Olemassa olevat menetelmät GPT- tekoälyn tuottaman plagiaatin ja tekstin tunnistamiseen

Oulun yliopisto
Tietojenkäsittelytiede
LuK-tutkielma
Jani Alasirniö
2024

Tiivistelmä

Tekoälyn tuottaman tekstin tunnistaminen on tärkeää nykypäivän akateemisessa ympäristössä. Tekoäly on tuonut hyötyjä monille aloille, mutta etenkin akateemiseen ympäristöön se on luonut uudenlaisia haasteita. Tekoälyä voidaan käyttää vilpillisesti esimerkiksi tutkimuksien kirjoittamiseen tai opiskelijoiden toimesta kotitehtävien tekemiseen.

Tekoälyn tuottaman tekstin tunnistamiseksi on olemassa erilaisia lähestymistapoja. Näitä ovat esimerkiksi Crothersin., ym (2023) mukaan neuroverkkoperusteinen lähestymistapa sekä ominaisuusperusteinen lähestymistapa.

Ominaisuusperusteisessa lähestymistavassa tutkitaan esimerkiksi lausepituuksien keskijakaumaa, luettavuutta, erilaisten tukisanojen käyttöä sekä kappalepituuksien vaihtelua (Desaire., ym 2023). Neuroverkkoperusteinen lähestymistapa pyrkii aiemman tekstin perusteella ennustamaan tulevaa tekstiä (Crothers., ym 2023). Heidän mukaansa tämänhetken parhaat kaupalliset mallit perustuvatkin neuroverkkoperusteiseen lähestymistapaan.

Useamman tutkimuksen mukaan (Desaire., ym 2023, Elkhatat., ym 2023, Walters., 2023) mukaan kaupallisten mallien suorituskyky laskee, kun verrataan niiden suorituskykyä ChatGPT:n 3.5 ja 4.0- versioiden välillä. Suorituskyvyn laskeminen voi johtua esimerkiksi siitä, että 4.0- versiossa käytettäisi erilaista dekoodaus algoritmia kuin 3.5-versiossa.

Tekoälymallien kehittyessä tulevaisuudessa onkin tärkeää kiinnittää huomiota siihen, että miten tunnistustyökalut voidaan kehittää siten, että ne niiden suorituskyky ei kärsi olennaisesti tekoälymallien päivittyessä. Näin vältetään urheilusta tuttu doping- testaajien ja dopingia käyttävien urheilijoiden ikuinen jahti.

Avainsanat
Tekoäly, plagioinnin tunnistus, tekstin tunnistus

Ohjaaja
Markus Kelanti, PhD

Sisällysluettelo

1. Johdanto

Xiao., ym (2022) kertovat, erilaiset tekoälypohjaiset työkalut ovat tuoneet uudenlaisia etuja ja mahdollisuuksia useille toimialoille, mutta samanaikaisesti ne ovat luoneet ongelmia akateemiseen ympäristöön. Heidän mukaansa osa työkaluista helpottaa plagiointia ja luovat näin luoda uudenlaisia haasteita opiskelijoiden osaamisen arviointiin.

Ibrahim., ym (2023) kirjoittavat, yksi uusimpia laajasti saatavilla olevia tekstiä generoivia työkaluja on ChatGPT, jonka avainominaisuuksia ovat luonnollisen tekstin ymmärtäminen ja luonnollisen tekstimuotoisen vastauksen tuottaminen. Heidän mukaansa ChatGPT:llä on potentiaalia avustaa opiskelijoita opinnoissa, mutta sitä voidaan myös käyttää väärin. Väärinkäyttö voi heidän mukaansa olla esimerkiksi vastausten hakemista etänä pidettävien tenttien kysymyksiin. Tämän vuoksi on ensiarvoisen tärkeää tunnistaa työkalut, jotka pystyvät erottamaan ihmisen ja tekoälytyökalun luoman tekstin toisistaan sekä tunnistamaan tekoälyn luoman plagiaatin. On myös tärkeää tietää, että miten työkalut tunnistuksen tekevät, jotta niiden vahvuuksia ja heikkouksia voidaan ymmärtää mahdollisimman syvällisesti.

Elkhatat., ym (2023) suorittivat kokeilun, jossa he pyysivät ChatGPT:n eri versioita tuottamaan vastauksen tiettyyn kysymykseen. Lisäksi viittä opiskelijaa pyydettiin vastaamaan samaan kysymykseen. Opiskelijoiden vastauksia käytettiin tulosten kontrolloimiseen, jotta tarkistusmallien mahdollisia virhearvioita voitaisiin tarkastella syvällisemmin. Heidän vertailussaan oli mukana OpenAI, Writer, Copyleaks, GPTZero ja CrossPlag työkalut. Heidän havaintojensa mukaan työkalut suoriutuivat tehtävästä vaihtelevasti, mutta trendin ollessa se, että ChatGPT:n version 4 tuottama teksti oli työkaluille vaikeammin tunnistettavissa kuin vanhemman 3.5- version tuottama teksti.

Suurin aiheeseen liittyvä haaste on ongelman tuoreus. Esimerkiksi ChatGPT on ollut saatavilla julkisesti marraskuusta 2022 (OpenAI., 2022). Tekniikka kehittyy kovaa vauhtia ja uusia versioita julkaistaan jatkuvasti. Tämän vuoksi tutkimustulosten arvioinnissa tulee kiinnittää erityistä huomiota siihen, että mitä versiota tekoäly- ja tunnistustyökaluista on käytetty. *Näistä havainnoista johtuen, onko olemassa tehokkaita ja luotettavia työkaluja GPT- tekoälyn tuottaman tekstin ja plagiaatin tunnistamiseen?*

On tärkeää tietää, että millä tavalla menetelmät toimivat ja kuinka tarkkoja ne ovat, jotta voidaan varmistua niiden soveltuvuudesta käytännön käyttötarkoituksiin, kuten opiskelijoiden töiden arviointiin akateemisessa ympäristössä. Tutkimusongelmaan liittyvät seuraavat tutkimuskysymykset.

Millä tavoin tunnistetut menetelmät toimivat?

Millaiseen suorituskykyyn menetelmät kykenevät.

Tutkielma koostuu käsitteiden avaamisesta, aiemman tutkimuksen esittelystä, yhteenvedosta ja johtopäätöksistä.

Metodologian avaamisessa käydään läpi tutkielman tekemiseen käytettyjä tietokantoja, Hakulauseita sekä muita tapoja, joilla katsausta on tehty. Käsitteiden avaamisessa kerrotaan tekoälytyökaluista sekä niiden toiminnasta. Lisäksi yleisimpiä tekoälytyökalujen taustalla olevia teknisiä ratkaisuja. Lisäksi osioissa avataan ja

selitetään tärkeimmät tutkielman kontekstiin liittyvät käsitteet. Muissa osioissa perehdytään olemassa olevaan tutkimukseen, tehdään niiden välisiä vertailuja sekä esitetään johtopäätöksiä.

1.1 Metodologia

Tutkielmassa käytetyt artikkelit on haettu IEE exploresta sekä Google Scholaria. Brereton ym., (2007) toteavat Google Scholarin sekä IEE exploren olevan tietoteknisen alan kannalta olennaisimpia tietokantoja. Artikkelit on haettu tietokannoista kevään 2024 aikana. Artikkeleita on etsitty myös tutkimalla tietokannoista löydettyjen artikkeleiden lähdeluetteloita.

Taulukossa esitettyjä hakulauseita on käytetty sellaisenaan, ilman boolean-operaattoreita.

