Kohdennettujen kalasteluviestien tunnistaminen koneoppimisen avulla

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Huhtikuu 2024
Timi Pietilä

TURUN YLIOPISTO
Tietotekniikan laitos

Timi Pietilä: Kohdennettujen kalasteluviestien tunnistaminen koneoppimisen avulla

LuK-tutkielma, 29 s.
Tietojenkäsittelytiede
Huhtikuu 2024

Kohdennetut kalasteluviestit ovat kalasteluviestien kehittyneempi versio, joka vaikeammin tunnistettava ja jatkuva uhka etenkin yrityksille. Kohdennetut kalasteluviestit toimivat usein hyökkäysvektorina, sillä tietoturvajärjestelmien välttäminen sosiaalisen manipuloinnin avulla on helpompaa kuin niiden murtaminen. Kohdennetujen kalasteluviestien toteutus on kyberrikollisille helppoa, sillä imitoitavista kohteista löytyy helposti julkista tietoa viestin muodostamiseksi. Turvamekanismeja löytyy kohdennettuja kalasteluviestejä vastaan, mutta ne eivät ole vielä laajassa käytössä. Tässä kirjallisuuskatsauksessa tutkitaan minkälaisia koneoppimisen keinoja hyödyntäviä ratkaisuja on kehitetty kohdennettujen kalasteluviestien tunnistamiseen, sekä mitä ominaisuuksia hyvällä tunnistamisjärjestelmällä on. Katsaus kokoaa yhteen uusinta tekniikkaa olevat ratkaisut ja vertaa niiden ominaisuuksia, sekä ehdottaa kehityskohteita.

Asiasanat: spearphishing, phishing, koneoppiminen, kalasteluviestit, tunnistaminen, luokittelu

Sisällys

Taulukot

1 Johdanto

Tietoturvahyökkäysten määrä on jo pitkään ollut kasvussa [1], tätä kasvua ei ole onnistunut hidastamaan kehittyneet suojauskeinot, sillä myös käytettävien teknologioiden määrä ja kompleksisuus on kasvanut tuoden uusia haavoittuvuuksia. Yleinen keino kyberrikollisille ohittaa kyberturvallisuustoimet on hyödyntää hyökkäyksessä kyberturvallisuuden heikointa lenkkiä, ihmistä. On helpompaa huijata käyttäjä lataamaan haitallinen tiedosto tai kertomaan salasanansa, kuin murtautua verkkoon kyberturvajärjestelmien läpi. Yleisin havaittu kyberhyökkäyksen muoto onkin phishing-viestit (tietojenkalastelu) [2], joissa käytetään sosiaalisen manipuloinnin keinoja tavoitteena saada kohde vaarantamaan oma tai organisaationsa kyberturvallisuus [3]. Yhdistyneessä kuningaskunnassa tehdyn kyselyn perusteella phishing-viestit on myös yritysten toimintaa häiritsevin hyökkäysmuoto [2].

Yksinkertaisin keino torjua phishing-viestit on olla reagoimatta epäilyttäviin viesteihin ilman varmuutta viestin alkuperästä. Organisaatioiden panostus kyberturvallisuuden ja työntekijöiden tietoisuuteen turvallisista toimenpiteistä parantamiseen on ollut viime vuodet nousussa [2], mutta Proofpointin *State of the Phish (2024)* raportin mukaan 71% käyttäjistä tekevät riskialttiita toimenpiteitä ja heistä 96% tiedostaa tekojensa riskialttiuden. On siis selvää että tietoturvallisuuskoulutus pelkästään ei auta torjumaan phishing-viestejä, vaan tarvitaan keino niiden tunnistamiseen ja estää niiden päätyminen käyttäjille. Phishing-viestien tunnista-

minen on laajalti tutkittu aihe, mutta kohdennettujen kalasteluviestien (jatkossa spearphishing-viestit) tunnistaminen on saanut vähemmän huomiota.

Paremmin väärennetyt ja tietyille kohteille kohdennetut spearphishing-viestit ovat tavallisia phishing-viestejä vaikeammin tunnistettavissa [4]. Niiden tunnistaminen on vaikeaa seuraavista syistä johtuen (1) viestit ovat räätälöityjä tietylle henkilölle tai kohderyhmälle, hyökkääjällä on tietoa kohteen domainista ja pystyy väärentämään viestin sisällön sekä otsaketiedot vaikuttamaan siltä että se tulisi kohteen organisaatiosta tai sidosryhmältä, (2) hyökkäysten viestien määrä on pieni, johtaen siihen että luokittelevien koneoppimismallien kouluttamiseen ei ole riittävää dataa. Näistä haasteista johtuen spearphishing-viestit ovat merkittävä haaste kyberturvallisuuden ylläpitämiselle, ja niiden tunnistamiseen tarvitaan tehokas ratkaisu.

