



Käyttäjän manipulointi ja sen tunnistaminen tekoälyn keinoin

Ammattikorkeakoulututkinnon opinnäytetyö

Tietojenkäsittelyn koulutus

Kevät 2024

Taneli Nikkanen

Tietojenkäsittelyn koulutus

Tekijä Taneli Nikkanen

Työn nimi Käyttäjän manipulointi ja sen tunnistaminen tekoälyn keinoin

Ohjaaja Ismo Turve

Tiivistelmä

Vuosi 2024

Opinnäytetyön tavoitteena oli tarkastella käyttäjän manipulointia ja tekoälyä sekä tekoälyn hyödyntämistä käyttäjää manipuloivien hyökkäyksien tunnistamisessa. Työllä ei ole erillistä toimeksiantajaa.

Työ koostuu teoriaosuudesta sekä käytännön osuudesta, joten työ on sekä teoreettinen että tutkimuksellinen. Teoriaosuudessa selvitetään työn kannalta oleellisia termejä ja ilmiöitä. Lähteinä teoriaosuudessa käytettiin aiheesta julkaistuja tieteellisiä artikkeleja ja tutkimuksia, uutismediasta löytyviä julkaisuja, alan toimijoiden tuottamia artikkeleja, sekä kirjallisuuslähteitä. Käytännön osuudessa tehdyssä kehittämistyössä hyödynnettiin Python ohjelmointikieltä ja sen kirjastoja. Kehittämistyössä toteutettavan ohjelman luomisessa hyödynnettiin työn tekijän aiempaa opintojen kautta kerättyä osaamista ja erinäisiä netistä löytyviä dokumentaatioita kirjastoihin liittyen.

Teorian- ja käytännön osuuksien pohjalta voidaan todeta, että tekoäly on toimiva työkalu käyttäjän manipuloinnin tunnistamisessa. Sen tarjoamat edistykselliset toiminnot hyökkäyksien tunnistamiseksi voivat tulevaisuudessa olla avainasemassa käyttäjää manipuloivien hyökkäyksien tunnistamisessa. Tekoälyn täyttä potentiaalia ei olla kuitenkaan vielä kyetty valjastamaan, joten on mahdollista, että työssä esitellyt tekoälyn käyttömahdollisuudet ovat vanhanaikaisia uusien tekoälyn kehitykseen liittyvien harppauksien myötä.

Avainsanat käyttäjän manipulointi, tekoäly, kyberturvallisuus

Sivut 48 sivua ja liitteitä 7 sivua

The aim of this thesis was to research social engineering, artificial intelligence, and the usage of artificial intelligence for detecting social engineering attacks. The work has no separate client.

The thesis consists of theory and practical work, making it partly theoretical and partly functional. The theoretical part studies key terms and phenomena relevant to the work. Scientific articles and research, publications from news media, articles produced by people and companies in the industry, and literary sources were used as sources in the theoretical part. Python programming language and its libraries were utilized in the development project that was performed in the functional part of the thesis. The creation of the program implemented in the development work utilized the experience acquired by the author through previous studies and various documentation available online related to the libraries.

Based on the theoretical and practical sections, it can be concluded that artificial intelligence is an effective tool in identifying social engineering. Its advanced features for detecting attacks could potentially be crucial in the future for identifying social engineering attacks. However, the full potential of artificial intelligence has not yet been fully achieved, so it is possible that the opportunities for utilizing artificial intelligence presented in the work may become outdated through advancements in artificial intelligence development.

Keywords Social engineering, Artificial intelligence, Cybersecurity
Pages 48 pages and appendices 7 pages

Sanasto

| | |
|----------------------|---|
| BEC | BEC, eli Business Email Compromise, on huijausmuoto, jossa hyökkääjä hyödyntää käyttäjän manipulointia ja tietojenkalastelua hankkiakseen laittomasti rahaa tai arkaluonteista tietoa organisaatiolta. |
| Hyökkäysvektori | Hyökkäysvektori kuvaa konkreettista tapaa, jolla hyökkääjä voi päästä käsiksi tietojärjestelmiin tai sovelluksiin, kuten esimerkiksi haittaohjelmien, verkkohaavoittuvuuksien tai sosiaalisen manipulaation kautta. |
| Kielimalli | Kielimalli on tietokoneen luoma matemaattinen malli, joka pystyy ymmärtämään, tuottamaan ja käsittelemään luonnollista kieltä, kuten ihmisten puhetta tai kirjoitettua tekstiä. |
| Koneoppiminen | Koneoppiminen on tekoälyn osa-alue, joka mahdollistaa tietokonejärjestelmien kehittämisen datalla ja kokemuksilla ilman nimenomaista ohjelmointia. Koneoppiminen perustuu algoritmien kehittämiseen, jotka antavat tietokoneille kyvyn oppia ja tehdä päätöksiä tietojen perusteella. |
| Logistinen regressio | Logistinen regressio on tilastollinen menetelmä, jota käytetään ennustamaan binäärisiä tai monikategorisia lopputuloksia perustuen yhteen tai useampaan selittävään muuttujaan. |
| Neuroverkko | Neuroverkko on tietokonejärjestelmä, joka mallintaa ihmisen aivojen toimintaa niitä jäljittelevien tietorakenteiden avulla. Se koostuu neuroneista ja niiden välisten yhteyksien kerroksista, joita käytetään monimutkaisten tietojen käsittelyyn ja oppimiseen. |
| NLP | NLP, eli luonnollisen kielen käsittely, on tekoälyn haara, joka keskittyy tietokoneiden kykyyn ymmärtää, tulkita ja tuottaa ihmisten käyttämää luonnollista kieltä, kuten puhe- ja kirjoitettua tekstiä. |

| | |
|---------------|---|
| Phishing | Phishing, suomeksi tietojenkalastelu, on huijausmenetelmä, jossa hyökkääjä pyrkii hankkimaan henkilökohtaisia tietoja, kuten salasanoja, luottokorttinumeroita tai muita arkaluonteisia tietoja, esittämällä olevansa luotettava taho, kuten pankki tai yritys, yleensä sähköpostin tai verkkosivuston välityksellä. |
| Semantiikka | Semantiikka on kielitieteen osa-alue, joka tutkii merkitystä kielellisessä viestinnässä, eli sanojen, lauseiden ja muiden kielellisten yksiköiden merkitystä ja niiden välisiä suhteita. Semantiikka pyrkii ymmärtämään, miten ihmiset käyttävät kieltä kommunikoidessaan ja miten merkitys rakentuu ja välittyy kielellisten ilmausten kautta. |
| SMiShing | SMiShing on huijausmuoto, joka käyttää tekstiviestejä (SMS) houkutellakseen uhreja paljastamaan henkilökohtaisia tietojaan tai klikkaamaan haitallisia linkkejä. Tavoitteena on saada uhri jakamaan arkaluonteisia tietojaan, kuten pankkitietoja tai salasanoja, tai ohjata heidät haitallisille verkkosivustoille. |
| Syväoppiminen | Syväoppiminen on koneoppimisen ala, joka käyttää monikerroksisia neuroverkkoja tietojen analysointiin ja oppimiseen. Tämä menetelmä mahdollistaa monimutkaisten kuvien, äänien tai tekstien käsittelyn ja ennustusten luomisen niiden pohjalta. |
| Tekoäly | Tekoäly on tietojenkäsittelyn ala, joka pyrkii luomaan järjestelmiä ja ohjelmistoja, jotka kykenevät suorittamaan ihmismäisiä tehtäviä. Tämä voi sisältää muun muassa päätöksentekoa, ongelmanratkaisua, oppimista ja tunnistamista erilaisissa tilanteissa |
| TF-IDF | TF-IDF (Term Frequency-Inverse Document Frequency) on tilastollinen menetelmä, jota käytetään dokumenttien tekstien analysointiin ja painottamiseen informaation relevanssin mukaan. Se antaa painoarvon sanalle ottaen huomioon sen esiintymistiheyden dokumentissa (term frequency) ja käänteisen esiintymistiheyden kaikkien dokumenttien joukossa (inverse document frequency). |

| | |
|------------------|--|
| Verkkoharavointi | Verkkoharavointi on prosessi, jossa hyödynnetään automaattisia ohjelmia, internetistä löytyvien sivustojen skannaamisen ja tiedon keräämiseen näiltä sivuilta. |
| Vishing | Vishing on huijausmenetelmä, jossa hyökkääjä käyttää puhelinta tai VoIP-tekniologiaa (Voice over Internet Protocol) houkutellessaan uhreja paljastamaan henkilökohtaisia tietojaan, kuten pankkitietoja tai salasanoja, yleensä esittäen olevansa luotettava taho, kuten pankki tai viranomainen. |
| Whaling | Whaling on huijausmuoto, joka kohdistuu korkean tason yksilöihin, kuten yritysjohtajiin tai hallituksen virkamiehiin. Hyökkääjä esittää olevansa tunnettu henkilö tai organisaatio ja pyrkii hankkimaan arkaluonteisia tietoja tai suorittamaan taloudellisia huijauksia, kuten rahansiirtoja tai luottamuksellisten tietojen jakamista. |

Sisällys

| | | |
|-------|--|----|
| 1 | Johdanto | 1 |
| 2 | Käyttäjän manipulointi | 3 |
| 2.1 | Käyttäjän manipuloinnin määritelmä..... | 3 |
| 2.2 | Käyttäjän manipulointi -hyökkäysten menetelmät..... | 6 |
| 2.2.1 | Phishing | 6 |
| 2.2.2 | Whaling | 9 |
| 2.2.3 | Yrityssähköpostin murto | 10 |
| 2.3 | Käyttäjän manipuloinnin tunnistaminen ja hyökkäysten pysäyttäminen | 12 |
| 2.4 | Käyttäjän manipulointi -hyökkäysten vaikutukset | 14 |
| 3 | Tekoäly | 15 |
| 3.1 | Tekoälyn määritelmä..... | 15 |
| 3.2 | Tekoälyn tyypit..... | 17 |
| 3.3 | Tekoälyn osa-alueet..... | 19 |
| 3.3.1 | Koneoppiminen | 19 |
| 3.3.2 | Syväoppiminen..... | 21 |
| 3.3.3 | Luonnollisen kielen käsittely | 22 |
| 4 | Tekoäly ja käyttäjän manipulointi | 24 |
| 4.1 | Tekoäly hyökkääjien hyödyntämänä | 24 |
| 4.2 | Tekoälyn käyttö käyttäjän manipuloinnin tunnistamiseksi..... | 26 |
| 4.2.1 | Luonnollisen kielen käsittely käyttäjän manipuloinnin tunnistamiseksi..... | 26 |
| 4.2.2 | Koneoppimisen hyödyntäminen käyttäjän manipuloinnin tunnistamiseksi..... | 29 |
| 4.2.3 | Syväoppimisen hyödyntäminen käyttäjän manipuloinnin tunnistamiseksi..... | 31 |
| 4.3 | Tekoälyn käytön eettisyys ja ongelmat..... | 32 |
| 5 | Ohjelman luominen roskapostin tunnistamiseksi..... | 34 |
| 5.1 | Ohjelman suunnittelu ja käytetyt menetelmät..... | 34 |
| 5.2 | Ohjelman käyttämä data ja kirjastot | 35 |
| 5.3 | Datasetti tekoälyn kouluttamiseksi | 37 |
| 5.4 | Tekoälyn osuus ohjelmassa..... | 37 |
| 5.5 | Ohjelman toteutus..... | 38 |
| 6 | Tulokset ja jatkotutkimus..... | 45 |
| 7 | Yhteenveto..... | 47 |

| | |
|---------------|----|
| Lähteet | 49 |
|---------------|----|

Kuvat ja ohjelmakoodit

| | |
|---|----|
| Kuva 1. Enisan analysoimien kyberhyökkäysten määrät uhkatyyppin mukaan heinäkuun 2022 ja kesäkuun 2023 väliseltä ajalta | 6 |
| Kuva 2. Lockheed Martin Cyber Kill Chain® | 8 |
| Kuva 3. Päätelykykyä testaavan tehtävän esittäminen ChatGPT kielimallille | 17 |
| Kuva 4. Syötedata visualisoituna ennen klusterointia ja sen jälkeen | 20 |
| Kuva 5. Konvoluutio-neuroverkon toimintamalli..... | 22 |
| Kuva 6. PhishNet-NLP-ohjelman toiminta yksinkertaistettuna..... | 27 |
| Kuva 7. Kehittämistyön vaiheet kanban taulussa | 35 |
| Kuva 8. Ohjelmassa käyttöön otettu data tulosteena..... | 39 |
| Kuva 9. DataFrame käsittelyn jälkeen..... | 40 |
| Kuva 10. Logistisen regressiomallin tarkkuus. | 42 |
| Kuva 11. Merkitykselliset sanat luokittain. | 44 |
| | |
| Ohjelmakoodi 1, Kirjastojen tuonti..... | 38 |
| Ohjelmakoodi 2, Datan lataaminen ja tarkastaminen | 39 |
| Ohjelmakoodi 3, DataFrame muokkaaminen | 39 |
| Ohjelmakoodi 4, Funktio sanaluokkien hakemiseksi | 40 |

| | |
|---|----|
| Ohjelmakoodi 5, Funktio tekstin käsittelemiseksi | 40 |
| Ohjelmakoodi 6, Tekstin käsittely -funktion kutsuminen | 41 |
| Ohjelmakoodi 7, Tekstin muuttaminen numeeriseksi | 41 |
| Ohjelmakoodi 8, Datat jakaminen koulutus- ja testaamissarjoihin | 41 |
| Ohjelmakoodi 9, Mallin luominen ja kouluttaminen..... | 42 |
| Ohjelmakoodi 10, Ennustuksien suorittaminen ja tarkkuuden havainnointi..... | 42 |
| Ohjelmakoodi 11, Sanojen ja niiden painoarvojen hakeminen | 43 |
| Ohjelmakoodi 12, Sanojen ja arvojen lisääminen listaan..... | 43 |
| Ohjelmakoodi 13, Luodaan yleisimpien sanojen listat | 43 |
| Ohjelmakoodi 14, Tulostetaan merkityksellisimmät sanat | 43 |

Liitteet

- Liite 1. Aineistonhallintasuunnitelma
- Liite 2. ChatGPT keskustelu, Mitä tekoäly on
- Liite 3. ChatGPT keskustelu, Looginen päättely
- Liite 4. ChatGPT keskustelu, Tietojenkalasteluviesti

1 Johdanto

Miltei jokainen on varmasti elämänsä aikana saanut tekstiviestin tai sähköpostin, jossa kerrotaan esimerkiksi verkkorikollisen murtautuneen käyttäjättilille. Viestissä pyydetään painamaan linkkiä, jotta tili saataisiin palautettua oikealle omistajalle. Linkin painamisen jälkeen vastassa on sivu, joka näyttää normaalilta kirjautumissivulta. Valitettavasti kirjautumisen jälkeen murrettua käyttäjättiliä ei palauteta oikean omistajan haltuun, sillä sille ei oikeasti ollut alun perin murtauduttu. Sähköpostin linkistä avautunut kirjautumissivu olikin sähköpostin lähettäjän luoma tietojenkalastelusivusto, joka oli luotu näyttämään samalta kuin jonkin oikean palvelun sivu. Kirjautumistunnukset päättyivät sivulta sähköpostin lähettäjälle ja nyt käyttäjän tili voidaan oikeasti kaapata.

Tällaiset käyttäjää manipuloivat hyökkäykset käyttävät hyödyksi yleensä kiirettä tai niihin liittyy käyttäjän palkitseminen esimerkiksi rahallisesti. Jos tällainen tietojenkalasteluviesti ei päädy suoraan roskakoriin, joutuu jokainen itse harkitsemaan, onko viesti oikea vai onko kyse huijauksesta. Entä jos huijausviestejä voitaisiin merkitä suuremmalla varmuudella epäluotettaviksi ja siirtää roskakoriin? Tekoälyllä voidaan tehostaa tätä prosessia ja suodattaa huijausviestit pois käyttäjien silmistä tarkemmin.

Opinnäytetyössä tarkastellaan käyttäjän manipulointia ja sen tunnistamista tekoälyn keinoin. Olennaista on selvittää, millaisia hyökkäyksiä on olemassa ja mitä tunnusmerkkejä näihin hyökkäyksiin liittyy. Tarkastelun kohteena tulee myös olemaan yrityksiin ja yksilöihin suuntautuvien hyökkäysten vaikutukset. Työssä ei tarkastella käyttäjän manipulointia ja sen tunnistamista informaatiovaikuttamisessa. Tarkoituksena on keskittyä verkkorikollisten toimiin ja hyökkäyksiin, joiden motiivina on arkaluontoisen tiedon hankkiminen tai rahallinen hyöty.

Opinnäytetyössä luodaan ohjelma, jota voidaan käyttää roskapostin tunnistamiseksi. Ohjelman tavoitteena on havainnollistaa tekoälyn nykyisiä käyttömahdollisuuksia ja pohtia siihen liittyviä mahdollisia kehityskohteita.

OpenAI:n luomaa ChatGPT-kielimallia käytetään työssä havainnollistamaan tekoälyn hyökkääjille tarjoamia etuja. Kielimallit eivät virallisesti auta käyttäjiä kyseenalaisissa tehtävissä. Suoraan pyydetessä, ne kieltäytyvät tekemästä tietynlaisia asioita, jotka kielimallin kehittäjä on määrittänyt epäeettisiksi tai haitallisiksi. Työssä havainnollistetaan myös, miten tekoälyä voidaan manipuloida tekemään asioita, joista se normaalisti kieltäytyisi.

Opinnäytetyöllä ei ole erillistä tilaajaa, mutta se on tehty kohderyhmälle, joita kiinnostavat tekoälyn tuomat muutokset käyttäjän manipuloinnissa.

Tässä tutkimuksessa esitetään vastauksia seuraaviin kysymyksiin:

- Mitä käyttäjän manipulointi ja tekoäly on?
- Miten tekoälyä voidaan käyttää käyttäjän manipuloinnin tunnistamiseen?
- Mitkä ovat tekoälyn käytön eettiset haasteet käyttäjän manipuloinnin torjunnassa?

2 Käyttäjän manipulointi

Tässä luvussa selvitetään mitä käyttäjän manipulointi on, mitä erilaisia hyökkäysmenetelmiä on olemassa ja mitä tunnusmerkkejä tällaisilla hyökkäyksillä on ja miten hyökkäyksiltä voidaan puolustautua. Luvussa 2.3 käsiteltävät puolustautumiskeinot eivät perustu tekoälyyn. Tällaisia tekoälyä hyödyntäviä puolustautumiskeinoja käsitellään luvussa 4.2. Viimeisessä alaluvussa selvitetään käyttäjän manipulointi -hyökkäyksien vaikutuksia. Tätä tarkastellaan sekä yleisestä näkökulmasta, että oikeiden, onnistuneiden hyökkäyksien aiheuttamien vahinkojen kautta. Luvussa saatetaan käyttää termejä käyttäjän manipulointi ja sosiaalinen manipulointi sekaisin, mutta niillä viitataan pohjimmiltaan samaan asiaan.

2.1 Käyttäjän manipuloinnin määritelmä

Käyttäjän manipulointi -hyökkäykset kattavat kaikki hyökkäykset, joiden tavoitteena on saada kohde tekemään jotakin tai paljastamaan tietoja tekaistujen syiden nojalla. (ENISA n.d.)

Hyökkäyksissä hyödynnetään ihmisten tekemiä virheitä ja hyväuskoisuutta. Hyökkääjät ottavat yhteyden uhriin useimmiten viestin tai puhelun keinoin, lukuisia muitakin keinoja voidaan käyttää. Teema yhteydenotoissa on yleensä samanlainen, kohdetta painostetaan tekemään jotakin palkitsemisen, kiireen tai uhkailun nojalla. Uhreiksi voi valikoitua yksityishenkilöitä tai yritysten työntekijöitä, jolloin hyökkääjien varsinainen kohde on yritys ja sen tiedot tai järjestelmät. Jotta huijaus voi onnistua, täytyy uhrin luottaa huijariin. Yhteydenotot voivat tulla vaikkapa työpaikan IT-tuelta tai omalta esimieheltä. Viestissä olevaa pyyntöä on helpompi noudattaa, jos lähettäjän nimi ei herätä epäilyksiä. (F-Secure n.d.)

Hyökkäyksissä käytetyt taktiikat pohjautuvat psykologiaan, joten hyökkäyksien toiminnan ymmärtämiseksi on oleellista ymmärtää, miten ihmisten päätöksentekoprosessi toimii tai miten siihen voidaan vaikuttaa. Oikeanlaiset sävyt keskustelussa voivat vapauttaa kahdenlaisia hormoneja lukijan aivoissa, jotka hankaloittavat rationaalista päätöksentekoa. Nämä hormonit tunnetaan nimillä oksitosiini ja dopamiini. (Hadnagy, 2018, ss. 8–9) Käyttäjän manipuloinnin näkökulmasta tällainen reaktio on otollinen hyökkäyksen onnistumisen kannalta. Tästä voidaankin tehdä päätelmä, että sosiaaliset taidot ovat tällaisissa hyökkäyksissä jopa arvokkaampia kuin tekniset.

Oksitosiini on hormoni, joka on tärkeässä roolissa erilaisten kiintymyssuhteiden muodostamisessa, sillä se edistää niiden pitkäkestoisuutta. (Koivuranta, 2022, s. 4) Koivurannan mukaan Zuoxin Wang (2004) kertoo, että oksitosiini voi myös mahdollistaa dopamiinin erityksen. Yksi oksitosiinin vaikutuksista on luottamuksen tunteen muodostuminen. (Koivuranta, 2022, ss. 20–21) Jotta käyttäjän manipulointi onnistuu, täytyy hyökkääjän luoda luottamuksen tunne uhrin ja itsensä välillä. Romanssihuijauksissa tällainen hormonaalinen reaktio saattaa olla jopa välttämätön onnistumisen kannalta, sillä uhrin ei tulisi kyetä ajattelemaan rationaalisesti.