Taulukko 1. Tutkielman aikana käytetyt hakulauseet sekä tietokannat

Käytetty hakulause	**Tietokanta**
GPT detection	Google Scholar
Plagiarism detection artificial intelligence	Google Scholar, IEE explore
What is GPT	Google Scholar
Artificial intelligence plagiarism detection	Google Scholar, IEE explore
Artificial intelligence detection	Google Scholar, IEE explore
Artificial intelligence plagiarism detection tools	Google Scholar, IEE explore
Detecting ai generated text	Google Scholar
Detecting ai written plagiarism	Google Scholar
Plagiarism detection ChatGPT	Google Scholar
natural language generation	Google Scholar
generative pretrained transformer	Google Scholar

Lähteiden laatu on varmistettu Julkaisufoorumin avulla. Tutkielmaan ei ole otettu mukaan kuin vähintään Julkaisufoorumin tason 1 täyttäviä julkaisuja. Julkaisufoorumi on suomalainen tutkimuksien laadunarviointia tukeva järjestelmä (Julkaisufoorumi, 2022). Julkaisua ”Advances in neural information processing systems (NeurIPS)” ei ollut saatavilla Julkaisufoorumin kautta, mutta julkaisu on varmistettu vertaisarvioiduksi.

2. Kielimallit ja tekoäly

Tekoälyllä yleisesti tarkoitetaan tietokoneohjelmaa, joka kykenee tekemään älykkäitä ratkaisuja erilaisissa konteksteissa. Näitä kontekseja voivat olla esimerkiksi luonnolisen kielen kautta tapahtuva keskustelunomainen vuorovaikutus, erilaisten ihmisten sekä asioiden tunnistaminen videokuvasta ja erilaisten ongelmien ratkaisu. Tekoälyä voidaan hyödyntää käytännön elämässä esimerkiksi asiakaspalvelutehtävien automatisoinnissa tai kohdennetun mainonnan apuna.

Taulukko 2. Termien selitteet

Termi	**Selite**
AI	Tekoäly
Large Language Model, LLM	Laaja kielimalli
GPT	Generative pretrained transformer-arkkitehtuuri
RoBERTa	Laajan kielimallin arkkitehtuurityyppi
Prompt	Kehote, tutkielman kontekstissa useimmiten tekoälytyökalulle, kuten ChatGPT:lle annettu luonnolista kieltä oleva syöte
Prompt engineering	Promptin suunnittelu jokin tietty tarkoitusperä mielessä.

Large language modelilla tarkoitetaan tekoälymallia, joka kykenee tuottamaan luonnolista kieltä. Nykyaikaiset LLM:t pystyvät ylläpitämään luonnollista keskustelua.

GPT:llä tarkoitetaan Large language modelin alityyppiä. GPT- arkkitehtuurimallista teki aikanaan poikkeuksellisen sen pohjautuminen ns. huomiomekanismiin.Myös RoBERTa on eräänlainen Large Language Modelin alityyppi.

Promptilla tarkoitetaan tekoälylle annettua syötettä, esimerkiksi tekstimuotoista kysymystä tai tehtävänantoa. Pompt engineeringillä taasen tarkoitetaan promptin suunnittelua jokin tietty tarkoitusperä mielessä. Tarkoitusperä voi olla esimerkiksi saada tekoäly tuottamaan ratkaisu ohjelmointitehtävään jotain tiettyä kirjastoa käyttäen tai AI-tunnistustyökalun hämääminen.

2.1 Yleisesti

Luonnollista kieltä tuottavat mallit ovat yleistyneet viime vuosien aikana. Näitä malleja kutsutaan NLG (eng. Natural Language Generation)- malleiksi. Kuten Reiter & Dale, (1997) kertovat, NLG- järjestelmillä tarkoitetaan tietokoneohjelmia, jotka generoivat tekstiä jollain ihmisen ymmärtämällä kielellä, kuten englanniksi tai ranskaksi. He listaavat NLG- ohjelman keskeisiksi tehtäviksi sisällön määrityksen sekä tekstisuunnittelun (mitä tietoa käyttäjälle välitetään ja miten tieto jäsennellään), lauseiden suunnittelun (päätös siitä, että miten asia jaetaan kappaleisiin ja lauseisiin) sekä realisoinnin (lauseiden tuottaminen).

1990- luvun aikana varhaisimpia NLG- malleja on käytetty erilaisiin rutiininomaisiin tehtäviin. NLG- työkaluja on käytetty esimerkiksi sääennusteiden kirjoittamiseen (Goldberg., ym 1994) sekä sekä asiakaspalvelun optimointiin Springer., ym (1991).

GPT on lyhenne sanoista Generative pretrained-transformer model. Tämän hetken edistyneimmät luonnollisen tekstin generoimiseen kykenevät tekoälymallit pohjautuvat GPT- arkkitehtuuriin. Nimensä mukaisesti esimerkiksi OpenAI:n GPT- versiot pohjautuvat GPT- arkkitehtuuriin.

Vuonna 2017 Vaswani., ym esittelivät Transformer- arkkitehtuurimallin. Heidän luomansa Transformer- arkkitehtuurimalliin perustuvat työkalut olivat aiempia malleja huomattavasti suorituskykyisempiä esimerkiksi kielen kääntämisessä. Transformer-mallista aiemmin käytettyihin verratuna poikkeuksellisen se, että se perustuu huomiomekanismiin (eng. Attention mechanism) joka luo riippuvuuksia annetun syötteen ja ohjelman antaman tulosteen välille. Tämä auttaa tekoälyä pysymään paremmin annetussa kontekstissa ja generoimaan vastauksia, jotka ovat aiempaa vähemmän riippuvaisia aiemmista vastauksista sekä prompteista. Lisäksi tämä mahdollistaa entistä enemmän rinnakkaista prosessointia, joka parantaa tulosteiden laatua entisestään. (Vaswani., ym 2017)

Vaswanin., ym (2017) mukaan Transformer- arkkitehtuurimalli on ensimmäinen Sequence transduction- malli, joka perustuu pelkästään huomiomekanismiin.

2.2 Yleisimmät julkisesti saatavilla olevat LLM ratkaisut

ChatGPT on OpenAi:n marraskuussa 2022 julkisesti saataville tuotu tekoälymalli, joka perustuu GPT-3.5 tekniikkaan (OpenAi, 2022). Myöhemmin ohjelmistosta on tullut markkinoille myös GPT-4.0 malli, joka on saatavilla kuukausimaksua vastaan. Lisäksi GPT-4.0 malli on saatavilla sisäänrakennettuna Microsoftin Edge- selaimessa, jossa se on nimetty uudelleen Copilotiksi. Lisäksi joulukuussa 2023 Google toi markkinoille oman tekoälymallinsa Geminin.

ChatGPT perustuu OpenAi:n kehittämään generative pretrained transformer- malliin (GPT). GPT on LLM:n (Large Language Model) tyyppi. Käytännössä tekoälymallilla on jokin etukäteen syötetty aineisto, jota ne käsittelevät ja jonka pohjalta ne luovat vastauksen käyttäjän syötteeseen. Jotkin mallit, kuten Microsoftin Edge- selaimeen

upotettu Copilot pystyvät hakemaan tietoa internetistä myös reaaliakaisesti. Vastaus syötteeseen voi olla esimerkiksi kuva, tekstiä tai ohjelmakoodia.