Tässä tutkielmassa kartoitan kuinka koneoppimista hyödynnetään spearphishing-viestien tunnistamisessa, mitä ominaisuuksia viesteistä tarvitaan tunnistamiseen, sekä millaisia apukeinoja käytetään koneoppimismallien tukena. Tunnistaminen on tärkeää, viestien korkean käytön ja mahdollisten vakavien seurasten johdosta. Työntekijöiden välinpitämättömyys tietoturvallisuudesta ja turhien riskien otto [2], tekevät tunnistamisesta ja torjunnasta entistä kriittisempää. Torjuntaan tarvitaan tehokkaita ja laajaan käyttöön soveltuvia ratkaisuja. Oletan soveltuvan ratkaisun löytyvän koneoppimisin hyödyntämisestä viestien luokittelussa.

Tämän työn **tutkimuskysymykset** ovat:

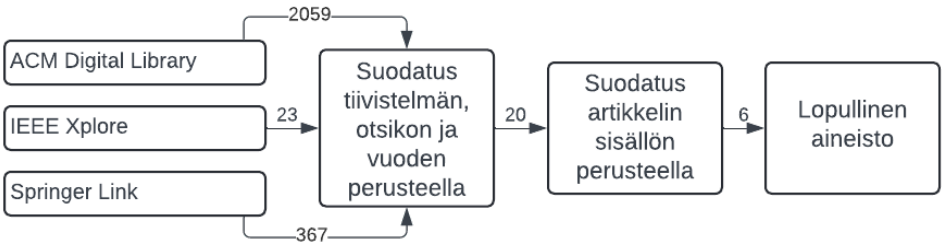
TK1. Mitkä ominaisuudet sähköpostiviestissä soveltuvat tunnistamiseen?

TK2. Miten koneoppimisalgoritmeja hyödynnetään tunnistamiseen?

TK3. Mitä muita ominaisuuksia hyvällä tunnistamisratkaisulla on?

Menetelmät ja Tiedonhaku

Tutkielma toteutettiin kirjallisuuskatsauksena, tietokannat joissa haku toteutettiin ovat IEEE, ACM ja Springer. Nämä valikoituivat tietokannoiksi, sillä alustava haku tuotti näissä tietokannoissa merkittävästi enemmän olennaisia artikkeleita kuin muissa tietokannoissa. Käytetty hakutermi oli *("spear phishing" OR spearphishing OR spear-phishing) AND ("machine learning" OR "artificial intelligence" OR AI) AND "detection"*. Seuloin haussa esiintyvistä artikkeleista otsikon ja abstraktin perusteella olennaiset artikkelit, joita löytyi IEEE:stä neljä (4), ACM:stä kahdeksan (8), ja Springeristä kahdeksan (8). Jo hakuvaiheessa jätin pois ennen vuotta 2015 julkaistut artikkelit rajatakseni katsauksen ajantasasiin artikkeleihin. Suodatusvaiheessa luin artikkelit läpi ja valitsin katsaukseeni koneoppimista hyödyntävää ratkaisua spearphishing-viestien tunnistamiseen esittävät artikkelit. Toisen suodatuksen jälkeen katsaukseen jäi kuusi (6) artikkelia. Tiedonhakuprosessi on esitetty kuvassa 1.1. Artikkelit, niissä esitellyt ratkaisut, ja niiden saavuttamat tulokset on esitetty taulukossa 3.1 sivulla 6.



Kuva 1.1: Tiedonhaku ja artikkeleiden suodatus

Tutkielman loppu on jaettu kolmeen osaan. Luvussa 2. käydään läpi taustatietoa. Luvussa 3. esitellään kirjallisuuskatsauksessa käsitellyt artikkelit ja vertaillaan niitä. Luvussa 4. muodostetaan yhteenveto tutkielman tuloksesta.

2 Taustatietoa

Phishing

Phishing eli tietojenkalastelu on yleisimpiä ja käyttäjille näkyvimpiä kyberhyökkäyksiä, siinä vastaanottaja yritetään saada avaaman haitallinen URL, lataamaan haitallinen tiedosto, tai luovuttamaan yksityistä tietoa. Phishingiä esiintyy useassa eri muodossa, sähköpostiviesteinä, puheluina, nettisivuina, tekstiviesteinä. Sähköpostiviesteillä hyökkääjät pystyvät helposti saavuttamaan laajan joukon käyttäjiä, ja se onkin käytetyin muoto. Phishing-viestit ovat usein alku laajemmalle hyökkäykselle, onnistuneen hyökkäyksen jälkeen hyökkääjä voi myydä saamansa tiedot eteenpäin tai käyttää niitä itse vakavamman hyökkäksen toteuttamiseen [5]. Phishing-viestien on todettu olevan pääasiallinen keino kiristyshaittaohjelmien (ransomware) levittämiseen [6].