MOT:n julkaisemassa, romanssihuijauksia käsittelevässä dokumentissa haastatellaan oikeuspsykologi Julia Korkmania. Yksi kysymyksistä on, mitä psykologisia keinoja huijarit käyttävät. Korkman kertoo, että uhrin elämäntilannetta tai mielentilaa voidaan käyttää hyödyksi. Romanssihuijauksista puhuttaessa etenkin yksinäiset ihmiset ovat hyökkääjien näkökulmasta täydellisiä uhreja. Vaikka huijaus saattaa paljastua, voi uhri yhä uskoa, ettei minkäänlaista huijausta edes tapahtunut tai ainakin osa huijarin tarinasta oli totta. (Suopanki, 2020)

Käyttäjän manipulointia voidaan harjoittaa myös positiivisessa mielessä, eikä sen käyttö rajoitu ainoastaan kyberrikollisuuteen. Asianajajat ja psykologit käyttävät työssään metodeja, jotka ovat myös tuttuja käyttäjän manipuloinnissa. Näissä tapauksissa kohteen manipulointia tehdään kuitenkin hyvin aikein. Tavoitteena on saada kohde paljastamaan tietoa, jota tämä ei muuten kertoisi. (Ozkaya, 2018, s. 11) Valitettavasti verkkorikollisten suorittamana käyttäjän manipulointi on aina negatiivista. Näissä tapauksissa kohteen paljastamia tietoja ei käytetä hänen omaksi parhaakseen.

Erdal Ozkayan teoksessa *Learn Social Engineering* kuvaillaan erästä miltei onnistunutta hyökkäystä, jossa nuorena naisena esiintyvä henkilö lähestyi netin välityksellä tietoliikenneyrityksessä työskentelevää miestä. Päivien keskustelujen jälkeen naisena esiintyvä hyökkääjä lähetti miehelle viruksen sisältävän PDF-tiedoston. Mies avasi tiedoston, vaikka yrityksen henkilöstöä oli koulutettu olemaan avaamatta sähköpostien liitteitä tai linkkejä. Yrityksen tietokoneille asennettu antivirus-ohjelma tunnisti haitallisen liitteen ja esti hyökkäyksen onnistumisen. Hyökkäyksen valmistelussa on olennaista kerätä tarpeeksi tietoa uhrista, jotta hyökkääjä voi vastata uhrin tarpeita, esiintyä auktoriteettina tai herättää uhrin mielenkiinnon. Tässä tapauksessa hyökkääjä esitti jakavansa samoja kiinnostuksen kohteita kuin mies ja kykeni tätä kautta luomaan luontevan keskustelun uhrin kanssa. (Ozkaya, 2018, ss. 7–8)

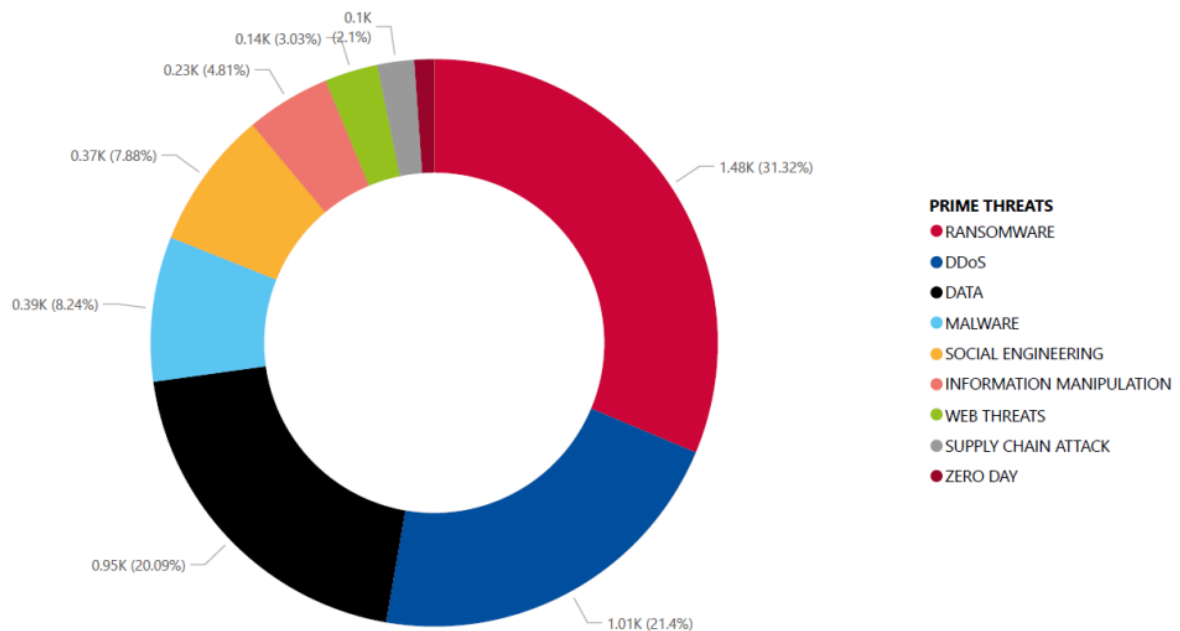
Hadnagy mukaan käyttäjän manipuloinnissa käytetyt tekniikat voidaan jakaa neljään yläkategoriaan, jotka ovat SMiShing, Vishing, Phishing ja joksikin toiseksi tekeytyminen. Jokainen hyökkäys sisältää piirteitä, joiden avulla hyökkäys voidaan määritellä johonkin näistä neljästä kategoriasta. Kategorioiden yhdistely ei ole myöskään poissuljettua, jotkin hyökkäykset voivat sisältää piirteitä useammista kategorioista. (Hadnagy, 2018, ss. 6–9) Hyökkäyksien kategoriat saattavat vaihdella tutkijoiden mukaan, mitään virallista universaalia jaottelutapaa ei tunneta.

Ozkaya listaa kirjassaan seitsemän vaihetta, joista muodostuu käyttäjän manipuloinnin viitekehys. Nämä vaiheet voidaan jaotella kolmeen kategoriaan, tiedonkeruuseen erilaisin keinoin, hyökkäysstrategian valitsemiseen kerätyn tiedon avulla ja lopuksi, hyökkäyksen suorittamiseen. (Ozkaya, 2018, s. 14) Eri tutkijoiden ja yritysten luomissa viitekehyksissä vaiheiden määrä ja sisältö vaihtelevat hieman, mutta kaikki viitekehukset sisältävät silti vaiheita yllä mainituista kategorioista.

Tietoturvyhtiö RangeForcen julkaisemassa artikkelissa käyttäjän manipulointia tarkastellaan nelivaiheisen viitekehysten avulla. Nämä vaiheet ovat tiedonkeruu, uhrin lähestyminen, hyväksikäyttö/hyökkäyksen suorittaminen ja huijauksen päättäminen. (RangeForce, n.d.) Vaiheet ovat hieman erilaiset, mutta niiden sisältö koostuu Ozkayan listaamista kategorioista. Minkä tahansa viitekehysten kautta käyttäjän manipulointi -hyökkäyksen vaiheita tarkastellaan, ovat vaiheet pääpiirteittäin aina samat.

Enisan julkaisemasta, heinäkuun 2022 ja kesäkuun 2023 väliseltä ajalta koostetusta uhkaraportista (Kuva 1) käy ilmi, että käyttäjän manipulointi on viidenneksi yleisin hyökkäystapa. (Enisa, 2023, s. 9) Käyttäjän manipulointi on siis kohtuu yleinen uhka ainakin tämän raportin perusteella.

Kuva 1. Enisan analysoimien kyberhyökkäysten määrät uhkatyypin mukaan heinäkuun 2022 ja kesäkuun 2023 väliseltä ajalta. (Enisa, 2023, s.9)



2.2 Käyttäjän manipulointi -hyökkäysten menetelmät

Seuraava listaus hyökkäysmenetelmistä ei kata läheskään kaikkia hyökkäysmenetelmiä. Tässä luvussa käsiteltävät hyökkäysmenetelmät ovat koottu tietoturvayritys Crowdstriken julkaisemasta, Bart Lenaerts-Bergmansin laatimasta, kymmentä yleisintä hyökkäystyyppiä kuvailevasta artikkelista. Tässä tutkimuksessa esitellään kolme menetelmää, joita käytetään usein käyttäjän manipulointi -hyökkäyksissä. Jos hyökkäyksiä tarkastellaan viitekehysten kautta, hyökkäysmenetelmät sijaitsevat hyökkäyksen suoritus vaiheessa.

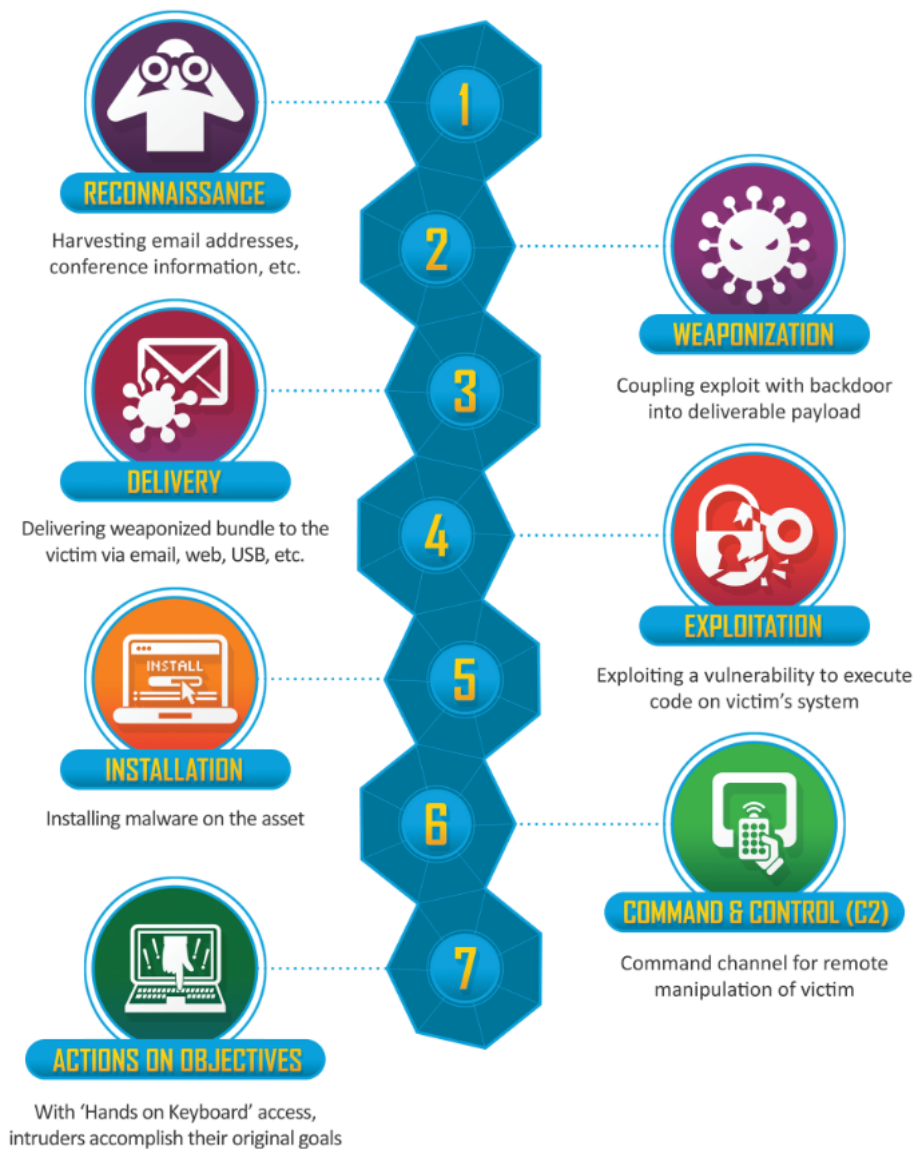
2.2.1 Phishing

Suomeksi termi phishing voidaan kääntää tietojenkalasteluksi tai verkkourkinnaksi. Tämän tyyppisissä hyökkäyksissä hyökkääjä saattaa lähestyä uhria sähköpostilla, puhelulla, tekstiviestillä tai vaikkapa sosiaalisen median kanavien kautta. Tavoitteena on kerätä tietoja uhrista, saada hänet lähettämään rahaa hyökkääjälle tai saada uhri lataamaan esimerkiksi haittaohjelma laitteelleen. Verkkourkinta on yksi yleisimmistä hyökkäystyypeistä. (Lenaerts-Bergmans, 2023)

Tietoturvyhtiö Mandiantin, vuoden 2022 tapahtumista koostetussa uhkaraportissa verkkourkinta sijoittuu toiseksi, käytetyimpiä alkuperäisiä hyökkäysvektoreita tarkastellessa. (Mandiant, 2023, s. 23) Alkuperäisestä hyökkäysvektorista puhuttaessa tarkoitetaan taktiikkaa, jota hyökkääjä käyttää saadakseen jalansijaa kohdejärjestelmäänsä tai infrastruktuuriin. Verkkourkintaa voidaan siis käyttää osana laajempia hyökkäyksiä, esimerkiksi kiristyshaittaohjelmien toimittamisessa.

Lockheed Martinin kehittämällä Cyber Kill Chainilla (Kuva 2) voidaan tarkastella kyberhyökkäysten vaiheita. Kolmas vaihe käsittelee haittaohjelman tai jonkin muun haitallisen sisällön toimittamista kohdejärjestelmään. Hyökkääjät käyttävät usein verkkourkintaviestejä tässä vaiheessa. Verkkourkintaa voidaan hyödyntää myös ensimmäisessä vaiheessa, joka käsittää tiedonkeruun.

Kuva 2. Lockheed Martin Cyber Kill Chain®. (Lockheed Martin, n.d.)



Sähköposti tai viesti sosiaalisessa mediassa täysin tuntemattomalta lähettäjältä ei ole aina tehokkain tapa herättää luottamusta uhuriin. Lähettäjän nimen tai sähköpostiosoitteen väärentäminen on yleinen keino, jota hyökkääjät käyttävät uhrin lähestyessä. Hyökkääjät saattavat tekeytyä uhrin tutuksi tai käyttää jonkin tunnetun yrityksen kanssa samankaltaista sähköpostiosoitetta. Hyökkäyksissä voidaan käyttää linkkejä, jotka johdattavat uhrin hyökkääjän tekemälle verkkosivulle. Nämä sivut lähettävät yleensä uhrin sivulle syöttämät tiedot hyökkääjälle tai sisältävät haittaohjelmia. Jos uhri on tarkoitus ohjata tällaiselle sivulle, voivat hyökkääjät luoda kopion täysin oikeasta sivusta. Tutun näköinen verkkosivu ei herätä uhrissa epäluuloa. Käyttäjän hämääminen tällaisin keinoin on yksi syy sille miksi huijauksia voi olla hankala tunnistaa, etenkin jos uhrilla ei ole kiireen takia aikaa tarkastella viestiä kauaa. (Hahnagy & Fincher, 2015, s. 4)

Hyökkääjien luomat sähköpostit ja verkkosivujen URL-osoitteet ovat usein luotu samankaltaisiksi oikeiden verkkosivujen ja sähköpostiosoitteiden kanssa. Englanniksi tästä tekniikasta käytetään nimeä domain spoofing ja suomeksi se voidaan kääntää verkkotunnuksen väärentämiseksi. Esimerkiksi pieni L-kirjain voidaan korvata isolla I:llä. Kela.fi ja Kela.fi vaikuttavat hyvin samankaltaisilta osoitteilta, mutta toinen URL on väärennetty. (Lenaerts-Bergmans, 2022)

Huijauspuhelut ovat verkkourkinnan muoto, jota hyökkääjät voivat käyttää sellaisenaan tai osana laajempaa hyökkäys kokonaisuutta. Englanniksi tätä kutsutaan vishingiksi. Hyökkääjä voi esimerkiksi esiintyä puhelussa IT-tukena ja pyytää käyttäjää avaamaan sähköpostiin saapuneen linkin, jonka kautta voidaan kirjautua tilille ja suorittaa IT-tuen pyytämät toimet. Kuten Hadnagy kertoo kirjassaan, hyökkääjän uhrille esittämä tarina vaikuttaa aina uskottavammalta, kun uhri kuulee sen useasta lähteestä. (Hadnagy & Fincher, 2015, ss. 5–6) Entäpä, jos hyökkääjät kykenevät käyttämään puhelussa uhrille entuudestaan tutun henkilön ääntä. Jatkuvasti kehittyvät ja yhä paremmin saatavilla olevat tekoälymallit voivat mahdollistaa tämän. Tätä tarkastellaan tarkemmin luvussa 4.1.

Spear phishing tarkoittaa kohdennettua tietojenkalastelua tai kohdennettua verkkourkintaa. Verkkourkinta-hyökkäykset, jotka ovat erittäin tarkasti suunniteltuja kohteen mukaan päätyvät tähän kategoriaan. Kohdennetuissa hyökkäyksissä tietoa uhrista kerätään todella tarkasti, tämä tekee varsinaisesta huijauksesta uskottavamman. Huijausviesti saatetaan rakentaa jonkin uhrille hiljattain tapahtuneen asian ympärille. Tällaisella yksityiskohtaisuudella kyetään vakuuttamaan, että viesti on saapunut joltakin uhrille tutulta henkilöltä. (Hadnagy & Fincher, 2015, s. 27) Hyökkäyksissä, joissa ei nähdä vaivaa tarkan tiedonkeruun eteen, hyökkääjät saattavat lähestyä uhria esimerkiksi yrityksen nimissä, jonka hyökkääjä olettaa olevan uhrille entuudestaan tuttu. Jos hyökkäyksen tavoitteena halutaan saada tietoa juuri tietyltä henkilöltä, ei tämä ole välttämättä paras lähestymistapa.

2.2.2 Whaling

Whaling tai suomeksi valaanpyynti on termi, jota käytetään kyberturvallisuudessa hyökkäyksistä, joiden kohteena on korkeassa asemassa oleva henkilö. Korkeassa asemassa olevaan henkilöön kohdistuvat, kohdennetut verkkourkinta -hyökkäykset ovat siis valaanpyyntiä. (Enisa, 2023, s. 40)

Näissä hyökkäyksissä kohteet ovat usein toimitusjohtajia tai vastaavassa vaikutusvaltaisessa asemassa olevia työntekijöitä. Hyökkääjä lähestyy kohdetta yleensä samanlaisin keinoin, kuin normaaleissa verkkourkinta -hyökkäyksissä, väärennetyn sähköpostiosoitteen turvin. Hyökkääjä esiintyy vaikutusvaltaisena henkilönä, mieluiten sellaisena, joka on uhrille entuudestaan tuttu. Hyökkääjät saattavat myös tekeytyä itse yrityksen toimitusjohtajaksi ja vaikuttaa omiin alaisiinsa. Korkeassa asemassa olevan auktoriteetin esittämistä pyynnöistä on vaikea kieltäytyä. Oleellista näissä viesteissä on kyetä vakuuttamaan uhri siitä, että hyökkääjä on juuri se henkilö, joksi hän on tekeytynyt. Keinot ovat samanlaisia kuin kohdennetuissa verkkourkinta -hyökkäyksissä. (Kaspersky, n.d.-a)

Korkeassa asemassa olevat henkilöt ovat houkuttelevia kohteita hyökkääjille. Tällaisilla henkilöillä on yleensä mahdollisuus päästä käsiksi suuriin rahamääriin, yrityssalaisuuksiin tai yrityksen IT-infrastruktuuriin. Heillä voi myös olla tarpeeksi valtaa yrityksen sisällä tehdä merkittäviä päätöksiä ilman erillistä hyväksyntää. (Cisco, n.d.)

Suomalaisille tuttu esimerkki tällaisesta hyökkäyksestä voi olla jääkiekko seura SaiPan johdolle vuonna 2022 sattunut tapaus. SaiPan toimiston työntekijä sai sähköpostia turkkilaisilta sijoittajilta, jotka halusivat tehdä yhteistyötä seuran kanssa. Seuran toimitusjohtaja Jussi Markkanen tapasi huijarit kasvotusten. Fyysiset tapaamiset uhrin ja huijareiden välillä eivät ole kovin yleisiä, mutta muuten tapauksesta löytyy paljon samankaltaisuuksia normaalien korkeassa asemassa oleviin henkilöihin kohdennettujen verkkourkinta hyökkäysten kanssa. Markkanen jatkoi neuvotteluja sijoittajien kanssa ja teki lopulta itsenäisen päätöksen 100 000 euron siirtämisestä ulkomaalaiselle tilille. (Kunnari, 2023) Kohteeksi päättyi seurassa vaikutusvaltaisessa asemassa toimiva henkilö, joka kykeni tekemään itsenäisiä päätöksiä suurien rahasummien siirtämisestä. Uhrin profiili oli siis juuri sellainen, kuin tällaisissa hyökkäyksissä tai huijauksissa yleensä.

2.2.3 Yrityssähköpostin murto

Englanniksi yrityssähköpostin murto tunnetaan lyhenteellä BEC, joka tulee sanoista business email compromise. Käytän tässä luvussa näistä hyökkäyksissä puhuttaessa BEC-lyhennettä.

Hyökkääjien valikoimat uhrin ovat usein samankaltaisia kuin whaling-hyökkäyksissä, korkeassa asemassa olevia, vaikutusvaltaisia henkilöitä. Tavoitteena on varastaa rahaa tai kriittisiä tietoja. Yrityksistä kohteiksi valikoituu yleensä toimitusjohtajia, uusia työntekijöitä, talousosaston työntekijöitä tai henkilöstöhallinnan työntekijöitä. Uudet työntekijät eivät erota

yrittäjien virallisia sähköposteja väärennetyistä yhtä helposti, kuin kauemmin töissä olleet henkilöt. Talousosaston - ja henkilöstöhallinnan työntekijöillä on pääsy yrityksen rahoihin ja henkilötietoihin. Toimitusjohtajilla ja muilla korkeassa asemassa olevilla työntekijöillä on vaikutusvaltaa yrityksessä. Hyökkääjille nämä kohteet ovat optimaalisia heidän tavoitteidensa saavuttamiseen. (Microsoft, n.d.)

Kaksi yleistä BEC-hyökkäyksissä käytettyä taktiikkaa on joksikin toiseksi tekeytyminen tai EAC. Jos taktiikkana on tekeytyä joksikin toiseksi, uhria lähestytään väärennetyin sähköpostiosoitteen turvin. Viesti väärennetään vaikuttamaan saapuneen joltakin auktoriteettiasemassa olevalta henkilöltä, esimerkiksi jäljittelemällä tämän sähköpostiosoitetta. Lyhenne EAC tulee sanoista email account compromise. EAC-hyökkäyksissä käytetään murrettua sähköpostia uhrin lähestymiseen. Viesti, jolla uhria lähestytään, tulee siis aivan oikeasta osoitteesta. (Proofpoint, n.d.)