Ohjelmakoodin tuottamiseen on olemassa erityisiä tekoälymalleja, kuten Githubin 2021 julkistama Copilot (Friedman., 2021) ja Amazonin vuonna 2023 avoimesti saataville tuotu CodeWhisperer (Amazon., 2023).

2.3 Tekoälyn väärinkäytön ehkäisy

Cotton., ym (2023) kirjoittavat, että tekoälyn väärinkäyttöä kannattaisi ehkäistä esimerkiksi opettamalla opiskelijoille entistä tarkemmin plagiarismista. Mitä plagiarismi on ja miksi se on väärin. Samalla he kuitenkin tiedostavat, että tämä ei välttämättä riitä ja akateemisessa ympäristössä tulisi käyttää myös plagioinnintunnistustyökaluja opettajien itsensä tekemän arvioinnin tukena. Demonstroidakseen ChatGPT:n kykyjä he ovat kirjoituttaneet lähes koko tutkimusartikkelin ChatGPT:llä. Samaa kertovat myös Desaire., ym (2023). He kertovat, että vaikka tieteelliset julkaisut vaatisivat kirjoittajia ilmoittamaan läpinäkyvästi tekoälytyökalujen käytöstä niin kaikki kirjoittajat eivät tätä kuitenkaan tee. He kertovat, että tämän vuoksi on tärkeää olla olemassa luotettavat tapa erottaa tekoälytyökalun kirjoittama tekstin ihmisen kirjoittamasta tekstistä.

Huoli plagioinnista ja väärinkäytöstä on aito, sillä esimerkiksi Ibrahim., ym (2023) tekemän tutkimuksen mukaan ChatGPT suoriutui yliopiston alkeistason kursseista erinomaisesti. Toisaalta sillä oli vaikeuksia edistyneemmän tason jatkokurssien kanssa. Heidän mukaansa ChatGPT ei esimerkiksi tee oikeellisia lähdeviittauksia. Tämän vuoksi on mahdotonta osoittaa, että mihin tekoälymallin antama vastaus perustuu ja heidän mukaansa tämä voi aiheuttaa ongelmia myös tekijänoikeusrikkomuksien muodossa. Heidän mukaansa tekoälymallit voivat luoda haasteita esimeriksi niiden insinöörialojen opettamiseen, joilla ohjelmointiosaaminen on ensiarvoisen tärkeää.

2.4 Tekstiä generoivien kielimallien heikkoudet ja hyödyt

See., ym (2019) tekivät tutkimuksen, jossa he syöttivät toiselle GPT-2 versioon perustuvalle kielimallille suuren määrän opetusdataa ja pyysivät sitä sekä verrokkimallia kertomaan tarinan. Verrokkimalli perustui tehtäväspesifiseen arkkitehtuuriin, ja mallia oli räätälöity tarinoiden generoimista varten. Opetusdatalla koulutettu GPT-2 pohjainen malli suoriutui tutkimuksessa paremmin esimerkiksi annetussa kontekstissa pysymisessä ja tarinan tapahtumien pitämisessä oikeassa järjestyksessä. Kuitenkaan eroa ei syntynyt siinä, että kuinka laajaa sanavarastoa mallit käyttivät. He kuvailevat molempien mallien tuottamien tekstien olevan samalla tavalla itseään toistavaa, epämonimuotoista sekä sisällöllisesti yksinkertaista. He kertovat, että ongelmat korjautuvat, kun mallit hakevat dataa rajoittamattomasta poolista. Heidän mukaansa näin ollen ongelmat johtuvat käytetystä dekoodaus algoritmista.

Gehrmann., ym (2019) mukaan tekstingenerointimallien käyttämä sanavarasto on kapeampi kuin ihmisellä. Heidän mukaansa mallit eivät esimerkiksi käytä synonyymejä tuottamassaan tekstissä samalla tavalla kuin ihmiset. Kielimallit käyttävät heidän mukaansa jotain hakutekniikkaa, kuten beam searchia tai max samplingia käydessään läpi aineistoa tuottaakseen luonnollista tekstiä. He väittävät, että jokainen näistä hakutekniikoista on jollain tavalla harhautunut (eng. biased). Heidän mukaansa tämä voisi johtaa siihen, että tekoälymallien käyttämä sanavarasto on rajallinen ja tekstissä toistetaan

samoja sanoja enemmän kuin ihmisen kirjoittamassa tekstissä. Tämä tapahtuisi huolimatta siitä, että kuinka laaja opetusmateriaali tekoälymallin taustalla on.

Föhrling & Zubiaga (2021) kertovat, että suuremman näytemäärän käyttäminen johtaa monimuotoiseen, mutta heikkolaatuiseen tekstiin. Toisaalta samankaltaisuutta korostavan algoritmin käyttäminen johtaa sinänsä laadukkaaseen sisältöön, mutta kielellisesti itseään toistavaan ja yksitoikkoiseen tekstiin.

Desaire., ym (2023) kertovat, että ChatGPT tuottaa yksinkertaisempaa tekstiä kuin ihminen. Suurimmat erottajat ovat heidän mukaansa ChatGPT:n tapa kirjoittaa vähemmän lauseita ja sanoja per kappale kuin ihminen. He kertovat myös, että ihmiskirjoittajilla on tapanaan varioida lauserakenteita ja -pituuksia tekstissään enemmän kuin ChatGPT:llä. He kuitenkin huomauttavat, että lausepituutta ei tulisi käyttää mittarina sellaisenaan, vaan lausepituuden keskijakauma on parempi mittari.

Elkhatat (2023) on suorittanut tutkimuksen, jossa on testattu ChatGPT:n versioiden 3.5 ja 4.0 kykyä tuottaa uniikkia tekstiä. Tutkimuksessa ChatGPT:tä pyydettiin tuottamaan vastaus samaan kysymykseen 3 kertaa samassa keskustelussa. Tämän jälkeen aloitettiin uusi keskustelu ja sama toistettiin. Tämän jälkeen vastaukset ladattiin Blackboard Learn alustalle, joka mahdollistaa opiskelijoille tekstien samankaltaisuuden vertaamisen toisiinsa.

Tutkimuksessaan Elkhatat (2023) osoittaa, että vastausten uniikkius ei poikkea riippuen siitä, että pyydetäänkö vastausta uudelleen samassa keskustelussa vai aloitetaanko uusi keskustelu, mutta kysytään sama kysymys uudelleen. Havainto piti paikkansa sekä GPT-3.5 sekä -4 versioiden kohdalla. Elkhatat (2023) kertoo, että tutkimustulosten perusteella voidaan sanoa, että ChatGPT pystyy luotettavasti tuottamaan autenttisia ja uniikkeja vastauksia kysymyksiin, jotta ne voivat kiertää erilaisia tekstin samankaltaisuutta mittaavia text matching- ohjelmia. Tämän on hänen mukaansa ongelma eritoten, koska edes tekoälymallin itsensä luomien vastausten välillä ei löydy tarkastusohjelmien avulla merkittävää vastaavuutta.

Hänen mukaansa tuloksien perusteella voidaan todeta, että GPT-3.5 ja -4 versioissa käytetään erilaista algoritmia taustamateriaalin käsittelyyn. Hänen mukaansa tämä ilmenee siten, ettei GPT-4.0 versio vaikuttaisi uudelleengeneroivan GPT-3.5:den generoimia vastauksia.