Phishing-viesteissä käytetään teknisiä ja sosiaalisen manipuloinnin keinoja kohteeseen vaikuttamiseksi. Teknisiä keinoja ovat viestien otsaketietojen muuttaminen luotettavan tahon imitoimiseksi, kuten lähettäjän osoitteen väärentäminen. Sosiaalisessa manipuloinnissa käytetään kohteesta saatavilla olevaa tietoa sekä useita eri manipuloinnin keinoja, tavoitteena saada kohde toimimaan tietyllä tavalla [7]. Tavallisissa phishing-viesteissä kohteesta usein ei ole paljoa tietoa, ja hyökkäykset tukeutuvat käyttämään sosiaalisen manipuloinnin keinoja. Yleinen käytetty sosiaalisen manipuloinnin keino on kiireellisyys, kohteelle luodaan kiireellisyyden tunne käskemällä toimimaan välittömästi ohjeen mukaan välttääkseen seuraksia.

Spearphishing

Spearphishing, eli kohdennetut kalasteluviestit, ovat phishing-viestien kehittyneempi muoto. Spearphishing-viestit ovat tietyille kohteille yksilöityjä kalasteluviestejä. Yksilöinnin tavoitteena on vähentää kohteen epäilyksiä viestin haitallisuudesta. Yksilöinti toteutetaan kohteesta saadun tiedon avulla, ja usein sisältää tietoa kohteesta sekä kohteen suhteesta imitoitavaan lähettäjään [8]. Hyökkääjä voi imitoida kohteen organisaation toimitusjohtajaa, ja lähettää talousosastolla työskentelevälle kohteelle sähköpostiviestin kehottaen tätä maksamaan kiireellisesti väärennetyn laskun. Tässä tilanteessa hyökkääjä hyödyntää kohteen ja imitoitavan lähettäjän työntekijä - työnantaja suhdetta, tietoa kohteen työtehtävistä, sekä kiireellisyyttä sosiaalisen manipuloinnin keinona.

Tietoturvayritys Barracuda Networksin raportin [9] mukaan tyypilliselle organisaatiolle saapuu päivässä 5 hienostonutta spearphishing-viestiä, ja 50% organisaatioista joutuu hyökkäysten uhriksi yleisimpänä seurauksena haittaohjelmatartunta ja arkaluontoisten tietojen menetys. Spearphishing-viestit ovat siis yleinen kyberhyökkäyksen keino, myös APT:eiden (advanced persisten threat) on todettu käyttävän niitä yhtenä ensisijaisista hyökkäysvektoreista niiden helppouden ja toimivuuden takia [10].

Tulosmittarit

Tutkimuksissa käytetään tulosmittareina kolmea arvoa, saanti (recall), tarkkuus (precision), sekä F1-tulos (F1-score). **Saanti** kertoo sen osuuden, kuinka monta spearphishing-viestiä onnistutaan tunnistamaan kaikista aineiston olevista spearphishing-viesteistä. **Tarkkuus** on oikein tunnistettujen spearphishing-viestien määrä, kaikista spearphishing-viesteiksi tunnistettujen joukosta. **F1-tulos** on saannin ja tarkkuuden harmoninen keskiarvo.

3 Artikkeleiden vertailu

Tässä luvussa esittelen ja vertailen kuutta kirjallisuuskatsaukseen valittua artikkelia. Käsittelen artikkeleiden esittelemät tunnistusratkaisut seuraavien osa-alueiden kautta: ratkaisujen rakenteet, analysoitavat ominaisuudet, käytetyt koneoppimisalgoritmit, sekä artikkeleiden tukimusmenetelmät ja -tulokset. Tarkastelemalla näitä osa-aluita saamme kattavan kuvan siitä mitä eri keinoja ratkaisuissa on käytetty paremman tuloksen saavuttamiseksi ja minkälaisiin tuloksiin on päädytty. Artikkeleiden painopisteet erosivat toisistaan selkeästi, ne sekä artikkeleiden käyttämät koneoppimisalgoritmit, ja niiden saavuttamat tulokset on esitetty taulukossa 3.1.

Taulukko 3.1: Käsitellyt artikkelit

Artikkeli	**Painopiste**	**Koneoppimisalgoritmit**	**Tulos**
Gascon ym.	Lähettäjän jättämät persoonalliset piirteet sähköpostin otsaketiedoissa	KNN, SVM	91% & 92,5% BS, 73% & 79% KD, 48% & 30% KS
Evans ym.	Vahvistettu oppiminen ominaisuuksien valinnassa	KNN	94% BS 83% KD 62% KS
Ding ym.	Mainetietojen hyödyntäminen, koulutusdatan laajennus	Random Forest	Recall 95.56% Precision 98.85% F1-Score 97.16%
Ling ym.	Sisäiset ja ulkoiset mainetiedot	Random Forest	Recall 95.08% Precision 93.55% F1-Score 94.31%
Arya & Chamotra	Hämäykseen perustuva hiekkalaatikointi	Useita eri	Recall 88.2 % Precision 75% F1-Score 81%
Han & Shen	Kampanjoiden tunnistaminen	KNN	Recall 97% FPR 0.02% Kampanjan tunnistus >90%

3.1 Aineiston esittely

Gascon ym. [11] esittävät ratkaisua, joka tunnistaa kohdennetut kalasteluviestit ainoastaan sähköpostin otsaketietojen (metadata) perusteella. Sähköpostin tekstisisältöä ei siis käytetä tunnistamiseen. Menetelmä perustuu havaintoon, jonka mukaan lähettäjät jättävät sähköpostin otsaketietoihin yksilöitäviä piirteitä, joita analysoimalla väärennetyt sähköpostiviestit pystytään tunnistamaan.