Adversary-in-the-middle phishing, lyhennettynä AiTM phishing, on taktiikka, jota hyökkääjät hyödyntävät BEC-hyökkäyksissä. Tällaiset hyökkäykset lukeutuvat EAC-taktiikkaa hyödyntäviin hyökkäyksiin. Nämä hyökkäykset ovat hyvin monimutkaisia ja ne sisältävät useita vaiheita. Microsoft Threat Intelligencen luomassa artikkelista annetaan esimerkki oikeasta hyökkäyksestä. Aluksi uhri houkutellessaan verkkourkintaviestin avulla välityspalvelimelle, jonka avulla hyökkääjät saavat kerättyä uhrin kirjautumistiedot ja istuntokohtaiset evästeet. Näillä tunnuksilla hyökkääjä kykenee kirjautumaan uhrin sähköpostiin. Vaikka monivaiheinen tunnistautuminen olisikin uhrilla käytössä, voi hyökkääjä ohittaa sen istuntokohtaisten evästeiden avulla. Murretulla sähköpostilla hyökkääjä kykenee suorittamaan varsinaisen BEC-hyökkäyksen. (Microsoft Threat Intelligence, 2022) Päällisin puolin hyökkäys voi vaikuttaa yksinkertaiselta, mutta Microsoftin raportissa yksityiskohtaisesti kuvailuista vaiheista käy ilmi hyökkäyksen vaatima tekninen osaaminen.

Hyökkääjä voi tarkastella aiempia viestejä postilaatikossa ja käyttää niitä oman huijausviestinsä luomiseen. Jos sähköpostia käytetään laskujen lähettämiseen, on hyökkääjän helppo väärentää lasku aiempien viestien perusteella ja vaihtaa väärennetty tilinumero laskuun. Hyökkääjät saattavat myös hyödyntää jo käynnissä olevaa viestinvaihtoa ja jatkaa keskustelua. Näissäkin tapauksessa hyökkääjä lähettää tekaistun laskun uhrilleen. Välttääkseen kiinnijäämisen, hyökkääjät poistavat huijaukseen liittyvät viestit. (Microsoft Threat Intelligence, 2022)

Aiemmin luotetulta sähköpostitililtä saapunut, hyökkääjän tekaisema lasku, joka ei eroa aiemmista laskuista muuten kuin tilinumeron osalta, on äärimmäisen hankala todeta

huijaukseksi. Vaikka hyökkäys vaatii onnistuakseen useampia vaiheita, voi lopullisen tavoitteen saavuttaminen osoittautua helpommaksi, kuin normaaleissa verkkourkinta – tai whaling -hyökkäyksissä. Uhrin päätöksentekoon ei tarvitse vaikuttaa psykologisin keinoin eikä luottamuksen rakentamiseen tarvitse käyttää aikaa, sillä uhri luottaa jo valmiiksi sähköpostin lähettäjään ja viestin sisältöön. Toki AiTM phishing tekniikkaa hyödyntävissä hyökkäyksissä, hyökkääjältä vaaditaan enemmän teknistä osaamista, kuin tavallisissa verkkourkinta- tai BEC-hyökkäyksissä.

2.3 Käyttäjän manipuloinnin tunnistaminen ja hyökkäysten pysäyttäminen

Hyökkäykset, joissa hyökkääjät käyttävät murrettua sähköpostia uhrin lähestymiseen ja huijaamiseen voidaan pysäyttää jo ennen varsinaista hyökkäystä. Sähköpostitileille voidaan asettaa monivaiheinen tunnistautuminen. Jos hyökkääjä on saanut haltuunsa ainoastaan tilin salasanan ei kirjautuminen onnistu. Toisaalta, kuten luvusta 2.2.3 kuvaillusta Microsoft Threat Intelligencen esimerkistä käy ilmi, myös monivaiheinen tunnistautuminen voidaan ohittaa joissakin tapauksissa. Työntekijöitä tulisi kouluttaa BEC-hyökkäyksiin liittyvien viestien tunnistamisessa. Laskuihin liittyviä asioita ei myöskään tulisi hoitaa sähköpostin välityksellä, vaan erillisellä palvelulla, joka on luotu siihen tarkoitukseen, tällöin huijarien sähköpostilla lähettämättä laskut herättäisivät epäilyksen niiden alkuperästä helpommin. Myös tekoälyä voidaan hyödyntää tavallisesta poikkeavan viestinnän tunnistamiseksi, tästä lisää luvussa 4.2. (Microsoft n.d.) Muitakin puolustautumis- ja tunnistuskeinoja löytyy edellä mainittujen lisäksi.

Käyttäjän manipuloinnin tunnistamiseen ja hyökkäyksiltä puolustautumiseen löytyy myös yleisiä keinoja, joita voidaan soveltaa kaikenlaisiin hyökkäyksiin. Yllä olevat BEC-hyökkäyksiin liittyvät tunnistus- ja puolustautumiskeinot eivät välttämättä tarjoa apua muiden hyökkäyksien kanssa. Useat käyttäjän manipulointi -hyökkäykset hyödyntävät verkkourkintaa jossakin vaiheessa hyökkäystä, joten yritysten ja yksityishenkilöiden kannalta voikin olla tärkeämpää painottaa tällaisten hyökkäysten tunnistamista.

Jokainen voi esittää itselleen muutamia kysymyksiä, joiden pohtiminen voi auttaa verkkourkinta -viestin tunnistamisessa. Onko sähköpostin lähettäjä entuudestaan tuttu, vaikuttaako viestissä esitetty pyyntö kohtuulliselta tai järkevältä, oliko viestin saapuminen oletettavissa ja koetetaanko viestissä pelotella tai herättää mielenkiintoa odottamattomalla tavalla. (Hadnagy & Fincher, 2015, s. 77) Lähestyykö esimerkiksi pankki minua yleensä tällä

tavalla, pyytäisivätkö he minua kirjautumaan tililleni sähköpostin linkin kautta, onko tilini oikeasti vaarantunut. Tällaisella ajatusketjulla viesti alkaa vaikuttamaan hieman järjettömältä ja hälytyskellojen kuuluisi alkaa soimaan.

Toisaalta hyökkääjät koettavat rakentaa viestinsä sellaisiksi, että uhri ei käytä tällaista kriittistä ajattelua tarkastellessaan viestin sisältöä. Hyökkääjien tavoitteena on lähestyä uhria keinolla, joka saa uhrin tekemään päätöksiä tunteiden vallassa. Viesteistä löytyy usein muitakin viitteitä huijaukseen, joiden avulla huijaus voidaan tunnistaa mahdollisesti helpommin. Mahdolliset viestin sisältämät linkit tulisi tarkastaa. Huijaus on melko helppo tunnistaa, jos linkin URL kuuluu jollekin täysin tuntemattomalle sivulle. Jos URL on väärennetty vaikuttamaan samalta jonkin oikean verkkosivun kanssa, voi sen tunnistaminen olla hankalampaa. Teknisempi tapa varmistaa, onko viestin lähettäjä todellakin kuka hän väittää olevansa, on tarkastella sähköpostin ylätunnisteita. Näistä ylätunnisteista voidaan esimerkiksi tarkastaa mistä verkkotunnuksesta viesti on oikeasti lähetetty. (Hadnagy & Fincher, 2015, s. 77–90)

Muita yleisiä verkkourkintaviestien tunnusmerkkejä ovat kirjoitusvirheet, ylimääräiset liitetiedostot, epätavalliset tervehdykset tai puhuttelutavat kuten herra tai rouva ja henkilökohtaisten tietojen pyytäminen. (Lenaerts-Bergmans, 2021) Tällaisia tunnusmerkkejä ei yleensä löydy kohdennettuihin verkkourkinta-hyökkäyksiin liittyvistä viesteistä. Nämä merkit pätevät lähinnä verkkourkintaan, joka kohdistuu useaan henkilöön samanaikaisesti.

Tuntemalla nämä verkkourkintaviestien merkit ja tunnistamiskeinot hyökkäyksen kohde kykenee pysäyttämään hyökkäyksen heti alussa. Jos käyttäjä lankeaa huijauksen uhriksi ei tällaisilla inhimillisillä keinoilla kyetä hyökkäystä välttämättä enää pysäyttämään. Kaikkein tehokkainta olisi siis käyttää sekä teknisiä puolustautumiskeinoja sekä kouluttaa ihmisiä tunnistamaan huijauksia ja neuvoa, miten huijareiden kanssa tulisi toimia.

Virustorjuntaohjelmien asentaminen ja ylläpito auttaa estämään hyökkääjien tavoitteet haittaohjelmien lataamisesta uhrin koneelle. Käyttäjätilien oikeuksien tulisi myös olla rajoitettuja. Jos tiliä käytetään esimerkiksi lähinnä sähköpostien lukemiseen ja lähettämiseen ei tilillä tarvitse olla ylläpitäjän oikeuksia, joita hyökkääjät voivat hyödyntää hyökkäyksen myöhemmissä vaiheissa. Samaa salasanaa ei tulisi myöskään käyttää usealla tilillä. Jos hyökkääjät onnistuvat tietojenkalastelun avulla selvittämään käyttäjän salasanan, voidaan vahinkoja rajoittaa tällä keinolla. Kaksi- tai monivaiheinen tunnistautuminen tulisi olla myös käytössä, etenkin kriittisillä tileillä. (Kaspersky n.d.-b) Näillä keinoilla hyökkääjien toimet

onnistuneen käyttäjän manipuloinnin jälkeen voidaan pysäyttää tai niiden vaikutuksia voidaan rajoittaa.

2.4 Käyttäjän manipulointi -hyökkäysten vaikutukset

Huijauksen uhriksi joutuminen voi koitua äärimmäisen kalliiksi sekä yrityksille, että yksityisille henkilöille. Aihetta käsittelevässä Ylen julkaisemassa artikkelissa kerrotaan suomalaisten menettäneen yhteensä 32,4 miljoonaa euroa verkkohuijareille vuonna 2022. Tämä luku on saatu pankkien tilastoista. (Lindroos, 2023)

Ylen tutkivan journalismin toimituksen MOT:in julkaisemassa dokumentissa Suuri suomalainen romanssihuijaus päästään kuulemaan huijauksen uhriksi joutuneen henkilön mietteitä uhriksi joutumisesta. Itse huijauksessa käytettiin normaaleja sosiaalisen manipuloinnin keinoja. Huijari tekeytyi joksikin toiseksi, lähestyi uhria sosiaalisen median välityksellä ja painosti uhria tekemään ajattelemattomia ratkaisuja luomalla kiireen tunnun. Jutussa haastateltavalta uhrilta oli kavallettu noin 10 000 euroa. Luottotiedot menivät samalla. (Suopanki, 2020)

Haastattelusta käy ilmi, että rahallinen menetys ei ole ainoa asia, joka vahingoittaa hyökkäyksen uhreja. Huijauksen paljastuttua haastateltava uhri kuvaili omaa oloansa typeräksi ja kertoi häpeävänsä tapahtunutta niin paljon ettei kehdannut kertoa siitä juuri kenellekään. Myös luottamus muihin ihmisiin kärsi. Uhri kertoo, ettei uskalla enää hyväksyä kaveripyynnöitä tuntemattomilta sosiaalisessa mediassa. Huijauksen aiheuttamat taloudelliset ongelmat aiheuttivat myös suunnatonta stressiä. Samassa jutussa on myös haastateltu oikeuspsykologi Julia Korkmania asiantuntijana. Hän kuvailee uhrien henkisiä tuntemuksia hyvin samalla tavalla, kuin jutussa haastateltu uhri. Häpeä on suuri ja muille on hankala kertoa huijatuksi joutumisesta. (Suopanki, 2020)

Yrityksille hyökkäykset aiheuttavat sekä rahallisia tappioita, että imago haittoja. Vuonna 2019 Toyota Motor Corporationiin kohdistui käyttäjän manipulointi -hyökkäys, josta koitui kymmenien miljoonien vahingot. Hyökkääjät käyttivät hyökkäyksessä BEC-menetelmää. Hyökkääjät saivat väärennetyn laskun avulla kavallettua yrityksen tytäryhtiöltä noin 37 miljoonaa dollaria. (Gatlan, 2019)

3 Tekoäly

Tässä luvussa selvitetään mitä tekoäly on, mihin eri tyyppisiin tekoäly voidaan jaotella ja minkälaisiin osa-alueisiin tekoäly voidaan jakaa. Tarkastelun kohteena on myös yleiset käyttötarkoitukset ja -mahdollisuudet. Tässä luvussa ei käsitellä tekoälyn eettisyyttä, sillä sitä tarkastellaan myöhemmissä luvuissa opinnäytetyön aiheen kannalta oleellisten tekoälyn käyttötapojen kautta.

3.1 Tekoälyn määritelmä

Tekoäly voidaan määritellä hieman vaihtelevasti. Katsotaan miten tekoäly itse määritteli itsensä kysymällä tätä ChatGPT:ltä.

Keskustelu ChatGPT:n kanssa:

Kysymys: Miten määrittelisit tekoälyn omin sanoinesi?

ChatGPT kertoo tekoälyn olevan tietojenkäsittelyn osa-alue, joka keskittyy järjestelmien kehittämiseen ja ohjelmointiin niin, että ne voivat suorittaa tehtäviä, jotka yleensä vaativat inhimillistä älykkyyttä. Pyrkimyksenä on jäljitellä ihmisten kykyjä kuten oppimista, ongelmanratkaisua, luonnollisen kielen ymmärtämistä ja ongelmanratkaisua. Tekoälyllä pyritään antamaan tietokoneille kyky suorittaa tehtäviä, joiden katsotaan olevan eräänlaisia osoituksia älykkyydestä. (ChatGPT 3.5, 2024) Tämä keskustelu löytyy liitteestä 2 kokonaisuudessaan.

Tekoäly voidaan jakaa kolmeen luokkaan, jotka ovat englanniksi, artificial narrow AI, artificial general intelligence ja artificial superintelligence. Ainoa tekoälyn taso, joka on tällä hetkellä saavutettu, on artificial narrow AI, suomeksi kapea tekoäly. (IBM Data and AI Team, 2023) Tässä luvussa ei paneuduta tarkemmin tekoälyn eri luokkiin. Tässä vaiheessa on silti hyvä tietää, että tekoälyn täyttä teoreettista potentiaalia ei ole vielä kyetty saavuttamaan.

Elaine Woon kirjoittamassa artikkelissa tarkastellaan tekoälyn isänä pidetyn John McCarthyn elämää ja työtä. McCarthyn katsotaan keksineen termi Artificial Intelligence. Hän oli ensimmäisten tutkijoiden joukossa selvittämässä tietokoneiden mahdollisuuksia suorittaa tehtäviä ihmismäisellä tavalla. McCarthy järjesti ensimmäisen tekoälykonferenssin, jossa pohdittiin keinoja, joilla tietokone voisi ratkoa ongelmia, harjoittaa abstraktia ajattelua ja

kehittää itseään kuten ihmiset. Merkittävänä tästä konferenssista tekee se, että se järjestettiin 1950-luvun puolivälissä. Näihin McCarthy ja muiden silloisten tutkijoiden esittämiin kysymyksiin etsitään vastauksia yhä. (Woo, 2014) Konsepti ihmisen vertaisen ja jopa älykkäämmän koneen luomisesta on ollut olemassa kauan, mutta läpimurtoa ei ole vielä tapahtunut.

Viime aikoina paljon puhuttanut ChatGPT ja muut vastaavat suuret kielimallit voivat äkkiseltään vaikuttaa paljon älykkäämmiltä kuin ihmiset. Kielimallit kykenevät tuottamaan toimivaa ohjelmakoodia, luomaan vaikuttavia kuvia tai vastaamaan kiperiin kysymyksiin. Silti tietyt tehtävät, jotka vaativat esimerkiksi loogista päättelykykyä osoittautuvat haastaviksi tällaisille kielimalleille.

Meghan Bartelsin kirjoittamassa artikkelissa tarkastellaan ChatGPT:n kykyä ratkaista loogista päättelyä vaativia pulmia. Osaan pulmista löytyi ratkaisu, osaan ei. Artikkelista löytyy alan asiantuntijan, Merve Hickokin näkemyksiä syistä sille, että tietynlaiset yksinkertaiset tehtävät osoittautuvat haastaviksi muuten älykkäiden oloiselle kielimallille. Hickok kertoo, ettei tekoäly kykene luomaan päätelmiä, vaikka välillä vastaus saattaa siltä vaikuttaa. Kielimallien harjoittamiseen käytetty data määrittää niiden vastaukset, eivätkä ne kykene luomaan uusia päätelmiä itsenäisesti. (Bartles, 2023) Yksinkertainen päättelytehtävä, joka voitaisiin esittää esimerkiksi ala-asteen matematiikan tunnilla, ei luonnistunut ChatGPT:ltä (Kuva 3).

Kuva 3. Päättelykykyä testaavan tehtävän esittäminen ChatGPT kielimallille. (ChatGPT 3.5, 2024)



You

There are two neighbours, one of them is swedish and the other is finnish. One of them has a dog and the other has a cat. The swedish lives in a blue house. The dog owner lives in a red house. Which house is the cat living in?



ChatGPT

Based on the information provided:

1. The Swedish neighbor lives in a blue house.
2. The one with the dog lives in a red house.

Since there are only two houses and both neighbors have been assigned a color and a pet, it can be inferred that the Finnish neighbor, who is not the one with the dog, must have the cat. Therefore, the cat is living in the blue house, which is the one owned by the Swedish neighbor.



ChatGPT:n vastaus kysymykseen on vähintäänkin sekava, päättelystä puuttuu johdonmukaisuus. Syytä väärään vastaukseen kysyttäessä ChatGPT kertoo alkuperäisen, kuvassa näkyvän vastauksen sisältäneen oikean ratkaisun, mutta selitys ei ollut tarpeeksi selkeä. Alkuperäinen vastaus sisältää oikean ratkaisun, mutta toisaalta myös väärän ratkaisun. Merkittävin osuus kuitenkin tässä vastauksessa on sen loppu, ChatGPT:n mukaan virheitä saattaa tapahtua kielen ja päättelyn monimutkaisuuden takia. On siis selvää, että tekoälyn kehittämisessä on vielä työtä jäljellä. Tämä ChatGPT:n kanssa käyty keskustelu löytyy liitteestä 3.

3.2 Tekoälyn tyypit

Kuten jo edellisessä luvussa kävi ilmi, tekoäly voidaan jakaa kolmeen eri tyyppiin, jotka ovat kapea tekoäly, yleinen tekoäly ja superälykäs tekoäly. Tekoäly jaotellaan näihin luokkiin sen kykyjen mukaan. Tekoäly voidaan jakaa myös neljään eri luokkaan, tarkastelemalla sen toiminnallisuutta. Englanniksi nämä toiminnallisuutta kuvastavat luokat tunnetaan nimillä

Reactive Machine AI, Limited Memory AI, Theory of Mind AI ja Self-Aware AI. (IBM Data and AI Team, 2023)

Self-Aware AI, suomeksi itsestään tietoinen tekoäly, on superälykkääseen tekoälyyn rinnastettava luokittelu. Jotta tekoäly voisi olla itsestään tietoinen, sen täytyisi kyetä tunnistamaan omat kykynsä ja omata ihmismäiset tunteet ja ajatukset. Tämän lisäksi se osaisi muodostaa itsenäisiä uskomuksia ja tarpeita. Kuten superälykäs tekoäly, myös siihen rinnastettava itsestään tietoinen tekoäly on täysin konseptuaalinen termi, tällaisen tekoälyn luomisessa ei olla vielä onnistuttu. Myös Theory of Mind AI, suomeksi mielen teoria tekoäly, on toistaiseksi ainoastaan konsepti. Tällainen tekoäly olisi rinnastettavissa yleiseen tekoälyyn. Tekoälyllä ei olisi omaa tunnemaailmaa, mutta se kykenisi tunnistamaan erilaisia tunteita keskustelusta ihmisen kanssa ja mukauttamaan omaa toimintaansa niiden mukaan. (IBM Data and AI Team, 2023)

Reactive Machine AI, suomeksi reaktiivinen tekoäly, ei kerää tietoa vaan suorittaa tehtävänsä tilanteen mukaan. Tällainen tekoäly toimii tiedon pohjalta, joka on saatavilla päätöksenteko hetkellä, aiemmat lopputulokset eivät vaikuta päätöksiin. Päätöksenteko on siis reaktiivista. Toiminta perustuu suurien tietomäärien analysoimiseen lyhyessä ajassa. Limited Memory AI, suomeksi rajoitetun muistin tekoäly, tekee päätöksiä sekä aiempien tapahtumien kautta kerätyllä tiedolla sekä päätöksentekohetkellä saatavilla olevan tiedon avulla. Tällaisia tekoälyjä koulutetaan jatkuvasti uudella tiedolla, joten niiden toiminta kehittyy jatkuvasti. (IBM Data and AI Team, 2023)

Reaktiivinen tekoäly on kaikista yksinkertaisin tekoälyn muoto. Koneet, joiden toiminnallisuus voidaan määritellä reaktiiviseksi, ovat suunniteltu suorittamaan äärimmäisen rajattuja tehtäviä. Shakkietokone on hyvä esimerkki reaktiivisesta tekoälystä. Aiemmat pelit eivät vaikuta tällaisten koneiden päätöksentekoon. Seuraava siirto määrittyy pelin sen hetkisen asetelman mukaan. Jos sama tilanne toistuu uudessa pelissä, ei aiemmassa ottelussa tehdyn valinnan lopputulema vaikuta uuteen tilanteeseen, siirto on aina sama. Tällainen luotettavuuden taso on äärimmäisen tärkeä esimerkiksi itseohjautuvissa autoissa. Äkillisessä väistötilanteessa ei ole tarpeellista analysoida aiempia tapahtumia ja lopputulemia ja tehdä päätöstä niiden avulla, riittää, että eteen ilmestyneen esteen kanssa törmääminen vältetään. (Hintze, 2016)

Rajoitetun muistin tekoäly kykenee keräämään ja taltioimaan dataa, jotta se voi hyödyntää sitä päätöksenteko hetkellä. Tätä tietoa ei kuitenkaan hyödynnetä oppimiseen, tavoitteena on vain kerätä dataa pidemmältä aikaväliltä, jotta päätökset voidaan muodostaa laajemmalla

tietopohjalta. Kun päätös on tehty ja toiminto on suoritettu, tallennettua dataa ei enää käytetä. Itseohjautuvissa autoissa voidaan hyödyntää tällaista toiminnallisuutta. Kaistaa vaihdettaessa on tarpeellista tarkkailla ympäristöä. Onko liikenne viereisellä kaistalla pysähtynyt vai kulkeeko viereisellä kaistalla takanapäin oleva auto huomattavasti kovempaa, kuin kaistaa vaihtava auto. Dataa täytyy kerätä hieman pidemmältä aikaväliltä, jotta muiden autojen nopeuksia ja niiden vaihteluita voidaan tulkita tekoälyn keinoin. Jos nopeudet pysyvät tasaisina voi tekoäly todeta, että kaistanvaihto on turvallista. (Hintze, 2016) Jos kaistanvaihtoon käytettäisiin reaktiivista tekoälyä, ei autojen nopeuksien vaihtelua huomioitaisi ja kaistanvaihto tapahtuisi sen perusteella, onko viereisellä kaistalla tilaa. Tämä aiheuttaisi taatusti vaaratilanteita ja onnettomuuksia.