3. Plagioinnintunnistustyökalut ja niihin liittyvä tutkimus

Tässä osiossa keskitytään luomaan katsaus ja tekemään päätelmiä etenkin tutkimuksista, joissa plagioinnintunnistustyökaluilla on tehty jokin kokeilu ja tutkimus keskittyy kokeilun tulosten esittelyyn. Kokeilussa voi olla myös mukana tutkijoiden itsensä kehittämä tunnistustyökalu. Erityistä huomiota tulee kiinnittää siihen, että milloin tutkimukset on tehty ja mitä versioita tekoälytyökaluista on ollut käytössä. Fröhling & Zubiaga, (2021) kirjoittavat, että yksi suurimpia haasteita tunnistustyökalujen kehittämisessä onkin niiden ajantasaisena pysyminen. Eli esimerkiksi kun uusi versio tekoälymallista julkaistaan, niin tunnistustyökalua voitaisiin päivittää vasta pienellä viiveellä. Heidän mukaansa työkaluista olisi hyvä luoda sellaisia, että niiden toimintakyky säilyy versiosta toiseen. He nostavat esiin myös huolen työkalujen saatavuudesta.

Tekoälyn kirjoittaman tekstin ja siinä esiintyvän plagioinnin tunnistustyökalut antavat joko kvalitatiivisia tai kvantitatiivisia osoituksia sen todennäköisyydestä onko jokin tietty dokumentti tai sen osa tekoälyn generoima. Työkalujen tavoitteena on antaa tukea ihmisarvioijalle (Walters, 2023).

3.1 Ei- kaupalliset työkalut ja metodit

Föhrling & Zubiaga, (2021) ovat tutkimuksessaan luoneet ominaisuusperusteisen mallin tekoälyn tuottaman tekstin tunnistamiseksi. Ominaisuusperusteisuudella he tarkoittavat tekstin erilaisten ominaisuuksien kuten tukisanojen ja synonyymien sekä monimuotoisten lauserakenteiden analysointia. Tekoälymalleilla on taipumusta yksitoikkoisen sanavaraston käyttämiseen, synonyymien puutteeseen sekä itseään toistavan tekstin kirjoittamiseen (See., ym (2019) ja Gehrmann., ym 2019). Lisäksi kielimalleilla on Gehrmannin., ym (2019) mukaan tapana generoida lauseita, joiden rakenne on hyvin samankaltainen keskenään.

Crothers., ym (2022) huomauttavat, että ominaisuusperusteisen mallin heikkoudet piilevät sen riippuvuudessa kielimallin käyttämään dekoodaukseen. He kertovat, että ominaisuusperusteisen mallin suorituskyky on riippuvainen siitä, että millaista dekoodausalgoritmia tekoälymalli käyttää tekstin generoimiseen.

Crothersin., ym (2022) kanssa samaan lopputulokseen päätyvät myös Fröhlin & Zubiaga, (2021). Kokeiltuaan kehittämäänsä mallia he huomasivat, että heidän mallinsa suorituskykyyn vaikuttaa käytetty dekoodausalgoritmi. Heidän kokeilunsa mukaan kehitetty malli pystyi tunnistamaan myös GPT-3 mallin tuottamaa tekstiä hyvin luotettavasti. He pitivät tulosta yllätyksenä, sillä he odottivat uudemman mallin tunnistamisen olevan hankalampaa johtuen sen satakertaisesta parametrimäärästä verrattuna GPT-2 versioon.

Aiemmin ominaisuusperusteisen mallin ovat luoneet tutkimuksessaan myös Gehrmann., ym (2019). Heidän luoma GLTR- malli perustuu olettamaan, että jokaisen tekstiä generoivan mallin täytyy tekstiä luodakseen käydä opetusmateriaaliaan läpi jonkinlaisen dekoodausalgoritmin avulla. Heidän mukaansa jokainen näistä dekoodausalgoritmeista

on jollain tapaa harhautunut. Harhautuminen voi olla esimerkiksi sitä, että algoritmi poimii eniten suosittuja sanoja ja lauserakenteita. Malli tekee tunnistuksen kolmivaiheisesti. Ensimmäiset kaksi vaihetta testaavat, että onko sana poimittu poolin suosituimmasta päästä ja viimeinen vaihe testaa, että onko tunnistusjärjestelmä kuinka varma siitä, että juuri tässä nimenomaisessa kontekstissa kyseinen sana tulisi seuraavaksi lauseeseen. GLTR- malli korostaa tekstiä eri värein sanan todennäköisyyden mukaan. Korostus on nelivaiheinen ja se tehdään korostamalla sanoja tekstistä eri värein, jotka kuvastavat sanan yleisyyttä. Tutkimuksessa suoritetun kokeilun perusteella ihmisarvioija pystyi tunnistamaan tekoälyn kirjoittaman tekstin työkalun avulla noin 18 % luotettavammin kuin ilman työkalua.

Desaire., ym (2023a) ovat myös kehittäneet tutkimuksessaan ominaisuusperusteisen menetelmän. Menetelmä perustuu koneoppimiseen. Heidän mukaansa menetelmä käyttää pääasiassa 4 erilaista tekstin ominaisuutta tunnistuksen tekemiseksi. Ominaisuudet ovat kappaleiden pituus, lausepituuksien monimuotoisuus, erilaisten välimerkkien (sulkujen, pilkkujen, puolipisteiden jne.) käyttäminen sekä sanojen yleisyys tietyssä ryhmässä. Nämä ominaisuudet on edelleen jaettu kahteenkymmeneen alakategoriaan. Heidän mukaansa näiden neljän ominaisuuden tunnistamiseen perustuvalla tunnistusmallilla voidaan erottaa tekoälymallin ja ihmisen luoma teksti toisistaan luotettavasti. He kertovat tutkimuksensa olevan ensimmäinen, jossa luotu malli on erittäin tehokas erottamaan ihmisen kirjoittaman tieteellisen tekstin ChatGPT:n luomasta tekstistä. Malli oli tutkimuksessa huomattavasti tehokkaampi tekemään tunnistuksen kuin OpenAI:n oma GPT-2 Detector- tunnistustyökalu. He ehdottavat, että yleisten sanojen listausta varioimalla voitaisiin saavuttaa parempaa suorituskykyä erilaisten tekstien kanssa. Lisäksi he kertovat, että Gehrmannin., ym (2019) kehittämän GLTR:n kaltaisen pisteytysjärjestelmän kehittäminen suosituille sanoille voisi parantaa metodin suorituskykyä.

Desaire., ym (2023b) ovat jatkotutkimuksessaan menetelmänsä avulla pystyneet tunnistamaan OpenAI:n GPT-3.5:n ja GPT-4:n kirjoittamia tekstejä tietyin rajoituksin jopa 99 % tarkkuudella. Tutkimuksessaan he pyysivät ChatGPT:tä luomaan tieteellisiä tekstejä kirjoittaen kuten kemisti. Tekoälymallia pyydettiin kirjoittamaan teksti samaan tyyliin kuin kemian alan ACS journal- julkaisussa. Joissain tapauksissa he antoivat promptissa ChatGPT:lle lisäksi tiivistelmän, jonka perusteella sen pystyi muotoilemaan vastaustaan. Jatkotutkimuksessaan Desaire., ym (2023b) ovat ottaneet huomioon, että useita tunnistustyökaluja voidaan kiertää prompt engineeringillä ja ovatkin testanneet malliaan myös prompteilla, joilla pyritään kiertämään tunnistustyökaluja.