Evansin ym. [8] pyrkivät kehittämään Gasconin ym. [11] ratkaisua lisäämällä analysoitavien otsaketietojen valitsemiseen vahvistetun oppimisen (reinforcement learning) algoritmin. Tarkoituksena on vähentää analysoitavien ominaisuuksien valitsemisen vaatimaa manuaalista työtä, ja parantaa tulosta valikoimalla sopivimmat ominaisuusosajoukot.

Dingin ym. esittämä ratkaisu [12] käyttää tunnistamisessa hyödyksi kolmannen osapuolen tarjoamia mainetietoja URL:eista, sekä välitysominaisuustietoja (forwarding features). Mainetietojen hyödyntäminen perustuu Dingin ym. havaintoon että spearphishing-viesteistä suurimassa osassa on sisällytettynä URL:eja. Epätasapainoisen datan vaikutuksen vähentämiseksi, Ding ym. kehittivät SMOTE metodiin pohjautuvaa KM-SMOTE metodia datan laajentamiseksi.

Lingin ym. [13] ratkaisu pohjautuu selkeästi Dingin ym. [12] tekemään työhön, Ding onkin mukana yhtenä tutkijoista. Dingiin ym. verrattuna kolmannen osapuolen tarjoamia mainetietoja hyödynnetään laajemmin, sekä analysoitavaksi tuodaan myös sisäisiä maineominaisuuksia. Myös Ling ym. käyttivät KM-SMOTE metodia datan laajentamiseksi.

Arya ja Chamotra [14] esittävät kahdesta analysoivasta komponentista koostuvan ratkaisun. Ensimmäinen komponentti analysoi sähköposteja käyttäen monikerroksista tunnistusmallia. Malli koostuu useasta koneoppimismallista, jotka analysoivat sähköpostista eri ominaisuuksia. Toinen komponentti käyttää hiekkalaatikointia

(sandboxing) sähköpostien liitteiden ja URLien käytöksen analysointiin eristetyssä ympäristössä.

Han ja Shen [4] keskittyivät spearphishing-hyökkäyskampanjoiden tunnistamiseen. Spearphishing-viestien luokittelun jälkeen Hanin ja Shenin ratkaisu pyrkii tunnistamaan hyökkäävän tahon käyttämiä menetelmiä. Luokitteluun sekä kampanjoiden tunnistamiseen Han ja Shen käyttivät attribuuttigraafin etenemiseen perustuvaa puolivalvottua oppimismenetelmää käyttäen.

Yleisellä tasolla ratkaisut koostuvat kolmesta komponentista: ominaisuustietojen kerääminen sähköpostista, datan käsittely, ja luokittelu koneoppimisalgoritmilla. Eroavaisuuksiin vaikuttaa oleellisesti se mitä ominaisuuksia sähköposteista valitaan analysoitavaksi, sekä kuinka niitä käsitellään ennen luokittelua.

3.2 Analysoitavat ominaisuudet

Taulukko 3.2 esittää mitä ominaisuuksia ratkaisut analysoivat sähköposteista luokittelua varten. Lähestymisessä ominaisuuksien valintaan on selkeitä eroja, joista merkittävimpänä Gascon ym. [11], sekä Evans ym. [8] käyttävät sähköposteista vain otsakedataa (header data). Ding ym. [12] ja Ling ym. [13] hyödynsivät otsakedatan ja tekstisisällön lisäksi välitystietoja sekä kolmannen osapuolen tarjoamia maineominaisuuksia,joiden lisäksi Ling ym. käyttävät sisäisiä maineominaisuuksia. Aryan ja Chamotra [14] analysoivat otsakedatan lisäksi viestin runkosta välittyviä tunnetiloja, sekä URLien ja liitteiden uhkia hiekkalaatikoinnin avulla. Han ja Shen [4] vahvistivat otsakedatan analyysia tarkastelemalla viestin aiheessa ja tekstisisällössä esiintyviä aihealueita sekä tekstin luettavuutta.