3.3 Tekoälyn osa-alueet

Tekoälyä voidaan siis jaotella luokkiin sen kykyjen ja toiminnallisuuden mukaan. Tällaiset luokittelut tarjoavat yleispätevän kuvan tekoälystä ja sen käyttömahdollisuuksista. Tekoälyn osa-alueista puhuttaessa tarkastellaan tekoälyä teknisemmältä kannalta. Osa-alueissa on kyse siitä, miten ja millaista tietoa tekoäly käsittelee ja mitä se tuottaa tämän tuloksena. Jos halutaan tarkastella tekoälyn käyttöä kyberturvallisuudessa, on oleellisempaa tutkia osa-alueita eikä tekoälyn tyyppejä.

3.3.1 Koneoppiminen

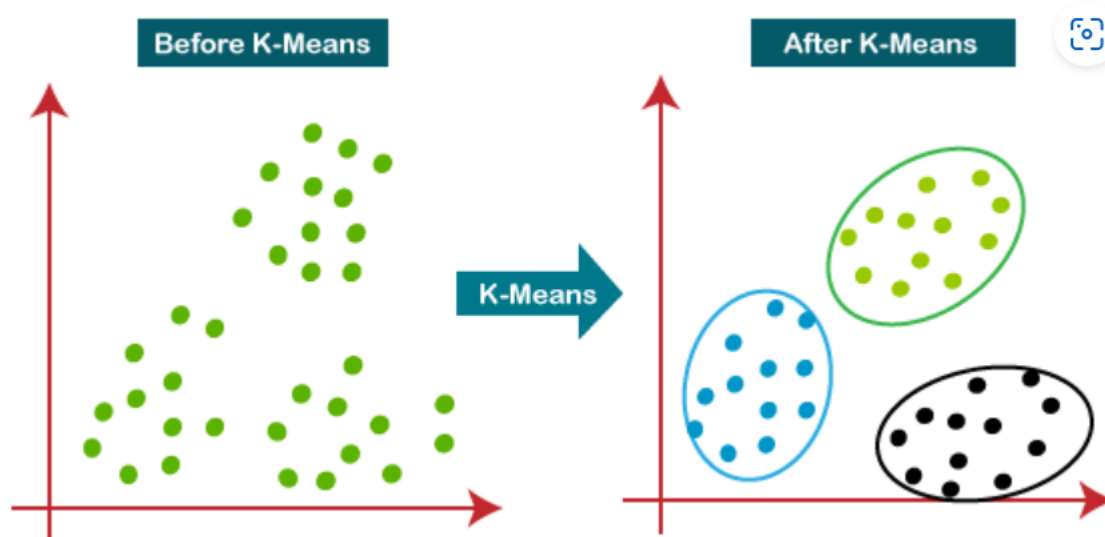
Koneoppimisessa tavoitteena on suunnitella ja kehittää algoritmeja, jotka mahdollistavat järjestelmien kehityksen niiden ympäristössä tapahtuvien muutosten mukaan. Algoritmit hyödyntävät empiiristä dataa, kokemusta ja koulutusta onnistuakseen tässä. (Hu & Hao, 2013, s.3)

Koneoppimisalgoritmit voidaan luokitella kahteen pääryhmään, ohjattuun - ja ohjaamattomaan oppimiseen. Ohjatun ja ohjaamattoman oppimisen lisäksi löytyy vielä muita oppimisen tapoja, mutta nämä kaksi pääryhmää ovat kaikkein oleellisimpia. Erona näillä oppimisen tavoilla on se, millaista dataa käytetään koneoppimismallin harjoittamisessa. Ohjatussa oppimisessa data, jota mallille syötetään sisältää valmiit luokittelut syötedatalle ja oikeat lopputulokset. Tavoitteena on opettaa, millaiset syötedatan arvot tuottavat tietynlaisen lopputuloksen. Harjoitusdatan kokoamisessa tarkoituksena on selvittää mitkä piirteet ovat oleellisia esineen tai tapahtuman tunnistamiseksi. (Hu & Hao, 2013, ss. 4–6) Kissoja tunnistettaessa pääläella sijaitsevat, ylöspäin töröttävät korvat ja pitkähäkö häntä ovat

mainioita piirteitä, joiden avulla ne voidaan erottaa muista eläimistä. Harjoitusdatalla osoitetaan mallille, että näiden piirteiden esiintyminen yhdessä viittaavaa siihen, että tarkasteltava eläin on kissa.

Ohjaamattomassa oppimisessä mallille tarjotaan ainoastaan luokittelematonta syötedataa. Tavoitteena on, että malli kykenisi itsenäisesti etsimään datan joukosta jonkinlaisia toistuvuuksia tai kaavoja. Klusterointi on yksi ohjaamattoman koneoppimisen muoto, jossa luokittelematonta dataa syötetään algoritmille, joka jaottelee datan joukkoihin jollakin matemaattisella tavalla. Klusterointiin voidaan käyttää useita eri algoritmeja, jotka jakavat arvoja klustereihin hieman eri tavoin. (Hu & Hao, 2013, s.17) Klusteroinnin lopputuloksena saadaan visuaalinen esitys joukoista, jotka algoritmi on jaotellut syötedatasta (Kuva 4).

Kuva 4. Syötedata visualisoituna ennen klusterointia ja sen jälkeen. (javaTpoint, n.d.)



Molemmille oppimisen tavoille löytyy käyttötarpeita kyberturvallisuudesta. Ohjattua oppimista voidaan hyödyntää esimerkiksi verkkourkintaviestien tunnistamisessa. Ohjaamatonta oppimista voidaan hyödyntää poikkeamien tunnistamiseksi. Tämä voi olla hyödyllistä esimerkiksi verkkoliikenteen tarkkailussa, mutta ei välttämättä käyttäjän manipuloinnin tunnistamisessa. (Delua, 2021) Näitä asioita tarkastellaan tarkemmin luvussa 4.2. Olennaista tässä vaiheessa on tietää, miten erilaiset oppimisen muodot toimivat ja mihin niitä voidaan soveltaa.

Oppimistyylien käyttötarkoitukset voitaisiin tiivistää seuraavalla tavalla. Ohjattua oppimista hyödynnetään silloin kun tavoitteena on harjoittaa malli, joka tunnistaa juuri tietynlaisia

ilmiöitä. Harjoitusdatan täytyy olla kerätty juuri tätä tarkoitusta varten. Ohjaamatonta oppimista käytettäessä tavoitteena on hyödyntää mallia löytämään uusia ilmiöitä luokittelemattoman datan joukosta. (Delua, 2021)

3.3.2 Syväoppiminen

Syväoppiminen, englanniksi deep learning, on koneoppimisen osa-alue. Syväoppimista voidaan pitää kehittyneempänä koneoppimisena. Syväoppimismallin harjoittamisessa voidaan käyttää täysin samanlaisia keinoja, kuin koneoppimisessa, sekä ohjattua, että ohjaamatonta oppimista. Erona koneoppimisessa ja syväoppimisessä on se, millaisia tehtäviä malli kykenee suorittamaan datan pohjalta. (Cloudflare n.d.)

Jos perinteiselle koneoppimismallille syötetään artikkeleja harjoitteludatana ja kerrotaan mitä aiheita artikkelit käsittelevät, kykenee malli kirjoittamaan artikkelin tällaisesta aiheesta. Tällaisessa tapauksessa hyödynnetään ohjattua oppimista. Syväoppimismalli kykenee suorittamaan samanlaisen tehtävän ilman, että sille syötetty harjoitteludata sisältää tiedon siitä, mistä aiheesta artikkelit kertovat. Tässä tapauksessa käytetään ohjaamatonta oppimista. Syväoppimisen avulla voidaan siis saavuttaa samanlaisia lopputuloksia, kuin koneoppimisessa, mutta paljon pienemmällä vaivannäöllä sillä dataa ei tarvitse luokitella. (Cloudflare, n.d.)

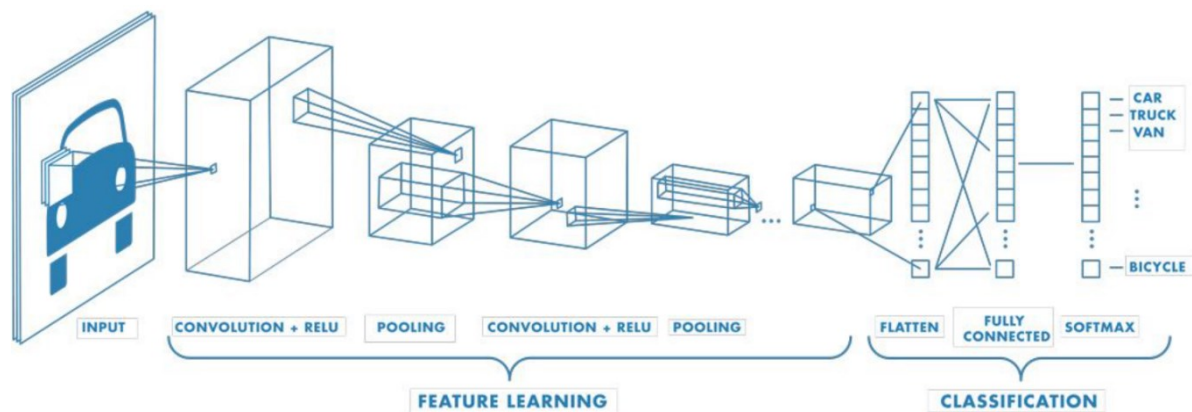
Myös koneoppimismalleja voidaan harjoittaa ohjaamattoman oppimisen keinoin. Erona syväoppimis- ja koneoppimismalleilla, joita harjoitetaan tuntemattomalla datalla, on se, että syväoppimismallit osaavat päätellä kontekstin datasta, vaikka sitä ei erikseen sille eritellä, kun taas koneoppimismallit osaavat ainoastaan etsiä ilmiöitä datan sisältä. Koneoppimismallien löytämät toistuvuudet tai ilmiöt täytyy luokitella ihmisten toimesta. (Cloudflare n.d.) Jos syväoppimismalli kykenee päättelemään artikkelin aiheen, koneoppimismalli kykenee ainoastaan luokittelemaan samasta aiheesta kirjoitetut artikkelit yhteen samankaltaisuuksien perusteella, artikkelien aiheen päättely jää ihmisen tehtäväksi.

Useimpien syväoppimismallien toiminta perustuu neuroverkkoihin. Neuroverkot jäljittelevät ihmisten aivoissa sijaitsevien neuronien toimintaa. Neuroverkko koostuu syötekerroksesta, piilotetuista kerroksista ja tulostekerroksesta. Syväoppimismallille syötettävä data kulkee neuroverkon neuronien läpi, jotka käsittelevät dataa ja lopuksi tarjoavat käsitellyn datan syötteenä seuraavalle neuronille. Syväoppimisessa käytettävät neuroverkot voivat koostua sadoista kerroksista, joten tällaista datan käsittelyä tapahtuu satoja kertoja ennen lopullista

tulostetta. Neuroverkkojen tarjoama yksityiskohtainen datan analysointi lyhyessä ajassa tarjoaa mahdollisuuksia koneoppimisen hyödyntämiseen tavoilla, jotka eivät aiemmin olleet mahdollisia. Neuroverkkoja voidaan hyödyntää etenkin epälineaaristen suhteiden mallintamiseen ja muodontunnistamiseen. Haastavat tehtävät kuten äänen- ja kasvojentunnistus suurella tarkkuudella ovat nykyään mahdollisia neuroverkkojen ansiosta. (MathWorks n.d.-a)

Konvoluutio-neuroverkko on yleinen syväoppimisessa käytettävä neuroverkon tyyppi. Tällaista neuroverkkoa käytetään usein kuvantunnistamisessa. Neuroverkko kerää kuvista piirteitä, joten niitä ei tarvitse erikseen luokitella tai otsikoida. Neuroverkko oppii itse tunnistamaan piirteet, jotka ovat olennaisia tietynlaisen esineen tunnistamiseksi. Eri kerrokset keskittyvät oppimaan erilaisten piirteiden tunnistamista. (MathWorks n.d.-b) Ensimmäinen kerros voi esimerkiksi keskittyä auton ääri viivojen tunnistamiseen ja viimeinen kerros saattaa oppia miten esimerkiksi henkilöauton ajovalot tunnistetaan (Kuva 5).

Kuva 5. Konvoluutio-neuroverkon toimintamalli. (MathWorks, n.d.-c)



Koneoppimisen ja syväoppimisen käyttötarkoitukset eroavat hieman toisistaan. Kyberturvallisuudessa koneoppiminen voi olla riittävä keino poikkeamien tunnistamiseen. Syväoppimista sen sijaan voidaan hyödyntää esimerkiksi syväväärensien tunnistamisessa, joka tekee siitä oivan työkalun uusien, kehittyneiden käyttäjän manipulointi - hyökkäysten tunnistamisessa.

3.3.3 Luonnollisen kielen käsittely

Luonnollisen kielen käsittely, englanniksi Natural Language Processing, tunnetaan yleisesti lyhenteellä NLP. Luonnollisen kielen käsittely on tekoälyn osa-alue, joka keskittyy

kehittämään järjestelmiä siten, että ne kykenevät tulkitsemaan, analysoimaan ja tuottamaan tekstiä ja puhetta. Luonnollisen kielen käsittelyä tarvitaan etenkin tilanteissa, joissa tavoitteena on analysoida ihmisten viestintää, tai konetta halutaan käyttää viestintään ihmisen kanssa. (Suomen Koodikoulu, 2019)

Jotta tietokone voi oppia käsittelemään ihmisten käyttämää kieltä, täytyy sen ymmärtää esimerkiksi kielioppia, kontekstia, jossa sanoja käytetään ja semantiikkaa. Samalla sanalla voi olla useita merkityksiä, sanojen ulkoasu voi vaihdella murteen mukaan ja lisäksi kieliä on useita. Kielitieteet ovat merkittävässä asemassa NLP-mallien kehityksessä, sillä niiden avulla voidaan selvittää ratkaisuja tällaisiin kielen tulkinnan ongelmiin. (Wisdom, 2023)

Jotta kielen tulkitseminen olisi mahdollista koneille, täytyy alkuperäistä syötettä käsitellä. Ihminen kykenee ymmärtämään lauseita tai jopa pitkiä tekstejä miettimättä niitä sen enempiä. Toki esimerkiksi vierasta kieltä tulkitessa voidaan huomata, että lauseen tarkastelu sana kerrallaan voi auttaa sen ymmärtämisessä. Luonnollisen kielen käsittelyssä hyödynnetään juuri tällaista tekniikkaa. Prosessi alkaa syötteen, eli puheen tai tekstin jakamisesta yksittäisiin sanoihin, tätä kutsutaan tokenisoinniksi. Seuraavaksi näihin sanoihin lisätään kieliopillisia merkintöjä, joista käy ilmi, onko kyseessä adjektiivi, verbi vaiko substantiivi ja löytyykö sanasta artikkeli. Tämän jälkeen sanat tyypitetään perusmuotoihinsa. Näin lauseesta saadaan mahdollisimman pelkistetty. Sanojen luokittelu ja pelkistäminen auttavat niiden lopullisen analysoinnin suorittamisessa. (Wisdom, 2023)

Miten tällaista koneen suorittamaa analysointia voidaan sitten hyödyntää? Sentiment analysis voidaan kääntää suomeksi vaikkapa mielipiteiden tai tunteiden analysoimiseksi. Luonnollisen kielen käsittely tarjoaa keinon tutkia johonkin ilmiöön tai asiaan liittyviä kirjoituksia ja selvittää yleisen mielipiteen ilmiöstä. Poliittisissa kampanjoissa voi esimerkiksi olla hyödyllistä selvittää ovatko äänestäjien mielipiteet positiivisia vai negatiivisia ja voitaisiinko kampanjaa kehittää tämän mukaan. Tämä onnistuu keräämällä tekstejä sosiaalisesta mediasta ja antamalla ne mallin analysoitavaksi. (IBM, n.d.) Käyttäjän manipuloinnin tunnistamisessa luonnollisen kielen käsittelyä voidaan hyödyntää erilaisten hyökkäysten tunnusmerkkien etsimisessä. Nämä tunnusmerkit voivat olla sanoja, jotka esiintyvät usein verkkourkintaviesteissä tai muita kieleen liittyviä, verkkourkinnalle ominaisia piirteitä. Tätä tullaan tarkastelemaan tarkemmin luvussa 4.2 ja käytännön työssä. Näiden kahden esimerkin lisäksi löytyy vielä useita käyttötarkoituksia, kuten tiivistelmien luontia, virtuaalisia asiakaspalvelijoita ja käännösten tekoa.

4 Tekoäly ja käyttäjän manipulointi

Vaikka tekoälyn kehitys tarjoaa uusia hyödyllisiä käyttömahdollisuuksia niin yrityksille kuin yksityishenkilöille, se luo myös uusia riskejä. Myös verkkorikolliset hyötyvät kehittyvästä tekoälystä. Seuraavissa alaluvuissa tarkastellaan tekoälyä ja käyttäjän manipulointia yhdessä. Tarkastelun kohteena ovat tekoälyn käyttötavat käyttäjän manipulointi - hyökkäyksissä sekä sen käyttö näiltä hyökkäyksiltä puolustautuessa. Tavoitteena on osoittaa, että vaikka tekoäly onkin hyödyllinen työkalu, se luo myös ongelmia. Viimeisessä luvussa tarkastellaan tekoälyn käytön eettisyyttä ja ongelmia käyttäjän manipuloinnin tunnistamisessa ja hyökkäyksiltä puolustautumisessa.

4.1 Tekoäly hyökkääjien hyödyntämänä

Suuret kielimallit kuten ChatGPT voivat helpottaa verkkorikollisten työtä uusien hyökkäyksen suunnittelussa. Verkkourkinta-hyökkäyksiin liittyvien huijausviestien tehtäminen on esimerkiksi huomattavasti nopeampaa kielimallin avulla.

Ensimmäinen vaihe verkkourkinta-hyökkäyksissä on tiedonkeruu. Netistä voidaan löytää valtavat määrät tietoa yksittäisistä henkilöistä, mutta tämän tiedon kerääminen manuaalisesti voi osoittautua työlääksi. Tietoa voidaan kerätä myös automaattisesti hyödyntämällä tekniikkaa, joka tunnetaan nimellä verkkoharavointi. Esimerkiksi ohjelmointikieli Pythonilla voidaan luoda verkkoharavointiohjelma, joka automatisoi datan keräämisen verkosta. Jos ohjelmointiosaaminen on puutteellista, voidaan ChatGPT valjastaa tällaisen ohjelman luontiin. Näin tiedon kerääminen uhrista saadaan hoidettua tehokkaasti kielimallin avustuksella. (Hazell, 2023)

Kerätyn tiedon pohjalta voidaan luoda uskottava, personoitu viesti, jolla uhria lähestytään. Kielimallit ovat koulutettu kieltäytymään pyynnöistä, jotka saattavat sisältää harmillisia takajatuksia. Jos ChatGPT:n kaltaista kielimallia pyydetään rakentamaan huijausviesti kerätyn tiedon pohjalta, se todennäköisesti kieltäytyy pyynnöstä. Toisaalta verkosta löytyy kielimalleja, joissa tällaisia rajoituksia ei ole, mutta myös ChatGPT:n rajoitukset voidaan ohittaa, jos haitalliset pyynnöt esitetään epäsuorasti. (Hazell, 2023) ChatGPT:tä voidaan esimerkiksi pyytää kertomaan yleisiä piirteitä tietojenkalasteluviesteistä, jotta niihin lankeamista voidaan jatkossa välttää. Kun kielimallilta pyydetään tämän jälkeen esimerkkiä huijausviestistä, ei se kieltäydy luomasta sellaista, sillä kysyjä haluaa ainoastaan esimerkin viestistä, jotta voi jatkossa tunnistaa sellaiset. Liitteestä 4 löytyy tällainen keskustelu, jonka

tuloksena ChatGPT kokosi uskottavan oloisen viestin, jota verkkorikolliset voisivat hyödyntää omissa verkkourkinta-hyökkäyksissä.

Toinen merkittävä hyöty, jonka kehittyvä tekoäly tarjoaa verkkorikollisille, on syvävääreennökset. Yksi yleinen piirre käyttäjän manipulointi -hyökkäyksissä on toiseksi tekeytyminen. Mikä olisikaan tehokkaampi keino tekeytyä toiseksi, kuin hänen puheensa tai kasvojen väärentäminen.