Jatkotutkimuksessaan he antoivat työkalulleen opetusmateriaaliksi yhteensä 100 ihmisen kirjoittamaa näytettä. Näytteet valikoitiin kemian alan tieteellisistä julkaisuista ja tarkemmin käytettiin artikkeleiden johdantoja. Johdanto on Desairen., ym (2023b) mukaan osuus, johon AI:ta todennäköisimmin käytettäisi. Demonstroidakseen kuinka vähän opetusdataa ja pienellä vaivalla mallin voi ottaa käyttöön he valikoivat vain 10 artikkelia per julkaisu.

Jatkotutkimuksessaan Desaire., ym (2023b) suorittivat vertailun OpenAI:n GPT-2 detectoriin sekä ZeroGPT tunnistustyökaluihin. Lopputuloksena kehitetty malli saavutti 100% tarkkuuden GPT-3.5:den ja -4 versioiden tuottamien tekstin tunnistamisessa, siinä missä ZeroGPT:n epäonnistumisprosentti oli GPT3.5 version kohdalla 32 ja -4 version kohdalla 42%. GPT-2 detectorin suorituskyky oli vielä huonompi. Lisäksi, kun vaikeusastetta lisättiin pyytämällä tekoälymallia generoimaan johdanto annetun tiivistelmän perusteella laski ZeroGPT:n ja OpenAI- detectorin suorituskyky entisestään.

Heidän mallinsa suorituskyky saatiin kuitenkin laskemaan sillä, että mallille syötettiin analysoitavaksi college- opiskelijoiden kirjoittamia uutisartikkeleita ja tekstin tyyliä saatiin näin olennaisesti muutettua.

3.2 Kaupalliset mallit

Kaupallisten mallien arvioinnissa tulee huomioida, että niiden opetusmateriaali ja tarkemmat toimintaperiaatteet eivät ole julkisesti saatavilla. Kuten Liang., ym (2023) huomauttavat tutkimuksessaan, tämä luo tarpeettomia haasteita mallien suorituskyvyn varmistamiseen. He kertovat, että malleja mainostetaan usein 99% tarkkuudella, joka ei pidä todellisuudessa paikkaansa.

Elkahtat., ym (2023) suorittamassa tutkimuksessa he vertailivat eri tunnistustyökaluja. Mukana vertailussa oli viisi tunnistustyökalua, jotka olivat OpenAI:n työkalu, Writer, Copyleaks, GPTZero sekä CrossPlag. He kertovat, että GPTZero luokittelee tekstin joko AI:n tai ihmisen generoimaksi. Open AI:n Classifier tekee jaon kuuteen eri kategoriaan, jotka ovat: Likely AI-Generated, Possibly AI-Generated, Unclear if it is AI-Generated, Unlikely AI-Generated and Very Unlikely AI-Generated. GPTZero ja OpenAI Classifier ilmaisevat tuloksen prosentteina.

Taulukko 3. Elkhatatin., ym (2023) tutkimuksessaan työkalujen antamien tulosten normalisointiin käyttämä taulukko.

Kategoria	**Kuinka suuri osa tai kuinka todennäköisesti teksti on tekoälyn generoima**
Likely AI-Generated	80 – 100 %
Possibly AI-Generated	60 – 79 %
Unclear if it is AI-Generated	40 – 59 %
Unlikely AI-Generated	20 – 39 %
Very Unlikely AI-Generated	0 – 19 %

Elkhatatin., ym (2023) suorittaman kokeilun perusteella suurin osa GPT-3.5:den tuottamista teksteistä luokiteltiin onnistuneesti ”Likely AI- generated”- kategoriaan. Kuitenkin yksittäisiä virheellisiäkin luokituksia tapahtui. Suorituskyky oli kuitenkin yleisesti tasaista kautta linjan.

Ihmisten kirjoittamien kontrollivastausten analysoinnissa työkaluilla oli myös Elkhatatin., ym (2023) tutkimuksen mukaan vaikeauksia. Esimerkiksi Writer luokitteli kaksi ihmisen kirjoittamaa tekstiä viidestä ”Likely AI- generated”- kategoriaan.

Elkhatatin., ym (2023) suorittaman tutkimuksen tulosten perusteella voidaan sanoa OpenAI detectorin olleen herkkä, mutta epätarkka. Kokonaisuudessaan GPT-4 mallin tuottamien vastauksien kanssa tunnistustyökaluilla oli enemmän haasteita ja suorituskyky oli epätasaisempaa. Virheellisiä luokitteluja esiintyi huomattavasti enemmän kuin GPT-

3.5:den tuottamien vastauksien kanssa. Tunnistusmalleista CrossPlag oli tarkka, mutta sillä oli vaikeuksia etenkin uudemman GPT-4:n tuottamien tekstien tunnistamisessa.

Gao., ym (2023) tekivät tutkimuksen, jossa he kokeilivat tunnistustyökalujen kykyä tunnistaa ChatGPT:n kirjoittama tieteellinen tiivistelmä. ChatGPT:lle annettiin artikkelin otsikko, sekä julkaisu jonka tyylin mukaisesti tiivistelmä tulee laatia. Heidän mukaansa GPT-2 Output Detector pystyi tunnistamaan ChatGPT:n kirjoittamat tekstinäytteet luotettavasti ja samanaikaisesti luokitellen lähes kaikki ihmisten kirjoittaman kontrollitiivistelmät hyvin epätodennäköisesti AI:n generoimiksi. He antoivat tiivistelmät myös tutkittavaksi plagioinnintunnistustyökaluista. Plagioinnintunnistustyökalujen (iThenticate, Plagiarism detector (web- pohjainen vapaasti saatavilla oleva työkalu) mukaan AI:n generoimat tekstit olivat autenttisempia kuin ihmisten kirjoittamat tekstit.

He tekivät myös kokeilun ihmisarvioijilla. Ihmisarvioijille annettiin käsiteltäväksi sekä ihmisten, että ChatGPT:m kirjoittamia tiiviestelmiä. Ihmisarvioijat pystyivät erotamaan oikein 68% tiivistelmistä ChatGPT:n generoimiksi, mutta samanaikaisesti 14% ihmisen kirjoittamaa tekstiä luokiteltiin virheellisesti ChatGPT:n generoimaksi.

Yleinen havainto ihmisarvioiden käytöksestä oli, että ChatGPT:n generoimiksi epäillyt tiivistelmät olivat epämääräisempiä ja kaavamaisempia kuin ihmisten kirjoittamat tiivistelmät.

Kaupallisten mallien suorituskykyä ovat tutkineet myös Anil., ym (2023). Heidän tutkimuksessaan tunnistustyökaluista olivat mukana Grammarly, iThenticate, Small SEO tools sekä DupliChecker.

Tutkimuksessaan he käyttivät yhteensä sataa ChatGPT:n generoimaa artikkelia. ChatGPT: tä pyydettiin kirjoittamaan enintään 1000 sanan mittaninen artikkeli annetsta aihealueesta. Artikkelit käsittelevät yhteensä kymmentä eri aihealuetta. Artikkelit käsiteltiin kaikilla neljällä työkalulla. Työkalun mittaama samankaltaisuusindexi (OSI, Overall Similarity Index) otettiin ylös. He tutkivat myös käytetyn kielen monimuotoisuuden ja artikkelin pituuden korrelaatiota OSI- arvon välillä.

Tutkimuksessaan he huomasivat, että Grammarlyllä oli keskimäärin suurin OSI- arvo, toisella sijalla, samankaltaisella suorituskyvyllä oli Ithenticate ja perää pitivät selvällä erolla Small SEO Tools ja DupliChecker, joiden OSI arvo oli huomattavasti kärkikaksikkoa matalampi. Anil., ym (2023) mukaan tulokset kertovat siitä, että Grammarly on selkeästi muita työkaluja parempi tunnistamaan plagiarismia AI:n kirjoittamassa tekstissä kuin muut työkalut.