Taulukko 3.2: Artikkeleissa käytetyt ominaisuudet

	Gascon	Evans	Ding	Ling	Arya	Han
Otsake						
Käytösominaisuudet	x	x	x	x	x	x
Rakenneominaisuudet	x	x	x	x	x	x
Kuljetusominaisuudet	x	x			x	x
Runko						
Tekstisisältö			x	x	x	x
URL			x	x	x	
Liitetiedostot					x	
Muut						
Välitysominaisuudet			x	x		
Sisäiset maine-ominaisuudet				x		
Kolmannen osapuolen maineominaisuudet			x	x		

3.2.1 Otsakedata

Kaikissa ratkaisuissa käytettiin otsakedatasta saatavia ominaisuuksia. Merkittävimmässä roolissa otsakeominaisuudet olivat Gasconilla ym. [11] ja Evansilla ym. [8], sillä he eivät käyttäneet muita ominaisuuksia. He osoittivat hyvin että pelkän otsakedatan avulla voidaan tunnistaa spearphishing-viestejä onnistuneesti. Ding ym. [12], Ling ym. [13], sekä Han ja Shen [4], käyttivät otsakeominaisuuksia yhdessä muista lähteistä saatavien ominaisuuksien kanssa, mistä johtuen he käyttivät huomattavasti vähemmän otsakeominaisuuksia Gasconiin ym. ja Evansiin ym. verrattuna. Arya ja Chamotra [14] eivät ilmoita tarkkaan mitä ominaisuuksia käyttivät, mutta mainitsevat ottaneensa huomioon Gasconin ym. tunnistamat ja käyttämät otsakeominaisuudet. Gasconin ym. tunnistamia otsakeominaisuuksia hyödynsivät myös Evans ym. ratkaisunsa lähtökohtana.

Gasconin ym. esittämät otsakeominaisuudet

Gascon ym. [11] tunnistivat sähköpostiviestin otsakedatasta 46 ominaisuutta, joita käyttäen heidän ratkaisunsa luokittelee sähköpostiviestin spearphishing-viestiksi tai

harmittomaksi. Ominaisuudet jakautuvat kolmeen kategoriaan: käytösominaisuudet (13 kpl), rakenneominaisuudet (22 kpl), sekä kuljetusominaisuudet (11 kpl).

Käytösominaisuuksien kautta tarkastellaan käyttäjille ominaisia tapoja, kuten liitteiden määrä ja tyyppi, epätyypillisten otsakkeiden esiintyminen, ja tyhjien otsakkeiden määrä. Käytösominaisuuksien käyttö perustuu [15] ja [16] havaintoon että sähköpostin lähettäjä jättää sähköpostin rakenteeseen henkilökohtaisia jälkiä, joiden avulla lähettäjät voidaan profiloida.

Rakenneominaisuudet ovat peräisin lähettäjän käyttämästä sähköpostiohjelmasta. Sähköpostien on alunperin täytynyt olla ASCII-merkkeinä, joten nykyään on olemassa lukuisia enkoodaus-skeemoja sähköpostin sisällön muuttamiseksi ASCII yhteensopivaksi. Skeemat ovat hieman toisistaan poikkeavia ja jättävät rakenteeseen ominaisuuksia, joita Gascon ym. [11] käyttävät profilontiin. Esimerkiksi Base64-koodaus ei vaadi kiinteää tekstin pituutta, mikä johtaa eroihin tekstilohkojen muotoilussa. Lisäksi moniosaisten MIME-viestien (Multipurpose Internet Mail Extensions) rakenteissa on pieniä poikkeamia, jotka kertovat sähköpostiohjelmasta ja sen konfiguraatiosta.

Kolmas ryhmä ominaisuuksia muodostuu **kuljetusominaisuuksista.** Kun sähköposti liikkuu verkossa lähettäjältä vastaanottajalle, se kulkee usein useiden välityspalvelimien kautta. Jokainen näistä hypyistä lisää uusia otsakkeita sähköpostin rakenteeseen. Nämä otsakkeet sisältävät tietoa sähköpostin kuljetusominaisuuksista kuten, sähköpostipalvelimen toimitusprotokollista, TLS:stä, ja aikavyöhykkeestä. Hyökkääjä ei pysty väärentämään kuljetuksen aikana syntyviä otsakkeita, joten ne soveltuvat hyvin lähettäjän tunnistamiseen. [11]

Evansin ym. [8] eivät käytä analysointiin ennalta määrättyjä ominaisuuksia. Koulutusvaiheessa he käyttävät vahvistetun oppimisen menetelmää luodakseen ominaisuusdatasetin, joka saavuttaa tarkimman tuloksen luokittelussa. Ratkaisu kerää sähköposteista kaiken otsakedatan raa'oiksi ominaisuuksiksi, jotka se arvioi vah-

vistetun oppimisen menetelmällä. Mitä tarkemmin ominaisuus pystyy yksilöimään lähettäjän, sitä hyödyllisempi se on. Arvioiduista ominaisuuksista valitaan hyödyllisimmät ja muodostetaan lopullinen ominaisuusdatasetti, joita analysoimalla sähköpostiviestit luokitellaan spearphishing-viesteiksi tai harmittomiksi. Tämän katsauksen mukaan kyseinen ratkaisu on ainoa, joka valitsee analysoitavat ominaisuudet dynaamisesti. Tämä ominaisuus tekee siitä tehokkaan työkalun myös uusien hyökkäysmuotojen tunnistamiseen.