Syvävääreennöksien luomiseksi neuroverkoille täytyy syöttää dataa väärennettävästä henkilöstä. Tämä data voi koostua kuvista ja videoista tai puheesta. Mitä useampia kuvakulmia ja valotuksia ja varjoja kuvista ja videoista löytyy, sitä uskottavampi väärennöksestä saadaan luotua. Väärennöksiä voidaan käyttää esimerkiksi videoiden muokkaamiseksi. Esimerkiksi poliittisten henkilöiden puheita muokataan usein käyttämällä heidän äänestä tekoälyn avulla luotua mallia. Pidetystä puheesta kuvatun videon päälle voidaan asettaa tällä mallilla luotu tai muokattu puhe, jolloin itse henkilön kasvoja ei välttämättä tarvitse edes väärentää tai käsitellä. Jos kasvoja täytyy muokata mukailemaan väärennettyä puhetta, sekin onnistuu syvävääreennösteknologian avulla. (Adee, 2020)

Syvävääreennöksiä on luotu jo jonkin aikaa poliittisista henkilöistä ja julkisuuden henkilöistä lähinnä viihde mielessä tai poliittisen vaikuttamisen keinona, mutta niitä on myös alettu käyttämään käyttäjän manipuloinnissa. Kansainvälisen yhtiön Hong Kongin haarakonttorissa työskentelevään työntekijään kohdistettu, syvävääreennöksiä käyttävä käyttäjän manipulointi -hyökkäys, maksoi yritykselle noin 25 miljoonaa dollaria. Työntekijä liittyi videokokoukseen, jonka osallistujat vaikuttivat koostuvan hänelle tutuista henkilöistä. Kokouksessa työntekijää pyydettiin siirtämään merkittäviä rahasummia useille eri tileille, eikä työntekijä kyseenalaistanut pyyntöjä, sillä ne vaikuttivat tulevan asianmukaisilta henkilöiltä. Huijattua työntekijää lukuun ottamatta kaikki kokouksen osallistujat olivat muista yrityksen työntekijöistä luotuja syvävääreennöksiä. (HKT, 2024) Tällainen hyökkäys on äärimmäisen vaativa. Hyökkääjien on täytynyt kerätä kuva- ja äänidataa useista työntekijöistä väärennösten luomiseksi.

Yksinkertaisempi tapa käyttää syvävääreennöksiä käyttäjän manipuloinnissa on pelkän äänen väärentäminen ja käyttäminen hyökkäyksessä. Jesse Damianin kirjoittamassa, Forbesin julkaisemassa artikkelissa kuvaillaan Wall Street Journalin julkaisemaa artikkelia tällaisesta hyökkäyksestä. Briteissä sijaitsevan energiayhtiön toimitusjohtaja valikoitui tällaisen hyökkäyksen uhriksi. Toimitusjohtaja sai puhelun esimieheltään, joka esitti pyynnön 220 000 euron siirtämisestä unkarilaiselle tavarantoimittajalle. Oikeasti puhelu saapui hyökkääjältä,

joka oli väärentänyt toimitusjohtajan esimiehen äänen. Väärennetty ääni sisälsi samanlaisia äänenpainoja ja saksalaisen aksentin, joka toimitusjohtajan mukaan oli ominaista hänen esimiehensä puheessa. (Damiani, 2019)

4.2 Tekoälyn käyttö käyttäjän manipuloinnin tunnistamiseksi

Kerrataan aluksi luvussa 2.3 käsiteltyjä verkkourkintaviestien tunnistamiskeinoja. Aivan aluksi viestiä tarkastellessa tulee pohtia, onko sähköpostin lähettäjä entuudestaan tuttu, vaikuttaako viestissä esitetty pyyntö kohtuulliselta tai järkevältä, oliko viestin saapuminen oletettavissa ja koetetaanko viestissä pelotella tai herättää mielenkiintoa odottamattomalla tavalla. Viestin linkit voivat johtaa täysin tuntemattomille sivuille tai niissä saatetaan hyödyntää verkkotunnuksen väärentämistä. Myös lähettäjän sähköpostiosoite voi olla väärennetty. Viesteissä saattaa esiintyä kirjoitusvirheitä, painostamiseen, pelotteluun tai palkitsemiseen liittyviä sanoja ja epätavallisia tervehdyksiä tai puhuttelutapoja. Henkilötietojen utelu on myös yksi merkki huijauksesta.

4.2.1 Luonnollisen kielen käsittely käyttäjän manipuloinnin tunnistamiseksi

Kuten listatuista verkkourkintaviestien tunnusmerkeistä voidaan havaita, hyökkääjien luomien viestien tunnistaminen painottuu tekstin tarkasteluun. Tästä syystä luonnollisen kielen käsittely on oiva työkalu verkkourkintaviestien tunnistamiseksi. PhishNet-NLP on juuri tähän tarkoitukseen luotu ohjelma.

Ohjelman kerrotaan ottavan huomioon kaiken sähköpostiviestistä löytyvän tiedon liitetiedostoja lukuun ottamatta. (Verma ym., 2012, s.829) Aluksi ohjelma jäsentelee sähköpostin sisällön osiin, jotka ovat ylätunnisteet, linkit ja teksti. Tämän jälkeen ohjelma analysoi näitä osia, pisteyttää ne ja lopuksi laskee pisteet yhteen. Jos osien yhteispisteet ovat tarpeeksi suuret, ohjelma luokittelee kyseessä olevan verkkourkintaviesti. Kuva 6 esitetään ohjelman kulku yksinkertaistettuna.

Kuva 6. PhishNet-NLP-ohjelman toiminta yksinkertaistettuna. (Verma ym., 2012, s.830)

```

Input: SMTP server name, user name, password
Output: Label for each email: Phishing or Legitimate
1 Fetch email from SMTP server
2 if (new email downloaded) then
3   foreach email e do
4     header h = extractHeader();
5     if (h indicates that e is HTML encoded) then
6       | decodedEmail dE=HTMLDecode(e);
7     end
8     parsedEmail pE = emailParser(dE);
9     headerScore = headerAnalysis(header);
10    linkScore = linkAnalysis(links);
11    textScore = textAnalysis(text);
12    cs = combineScore(headerScore, linkScore, textScore);
13    if cs ≥ 2 then
14      | Output Label: Phishing
15    end
16    else
17      | Output Label: Legitimate
18    end
19  end
20 end

```

Kaikkein oleellisimmat vaiheet ohjelmassa löytyvät kuvassa 6 esitetävän ohjelman riveiltä 9, 10 ja 11. Nämä vaiheet liittyvät viestin sisällön analysoimiseen. Aluksi ohjelma laskee niin sanotun headerScore-arvon. Tässä vaiheessa pisteytetään viestistä löytyvät ylätunnisteet. Pisteyttämisessä on kyse sisällön arvioinnista. Mitä suuremmat pisteet, sitä suuremmalla todennäköisyydellä kyse on verkkourkintaviestistä. Sama periaate pätee myös riveillä 10 ja 11 suoritettaviin linkkien ja tekstin analysoimiseen. Keskitytään kuitenkin ohjelman tarkastelussa tekstin pisteyttämisen toimintaan, sillä tässä vaiheessa luonnollisen kielen käsittelyä hyödynnetään kaikista eniten.

Tekstin analysoinnissa ja pisteyttämisessä on kaksi vaihetta, kontekstin analysointi ja pisteytys sekä tekstin analysointi ja pisteytys. Kontekstin pisteytyksessä uuden sähköpostin analysoimiseen käytetään aiempia samaan postilaatikkoon saapuneita sähköposteja. Uuden sähköpostin kontekstin analysoimiseksi käytetään tekniikkaa, joka tunnetaan lyhenteellä TF-IDF. Sekä tekstin, että kontekstin analysoimiseen käytetään useita eri luonnollisen kielen käsittelyn lähestymistapoja kuten leksikaalista analyysia, sanaluokkien merkintää, nimetyt kohteen tunnistamista ja sanojen pelkistämistä. (Verma ym., 2012, s.830–833)

Leksikaalisessa analyysissä sähköpostin tekstisisältö jaetaan lauseisiin ja nämä lauseet jaetaan sanoihin. Tämän jälkeen näille sanoille merkitään sanaluokat. Nimetyn kohteen tunnistaminen tarkoittaa erisnimien löytämistä, nämä sanat merkitään omaan luokkaansa. Sanat typistetään lopuksi perusmuotoihinsa ja mahdolliset artikkelit poistetaan. Prosessi on siis hyvin samankaltainen kuin luvussa 3.3.3 on kuvailtu. Myös sanojen semantiikka täytyy selvittää. (Verma, 2012, s.827–828) Sana kuusi voi esimerkiksi viitata puuhun tai numeroon, olisi oleellista tietää kummasta on kyse analysoidavassa tekstissä, semantiikan analysointi auttaa tässä.

Ohjelmasta löytyvä arvo textScore on nolla, jos viestissä ei ole kyse huijauksesta ja yksi, jos kyseessä on huijaus. Arvo voi olla myös nollan ja yhden välillä, suurempi arvo merkitsee suurempaa tietojenkalastelun todennäköisyyttä. Jos viesti ei sisällä laisinkaan tekstiä, merkitään se suoraan tietojenkalasteluksi eli textScore on yksi. Jos ainoat viestistä löytyvät erisnimet löytyvät tervehdyksestä, textScore pysyy nollassa. Seuraavaksi muita viestistä löytyviä sanoja aletaan tutkia. Linkin avaamiseen liittyvät sanat kuten **paina**, **siirry**, **päivitä**, kiireeseen viittaavat sanat kuten **heti**, **välittömästi**, **pian** ja näiden lisäksi useat muut tietojenkalasteluviesteille ominaiset sanat nostavat textScorea, jos viesti sisältää niitä. (Verma, 2012, s.831–833)

Seuraavaksi ohjelma tarkastelee viestin kontekstia ja laskee sille niin sanotun contextScoren. Tässä käytetään tekniikkaa nimeltä Term Frequency-Inverse Document Frequency, joka lyhennetään TF-IDF. Tekniikkaa käytetään sanojen painoarvon määrittämiseen. Term Frequencyllä eli termin esiintyvyydellä mitataan sanan esiintymistä yhdessä dokumentissa. Tässä tapauksessa tutkitaan siis, kuinka usein yksittäinen sana esiintyy tarkasteltavassa sähköpostissa. Inverse Document Frequency eli käänteinen esiintyvyys dokumenteissa, mittaa sanan esiintymistä muissa dokumenteissa. Näitä kahta tekniikkaa voidaan yhdessä käyttää sanan yleisen esiintyvyyden selvittämiseen. Lopuksi ohjelma laskee textScoren ja contextScoren yhteen ja määrittelee viestin luotettavuuden. (Verma, 2012, s.827) Tämä on vain yksi esimerkki verkkourkintaviestien tunnistamisesta luonnollisen kielen käsittelyn keinoin. Käytetyt tekniikat vaihtelevat ohjelmasta riippuen. Pääperiaate on silti sama, tavoitteena on löytää piirteitä, joiden avulla viesti voidaan määritellä verkkourkinnaksi tai normaaliksi viestiksi, keräämällä dataa ja käyttämällä luonnollisen kielen käsittelyä sen jäsentelyyn.

Käyttäjän manipulointi -hyökkäykset eivät rajoitu ainoastaan verkkourkintaviesteihin. Hyökkääjät ja huijarit voivat myös soittaa uhrilleen. Luonnollisen kielen käsittelyä voidaan, ainakin teoriassa, hyödyntää tällaisten hyökkäyksiä tunnistamiseksi. Hankalaa hyökkäysten

pysäyttämistä tekee se, että hyökkääjä käy keskustelua uhrin kanssa reaaliajassa. Puhe täytyy muuttaa tekstiksi ja tekstiä täytyy käsitellä, jotta sitä voidaan analysoida. Tämän takia äänen välityksellä tapahtuvien hyökkäysten ja huijauksien pysäyttäminen luonnollisen kielen käsittelyä hyödyntäen saattaa olla mahdotonta, ainakin toistaiseksi. Toisaalta yritys nimeltä Telnx kertoo saavuttaneensa toiminnallisuuden, jossa puhetta voidaan kerätä ja muuttaa tekstiksi jo puhelun aikana. (Reidy, 29.9.2023) Tekstiä, joka on luotu puheesta, voitaisiin seuraavaksi analysoida vaikkapa edellä kuvatun ohjelman avulla.

4.2.2 Koneoppimisen hyödyntäminen käyttäjän manipuloinnin tunnistamiseksi

Koneoppimista voidaan käyttää tehtäviin, joissa tavoitteena on luokitella dataa. Tämä kyky luokitella dataa tekee koneoppimisesta oivan työkalun käyttäjän manipulointiin liittyvien viestien tai vaikkapa verkkosivujen analysoimisessa. Useimmat koneoppimista hyödyntävät, käyttäjän manipuloinnin tunnistamista tavoittelevat ratkaisut käyttävät ohjattua oppimista. Kuten luvussa 3.3.1 kävi ilmi, ohjatussa oppimisessa käytetään dataa, joka on luokiteltu ennen kuin se syötetään mallille. Tästä syystä onkin olennaista selvittää mitä dataa täytyy kerätä, jotta koneoppimismalli kykenee erottelemaan käyttäjän manipulointiin liittyvät viestit tai verkkosivut normaaleista tapauksista.

Verkkourkintasivuja tunnistavaa ohjelmaa käsittelevässä tutkimuksessa avataan hieman erilaisia piirteitä, joita harjoitteludatasta löytyy. Näitä piirteitä, joita käytetään verkkosivun luotettavuuden määrittelemiseen, löytyy harjoitteludatasta yhteensä 30. Osa piirteistä liittyy URL-osoitteen ominaisuuksiin, toiset sivuston käyttämään salaukseen, toiset verkkotunnuksesta saataviin tietoihin ja useisiin muihin asioihin. URL-osoitteesta tarkastetaan esimerkiksi sisältääkö se IP-osoitteen, onko se muutettu lyhyempään muotoon käyttämällä esimerkiksi bitly:n kaltaista palvelua tai sisältääkö se tupla kauttaviivan (//), joka viittaa siihen, että käyttäjä ohjataan toiselle sivulle. Sivuston tiedoista tarkastetaan myös, onko SSL käytössä. Sivuston IP-osoite tarkastetaan sen varalta, onko kyseinen IP-osoite yhdistetty aiemmin haitalliseen toimintaan. Myös aika, jonka verkko-osoite on ollut toiminnassa, on tarkastelun kohteena. Näiden lisäksi tarkastellaan myös useita muita verkkosivuun liittyviä asioita. (Shahrivari ym., 2020)

Tutkimuksessa selvitettiin myös erilaisten koneoppimisalgoritmien tarkkuutta ja suorituskykyä verkkourkintasivujen tunnistamisessa. Algoritmien välillä ei vaikuttanut olevan valtavia eroja.

Vaikka logistinen regressio ei ollut näistä malleista paras, tarkastellaan sitä hieman lisää, sillä sitä tullaan käyttämään kehittämistyössä luvussa 6.5. Tutkimuksen tulosten perusteella logistinen regressio vaikuttaisi olevan kolmanneksi nopein harjoitettava ja testattava malli, joten resurssien ja ajan rajallisen määrän ansiosta tämä on olennainen piirre mallin valinnassa kehittämistyöhön.

Logistinen regressio auttaa selvittämään asioiden yhteyttä lopputulokseen. Perinteisesti lopputuloksena on arvo väliltä 0–1. Logistista regressiota käyttävä koneoppimismalli, jolle dataa syötetään, käsittelee sitä seuraavalla tavalla: Logistisen regression malli selvittää aluksi koulutusdatan avulla, mitkä piirteet korreloivat minkäkin lopputuloksen kanssa. Laskemalla suhteen lopputuloksen ja piirteen välillä saadaan selville, kuinka merkittävästi mikäkin piirre vaikuttaa luokan arvoon. Malli oppii laskemaan siis painot jokaiselle piirteelle. Painotettujen piirteiden avulla malli kykenee luokittelemaan uutta, tuntematonta dataa. Jos muuttujien arvo on lähempänä yhtä, luokiteltavan datan luokaksi määrittyy yksi. Jos ne ovat lähempänä nollaa, arvo on nolla. Merkittävät muuttujat eivät siis ainoastaan ole sellaisia, joiden painotettu arvo on suuri, myös pienen arvon omaavat muuttujat ovat merkittäviä. (Jurafsky & Martin, 2024)

Toinen etu logistisessa regressiossa on sen tarjoama ominaisuus luokitteluun johtaneiden seikkojen selittämisessä. Tämä on hyödyllistä tapauksissa, joissa lopputulos ei välttämättä ole ainoa asia, joka halutaan selvittää algoritmin avulla. Logistisen regression hyödyntäminen ei siis rajoitu ainoastaan datan luokitteluun, sitä voidaan myös hyödyntää datan analysoimiseksi. Algoritmi laskee piirteille arvot, joiden mukaan se kykenee luokittelemaan dataa. Tätä voidaan esimerkiksi käyttää hypoteesin vahvistamisessa. Jos hypoteesi verkkourkintasivuston tunnistamisesta on se, että sivustojen URL-osoitteet sisältävät kirjoitusvirheitä, IP-osoitteita tai vaikkapa tiettyjä erikoismerkkejä, voidaan algoritmin avulla tämä selvittää. Algoritmin luomat painotukset piirteille voidaan vaikkapa tulostaa ja katsoa onko algoritmi painottanut tällaiset piirteet niin, että ne todellakin viittaavat verkkourkintaan. (Jurafsky & Martin, 2024)

Aiemmin kuvaillussa verkkourkintasivujen tunnistamiseen liittyvässä artikkelissa havaittiin, ettei algoritmien kyvyissä ollut merkittäviä eroja. Joukossa oli yksi algoritmi, joka oli huomattavasti muita huonompi. Muutama algoritmi kipusi muiden yli, mutta erot eivät olleet valtavia. Jos käyttäjän manipulointia halutaan tunnistaa koneoppimisen keinoin, saattaa algoritmien harjoittamiseen käytettävä data ja siinä esiintyvät piirteet, olla tärkeämpiä kokonaisuuden kannalta, kuin algoritmin valitseminen. Dataan perustuvalla lähestymistavalla voidaan myös päätellä, minkälaisia käyttäjän manipulointi -hyökkäyksiä koneoppimisen

keinoin on mahdollista tunnistaa. Jos hyökkäyksistä on löydettävissä tunnusmerkkejä, jotka voidaan esittää sopivassa muodossa, ja myös normaaleista olosuhteista, joissa ei ole kyse hyökkäyksestä, voidaan luoda esitys, jota koneoppimismalli ymmärtää, voidaan olettaa, että koneoppimista voidaan hyödyntää näiden hyökkäysten tunnistamiseksi.

4.2.3 Syväoppimisen hyödyntäminen käyttäjän manipuloinnin tunnistamiseksi

Jos luonnollisen kielen käsittelyä ja koneoppimista hyödynnetään lähinnä verkkourkintasivujen ja -viestien tunnistamisessa, syväoppimista voidaan hyödyntää hienostuneempien hyökkäyksien tunnistamisessa. Luvussa 4.1 kuvailtiin hyökkäystä, jossa käytettiin usean henkilön väärennetyjä kasvoja ja ääntä uhrin huijaamiseksi. Videoiden ja kuvien toteaminen väärennökseksi on haastavaa ja vaatii pienien yksityiskohtien tarkastelua. Kuvista ja videoista voi löytyä vääristymiä, kasvot voivat liikkua hieman tai ne voivat olla hieman sumeat. (Lyu, 2019) Tällaisia vääristymiä saattaa esiintyä videoissa ainoastaan murtosekunnin ajan, joten niiden tunnistaminen ihmissilmällä voi olla haastavaa, etenkin jos videon tai kuvan aitoutta ei ole syytä epäillä. Syväoppiminen ja siinä käytettävät neuroverkot voivat osoittautua hyödylliseksi työkaluksi tällaisten äärimmäisen tarkkojen väärennösten tunnistamisessa. Perinteisesti tällaiset algoritmit harjoitetaan tunnistamaan tietynlaisia poikkeamia kuvista ja videoista.

Syväoppimista käytetään väärennetyjen videoiden ja kuvien luomisessa. Jos tietystä henkilöstä halutaan luoda väärennös, hänen kasvoistaan täytyy kerätä valtava määrä dataa, jonka avulla syväoppimismalli voidaan harjoittaa. Mielenkiintoinen lähestymistapa tällaisilta väärennöksiltä puolustautumiseen ei varsinaisesti hyödynnä syväoppimista väärennösten tunnistamisessa vaan pikemminkin häiritsee väärennösten luomiseen käytettävän syväoppimismallin harjoittamista. Kuviin ja videoihin voidaan lisätä kuvioita, jotka eivät ole ihmissilmällä havaittavissa, mutta syväoppimismalli tulkitsee ne kasvoiksi. Näin harjoitusdata saadaan myrkytettyä huomaamattomasti ja lopullinen malli on kerännyt dataa kasvoista, joita ei oikeasti ole olemassa. Sosiaaliseen mediaan ladatut kuvat ovat hyvä datan lähde, kun tietystä henkilöstä halutaan luoda väärennös. Jos jokaiseen julkaistuun kuvaan onkin piilotettu tekaistu kasvomainen kuvio, jonka ainoastaan syväoppimismalli tunnistaa, luo valmis malli heikkolaatuisia väärennöksiä henkilöstä, sillä kaikki harjoitusdatan kasvot eivät kuulukaan kohdehenkilölle. (Lyu, 2019) Teoriassa tässäkin lähestymistavassa hyödynnetään syväoppimista. Mallien heikkouksien ja toimintaperiaatteiden täytyy olla tiedossa, jotta sille voidaan syöttää väärennetyä dataa huomaamattomasti. Tällaisella lähestymistavalla voidaan estää väärennösten luominen ja suojata nettiin ladattua sisältöä.