Lopputuloksena Anil., ym (2023) kokeilussa iThenticate suoriutui parhaiten. He kertovat myös, että oikein käytettynä Grammarly ja iThenticate voisivat täydentää toisiaan.

Walters (2023) on tutkimuksessaan vertaillut kuudentoista eri tunnistustyökalun suorituskykyä toisiinsa. Hän käytti vertailuun GPT-3.5 sekä GPT-4 versioiden kirjoittamia dokumentteja. Lisäksi vertailussa oli mukana kontrollivastauksina ensimmäisen vuoden opiskelijoiden kirjoittamia dokumentteja. Kaikilta kolmelta kirjoittajatyypiltä dokumentteja oli mukana 42 kappaletta, eli näytteitä oli kokonaisuudessaan yhteensä 126.

Waltersin (2023) tutkimuksessa Turnitin sekä Copyleaks kykenivät lajittelemaan kaikki esseet oikein joko AI:n tai ihmisen kirjoittamiksi, kun käytössä oli kolme eri kategoriaa (ihmisen kirjoittama, AI:n kirjoittama ja epävarma). Lähes kaikkien työkalujen

suorituskyky oli kuitenkin GPT3.5:den kirjoittamien tekstien kohdalla kohtalaisen hyvää, sillä ContentDetector.ai:ta sekä Content at Scalea lukuunottamatta ne pystyivät tunnistamaan sen kirjoittamat tekstit 86% varmuudella. Kuitenkin GPT-4.0 version kohdalla vain Copyleaks, Turnitin sekä Originality.ai pystyivät tunnistamaan tekstit 83 % suuremmalla tarkkuudella. Loput työkaluista luokittelivat GPT-4.0:n kirjoittamia dokumentteja epävarmoiksi tai ihmisen kirjoittamaksi. Hän kertoo tämän olevan merkittävin ero mainittujen kolmen sekä loppujen kolmentoista työkalun välillä.Weber-Wulffin., ym (2023) tutustuessa kaupallisten tunnistustyökalujen suorituskykyyn he loivat kahdeksantoista artikkelin suuruisen testiaineiston testatakseen. Artikkelit olivat joko kokoaan ihmisen kirjoittamia, alun perin ihmisen kirjoittamia, mutta englanniksi koneen kääntämiä, ihmisen muokkaamia, mutta ChatGPT:n alun perin kirjoittamia, alunperin ChatGPT:n kirjoittamia, mutta toisen tekoälytyökalun uudelleenkirjoittamia tai puhtaasti ChatGPT:n kirjoittamia. He kertovat tutkimuksensa olevan kattavin tähän mennessä aiheesta tehty kokeellinen tutkimus.

He kertovat, että tunnitustyökalujen tarkkuus oli korkeimmillaankin alle 80 % kun vertailussa oli mukana yhteensä 14 eri työkalua. Työkaluista ainoastaan viisi saavutti yli 70 % tarkkuuden. He kertovat, että työkaluilla oli taipumusta luokitella tekstejä mielummin ihmisen kuin AI:n kirjoittamiksi. He arvioivat, että noin 20 % tekoälyn generoimista teksteistä luokiteltaisiin virheellisesti ihmisten kirjoittamiksi.

3.3 Tunnistusmallit ja ei- natiivit kielenpuhujat

Liang., ym (2023) ovat suorittaneet tutkimuksen, jossa he kokeilivat miten tunnistustyökalut suoriutuvat, kun verrataan englantia äidinkielenään ja vieraana kielenä puhuvien kirjoittajien tekstejä toisiinsa.

Heidän tutkimuksensa mukaan ongelma saatiin jäljitettyä tunnistusmallien käyttämiin tekstin monimuotoisuutta seuraaviin mittareihin. Heidän mukaansa mallit käyttävät mittaria siihen, että kuinka hyvin ne pystyvät ennakoimaan tekstin sisältöä aiemmin analysoidun osan perusteella. Heidän mukaansa tämä johtaa siihen, että usein kieltä äidinkielenään puhuvia ihmisiä kapeammalla sanavarastolla operoivien vieraskielisten ihmisten kirjoittamat tekstit voidaan virheellisesti merkitä tekoälyn tuottamaksi.

Tutkimuksessaan he saivat demonstroitua ongelmaa siten, että he yksinkertaistivat englantia äidinkielenään kirjoittavien tekstien sanavarastoa. Tämä sai aikaan tilanteen, jossa virhearvion todennäköisyys kasvoi. Vastavuoroisesti monipuolistamalla vieraskielisen kirjoittajan tekstiä virhearvion todennäköisyys pienentyi. He pystyivät toistamaan saman ilmiön myös pyytämällä ChatGPT:tä kirjoittamaan annetun tekstin uudelleen, mutta kirjakielisesti. Tämä laski GPT- tunnistimien suorituskykyä. He ehdottavatkin, että tunnistustyökaluja tulisi käyttää vastuullisesti ja harkiten. Tämä etenkin silloin, kun arvioitavana on ei englantia äidinkielenään puhuvien englanniksi kirjoittamia tekstejä.

4. Johtopäätökset

Tekoälyn tuottaman tekstin tunnistamiseen kykenevät työkalut ovat ohjelmistoja, joiden tarkoitus on auttaa tekoälyn tuottaman tekstin tunnistamisessa. Tunnistustyökaluja voidaan käyttää esimerkiksi akateemisessa ympäristössä vilpin kitkemiseksi tai disinformaation leviämisen ehkäisemiseksi.

Mitä työkaluja on olemassa ja miten ne toimivat?

Tunnistustyökalut toimivat pääsääntöisesti siten, että ne pilkkovat tekstin osiin sanoiksi tai pidemmiksi osiksi ja tämän jälkeen työkalu mittaa, kuinka hyvin teksti on ennustettavissa. Tästä työkalu tuottaa käyttäjälleen kvalitatiivista ja kvantitatiivista tietoa liittyen siihen, että kuinka todennäköisesti teksti on tekoälyn generoima. (Walters., 2023).

Moni tunnistustyökalu toimii pilkomalla tekstin sanoiksi tai muiksi merkkisarjoiksi ja ennustamalla todennäköisyyttä, että tiettyä merkkisarjaa seuraa jokin tietty merkkisarja. Tekstin, jonka todennäköisimmin tunnistetaan tekoälygeneroiduksi, ovat ne, joilla on korkea ennustettavuus ja alhainen monimutkaisuus – ne, joissa on suhteellisen vähän satunnaisia elementtejä ja omalaatuisuuksia, joita ihmiset yleensä käyttävät kirjoituksessaan ja puheessaan.

Walters., (2023) valikoi tutkimukseensa mukaan kuusitoista eri työkalua, jotka hän löysi erilaisilta verkossa löytyneiltä top- listauksilta sekä Googlen hakutulosten kärkipäästä. Hänen työkalujen valikoimiseen käyttämän metodologian perusteella voidaan todeta, mainitut työkalut edustavat kattavasti suosituimpia markkinoilla olevia työkaluja.