3.2.2 Sähköpostiviestin runko

Sähköpostiviestin rungosta ominaisuuksia käyttivät Ding ym. [12], Ling ym.[13], Arya ja Chamotra [14], sekä Han ja Shen [4]. Ominaisuuksien lähteitä rungossa on viestin aihe, tekstisisältö, URLit, ja liitetiedostojen sisältö.

Tekstisisältö

Tekstisisällöstä Ding ym. [12], Ling ym. [13], sekä Han ja Shen [4], keräsivät ominaisuuksia, jotka kertovat sähköpostin tekstuaalisista ja rakenteellisista piirteistä. Ding ym. ja Ling ym. keskittyivät Hania ja Sheniä enemmän rakenteellisten piirteiden kuten, sanojen, URLien, ja domainien määrään analysointiin. Lisäksi Ding ym. ja Ling ym. perehtyivät ominaisuuksiin, jotka kertovat uniikkien sanojen määrästä, funktionaalisten sanojen määrästä, tekstin rikkaudesta, tiettyjen sanojen esiintymisestä (”verification”, ”attachment”), sekä puhelinnumeron esiintymisestä.

Hanin ja Shenin [4] käyttämät ominaisuudet kertovat tekstin rakenteesta, niiden lisäksi he käyttivät luonnollisen kielen käsittelyyn tarkoitettua metodia LSA:ta (Latent semantic analysis) analysoimaan tekstissä esiintyviä aiheita. Analyysin tuloksena he nostavat kymmenen useimmin esiintyvää aihetta ja määrittelivät ne ominaisuuksiksi. Tämä lähestymistapa tarjoaa kattavamman kuvan sähköpostiviesteissä esiintyvistä aiheista, verrattuna tiettyjen sanojen etsimiseen.

Han ja Shen [4] tutkivat Dingiä ym. [12] ja Lingiä ym. [13] tarkemmin myös tekstin luettavuutta. He määrittelivät kahdeksan ominaisuutta kertomaan tekstin luettavuudesta. Ominaisuuksiin kuuluu funkitonaalisten sanojen määrä, kompleksisten ja yksinkertaisten sanojen määrä, sekä sanojen keskimääräinen pituus. Tekstille he myös laskivat Fog, SMOG, ja Flesch-Kincaid-indeksit, mitkä kaikki kertovat tekstin luettavuudesta.

Arya ja Chamotra [14] analysoivat tekstisisällössä olevia aihealueita, tavoitteena tarkastella esiintyykö tekstissä spearphishing-viesteille tunnusomaisia aiheita. Aihealueiksi Arya ja Chamotra valitsivat seuraavat: kiireellisyys, pyyntö, maksu, lasku, lainvalvonta, ilo, suru, skannattu dokumentti, ja epäonnistunut sähköpostiviestin lähetys. Tunnistamiseen Arya ja Chamotra käyttivät LDA:han (Latent Dirichlet Allocation) perustuvaa todennäköisyysmallia.

URL

Ding ym. [12], Ling ym. [13], sekä Arya ja Chamotra [14] olivat ainoat jotka analysoivat viestissä olevia URLeja. Ding ym. ja Ling ym. käyttivät niitä maineominaisuuksiensa lähteinä, sekä huomioivat ne tekstin rakenteellisten piirteiden analysoinnissa. Arya ja Chamotra analysoivat sähköpostissa olevia URLeja käyttäen koneoppinutta logistista regressio-algoritmia. URLista muodostetaan ominaisuusvektori, joka sisältää kahdenlaisia ominaisuuksia: host-pohjaiset ominaisuudet, ja leksikaaliset ominaisuudet. Host-pohjaiset ominaisuudet kertovat nettisivun maineesta ja domainin iästä. Leksikaaliset ominaisuudet kertovat URLin tekstuaalisista piirteistä kuten, erikoismerkkien esiintymisestä, host-nimen pituudesta, URLin pituudesta ja enkoodauksesta. Algoritmi laskee ominaisuuksille pisteet, ja yhteispisteiden ylittäessä asetetun rajan todetaan URL haitalliseksi.

Liitetiedostot

Kaikki ratkaisut huomioivat liitetiedostojen määrän ja tyypin ominaisuuksissaan. Ainoastaan Arya ja Chamotra [14] analysoivat liitetiedostojen sisältöä. Heidän ratkaisussaan on komponentti, joka tunnistaa haitalliset liitetiedostot staattisen analyysin avulla. Alustavassa sääntöpohjaisessa suodatuksessa haitalliseksi tunnistettavat liitetyypit, kuten ohjelmatiedostot (executable files), suodatetaan pois tarkemmasta analyysista. PDF-tiedostot analysoidaan yksiluokkaista tukivektorikone (One-class Support Vector Machine) koneoppimisalgoritmia käyttäen, mikä luokittelee tiedostot haitallisiksi tiedostojen sisältämälle JavaScript-koodille tehdyn leksikaalisen analyysin perusteella.