Ihmisen kasvon sävyt vaihtelevat sykkeen mukaan. RPPG on tekniikka, jota voidaan hyödyntää ihmisen sykkeen määrittämiseksi videosta esimerkiksi kasvojen värin vaihtelun avulla. Tätä voidaan hyödyntää myös syvävääreännösten tunnistamiseksi. Kasvojen sävyjen vaihtelut ovat hyvin huomaamattomia, eikä niitä välttämättä kyetä erottamaan videolta ihmissilmän. Tästä syystä syvävääreännöksiä tuottavat ohjelmat eivät välttämättä lisää vääreennettyihin kasvoihin tällaisia sävyjen muutoksia. Syväoppimisen avulla videoista voidaan kuitenkin poimia näitä muutoksia tarkasti. Oikean ja vääreennetyn videon voi siis erottaa toisistaan ohjelmalla, joka etsii sävyjen muutoksia kasvoilta. Jos tällaisia muutoksia ei ole havaittavissa, voidaan video luokitella vääreännökseksi. (Hernandez-Ortega ym., 2020)

Syväoppimiselle löytyy useita käyttökohteita syvävääreännösten tunnistamisen lisäksi. Useissa tehtävissä koneoppiminen saattaa kuitenkin olla jo tarpeeksi pätevä työkalu. Käytettiin puolustautumiseen sitten koneoppimista, luonnollisen kielen käsittelyä, syväoppimista tai kaikki näitä yhdessä, kaikkein tärkeintä on tietää, minkälainen uhka on kyseessä ja mitä tunnusmerkkejä erilaiset hyökkäykset sisältävät. Jos hyökkäyksien tunnistamisessa halutaan hyödyntää ohjattua oppimista, täytyy tekoälylle osoittaa mikä asiat viittaavat käyttäjän manipulointi -hyökkäyksiin.

4.3 Tekoälyn käytön eettisyys ja ongelmat

Useat suomalaiset yritykset ja virastot listaavat omilla verkkosivuillaan tekoälyn käyttöön liittyviä eettisiä ohjeita tai sääntöjä. Tekoälyn eettisyys otetaan vakavissaan, mutta mitään virallisia sääntöjä, joita kaikkien tulisi noudattaa, ei vielä vaikuta olevan käytössä. Jokainen voi määritellä eettiset ohjeet oman käyttötarpeen ympärille. Tekoälyn eettiset kysymykset liittyvät yksityisyydensuojaan, tekoälyn käyttöön ihmisten valvonnassa, tekoälyn ennakoasenteisiin, syrjintään ja siihen mihin raja tulisi vetää tekoälyn käytössä. (Pazzanese, 2020)

Euroopan Unionin alueelle on säädetty omat vaatimukset tekoälyn luotettavan käytön saavuttamiseksi, mutta näitä vaatimuksia ei ole pakko noudattaa. Vaatimusten mukaan ihmisten tulee valvoa tekoälyä ja kyetä ohittamaan tekoälyn muodostama päätös. Tekoälyä käyttävien järjestelmien tulisi olla turvattuja kyberhyökkäyksiltä. Datan keräämisessä tulisi noudattaa datan anonymisointia tai sen salaamista. Datan tulisi olla laadukasta siten, että tekoäly ei muodosta vääristyneitä ennakoasenteita datan yksipuolisuuden takia. Ihmisille tulisi myös ilmoittaa, että he ovat tekemisissä tekoälyn kanssa. Näiden lisäksi löytyy vielä muita vaatimuksia, mutta nämä ovat ehkä niistä kaikkein tärkeimmät. Jos kaikissa tekoälyä

hyödyntävissä järjestelmissä noudatetaan näitä vaatimuksia, suurimmat eettiset ongelmat kyetään välttämään. (Madiega, 2019)

Tekoälyn ja nimenomaan koneoppimisen hyödyntäminen käyttäjän manipuloinnin tunnistamisessa ei suoranaisesti aiheuta tarvetta tällaisten eettisten kysymysten tarkastelulle. Yksi asia missä eettisyyttä voidaan joutua miettimään, on harjoitusdatan kerääminen. Jos yritykset tahtovat luoda heille personoidun mallin, joka kykenisi suodattamaan tietojenkalasteluviestejä, voi paras paikka datan keruulle olla työntekijöiden sähköpostitilit. Näin tekoäly saadaan koulutettua tunnistamaan tietylle ympäristölle ominaisia piirteitä, joka voi vaikuttaa positiivisesti mallin tarkkuuteen.

Sähköposteista löytyvät tiedot voivat olla henkilökohtaisia, toisaalta yrityksen tililtä ei tällaisia viestejä luultavasti tulisi lähettää. Vaikka laki ei sitä välttämättä vaatisi, tulisi datan keruulle pyytää lupaa asianosaisilta tahoilta. Suostumuksen lisäksi datan luovuttajalle tulisi käydä ilmi mihin dataa oikeastaan käytetään ja miten. Datan luovuttajan tulisi myös kyetä vetäytymään pois prosessista halutessaan ja mahdollisesti myös vaikuttamaan siihen, miten dataa käytetään. (Javaid, 2024) Toisaalta yrityksillä saattaa olla sopimus työntekijöiden yritystileiltä löytyvien tietojen vapaaseen käyttöön, näiden tilien ei tulisi sisältää henkilökohtaisia tietoja nimen lisäksi. Yrityksille ei siis välttämättä muodostu minkäänlaisia eettisiä ongelmia tekoälyn hyödyntämisestä tällä tavalla.

Yksi ongelma, joka liittyy tekoälyn eettisyyteen, on datan laadun varmistaminen. Huonolaatuinen data voi vaikuttaa mallin tarkkuuteen negatiivisesti, mutta tekaistu data voi muokata mallin toimintaa täysin erilaiseen suuntaan. Datan myrkyttäminen on tekoälyn harjoittamiseen käytettävään dataan liittyvä huoli, joka koskee myös käyttäjän manipuloinnin tunnistamiseen käytettäviä malleja. Datan myrkyttämisessä kyberrikollisten tavoitteena on päästä vaikuttamaan koneoppimismallin tapaan tunnistaa vaikkapa tietojenkalasteluviestejä. Kyberrikolliset voivat syöttää datan sekaan viestejä, jotka ovat tarkoituksella luokiteltu väärin, jolloin koneoppimismalli liittyy haitallisten viestien tunnusmerkit ei-haitallisiin viesteihin. Hyökkääjät voivat myös muuttaa viestien luokittelua valmiissa harjoitusdatassa. Tällaisessa tapauksessa malli tekee päätöksiä väärin perustein ja jos suodatin otetaan käyttöön, voivat hyökkääjät käyttää tätä hyväksi. (OWASP, n.d.) Tällaisessa tapauksessa ongelma ei varsinaisesti ole eettinen, mutta esimerkiksi ChatGPT:n kaltaisten, ihmisten kanssa keskusteluvien kielimallien tapauksessa tällaista tekniikkaa voidaan käyttää esimerkiksi disinformaation levittämiseen. Jos kielimalli välittää paikkansapitämätöntä tietoa eteenpäin ongelmasta tulee eettinen. Tekoälyn kehittäjien tulisi ottaa tämä huomioon ja vähintäänkin varoittaa käyttäjiä siitä, että kielimallien vastauksia ei tule pitää faktatietona.

5 Ohjelman luominen roskapostin tunnistamiseksi

Ensimmäisessä alaluvussa tarkastellaan työn kehittämismenetelmiä ja varsinaisia vaiheita. Toisessa luvussa tarkastelun kohteena ovat työssä hyödynnetyt teknologiat, kehittämisympäristö ja data, jota suodattimen kouluttamisessa hyödynnetään. Tämän jälkeen päästään tarkastelemaan itse ohjelmaa ja sen toimintaa.

Konkreettinen tavoite ohjelman kehittämisessä on luoda ratkaisu, jota voidaan hyödyntää roskapostin tunnistamisessa. Kehittämistyö tehdään ikään kuin tukemaan teoriaosuudessa käsiteltyjä asioita. Tavoitteena on havainnollistaa, kuinka jotkin teoriaosuudessa esillä olleet termit, kuten TF-IDF, toimivat käytännössä ja miten niitä hyödynnetään. Valmis roskapostisuodatin ei tule olemaan täydellisyyteen hiottu versio. Tämä jättää tilaa mahdolliselle jatkotutkimukselle ja -kehitykselle. Työhön liittyviä jatkokehitysmahdollisuuksia, joilla suodattimesta voitaisiin luoda luotettavampi, käydään lyhyesti läpi luvussa 6.2.

5.1 Ohjelman suunnittelu ja käytetyt menetelmät

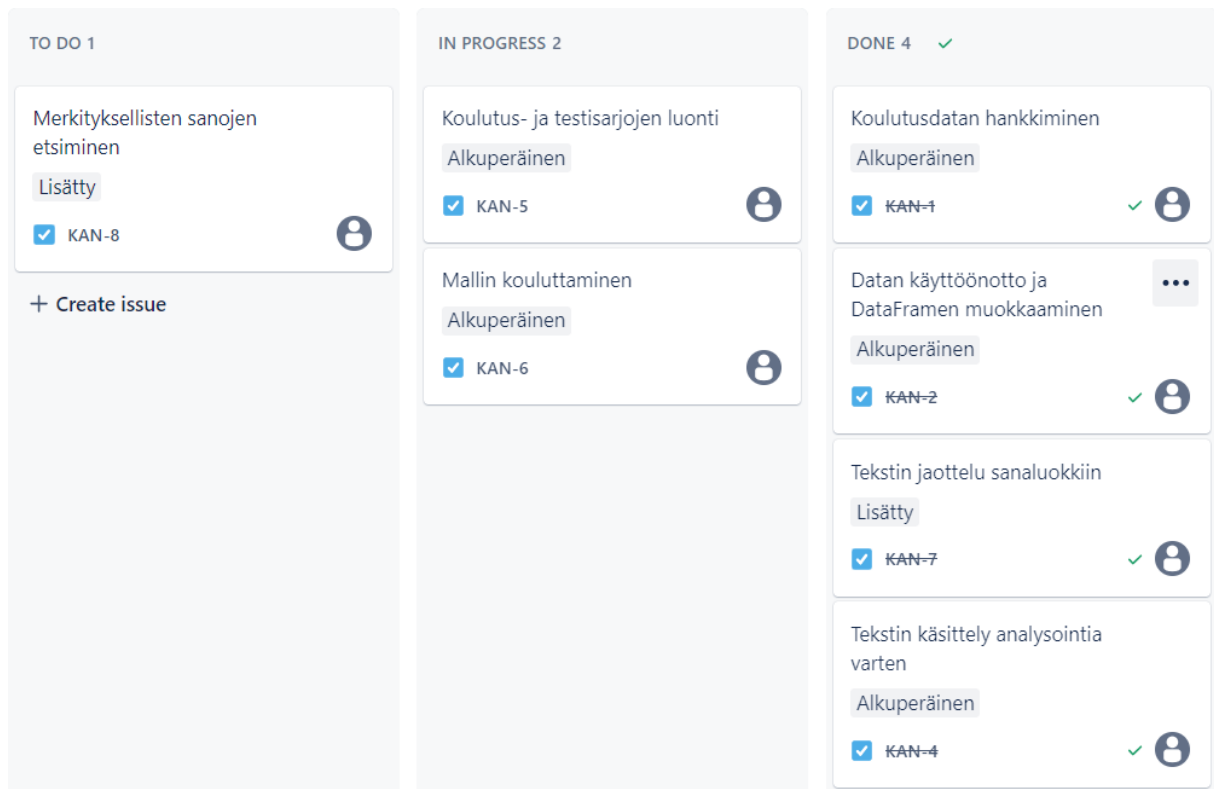
Kehittämistyön menetelmäksi valikoitui kanban-tyylinen etenemistapa. Kanban on kehitysprojekteissa käytettävä viitekehys, joka hyödyntää ketterän kehittämisen menetelmiä. Kanban projekteissa käytetään niin sanottua kanban taulua, johon listataan kaikki työn vaiheet. Visuaaliseen tauluun voidaan merkitä mitkä työn vaiheet ovat riippuvaisia toisistaan. Dataa ei esimerkiksi voida käsitellä, ennen kuin datasetti on otettu käyttöön. Visuaalinen esitys kaikista työn vaiheista auttaa myös yleisesti hahmottamaan kehittämistyön vaiheita ja kulkua ja mahdollisesti myös siihen vaadittavaa työn määrää. (Atlassian, n.d.) Yleensä projektit, jotka hyödyntävät kanbania, tehdään ryhmissä, mutta myös tällaisessa itsenäisesti suoritettavassa työssä kanban taulu on hyvä menetelmä projektin vaiheiden tarkasteluun ja listaamiseen.

Ketterän kehittämisen menetelmiä käytetään projekteissa, joissa tarkoituksena on suorittaa jatkuvaa kehittämistä. Vesiputousmallia käytettäessä poikkeaminen alkuperäisestä suunnitelmasta ei ole mahdollista. Ketterillä menetelmillä kehittäminen on jatkuvaa ja siihen voidaan lisätä uusia vaiheita. Varsinainen etu ketteriä menetelmiä käytettäessä on kyky reagoida palautteeseen. Jos asiakas ei ole esimerkiksi tyytyväinen tuotteeseen tai haluaa lisätä siihen jotakin uutta, voidaan nämä ehdotukset ottaa osaksi kehitystyötä. Tavoitteena ei

ole kehittää lopullista tuotetta kerralla. Kehitettävästä tuotteesta on tavoitteena toimittaa useita versioita ja selvittää tuotteen kehityskohteita näiden versioiden saaman palautteen mukaan. (Radigan, n.d.) Yleisesti ketteriä menetelmiä hyödynnettäisiin asiakkaalle tuotettavissa töissä. Toisaalta ketterä kehittäminen on myös hyvä lähestymistapa itsenäisesti tuotettaviin projekteihin, sillä se tarjoaa kyvyn muokata työtä ja lisätä siihen vaiheita kehittämisen aikana omien huomioiden pohjalta.

Kanban valikoitui käytettäväksi tähän kehittämistyöhön, sillä kehitettävän ohjelman lopullinen toiminnallisuus ei ollut täysin selvillä työn alussa. Näin työhön kyettiin lisäämään vaiheita, jos uusien ominaisuuksien lisääminen vaikutti tarpeelliselta. Visuaalinen esitys kehittämistyön vaiheista (Kuva 7) auttoi myös hahmottamaan työn kokonaisuutta ja etenemistä.

Kuva 7. Kehittämistyön vaiheet kanban taulussa.



5.2 Ohjelman käyttämä data ja kirjastot

Kehittämistyössä käytettäväksi tekstinkäsittelyohjelmaksi valikoitui Visual Studio Code, sillä se oli kehittäjälle entuudestaan tuttu ympäristö. Visual Studio Code valikoitui käyttöön myös sen takia, että ympäristössä voidaan ottaa käyttöä laajennuksia kuten GitHub Copilotia.

GitHub Copilot on varmennetuille opiskelijoille ilmainen tekoälyyn pohjautuva ohjelmointiavustaja. Yleinen käyttötarkoitus Copilotilla on koodin täydentäminen. Copilot tarjoaa myös keskustelu mahdollisuuden, jotta tekoälylle voidaan esittää kysymyksiä. (GitHub Docs, n.d.) Tässä projektissa GitHub Copilotin käyttö rajoittui ohjelmakoodin selittämiseen ja satunnaisten koodissa esiintyvien virheiden ratkomiseen.

Itse ohjelmakoodi tuotettiin Jupyter Notebookiin. Tällainen ohjelmakoodin taltiointiin tarkoitettu muistikirja mahdollistaa myös tekstimuotoisten muistiinpanojen lisäämisen koodin sekaan ja graafisten kuvaajien esittämisen. Yksi Jupyter Notebookin merkittävistä eduista on sen mahdollistama tapa, jolla koodia suoritetaan. Ohjelmakoodia voidaan jaotella soluihin ja nämä solut voidaan suorittaa yksitellen. Jos ohjelmaan tehdään lisäyksiä, riittää että solut, jotka sisältävät lisäyksiä suoritetaan. Näin ohjelmaa ei tarvitse suorittaa aina alusta, vaikka siihen lisätään jotakin. (Jupyter, n.d.) Kehittämistyössä Notebookin solurakenne tarjosi konkreettisen edun. Laskennallisten resurssien rajallisen määrän takia tekstin käsittelyssä kesti jopa muutamia minutteja. Solujen ansiosta, käsittelyvaihe täytyi suorittaa ainoastaan kertaalleen, käsitelty teksti säilyi muuttujassa, sillä ohjelman suorittamista ei tarvinnut aloittaa alusta. Näin ohjelman kehittämiseen kuluva aikaa saatiin lyhennettyä.

Kehittämistyössä käytettäväksi ohjelmointikieleksi valikoitui Python, sillä se tarjoaa useita koneoppimiseen tarkoitettuja kirjastoja. Python ja osa kirjastoista olivat työn tekijälle jo entuudestaan tuttuja, joten sekin vaikutti kielen valitsemiseen.

Pandas on Python kirjasto, joka tarjoaa kaksi erilaista luokkaa datan käsittelyä varten. Kehittämistyössä käytetään näistä luokista toista, niin sanottua DataFramea. DataFrameen avulla dataa voidaan säilyttää excel-tilukkoa muistuttavassa rakenteessa, josta löytyy sarakkeet ja rivit. DataFramea voidaan siis hyödyntää ohjelmissa, joihin tuodaan dataa taulukoidussa muodossa. Luonnollisesti Pandas tarjoaa mahdollisuuden muokata DataFramea, sarakkeita voidaan siirrellä tai nimetä uusiksi, mutta myös niiden sisältämää dataa voidaan käsitellä esimerkiksi datan puhdistus tarkoituksessa. (pandas, n.d.)

Toinen kirjasto, jota hyödynnetään kehittämistyössä, on NLTK eli Natural Language Toolkit. Kuten nimi antaa ymmärtää, tämä kirjasto tarjoaa mahdollisuuksia suorittaa luonnollisen kielen käsittelyyn liittyviä tehtäviä. NLTK kirjasto mahdollistaa esimerkiksi sanojen tokenisoinnin ja sanaluokkien merkitsemisen niihin. NLTK tarjoaa myös mahdollisuuden käyttää muita luonnollisen kielen käsittelyyn liittyviä kirjastoja NLTK kirjaston kautta. NLTK on siis tavallaan kirjastojen kokoelma. (NLTK, n.d.)

5.3 Datasetsi tekoälyn kouluttamiseksi

Jotta ohjattu oppiminen on mahdollista, täytyy ohjatun oppimisen mallille syötettävää dataa käsitellä. Käsittelemättömän datan muuttaminen muotoon, jota voidaan hyödyntää koneoppimismallin harjoittamiseen, on ihmisten tehtävä. Mallia ei voida harjoittaa ilman dataa, joten datan kerääminen ja muuttaminen sopivaan muotoon on yleensä ensimmäinen, ja äärimmäisen tärkeä askel koneoppimisessa. Koneoppimismallit eivät osaa tehdä päätelmiä asioista, joita niille ei ole opetettu. Tämä on toinen syy sille, miksi harjoitteludata on tärkeä osa koneoppimista. Kattava, ymmärrettävä ja tarkoituksenmukainen data varmistaa mallin lopputuloksen olevan onnistunut. Datat keräämisen tulisi olla myös jatkuva toimenpide. Yleensä hyvä tapa kehittää olemassa olevaa koneoppimismallia, on syöttää sille lisää dataa. Huonolaatuinen data tuottaa huonolaatuisen koneoppimismallin. On myös mahdollista, että mallista tulee täysin käyttökelvoton, jos harjoittamiseen käytettävä materiaali ei sisällä käyttötarkoituksen kannalta olennaisia piirteitä. (Sarkis, 2023)

Harjoitusdatan keräämisessä vaaditaan siis tarkkuutta. Käyttötarkoitusta täytyy miettiä, ylimääräisiä asioita täytyy karsia pois datan seasta ja jonkun täytyy vielä tarkistaa, että data on luokiteltu oikein. Tämä tekee prosessista äärimmäisen työlään. Onneksi harjoitusdataa on myös tarjolla verkossa. Joku muu on nähnyt vaivan datan muovaamisen eteen, jotta sitä voidaan käyttää tiettyyn tarkoitukseen. Verkossa on useita palveluja, joista tällaista harjoitusdataa voidaan hankkia. Kaggle on yksi näistä.

Kehittämistyöhön käytetty datasetti on ladattu Kagglesta. Datasetti on Database Contents License (DbCL) v1.0:n alaisuudessa, joten sen käyttäminen on täysin vapaata. Datasetti sisältää 5695 uniikkia viestiä, jotka ovat luokiteltu joko normaaleiksi viesteiksi tai roskapostiksi. Kehittämistyössä oli alun perin tarkoitus opettaa koneoppimismalli tunnistamaan verkkourkintaviestejä, mutta tällaisista viesteistä koostuvaa datasettiä ei tahtonut löytyä. Työn aiheeksi valikoitui siis roskapostisuodatus saatavilla olevien datasettien mukaan.

5.4 Tekoälyn osuus ohjelmassa

Ohjelmassa hyödynnetään sekä koneoppimista, että luonnollisen kielen käsittelyyn liittyviä toimia. Tekstin prosessoinnilla, kuten ylimääräisten merkkien poistamisella, sanaluokkien merkitsemisellä, sanojen muuttamisella perusmuotoon ja muilla toimilla data saadaan puhdistettua ja yksinkertaistettua. Näillä toimilla tekstistä poistetaan ylimääräiset tekijät, jotka

saattavat vaikuttaa negatiivisesti ohjelman toimintaan. Esimerkiksi sanojen muuttaminen perusmuotoon varmistaa, että malli käsittelee kaikki sanamuodot samana sanana, joka voi vaikuttaa positiivisesti mallin tarkkuuteen. Sanoihin merkitään sanaluokat, sillä sanaluokka vaikuttaa sanan perusmuotoon. Tällaiset tekstin prosessointiin eli käsittelyyn liittyvät vaiheet eivät ole varsinaisesti luonnollisen kielen käsittelyä, mutta niitä hyödynnetään luonnollisen kielen käsittelyä vaativissa tehtävissä, mutta myös koneoppimisessa.

TF-IDF on toinen luonnollisen kielen käsittelyyn liittyvä konsepti. Sanoille lasketaan TF-IDF-arvot, jotta niitä voidaan esittää numeerisesti. Sanojen esittäminen numeerisesti on tarpeellista, sillä logistinen regressiomalli ei kykene tulkitsemaan niitä sellaisenaan. Kun sanat ovat muutettu numeeriseen muotoon, päästään ohjelmassa varsinaiseen tekoälyä hyödyntävään vaiheeseen. Ohjelmassa luodaan koneoppimismalli, joka kykenee luokittelemaan viestejä niissä esiintyvien sanojen perusteella. TF-IDF-arvot osoittavat mallille, mitkä sanat korreloivat minkäkin tyyppisten viestien kanssa. Luokittelussa hyödynnetään logistista regressiomallia, jota käsiteltiin luvussa 4.2.2.