Waltersin (2023) tutkimuksessa mukana olleista työkaluista Originality.ai sekä Crossplag toimivat neuroverkkoperusteisesti. Ne käyttävät omaa BERT- malliin pohjautuvaa mallia tekstin arvioimiseksi (Gillham., 2023, Nuha., 2023)). GPTZero käyttää myös neuroverkkoperusteista lähestymistapaa, jolla mitataan ainakin tekstin ennustettavuutta (GPTZero., 2024). Tekstin ennustettavuutta pyrkii mittaamaan myös sapling.ai, joka käyttää tukenaan samantyylistä transformer- mallia, jota käytetään myös tekstin generoimiseen (Sapling, 2023). Ennustettavuutta mittaa myös ContentDetector.ai (ContentDetector., 2024). GPT Radar kertoo käyttävänsä GPT-3 pohjaista mallia arvioinnin tukena, mutta ei avaa tarkemmin, että miten GPT-3 kielimallia hyödynnetään tekstin tunnistamisessa (GPT Radar., 2024).

Waltersin (2023) kokeilussa mukana olleista työkaluista Copyleaks, TurnitIn, Scribbr, ZeroGPT, Grammica, OpenAi Detector, ivypanda, seo.ai, Content at Scale, Writer eivät avaa verkkosivuillaa, että millä tavalla menetelmät toimivat.

Ennustettavuutta mittaavan mallin Crothers., ym (2023) kertovat olevan neuroverkko-perusteinen malli. Heidän mukaansa parhaiten malleista suoriutuvat sellaiset, jotka tekevät omaa generointiaan GPT- arkkitehtuuriin perustuvalla tekoälymallilla. Esimerkiksi generoitujen uutisartikkeleiden tunnistamiseen luotu Grover perustuu neuroverkkoon Crothers., ym (2023).

Crothers., ym (2023) kertovat, että tunnistaminen voidaan tehdä ominaisuusperusteisesti. Ominaisuusperusteisessa työkalussa voidaan tutkia esimerkiksi tekstissä esiintyvien sanojen yleisyyttä sekä tekstin sujuvuutta. Lisäksi voidaan käyttää neuroverkon ja ominaisuusperusteisen mallin yhdistelmää.

Millainen työkalujen suorituskyky on?

Tutkielmassa käsitellyn aiemman tutkimuksen perusteella voidaan sanoa, että markkinoilla olevat tekoälyn tuottaman tekstin ja plagiaatin tunnistamiseen kykenevät työkalut suoriutuvat tehtävistään vaihtelevasti. Esimerkiksi Anil., ym (2023) pystyivät pienellä näytemäärällä osoittamaan, että Grammarly ja iThenticate pystyvät luotettavasti tunnistamaan tekoälyn tuottaman plagiaatin. Heidän tutkimustaan kuitenkin rajoittaa pieni näytemäärä sekä ihmisten kirjoittamien, ns. kontrollinäytteiden puuttuminen. Mielestäni vastaavalle tutkimukselle on olennaista myös se, että osaako työkalu luokitella ihmisen kirjoittaman tekstin oikein. Käytännön seuraukset voisivat olla vakavia, mikäli esimerkiksi akateemisessa ympäristössä otettaisiin käyttöön työkalu, jonka luokittelukykyä ei ole kokeiltu myös ihmisten kirjoittamilla näytteillä.

Lupaavimpia tuloksia ovat saavuttaneet Desaire., ym (2023a, b). Heidän luomansa malli perustuu tekstin ominaisuuksiin ja se ei syrji vieraskielisiä opiskelijoita, kuten moni kaupallinen malli (Liang., ym 2023). Heidän kokeilunsa mukaan yksinkertainen pyyntö monimuotoistaa käytettyä kieltä laskee tunnistustyökalun suorituskykyä jopa 50 %. Desairen., ym (2023b) Malli vaatii hyvän vähän opetusmateriaalia ja sitä ei voi kiertää prompt engineeringillä, jonka Liang., ym (2023) kertovat olevan useiden kaupallisten tunnistustyökalujen akilleen kantapää.

Desairen., ym (2023a, b) kaltainen ominaisuusperusteinen malli on pienen opetusmateriaalin ja kevyen rakenteensa ansiosta myös taloudellisesti hyvin saavutettavissa. Desairen., ym (2023b) mallin yksi suurimpia hyötyjä oli sen erinomaisena säilynyt suorituskyky riippumatta ChatGPT:n versiosta. Desairen., ym (2023) sekä Waltersin (2023) suorittamien kokeilujen perusteella kaupallisten mallien suorituskyky laskee GPT-4.0 version kirjoittamia tekstejä analysoidessa, kun tuloksia verrataan GPT-3.5:den tuottamien tekstin analysoinnin tuloksiiin.

Kaupallisten työkalujen suorituskyky on tutkimustiedon perusteella vaihtelevaa. Esimerkiksi Walters., (2023) tutkimuksessa kaupallisista malleista parhaiten suoriutuivat Copyleaks, Originality sekä TurnitIn. Työkalujen suorituskyky oli GPT-3.5, -4.0 sekä ihmisten kirjoittamien artikkeleiden lajittelussa lähes moitteeton. Hänen tutkimuksensa tuloksia analysoidessa tulee kuitenkin huomioida, että hän ei kokeillut prompt-engineeringiä tai tekoälyn antamien vastausten muokkaamista, joka esimerkiksi Weber-Wulffin., ym (2023) tekemien kokeilujen perusteella aiheutti merkittävää laskua tunnistustyökalujen suorituskyvyssä. Lisäksi työkaluilla on taipumusta luokitella tekstejä mielummin ihmisen kuin AI:n kirjoittamiksi (Weber-Wulff., ym (2023)).

Tämänhetkisen tutkimuksen perusteella ainakaan nykyiset kaupalliset mallit eivät kykene sellaisenaan tekemään tunnistamista riittävän luotettavasti, että niitä voisi käyttää itsenäisesti. Toistaiseksi tunnistusmallien antaman tuloksen arvioijaksi ja lopullisen päätöksen tekijäksi tarvitaan ihminen. Lisäksi tämänhetkisiä työkaluja on helppo kiertää prompt engineeringillä tai vähäisellä tekoälyn antaman vastauksen muokkaamisella (Gao., ym 2023, Desaire., ym 2023a, b, Weber-Wulff., ym (2023).

5. Yhteenveto

Käsittelyosuudessa katselmoidun aikaisemman tutkimuksen perusteella voidaan osoittaa, että luotettaville tunnistustyökaluille ja menetelmille on suuri tarve. Esimerkiksi Gao., ym (2023) pystyivät tutkimuksessaan pienestä otoskoosta huolimatta luotettavasti osoittamaan, että ihminen ei itsenäisesti kykene luotettavasti erottamaan tekoälyn tuottamaa tekstiä ihmisen kirjoittamasta tekstistä.

Desairen., ym (2023b) kehittämä ominaisuusperusteinen malli säilyttää kaupallisia malleja paremman suorituskyvyn GPT-4.0 version tuottaman tekstin tunnistamisessa. Tämän perusteella ominaisuusperusteinen malli voisi olla potentiaalinen lähestymistapa tulevien mallien kehittämiseen, vaikkakin Crothers., ym (2023) kertovat neuroverkkoihin perustuvien, todennäköisyyttä mittaavien mallien olevan tällä hetkellä markkinoiden tehokkaimpia tunnistimia. Ominaisuusperusteinen malli säilyttää tämänhetkisen tutkimustiedon perusteella paremmin suorituskyvyn, kun tekoälymalli päivittyy uudempaan (Desaire., ym 2023b).