Välitysominaisuudet

Dingin ym. [12] ratkaisu käyttää analysoinnissa välitysominaisuuksia vedoten havaintoon että spearphishing-viestejä ei yleensä välitetä eteenpäin, ja että vastaanottajilla on usein jokin yhteys. Ding ym. tekevät olettamuksen että spearphishing-viestin välitysmäärä on pieni ja kaikilla vastaanottajilla on yhteinen tekijä. Ding ym. käyttävät kolmea välitysominaisuutta: välitysten määrä, lähettäjän ja vastaanottajan välinen yhteys, sekä vastaanottajien välinen yhteys. Yhteyden he määritelevät sähköpostipalvelimen nimen perusteella. Ling ym. [13] käyttävät ratkaisussaan samoja välitysominaisuuksia, ja lisäävät joukkoon vastaanottajan esiintyvyyden lähettäjän aiemmissa sähköposteissa. Ominaisuuden käyttö perustuu olettamukseen että viesti on epäilyttävämpi jos kyseinen lähettäjä ei ole aiemmin lähettänyt sähköposteja vastaanottajalle.

Maineominaisuudet

Dingin ym. [12] tutkimuksessa käytetyssä datasetissä 90% spearphishing-viesteistä sisältävät URL:in, tähän havaintoon nojaten Ding ym. käyttivät ulkoisia mainepal-

veluja saadakseen tietoa sähköposteissa olevien URLien vaarallisuudesta. He käyttivät *'VirusTotalin'* ja *'PhishTankin'* tarjoamia rajapintoja URLien, IP-osoitteen, sekä domainin analysointiin, jotka he yhdistivät analysoitaviksi ominaisuuksiksi. Ding ym. olivat tiettävästi ensimmäisiä ulkopuolisten maineominaisuuksien hyödyntäjiä.

Ling ym. [13] lisäävät Dingin ym. [12] käyttämien ulkoisten maineominaisuuksien joukkoon domainin suosion ja rekisteröintipäivän. Nämä ominaisuudet Ling ym. lisäävät perustaen havaintoihin että phishing-sivut eivät saa paljoa liikennettä eivätkä ole kovin vanhoja. Ulkoisten maineominaisuuksien lisäksi Ling ym. [13] sisällyttivät ratkaisuunsa **sisäisiä maineominaisuuksia**, jotka kertovat kuinka usein täysin määritelty verkkotunnus (fully qualified domain name), lähettäjän nimi ja sähköpostiosoite, ovat esiintyneet aiemmissa sähköposteissa. Oletus on että mitä suurempi esiintyvyys sen luotettavampia ne ovat.

3.3 Jatkokäsittely

Ratkaisuista Ding ym. [12], Ling ym. [13], sekä Arya ja Chamotra [14] käsittelivät dataansa vielä ennen koneoppimisalgoritmin suorittamaa luokittelua. Hanin ja Shenin [4] ratkaisussa puolestaan on komponentti hyökkäyskampanjoiden tunnistamista varten, joka suoritetaan spearphishing-viesteiksi luokitelluille viesteille. Ding ym., Ling ym., sekä Arya ja Chamotra käsittelevät dataa tavoitteena saada parempi tulos luokittelusta, Hanin ja Shenin kampanjoiden tunnistus puolestaan pyrkii helpottamaan tulevien viestien luokittelua yhdistämällä piirteitä hyökkäyskampanjoihin.

Datan laajennus

Koska koneoppimisalgoritmin koulutusta varten olemassa olevien spearphishing-viestien määrä on pieni, Ding ym. [12] esittävät SMOTE metodiin pohjautuvaa KM-SMOTEa datan laajentamiseen. SMOTE (Synthetic Minority Oversampling Tech-

nique) luo synteettisiä vähemmistöluokan näytteitä interpoloimalla ominaisuusarvoja valitun näytteen ja sen lähimpien naapurien välillä [17]. SMOTEn tarkoituksena on parantaa luokittelijan suorituskykyä vähemmistöluokan ennustamisessa. Dingin ym. esittämä KM-SMOTE parantaaa SMOTE algoritmiä hyödyntämällä siinä K-keskiarvo klusterointia. KM-SMOTEssa luodaan ainoastaan klustereiden sisälle synteettisiä näytteitä.

Ding ym. [12] ja Ling ym. [13] todistivat datan laajennuksella olevan positiivinen vaikutus luokittelun tulokseen (Taulukko 3.7). Molemmissa artikkeleissa esitettiin tulokset eri ominaisuusryhmien vaikutukselle tulokseen, sekä KM-SMOTEn ja kaikkien ominaisuusryhmien käytön vaikutus tulokseen. Puuttumaan kuitenkin jää kaikkien ominaisuusryhmien käyttö ilman KM-SMOTEa, joten emme tiedä tarkkaan kuinka merkittävä vaikutus KM-SMOTEn käytöllä on.