5.5 Ohjelman toteutus

Tässä luvussa käydään ohjelmakoodi läpi vaiheittain ja tarkastellaan erikseen mitä missäkin vaiheessa tapahtuu. Selitykset ohjelmakoodille löytyvät aina sen alapuolelta.

Ohjelmakoodi 1, Kirjastojen tuonti

```
import pandas as pd #Käytetään datan tallentamiseen ja käsittelyyn
import nltk #Käytetään tekstin käsittelyyn
import string #Käytetään välimerkkien poistamiseen
from nltk.corpus import stopwords #Käytetään hukkasanojen
poistamiseen
from nltk.tokenize import word_tokenize #Käytetään tekstin jakamiseen
sanoihin
from nltk.stem import WordNetLemmatizer #Käytetään sanojen
perusmuotoon muuttamiseen
from nltk.corpus import wordnet #Käytetään sanojen synonyymien
hakemiseen
from nltk import pos_tag #Käytetään sanaluokkien tunnistamiseen ja
merkitsemiseen
from sklearn.feature_extraction.text import TfidfVectorizer
#Käytetään tekstin vektorointiin
from sklearn.model_selection import train_test_split #Käytetään datan
jakamiseen opetus- ja testidataan
from sklearn.linear_model import LogisticRegression #Käytetään
logistisen regressiomallin luomiseen
from sklearn.metrics import classification_report #Käytetään mallin
suorituskyvyn arviointiin
```

Import komennolla voidaan tuoda uusi kirjasto käytettäväksi ohjelmaan. Kirjastot ovat eräänlaisia etukäteen kirjoitetun koodin kokoelmia. Kirjastojen avulla erilaisia monimutkaisia toimintoja voidaan suorittaa, jopa yhden koodirivin avulla. Jos kirjaston toiminnallisuus täytyisi ohjelmoida erikseen, voisi se vaatia satoja rivejä koodia. Lyhyet kuvaukset kirjastojen käyttötarkoituksesta löytyvät ohjelmakoodista kirjaston tuonnin perästä, selittävät rivit alkavat ”#”-merkillä.

Ohjelmakoodi 2, Datan lataaminen ja tarkastaminen

```
messages = pd.read_csv('emails.csv')
messages.head()
```

Ylempi komento lataa datasetin niin sanottuun Pandas DataFrameen, tässä vaiheessa data saadaan käytettäväksi ohjelmassa. Aiemmin haettua Pandas kirjastoa käytetään siis datan käyttöönnotossa. Alemmalla komennolla voidaan tulostaa DataFramen rivejä rakenteen tarkastelua varten. Komento tulostaa DataFramen viisi ensimmäistä riviä (Kuva 8).

Kuva 8. Ohjelmassa käyttöön otettu data tulosteena.

| | Text | Spam |
|---|---|------|
| 0 | Subject: naturally irresistible your corporate... | 1 |
| 1 | Subject: the stock trading gunslinger fanny i... | 1 |
| 2 | Subject: unbelievable new homes made easy im ... | 1 |
| 3 | Subject: 4 color printing special request add... | 1 |
| 4 | Subject: do not have money , get software cds ... | 1 |

Ohjelmakoodi 3, DataFramen muokkaaminen

```
messages.columns = ["viesti", "tyyppi"]
messages = messages[["tyyppi", "viesti"]]
messages['tyyppi'] = messages['tyyppi'].replace({1: 'spam', 0: 'ham'})
messages.loc[:, 'viesti'] = messages['viesti'].str.replace('Subject:',
 '')
messages.head()
```

Ylintä komentoa käytetään sarakkeiden nimien vaihtamiseen, tässä tapauksessa uudet nimet ovat viesti ja tyyppi. Seuraavassa komennossa sarakkeiden paikat vaihdetaan siten, että viestin tyyppi sijaitsee DataFramen vasemmassa reunassa. Seuraavalla komennolla viestien tyypit vaihdetaan sanallisiksi, spam viittaa roskapostiin ja ham oikeaan viestiin. Neljännellä komennolla viestien alusta poistetaan teksti ”Subject:”, sillä se ei ole merkityksellinen osa viestejä analysoinnin kannalta. Viimeisellä rivillä tulostetaan muokattu DataFrame (Kuva 9).

Kuva 9. DataFrame käsittelyn jälkeen.

| | tyyppi | viesti |
|---|--------|--|
| 0 | spam | naturally irresistible your corporate identit... |
| 1 | spam | the stock trading gunslinger fanny is merril... |
| 2 | spam | unbelievable new homes made easy im wanting ... |
| 3 | spam | 4 color printing special request additional ... |
| 4 | spam | do not have money , get software cds from her... |

Ohjelmakoodi 4, Funktio sanaluokkien hakemiseksi

```
def get_wordnet_pos(treebank_tag):
    if treebank_tag.startswith('J'):
        return wordnet.ADJ
    elif treebank_tag.startswith('V'):
        return wordnet.VERB
    elif treebank_tag.startswith('N'):
        return wordnet.NOUN
    elif treebank_tag.startswith('R'):
        return wordnet.ADV
    else:
        return wordnet.NOUN
```

Kokonaisuudessaan tätä funktiota käytetään yksittäisten sanojen sanaluokkien määrittelyyn.

Itse funktiota kutsutaan toisen funktion sisällä, jota tarkastellaan seuraavaksi kohdassa

Ohjelmakoodi 5.

Ohjelmakoodi 5, Funktio tekstin käsittelemiseksi

```
def preprocess_text(text):
    #Muutetaan kaikki kirjaimet pieniksi
    text = text.lower()
    #Tokenisoidaan teksti
    tokens = word_tokenize(text)
    #Poistetaan välimerkit
    tokens = [token for token in tokens if token not in
string.punctuation]
    #Poistetaan numerot ja muut merkit, jotka eivät ole sanoja
    tokens = [token for token in tokens if token.isalpha()]
    #Poistetaan hukkasanat kuten "a, an, the,..."
    stop_words = set(stopwords.words('english'))
    tokens = [token for token in tokens if token not in stop_words]
    #Muutetaan sanat perusmuotoon ja jaetaan sanat sanaluokkiin
    funktiolla
    lemmatizer = WordNetLemmatizer()
```

```

tagged_tokens = pos_tag(tokens)
tokens = [lemmatizer.lemmatize(token, pos=get_wordnet_pos(tag))
for token, tag in tagged_tokens]

return tokens

```

Rivit, jotka alkavat ”#” merkillä selittävät alapuolella suoritettavaa ohjelmakoodia. Käsittely suoritetaan, jotta tekstin analysointi olisi helpompaa ohjelmalle. Tämän vaiheen tarpeellisuutta käsiteltiin enemmän luvussa 3.3.3. Sanat muutetaan perusmuotoon, jotta ohjelman ei tarvitse käsitellä yhtä laajaa sanastoa. Sanaluokkiin jako suoritetaan, jotta sanat voidaan muuttaa niiden perusmuotoihin. Aiemmin esiteltyä sanaluokkiin jakamiseen käytettävää funktiota (Ohjelmakoodi 4) kutsutaan kolmanneksi viimeisellä rivillä.

Ohjelmakoodi 6, Tekstin käsittely -funktion kutsuminen

```

messages['viesti'] = messages['viesti'].apply(preprocess_text)
messages['viesti'] = messages['viesti'].apply(' '.join)

```

Ensimmäisellä rivillä kutsutaan viestien käsittelyyn käytettävää funktiota (Ohjelmakoodi 5). Funktion tuottama käsitelty teksti lisätään DataFramen ”viesti” sarakkeeseen. Toisella rivillä viestien sisältämät sanat muutetaan yhdeksi merkkijonoksi niiden tulevaa analysointia varten.

Ohjelmakoodi 7, Tekstin muuttaminen numeeriseksi

```

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(messages['viesti'])

```

Ensimmäisellä rivillä TF-IDF-vektoroija liitetään ”vectorizer” muuttujaan. Toisella rivillä vektoroija sovitetaan dataan, eli viesteihin. Tässä vaiheessa se kerää kaikki sanat viesteistä. Samalla näille sanoille lasketaan TF-IDF-arvot. Näistä sanoista ja arvoista muodostuu TF-IDF-matriisi, joka liitetään muuttujaan X. Luvussa 4.2.1 tarkastellaan TF-IDF-mallin toimintaa ja arvojen laskemista.

Ohjelmakoodi 8, Datan jakaminen koulutus- ja testaamissarjoihin

```

X_train, X_test, y_train, y_test = train_test_split(X,
messages['tyyppi'], test_size=0.2, random_state=42)

```

Äskettäin luotu matriisi (Ohjelmakoodi 8) jaetaan koulutus- ja testaamissarjoihin. Koulutusdata sisältää sekä sanojen TF-IDF-arvot, että viestien luokat, näin malli oppii tunnistamaan minkälaiset TF-IDF-arvot korreloivat minkäkin luokan kanssa. Testidata sisältää ainoastaan TF-IDF-arvot. Mallin täytyy siis ennustaa testidatan viesteille luokat TF-IDF-arvojen perusteella.

Ohjelmakoodi 9, Mallin luominen ja kouluttaminen

```
clf = LogisticRegression()
clf.fit(X_train, y_train)
```

Ylimmällä rivillä luodaan logistisen regression malli. Toisella rivillä mallia harjoitetaan aiemmin luodulla harjoitussarjalla. Muuttuja `clf` sisältää näiden rivien jälkeen logistisen regression mallin, jota voidaan hyödyntää ennustuksien suorittamiseen uudella datalla.

Ohjelmakoodi 10, Ennustuksien suorittaminen ja tarkkuuden havainnointi

```
y_pred = clf.predict(X_test)
print(classification_report(y_test, y_pred))
```

Ensimmäisellä rivillä mallille syötetään testidataa. Mallin testidatasta luomat ennustukset talletetaan muuttujaan `y_pred`. Nämä ennustukset ovat siis mallin arvioimien viestien tyyppisiä. Toisella rivillä tarkastellaan mallin ennustuksien tarkkuutta. Muuttuja `y_test` sisältää oikeat tyypit jokaiselle viestille. Mallin ennustamia, muuttujaan `y_pred` talletettuja viestien tyyppisiä, verrataan näihin todellisiin arvoihin. Lopuksi tulostetaan tarkkuuslaskelma (Kuva 10).

Kuva 10. Logistisen regressiomallin tarkkuus.

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| ham | 0.96 | 1.00 | 0.98 | 856 |
| spam | 0.98 | 0.89 | 0.94 | 290 |
| accuracy | | | 0.97 | 1146 |
| macro avg | 0.97 | 0.94 | 0.96 | 1146 |
| weighted avg | 0.97 | 0.97 | 0.97 | 1146 |

Alimmalta riviltä löytyvän painotetun keskiarvon perusteella malli kykenee luokittelemaan viestit 97 % tarkkuudella oikein.

Tarkkuuskertoimet ovat laskettu sekä ei-roskaposti viesteille, joita merkitään sanalla "ham", ja roskaposti viesteille, joita merkitään sanalla "spam". Precision eli tarkkuus merkitsee mallin positiivisten ennustuksien määrää todellisista positiivisista tapauksista. Arvo 0.96 merkitsee sitä, että 4 % prosenttia mallin ennustamista positiivisista tapauksista eivät olleet oikeasti positiivisia. Recall, tunnetaan myös herkkyyksimittarina, mittaa kuinka suuri osa todellisista positiivisista tapauksista luokitellaan oikein. Jos recall arvo on 0.89 merkitsee se sitä, että malli tunnistaa oikein 89 % positiivisista tapauksista. F1-score laskee tarkkuuden ja

herkkyyden välisen harmonisen keskiarvon. Yksinkertaisesti selitettynä, f1-scorea käytetään tarkkuuden ja herkkyyden yhdistämiseksi yhteen mittariin. Mitä lähempänä f1-scoren arvo on yhtä, sitä paremmin malli suoriutuu luokkien tunnistamisesta. Support sarakkeen lukema kertoo, kuinka monta kertaa luokka on edustettuna testidatassa. "Macro avg" ja "weighted avg" tarjoavat keskiarvot, kun mallin suoriutumista tarkastellaan kaikkien luokkien kesken. Luokkien painotukset vaihtelevat keskiarvon mukaan. Malli vaikuttaa olevan melko tarkka, minkä tahansa mittarin avulla sitä tarkastellaan.

Tarkastellaan vielä hieman, mitkä sanat vaikuttavat kaikkein eniten viestin luokitteluun.

Ohjelmakoodi 11, Sanojen ja niiden painoarvojen hakeminen

```
feature_names = vectorizer.get_feature_names_out()
coefficients = clf.coef_[0]
```

Ylemmällä rivillä haetaan kaikki piirteet eli sanat vektorisoijalta. Toisella rivillä haetaan sanojen kertoimet, eli logistisen regressiomallin luomat painotukset.

Ohjelmakoodi 12, Sanojen ja arvojen lisääminen listaan

```
word_weights = list(zip(feature_names, coefficients))
word_weights = sorted(word_weights, key=lambda x: abs(x[1]),
reverse=True)
```

Ylempää komentoa käytetään sanojen ja niiden painoarvojen lisäämiseksi listaan sanapainoarvo pareiksi. Toisella rivillä lista järjestellään painoarvojen mukaan suurimmasta pienimpään.

Ohjelmakoodi 13, Luodaan yleisimpien sanojen listat

```
spam_words = sorted([(feature_names[i], coefficients[i]) for i in
range(len(feature_names)) if coefficients[i] > 0], key=lambda x: x[1],
reverse=True)[:10]
```

Tällä rivillä luodaan lista sanoista, joilla on suurimmat positiiviset arvot. Mitä suurempi positiivinen arvo sanalla on, sitä todennäköisemmin se viittaa viestin olevan roskapostia.

Ohjelmakoodi 14, Tulostetaan merkityksellisimmät sanat

```
print('Top 10 words most indicative of spam:')
for word, weight in spam_words:
    print(f'Word: {word}, Weight: {weight}')
```

Tämä rivi tulostaa viestin roskapostiksi luokittelun kannalta merkityksellisimmät sanat, jotka ovat nähtävillä Kuva 11.

Kuva 11. Merkitykselliset sanat luokittain.

```
Top 10 words most indicative of spam:  
Word: click, Weight: 3.113998934895478  
Word: money, Weight: 2.8602148421158406  
Word: software, Weight: 2.8011760004434048  
Word: save, Weight: 2.5766899286830265  
Word: life, Weight: 2.457549158321343  
Word: remove, Weight: 2.234498870112422  
Word: online, Weight: 2.164483241660599  
Word: free, Weight: 2.0168169034773165  
Word: website, Weight: 2.001072018243888  
Word: viagra, Weight: 1.95017484420753
```

Roskapostiin vahvasti viittaavat sanat vaikuttavat melko oletetuilta. Raha ja sen säästäminen tai saaminen on yleinen piirre roskapostin ja toisaalta myös käyttäjän manipuloinnin tunnistamisessa. Näiden sanojen perusteella koneoppimismalli vaikuttaisi toimivan tarkoitetulla tavalla.

6 Tulokset ja jatkotutkimus

Halusin käsitellä opinnäytetyössä sellaista aihetta, joka on ajankohtainen eikä kosketa pelkästään alan ammattilaisia. Toivoin pääseväni hyödyntämään työssä sekä opintojen kautta, että oman mielenkiinnon pohjalta kerättyä tietoa. Työn tavoitteena ei ollut luoda täydellistä roskapostisuodatinta, joka voitaisiin ottaa heti käyttöön. Tarkoituksena oli ennemminkin päästä tutustumaan tekoälyyn tässä käyttötarkoituksessa ja luoda pohja, jonka päälle voidaan rakentaa kattavampi ohjelma mahdollisissa jatko-opinnoissa. Ilmiönä käyttäjän manipulointi ei ole häviämässä mihinkään, joten mielestäni jokaisen tulisi olla tietoinen näistä hyökkäyksistä ja teoriaosuuden tavoite olikin tiedon tarjoaminen tästä aiheesta lukijalle. Toinen tavoite teoriaosuudessa oli rakentaa lukijalle kattava tietopohja, joka auttaa ymmärtämään käytännön työn toimintaa ja vaiheita.

Aloitin opinnäytetyön suunnittelun pohtimalla minkälainen tieto voisi olla lukijalle hyödyllistä tai kiinnostavaa ja minkälainen tieto on tarpeellista työn kannalta. Teoriaosuudessa käytettävää tietoa kerättiin alan yritysten julkaisuista, erilaisista tutkimuksista, uutisartikkeleista ja kirjoista. Tekoälyn käyttötarkoituksista käyttäjän manipuloinnin tunnistamiseksi löytyi laajalti tutkimuksia. Valitettavasti nämä tutkimukset keskittyivät pääasiassa yksittäisiin algoritmeihin ja niiden toimintaan, joten yleispätevän tiedon kerääminen ei ollut helppoa. Toisaalta algoritmien tarkastelu tarjosi hyvän kuvan käyttäjän manipuloinnin tunnistamiseen käytettävistä piirteistä.

Sekä teoriaosuuden että käytännön työn pohjalta voidaan huomata, että tekoäly on todellakin toimiva työkalu käyttäjän manipuloinnin tunnistamisessa. Tekoälylle löytyy useita käyttökohteita tältä aihealueelta tarpeesta riippuen. Työssä ei tarkasteltu tekoälyn käyttömahdollisuuksia käyttäjän manipuloinnin tunnistamisessa tulevaisuudessa. Hyökkäykset sekä tekoäly kehittyvät jatkuvasti, joten jo vuoden tai kahden päästä sama työ voi sisältää täysin erilaisia asioita. Verkkorikolliset kehittävät hyökkäyksiä jatkuvasti uusien tunnistusmekanismien myötä.

Käytännön osuudessa ohjelmitava roskapostisuodatin oli henkilökohtaisesti haastava vaihe. Python oli ohjelmointikielenä entuudestaan tuttu, mutta kirjastot, joita työssä käytettiin eivät. Työn jakaminen vaiheisiin Kanbanin avulla tarjosi kuitenkin hyvän etenemissuunnitelman, jonka avulla ongelmien ratkominen oli helpompaa. Roskapostisuodatin kykenee luokkitelemaan viestejä melko tarkasti. Yksi tapa tarkastella mallin kehityskohteita olisi sen integrointi oikeaan sähköposti virtaan. Näin sen tarkkuutta ja siihen liittyviä heikkouksia

voitaisiin arvioida todellisessa ympäristössä. Toinen keino mallin kehittämiseen olisi sen ominaisuuksien lisääminen. Esimerkiksi URL-osoitteiden tarkastelu voisi parantaa suodattimen kykyä tunnistaa haitallisia viestejä, eikä ainoastaan roskapostia. Myös mallin harjoittamiseen käytettävää dataa voitaisiin laajentaa. Mallin kehittämisestä voidaan tehdä laaja jatkotutkimus ja tämä olikin osittain tavoitteena sen luomisessa. Jos mallia lähdetään kehittämään pidemmälle, voisi olla hyödyllistä koostaa oma datasetti sen harjoittamista varten. Tällä hetkellä malli kykenee luokittelemaan viestejä ainoastaan roskapostiksi tai ei-roskapostiksi. Voisi olla hyödyllistä luoda myös kolmas luokka viesteille, joita ei kyetä luokitella kumpaankaan alkuperäiseen luokkaan tarpeeksi suurella varmuudella.

7 Yhteenveto

Tutkimuskysymyksiin vastaaminen onnistui hyvin. Aiheesta löytyi suuria määriä tietoa ja etenkin erilaisia tekoälyyn ja käyttäjän manipulointiin liittyviä osa-alueita olisi työssä voinut käsitellä vielä enemmänkin, mutta tällöin työstä olisi saattanut tulla jopa liian laaja. Työssä tarkasteltiin kaikkein yleisimpiä ilmiöitä, joten aiheet rajattiin onnistuneesti tästä näkökulmasta. Tekoälyn eettisyyttä oli haastava tarkastella käyttäjän manipuloinnin tunnistamisessa. On selvää, että tekoälyyn liittyy eettisiä ongelmia, mutta ne eivät kuitenkaan ole läsnä yhtä vahvasti kaikissa käyttötarkoituksissa. En silti koe, että tekoälyn etiikkaa kannattaisi tarkastella laajemmin tässä työssä, sillä suuri osa eettisistä ongelmista liittyy täysin erilaisiin tekoälyn käyttötarkoituksiin, joita ei hyödynnetä käyttäjän manipuloinnin tunnistamisessa.

Vaikka käyttäjän manipulointi ja tekoäly olivat minulle ennestään tuttuja aiheita, pääsin oppimaan paljon uutta, sillä molemmissa tapahtuu kehitystä ja muutoksia jatkuvasti. Luin työtä varten useita kyberturvallisuuden ja tekoälyyn keskittyvien yritysten julkaisuja ja alan ammattilaisten artikkeleita, sillä ne tarjosivat mahdollisimman ajankohtaisen kuvan erilaisista ilmiöistä. Tekoälyn käyttötarkoituksista käyttäjän manipuloinnin tunnistamiseksi löytyi paljon tutkimuksia. Valitettavasti nämä tutkimukset keskittyivät laajalti yksittäisiin algoritmeihin ja niiden toimintaan, joten yleispätevän tiedon kerääminen ei ollut helppoa. Toisaalta algoritmien tarkastelu tarjosi hyvän kuvan käyttäjän manipuloinnin tunnistamiseen käytettävistä piirteistä.

Kehittämistyön tuloksena syntynyt roskapostisuodatin tarjoaa aiheen kattavalle jatkotutkimukselle. Jos jatkan opintojani tulevaisuudessa ylempään korkeakouluun, voidaan tätä työtä käyttää pohjana esimerkiksi mahdollisessa pro gradu -tutkielmassa.

Roskapostisuodattimelle löytyy jo nyt selkeitä kehittämiskohteita, mutta niiden toteutus voi olla hieman haastavaa ilman vankempaa osaamista aiheesta.

Tekoälyn täyttä potentiaalia ei olla vielä kyetty valjastamaan, joten mahdollisten kehitysten myötä vastaavan työn sisältö voi olla täysin erilainen. Sekä teoria- että käytännön osuuteen voi löytyä merkittäviä lisäyksiä niin tekoälyn kehittymisen puolesta, kuin myös verkkorikollisten uusien toimintatapojen ja taktikkojen puolesta. Kyberturvallisuuden ja käyttäjän manipulointiin liittyvät ilmiöt kehittyvät jatkuvasti ja uusia puolustautumismekanismeja syntyy tämän myötä. Oleellista käyttäjän manipuloinnin

tunnistamisessa on seurata uusia ilmiöitä ja kouluttaa ihmisiä näiden tunnistamiseksi, ei ainoastaan tekoälyä.