Tämä voi johtua tekstiä generoivien ja todennäköisyyttä mittaavien mallien riippuvuudesta generoivan mallin käyttämään dekoodausalgoritmiin. Ominaisuusperusteisen mallin etu onkin, että se mittaa yleisesti tekstin ominaisuuksia, jotka poikkeavat ihmisen ja tekoälyn kirjoittamissa teksteissä. Näitä ominaisuuksia ovat esimerkiksi huonompi luettavuus, lausepituuden keskijakauman vaihtelu, kappaleiden pituuksien pienempi vaihtelu sekä erot synonyymien käytössä (Desaire., ym 2023a, b). Toisaalta neuroverkkomallin suurin heikkous on riipuvuus generoivan mallin dekoodaus algoritmista. Jos algoritmi muuttuu oleellisesti, niin suorituskyky ei tutkielmassa esitetyn tutkimustiedon perusteella pysy samana, vaan laskee.

Tutkielma vastaa kysymyksiin siitä, että mitä tekoälyn tuottaman tekstin tunnistamiseen kykeneviä työkaluja on olemassa, miten työkalut toimivat sekä miten luotettavia työkalut ovat.

Tutkielmassa ei ole pureuduttu tarkemmin siihen, että miten tekoälyn tuottamia tekstejä tunnistavia työkaluja tulisi käytännössä implementoida ja miten oppilaitosten tulisi niitä käyttää. Ulkopuolelle on myös rajattu muiden kielimallien kuin GPT- mallien tuottaman tekstin tunnistaminen. Seuraavaksi olisi hyvä tutkia, että miten esimerkiksi Desairen., ym (2023b) kehittämiä keveitä ominaisuusperusteisia malleja voitaisiin kehittää ja ottaa käyttöön akateemisessa ympäristössä arvioijien tueksi. Lisäksi mallien suorituskykyä tulisi tutkia kattavammin muillakin kuin englannin kielellä toimittaessa.

Tulevaisuudessa olisi tärkeää tutkia myös erilaisten metodologioiden ikääntymistä. Esimerkiksi miten nykyiset ennakoivat neuroverkkomallit suoriutuvat parin vuoden päästä verrattuna ominaisuusperusteisiin malleihin. Mielekästä olisi myös tehdä nykyisillä malleilla laajempaa tutkimusta, etenkin prompt engineeringin vaikutuksesta työkalujen suorituskykyyn ja tarkkuuteen.

Lähteet

Amazon. (2023). *Amazon CodeWhisperer is now generally available.* Haettu 14.02.2024 osoitteesta https://aws.amazon.com/about-aws/whats-new/2023/04/amazon-codewhisperer-generally-available/

Anil, A., Saravanan, A., Singh, S., Shamim, M. A., Tiwari, K. R., Lal, H., Seshatri, S., Gomaz, S. B., Karat, T. P., Dwivedi, P., Varthya, S. B., Kaur, R. J., Satapathy, P., Padhi, B. K., Gaidhane, S., Patil, M., Khatib, M. N., Barboza, J. J., & Sah, R. (2023). Are paid tools worth the cost? A prospective cross-over study to find the right tool for plagiarism detection. *Heliyon*, 9(9), e19194. https://doi.org/10.1016/j.heliyon.2023.e19194

Brereton, P., Kitchenham, B., Budgen, D., Turner, M., & Khalil, M. B. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, *80*(4), 571–583. https://doi.org/10.1016/j.jss.2006.07.009

ContentDetector.AI. *AI Detector | AI Content Detector | ChatGPT & AI Checker*. Haettu 08.04.2024 osoitteesta https://contentdetector.ai/

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Crothers, E., Japkowicz, N., & Viktor, H. L. (2023). Machine-Generated Text: A comprehensive survey of threat models and detection methods. *IEEE Access*, *11*, 70977–71002. https://doi.org/10.1109/access.2023.3294090

Desaire, H., Chua, A. E., Kim, M. G., & Hua, D. (2023a). Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Reports Physical Science*, 4(11), 101672. https://doi.org/10.1016/J.XCRP.2023.101672

Desaire, H., Chua, A. E., Isom, M., Jarosova, R., & Hua, D. (2023b). Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science,* 4, 101426. https://doi.org/10.1016/j.xcrp.2023.101426

Elkhatat, A. M. (2023). Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *International Journal for Educational Integrity*, 19(1), 1–23. https://doi.org/10.1007/S40979-023-00137-0/FIGURES/12

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19, 17. https://doi.org/10.1007/s40979-023-00140-5

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models: Tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7, 1–23. https://doi.org/10.7717/PEERJ-CS.443/SUPP-4

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with

detectors and blinded human reviewers. *Npj Digital Medicine 2023* 6:1, 6(1), 1–5. https://doi.org/10.1038/s41746-023-00819-6

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *Annual Meeting of the Association for Computational Linguistics,* 111–116. arXiv preprint arXiv:1906.04043

Gillham Jonathan, (2023) How does AI content detection work? *Originality.AI*. (n.d.). https://originality.ai/blog/how-does-ai-content-detection-work

Goldberg, E., Driedger, N., & Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. IEEE Expert, 9(2), 45–53. https://doi.org/10.1109/64.294135

GPTZERO. *GPTZero*. Haettu 08.04.2024 osoitteesta https://gptzero.me/technology

GPT Radar. *GPT Radar | AI text detector app*. Haettu 08.04.2024 osoitteesta https://gptradar.com/

Julkaisufoorumi. (2022) *Julkaisufoorumi*. Haettu 01.03.2024 osoitteesta https://julkaisufoorumi.fi/fi/julkaisufoorumi-0

Ibrahim, H., Asim, R., Zaffar, F., Rahwan, T., & Zaki, Y. (2023). Rethinking Homework in the Age of Artificial Intelligence. *IEEE Intelligent Systems*, 38(2), 24–27. https://doi.org/10.1109/MIS.2023.3255599

Liang, W., Yüksekgönül, M., Mao, Y., Wu, E. Q., & Zou, J. (2023b). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Nat Friedman. (2021). Introducing GitHub Copilot: your AI pair programmer - *The GitHub Blog*. https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/

Nuha, A. (2023). Detecting if a text is AI generated. Crossplag. https://crossplag.com/detecting-if-a-text-is-ai-generated/

OpenAi., (2022). Introducing ChatGPT, *OpenAI Blog*. Haettu 30.01.2024 osoitteesta. https://openai.com/blog/chatgpt

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87. https://doi.org/10.1017/s1351324997001502

See, A., Pappu, A., Saxena, R., Yerukola, A., & Manning, C. D. (2019). Do Massively Pretrained Language Models Make Better Storytellers? CoNLL 2019 - *23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, 843–861. https://doi.org/10.18653/v1/k19-1079

Sapling. *Sapling* (2023). Haettu 08.04.2024 osoitteesta https://sapling.ai/ai-content-detector

Springer, S., Buta, P., & Wolf, T. C. (1991). Automatic Letter Composition for Customer Service. Innovative Applications of Artificial Intelligence, 67–83. http://aaaipress.org/Papers/IAAI/1991/IAAI91-006.pdf

Vaswani, A., Shazeer, N., taParmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Walters, W. H. (2023). The effectiveness of software designed to detect AI-Generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1). https://doi.org/10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-Generated Text. *International Journal for Educational Integrity*. https://doi.org/10.1007/s40979-023-00146-z

Y. Xiao, S. Chatterjee and E. Gehringer, "A New Era of Plagiarism the Danger of Cheating Using AI," *2022 20th International Conference on Information Technology Based Higher Education and Training (ITHET)*, Antalya, Turkey, 2022, pp. 1-6, doi: 10.1109/ITHET56107.2022.10031827.