). Medical practitioner's adoption of intelligent clinical diagnostic decision support systems: A mixed-methods study. *Information & Management*, *58*(7), 103524. https://doi.org/10.1016/j.im.2021.103524

Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing & Applications*, *32*(10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

*Reddit Rules*. (ei pvm.). Noudettu 9. huhtikuuta 2025, osoitteesta https://redditinc.com/policies/reddit-rules

*Rules—4chan*. (ei pvm.). Noudettu 9. huhtikuuta 2025, osoitteesta https://www.4chan.org/rules

Saleem, M. A., Senan, N., Wahid, F., Aamir, M., Samad, A., & Khan, M. (2022). Comparative Analysis of Recent Architecture of Convolutional Neural Network. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/7313612

Scheuerman, M. K., Jiang, J. A., Fiesler, C., & Brubaker, J. R. (2021). A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–33. https://doi.org/10.1145/3479512

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. Teoksessa L.-W. Ku & C.-T. Li (Toim.), *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (ss. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-1101

Shilpashree, S., & Ashoka, D. V. (2024). F-DenseCNN: Feature-based dense convolutional neural networks and swift text word embeddings for enhanced hate speech prediction. *Social Network Analysis and Mining*, *14*(1), 192-. https://doi.org/10.1007/s13278-024-01345-3

Sipior, J. C., Ward, B. T., Rusinko, C. A., & Lombardi, D. R. (2024). Bias in Using AI for Recruiting: Legal Considerations. *Information Systems Management*, *41*(4), 399–412. https://doi.org/10.1080/10580530.2023.2294453

Smith, J. (2025). Daily briefing: The pros and cons of DeepSeek. *Nature (London)*. https://doi.org/10.1038/d41586-025-00330-w

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining “Fake News”: A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Thorleifsson, C. (2022, tammikuuta). *From cyberfascism to terrorism: On 4chan/pol/ culture and the transnational production of memetic violence—Turku University*. https://utuvolter.fi/discovery/fulldisplay?docid=cdi_proquest_journals_2632054851&context=PC&vid=358FIN_UTUR:VU1&lang=fi&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,4chan&offset=0

Wang, S., & Kim, K. J. (2023). Content Moderation on Social Media: Does It Matter Who and Why Moderates Hate Speech? *Cyberpsychology, Behavior, and Social Networking*, *26*(7), 527–534. https://doi.org/10.1089/cyber.2022.0158

*Yhteisön Säännöt | Jodel Support Hub*. (ei pvm.). Noudettu 9. huhtikuuta 2025, osoitteesta https://support.jodel.com/fi/articles/83058-yhteison-saannot