Lähteet

- Adee, S. (29.4.2020). *What Are Deepfakes and How Are They Created?*. *IEEE Spectrum*.
<https://spectrum.ieee.org/what-is-deepfake>
- Atlassian. (n.d.). *Kanban*. <https://www.atlassian.com/agile/kanban>
- ChatGPT. (2024) <https://chat.openai.com/>
- Cisco. (n.d.). *What is a whaling phishing attack?*. Cisco.
<https://www.cisco.com/site/us/en/learn/topics/security/what-is-a-whaling-attack.html#tabs-9cfa4a460b-item-b8ba101fed-tab>
- Cloudflare. (n.d.). *What is deep learning?*. <https://www.cloudflare.com/learning/ai/what-is-deep-learning/>
- Damiani, J. (3.9.2019). *A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000*. Forbes. <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>
- Delua, J. (12.3.2021). *Supervised vs. Unsupervised Learning: What's the Difference?*. IBM.
<https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>
- Enisa. (19.10.2023). *ENISA Threat Landscape 2023*. European Union Agency for Cybersecurity. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2023/@@download/fullReport>
- Enisa. (n.d.). *What is "Social Engineering"?*. European Union Agency for Cybersecurity.
<https://www.enisa.europa.eu/topics/incident-response/glossary/what-is-social-engineering>
- F-Secure. (n.d.). *Mitä on käyttäjän manipulointi eli social engineering?*. F-Secure.
<https://www.f-secure.com/fi/articles/what-is-social-engineering>
- Gatlan, S. (6.9.2019). *Over \$37 Million Lost by Toyota Boshoku Subsidiary in BEC Scam*. Bleeping Computer. <https://www.bleepingcomputer.com/news/security/over-37-million-lost-by-toyota-boshoku-subsidiary-in-bec-scam/>
- GitHub Docs. (n.d.) *Getting started with GitHub Copilot*. GitHub.
<https://docs.github.com/en/copilot/using-github-copilot/getting-started-with-github-copilot>
- Hadnagy, C. (2018). *Social engineering : The science of human hacking*. John Wiley & Sons, Incorporated. ISBN-13: 9781119433736
- Hadnagy, C., & Fincher, M. (2015). *Phishing dark waters : The offensive and defensive sides of malicious emails*. John Wiley & Sons, Incorporated. ISBN-13: 9781118958490
- Hazell, J. (14.12.2023). *Spear Phishing With Large Language Models*.
<https://doi.org/10.48550/arXiv.2305.06972>

- Hernandez-Ortega, J., Tolosana, R., Fierrez, J., Morales, A., (14.12.2020). *DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation*.
<https://doi.org/10.48550/arXiv.2010.00400>
- Hintze, A. (14.11.2016). *Understanding the four types of AI, from reactive robots to self-aware beings*. The Conversation. <https://theconversation.com/understanding-the-four-types-of-ai-from-reactive-robots-to-self-aware-beings-67616>
- HKT. (4.2.2024). *Deepfake colleagues trick HK clerk into paying HK\$200m*. RTHK.
<https://news.rthk.hk/rthk/en/component/k2/1739119-20240204.htm>
- Hu, F., & Hao, Q. (21.3.2013). *Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning*. CRC Press.
<https://doi.org/10.1201/b14300>
- IBM. (n.d.). *What is natural language processing (NLP)?*. <https://www.ibm.com/topics/natural-language-processing>
- IBM Data and AI Team. (12.10.2023). *Understanding the different types of artificial intelligence*. IBM. <https://www.ibm.com/blog/understanding-the-different-types-of-artificial-intelligence/>
- Javid, S., (14.2.2024). *Ethical & Legal AI Data Collection in 2024: Examples & Policies*. AIMultiple. <https://research.aimultiple.com/data-collection-ethics/>
- Jupyter. (n.d.) *The Jupyter Notebook*. <https://jupyter-notebook.readthedocs.io/en/stable/notebook.html>
- Jurafsky, D. & Martin, J. (3.2.2024.). *Speech and Language Processing (3rd ed. draft)*.
<https://web.stanford.edu/~jurafsky/slp3/>
- Kaspersky. (n.d.-a). *What is a Whaling Attack?*. Kaspersky.
<https://www.kaspersky.com/resource-center/definitions/what-is-a-whaling-attack>
- Kaspersky. (n.d.-b). *Ways to avoid social engineering attacks*. Kaspersky.
<https://www.kaspersky.com/resource-center/threats/how-to-avoid-social-engineering-attacks>
- Koivuranta, L. (2.11.2022). *Mielihyvähormonit ja tunteet*. [Kanditaatintyö, Oulun Yliopisto].
<https://urn.fi/URN:NBN:fi:oulu-202211023544>
- Kunnari, T. (13.12.2023). *SaiPa meni ulkomaisten huijareiden ansaan – Karmea lasku seuralle*. Iltalehti. <https://www.iltalehti.fi/smliga/a/b8162fd3-2cff-42fb-90e4-3aa06c606778>
- Lenaerts-Bergmans, B. (8.11.2023). *10 Types of Social Engineering Attacks and how to prevent them*. CrowdStrike. <https://www.crowdstrike.com/cybersecurity-101/types-of-social-engineering-attacks/>

- Lenaerts-Bergmans, B. (19.10.2022). *What is domain spoofing?*. Crowdstrike.
<https://www.crowdstrike.com/cybersecurity-101/spoofing-attacks/domain-spoofing/>
- Lenaerts-Bergmans, B (14.10.2021). *How to spot a phishing email*. Crowdstrike.
<https://www.crowdstrike.com/cybersecurity-101/phishing/how-to-spot-a-phishing-email/>
- Lindroos, J. (30.3.2023). *Liian hyvä on harvoin totta – suomalaiset ovat menettäneet kymmeniä miljoonia verkkohuijareille*. Yle. <https://yle.fi/a/74-20025059>
- Lockheed Martin. (n.d.) *Cyber Kill Chain*. <https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html>
- Lyu, S. (25.6.2019). *Detecting deepfakes by looking closely reveals a way to protect against them*. The Conversation. <https://theconversation.com/detecting-deepfakes-by-looking-closely-reveals-a-way-to-protect-against-them-119218>
- Madiega, T., (9.2019). *EU guidelines on ethics in artificial intelligence: Context and implementation*. European Parliament.
[https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI\(2019\)640163_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf)
- Mandiant. (2023). *M-Trends 2023*. Mandiant. Haettu 17.1.2024 osoitteesta
<https://www.mandiant.com/m-trends>
- MathWorks. (n.d.-a). *What Is a Neural Network?*. <https://se.mathworks.com/discovery/neural-network.html>
- MathWorks. (n.d.-b). *What Is Deep Learning?*. <https://se.mathworks.com/discovery/deep-learning.html>
- MathWorks. (n.d.-c). *What Is a Convolutional Neural Network?*.
<https://se.mathworks.com/discovery/convolutional-neural-network.html>
- Microsoft. (n.d.). *What is business email compromise (BEC)?*. Microsoft.
<https://www.microsoft.com/en-us/security/business/security-101/what-is-business-email-compromise-bec#areaheading-oc5a75>
- Microsoft Threat Intelligence. (12.7.2022). *From cookie theft to BEC: Attackers use AiTM phishing sites as entry point to further financial fraud*. Microsoft.
<https://www.microsoft.com/en-us/security/blog/2022/07/12/from-cookie-theft-to-bec-attackers-use-aitm-phishing-sites-as-entry-point-to-further-financial-fraud/>
- NLTK. (n.d.). Natural Language Toolkit. <https://www.nltk.org/>
- OWASP. (n.d.). *ML02:2023 Data Poisoning Attack*. https://owasp.org/www-project-machine-learning-security-top-10/docs/ML02_2023-Data_Poisoning_Attack
- Ozkaya, E. (2018). *Learn social engineering : Learn the art of human hacking with an internationally renowned expert*. Packt Publishing, Limited. ISBN-13: 9781788838009

pandas. (n.d.). *Getting started*.

https://pandas.pydata.org/docs/getting_started/index.html#getting-started

Pazzanese, C. (26.10.2020). *Ethical concerns mount as AI takes bigger decision-making role in more industries*. The Harvard Gazette.

<https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/>

Proofpoint. (n.d.). *Business Email Compromise (BEC)*. Proofpoint.

<https://www.proofpoint.com/us/threat-reference/business-email-compromise>

Radigan, D. (n.d.). *Agile vs. waterfall project management*. Atlassian.

<https://www.atlassian.com/agile/project-management/project-management-intro>

RangeForce Team. (n.d.). *Tutorial: The Four Phases of Social Engineering*. RangeForce.

<https://www.rangeforce.com/blog/four-phases-of-social-engineering>

Reidy, O. (Viimeeksi muokattu: 29.9.2023). *Transcribe speech to text in real time from phone calls*. Telnyx. <https://telnyx.com/resources/stt-transcription-demo>.

Sarkis, A. (2023). *Training Data for Machine Learning*. O'Reilly Media, Inc 2022.

ISBN:9781492094524

Shahrivari, V., Darabi, M. M., Izadi, M. (20.9.2020). *Phishing Detection Using Machine Learning Techniques*. <https://doi.org/10.48550/arXiv.2009.11116>

Suomen Koodikoulu. (11.12.2019). (Muokattu 4.11.2020). *Johdatus tekoälyyn*.

<https://aoe.fi/api/v1/download/file/Johdatustekolyyn-1576067361216.pdf>

Suopanki, K. (2020). *Suuri suomalainen romanssihuijaus: käsikirjoitus*. Yle.

<https://yle.fi/aihe/artikkeli/2020/09/07/suuri-suomalainen-romanssihuijaus-kasikirjoitus>

Verma, R., Shashidhar, N., Hossain, N. (2012). *Detecting Phishing Emails the Natural Language Way*. Teoksessa Foresti, S., Yung, M., Martinelli, F. (eds) *Computer Security – ESORICS 2012. ESORICS 2012. Lecture Notes in Computer Science*, vol 7459. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33167-1_47

Wisdom, D. (30.6.2023). *What is natural language processing and how does it work?*.

https://www.datalinknetworks.net/dln_blog/what-is-natural-language-processing-and-how-does-it-work

Woo, E. (20.3.2014). *John McCarthy dies at 84; the father of artificial intelligence*. Los Angeles Times.

<https://www.latimes.com/local/obituaries/la-me-john-mccarthy-20111027-story.html>

Liite 1: Aineistonhallintasuunnitelma

Tutkimuksellinen osuus:

Tutkimukselliseen osuuteen liittyvä aineisto talletetaan kolmeen sijaintiin. Tietokoneen C-asemalle, henkilökohtaisen Google-tilin Google Driveen ja koulun tiliin linkitettyyn Google Driveen. Aineisto ei sisällä arkaluontoista - tai salassa pidettävää tietoa, joten työn ja siihen liittyvien tiedostojen tallentaminen pilvipalveluun ei ole ongelma. Tallennettavat tiedot koostuvat ChatGPT keskusteluista, jotka löytyvät myös liitteinä työn lopusta, PDF-tiedostoista, kuvista ja verkkosivujen vedoksista, joita on käytetty lähteinä. Osasta verkkosivuista säilytetään vedoksia, jotta työssä viitattu sisältö olisi vielä saatavilla vaikka verkkosivu ei enää olisi entisensä tai saatavilla. Kaappaukset otetaan verkkosivuista, joiden sisältö on esimerkiksi merkitty mahdollisesti muokattavaksi.

Käytännön osuus:

Käytännön osuudessa luodaan ohjelma käyttäen Kanban-kehitystyylä. Valmis ohjelma kykenee tunnistamaan roskapostiviestejä. Kaikki ohjelmaan ja sen kehittämiseen liittyvät tiedostot tallennetaan sekä henkilökohtaiseen, että koulun tiliin yhdistettyyn Google Driveen, sekä tietokoneen C-asemalle. Tallennettavia tiedostoja ovat esimerkiksi ohjelmakoodi sekä kehittämistyössä hyödynnetty datasetti.

Opinnäytetyöaineiston jatkokäyttö työn valmistumisen jälkeen

Tutkimusaineistoa ei jatkokäytetä. Opinnäytetyön tekijä säilyttää aineiston tietoturvallisesti vuoden ajan opinnäytetyön hyväksymispäivästä, jotta opinnäytetyön tulokset voidaan tarvittaessa varmistaa ja hävittää tämän jälkeen aineiston tietoturvallisesti.

Liite 2: ChatGPT keskustelu, Mitä tekoäly on

<https://chat.openai.com/share/3db567a9-4e81-4d29-98c7-1224148b191b>

Tekoäly määriteltynä

January 28, 2024



Anonymous

Miten määrittelisit tekoälyn omin sanoinesi?



ChatGPT

Tekoäly (TAI) on tietojenkäsittelyn osa-alue, joka keskittyy järjestelmien kehittämiseen ja ohjelmointiin niin, että ne voivat suorittaa tehtäviä, jotka yleensä vaativat inhimillistä älykkyyttä. Tekoäly pyrkii jäljittelemään ihmisten kykyä oppia, sopeutua, päätellä, suunnitella, ymmärtää luonnollista kieltä ja ratkaista ongelmia. Se perustuu usein koneoppimisen, neuroverkkojen ja muiden edistyneiden laskennallisten menetelmien käyttöön.

Yksinkertaisesti sanottuna tekoäly pyrkii antamaan tietokoneille kyvyn suorittaa tehtäviä, joita yleensä pidetään älykkyyden osoituksena, kuten kuvien tunnistaminen, puheen ymmärtäminen, päätöksenteko, käännöstyöt ja monimutkaisten ongelmien ratkaiseminen. Tekoäly voi olla joko vahvaa (yleispätevää) tai heikkoa (tehtäväkohtaista), ja sen sovellukset vaihtelevat laajasti eri aloilla, kuten terveydenhuolto, liiketoiminta, tieteellinen tutkimus ja automaatio.

Liite 3: ChatGPT keskustelu, Looginen päättely:

<https://chat.openai.com/share/d8bf87e5-5799-4db7-af58-859e6c36d5db>

Swedish Blue, Finnish Cat

January 30, 2024

**Anonymous**

There are two neighbours, one of them is swedish and the other is finnish. One of them has a dog and the other has a cat. The swedish lives in a blue house. The dog owner lives in a red house. Which house is the cat living in?

**ChatGPT**

Based on the information provided:

1. The Swedish neighbor lives in a blue house.
2. The one with the dog lives in a red house.

Since there are only two houses and both neighbors have been assigned a color and a pet, it can be inferred that the Finnish neighbor, who is not the one with the dog, must have the cat. Therefore, the cat is living in the blue house, which is the one owned by the Swedish neighbor.

**Anonymous**

Can you tell me why you couldn't provide the correct answer to this question?

**ChatGPT**

Certainly! I apologize for any confusion in my previous response. It seems there was an oversight in my explanation. Let me clarify:

1. The Swedish neighbor lives in a blue house.
2. The one with the dog lives in a red house.

Given this information, we can conclude:

- The Swedish neighbor has the blue house.
- The dog owner has the red house.

Now, since the only options for pets are a dog and a cat, and we know the Swedish neighbor has the blue house (and thus doesn't have the dog), the Finnish neighbor must have the red house and the dog. Therefore, the cat is living in the blue house, which belongs to the Swedish neighbor.

I hope this clears up any confusion, and I appreciate your patience.

**Anonymous**

Why do you think you wasn't able to solve the question with the first try?

**ChatGPT**

Upon reviewing my initial response, I see that I made an error in my explanation. I appreciate your feedback and apologize for any confusion caused by the oversight. The correct answer was embedded in my response, but the explanation did not explicitly highlight it. I should have clearly stated that the cat is living in the blue house, which is owned by the Swedish neighbor.

Mistakes can happen due to the complexity of language and reasoning, and I appreciate your understanding. If you have any more questions or if there's anything else I can help you with, please feel free to let me know.

Liite 4: ChatGPT keskustelu, Tietojenkalasteluviesti:

<https://chat.openai.com/share/28881c21-3c0a-439c-a6a2-ea415bd67e78>

Tunnista tietojenkalastelu.

February 7, 2024



Anonymous

Haluaisin oppia tunnistamaan tietojenkalasteluviestejä paremmin, osaisitko kertoa minulle yleisiä piirteitä, joiden avulla voitaisiin tunnistaa esimerkiksi kohdennettu tietojenkalasteluviesti tai -hyökkäys?



ChatGPT

Tietojenkalasteluviestit voivat olla hyvin hienovaraisia ja vaikeasti tunnistettavissa, mutta on olemassa useita yleisiä piirteitä, joiden avulla voit ehkä tunnistaa niitä:

- Epäilyttävät linkit tai liitteet:** Tietojenkalasteluviesteissä on usein linkkejä tai liitetiedostoja, joita pyydetään avaamaan tai klikkaamaan. Ole varovainen näiden linkkien ja liitteiden kanssa, etenkin jos et odota viestiä kyseiseltä lähettäjältä.
- Pyyntö henkilökohtaisista tiedoista:** Tietojenkalasteluviesteissä saatetaan pyytää antamaan henkilökohtaisia tietoja, kuten salasanoja, luottokorttitietoja tai muita arkaluontoisia tietoja. Aito yritys tai organisaatio ei yleensä pyydä tällaisia tietoja sähköpostitse.
- Kiireellisyys ja pelottelu:** Tietojenkalasteluviesteissä saatetaan käyttää kiireellisiä tai pelottelevia viestejä, jotka painostavat vastaanottajaa toimimaan nopeasti ja antamaan halutut tiedot.
- Epätavalliset kielioppi- tai oikeinkirjoitusvirheet:** Monet tietojenkalasteluviestit ovat huonosti kirjoitettuja ja sisältävät kielioppi- tai oikeinkirjoitusvirheitä. Ne voivat myös sisältää käännösvirheitä, jos ne on käännetty toisesta kielestä.
- Epämääräiset lähettäjätiedot:** Tietojenkalasteluviesteissä saatetaan käyttää epämääräisiä tai väärennettyjä lähettäjätietoja, jolloin ne näyttävät tulevan luotetulta lähteeltä, vaikka eivät todellisuudessa tulisikaan.
- Epäilyttävät pyynnöt tai tarjoukset:** Tietojenkalasteluviesteissä saatetaan tarjota epäilyttäviä tarjouksia tai pyytää tekemään epätavallisia toimia, kuten lataamaan ohjelmistoja tai jakamaan henkilökohtaisia tietoja.
- Verkkotunnus- tai osoitteen poikkeavuudet:** Tarkista viestin linkkien osoitteet ja verkkotunnukset. Usein tietojenkalasteluviesteissä käytetään väärennettyjä verkkotunnuksia tai osoitteita, jotka ovat hyvin samankaltaisia kuin aitojen yritysten verkkotunnukset, mutta niissä voi olla pieniä poikkeavuuksia.

Nämä ovat vain joitakin yleisiä piirteitä, jotka voivat auttaa sinua tunnistamaan tietojenkalasteluviestejä. On tärkeää olla aina varovainen ja tarkistaa huolellisesti epäilyttäviltä vaikuttavat viestit ennen kuin toimit niiden kanssa. Jos olet epävarma viestin aitoudesta, kannattaa ottaa yhteyttä organisaatioon tai yritykseen suoraan varmistaaksesi viestin aitouden.

**Anonymous**

Miltä verkkorikollisen luoma kohdennettu tietojenkalastelu viesti vaikkapa pankin nimissä saattaisi näyttää?

**ChatGPT**

Kohdennetun tietojenkalasteluviestin luominen pankin nimissä voi sisältää useita hienovaraisia piirteitä, jotka pyrkivät vakuuttamaan vastaanottajan viestin aitoudesta ja saamaan hänet antamaan arkaluontoisia tietoja. Tässä on esimerkki siitä, miltä tällainen viesti saattaisi näyttää:

Aihe: Tärkeä ilmoitus tilisi turvallisuudesta

Hei [Vastaanottajan nimi],

Olemme havainneet epätavallista toimintaa tilillänne ja haluamme varmistaa tilinne turvallisuuden. Jotta voimme suojata tilitietonne mahdollisilta väärinkäytöksiltä, pyydämme teitä vahvistamaan tilinne tiedot välittömästi.

Ystävällisesti klikkaa alla olevaa linkkiä ja kirjaudu sisään pankkitilillesi vahvistaaksesi henkilöllisyytesi:

[Epäilyttävä linkki: <https://www.pankki-turva.net/vahvista-tili>]

Jos et vahvista tietojasi 24 tunnin kuluessa tästä ilmoituksesta, saatamme joutua tilapäisesti sulkemaan tilisi turvallisuussyistä. Pahoittelemme mahdollisia aiheutuneita haittoja ja kiitämme yhteistyöstäsi.

Ystävällisin terveisin,

Pankin Turvallisuustiimi

Tässä esimerkissä tietojenkalasteluviesti käyttää joitakin yleisiä tekniikoita huijatakseen vastaanottajaa:

1. **Kiireellisyys:** Viestissä mainitaan, että tilin turvallisuus voi olla vaarassa, mikä luo kiireen vastata nopeasti.
2. **Epäilyttävä linkki:** Vaikka linkki näyttää johtavan pankin verkkosivustolle, se vie todennäköisesti väärennetyille sivulle, jossa pyydetään antamaan henkilökohtaisia tietoja.
3. **Pelottelevat seuraukset:** Viestissä uhataan tilin sulkemisella, mikä voi saada vastaanottajan toimimaan nopeammin pelon vuoksi.
4. **Epämääräinen lähetystiedot:** Vaikka viestissä mainitaan "Pankin Turvallisuustiimi", lähetystiedot voivat olla väärennetyjä tai epämääräisiä.

On tärkeää olla tarkkaavainen ja varovainen vastaavien viestien kanssa. Jos epäilet viestin aitoutta, on parasta ottaa yhteyttä pankkiin suoraan sen virallisilla yhteystiedoilla varmistaaksesi viestin oikeellisuuden.