Hämäys-hiekkalaatikointi, käytösanalyysi

Arya ja Chamotra [14] esittävät ratkaisussaan käytösanalyysikomponenttia, joka käyttää hämäys-hiekkalaatikointia (deception-sandboxing) URLien ja liitteiden käytöksen analysointiin sekä väärien positiivisten havaintojen karsimiseksi. Hiekkalaatikoinnissa luodaan eristetty ympäristö, joka ei ole yhteydessä muihin verkkoihin. Hämäystekniikat auttavat tunnistamaan hyökkääjien yrityksiä tunnusten keräämiseen, käyttöoikeuksien laajentamiseen, ja sivuttaiseen liikkeeseen. Hiekkalaatikon sisälle on asetettu useita hämäyselementtejä houkuttimiksi hyökkääjille, sekä tiedonkeruuvälineitä datan saamiseksi. URLit ja liitteet avataan hiekkalaatikon sisällä ja tiedonkeruuvälineiden avulla seurataan johtaako URLien tai liitteiden avaaminen hyökkäysyrityksiin. Hämäys-hiekkalaatikointi vaatii huomattavan määrän resursseja, joten se suoritetaan vain alustavien analyysien perusteella epäilyttävimmille viesteille.

Hyökkäyskampanjoiden tunnistus

Han ja Shen [4] sisällyttivät ratkaisuunsa jatkokäsittelykomponentin luokittelun jälkeen. Komponentin tehtävänä on tunnistaa todetuista spearphishing-viesteistä kaavoja hyökkäyskampanjoiden tunnistamiseksi, ja siten parantaa kykyä tunnistaa samaan kampanjaan kuuluvia viestejä. Kampanjoiden tunnistus tekee ratkaisusta vankemman, kykenee tunnistamaan spearphishing-viestit tehokkaammin hyökkäyskeinojen muuttuessa.

3.4 Käytetyt koneoppimisalgoritmit

Pääasiallinen käyttökohde koneoppimisalgoritmeille ratkaisuissa on sähköpostiviestien luokittelu haitallisiksi tai hyvänlaatuisiksi viesteiksi. Taulukosta 3.3 huomataan useassa ratkaisussa viestien luokittelun olevan ainoa koneoppimista hyödynnetävä osio. Poikkeuksena tähän Evans ym. [8] käyttävät KNN-algoritmia myös vahvistetun oppimisen agentin toteutuksessa, sekä Arya ja Chamotra jotka käyttivät useita koneoppimisalgoritmeja eri komponenteissa.

Taulukko 3.3: Käytetyt koneoppimisalgoritmit

Kohde	**Koneoppimisalgoritmi**	**Artikkeli**
Viestien luokittelu	KNN	Gascon ym. Evans ym. Han ja Shen
	Tukivektorikone	Gascon ym.
	Random Forest	Ding ym. Ling ym.
Ominaisuuksien arviointi	KNN	Evans ym.
Impersonaatioanalyysi	Lineaariaikainen algoritmi, Gonzales ym.	Arya ja Chamotra
Aiheiden tunnistus	LDA:iin perustuva todennäköisyysmalli	Arya ja Chamotra
URL -analyysi	Logistinen regressio	Arya ja Chamotra
Liiteanalyysi	Tukivektorikone	Arya ja Chamotra
Analyysien yhdistys	Directed Anomaly Scoring	Arya ja Chamotra

Viestien luokittelu

Viestien luokittelussa ratkaisut käyttävät koneoppimisalgoritmeja viesteistä otettujen ominaisuuksien arviointiin, jonka perusteella viestit luokittellaan haitallisiksi tai harmittomiksi viesteiksi. Luokitteluissa käytetyin algoritmi on KNN (k-nearest neighbors). Se soveltuvus luokitteluun hyvin, sillä se kykenee tuottamaan tarkan tuloksen myös vähäisellä koulutusdatalla [11]. Gascon ym. käyttivät KNN:n rinnalla myös **tukivektorikonetta (Support Vector Machine)**, sillä sen on todettu suoriutuvan hyvin korkea-ulotteisissa vektoriavaruuksissa [11]. Han ja Shen [4] käyttivät ratkaisunsa analyyttisissä komponenteissä KNN:ää käyttävää attribuuttigraafin etenemiseen perustuvaa puolivalvottua (semi-supervised) oppimismenetelmää. Ding ym. ja Ling ym. käyttivät puolestaan random forest algoritmiä, johon he päätyivät vertailtuaan eri algoritmien saavuttamia tuloksia.

Ominaisuuksien valinta

Evansin ym. [8] ratkaisun ytimenä toimiva vahvistetun oppimisen agentti muodostaa viestien ominaisuuksia arvioimalla ominaisuusosajoukon, joka saavuttaa tarkimman mahdollisen tuloksen. Koulutusvaiheessa agentti muodostaa useita ominaisuusosajoukkoja vaihtaen niissä esiintyviä ominaisuuksia, näin agentti saa lajiteltua ne järjestykseen niiden vaikutuksen mukaan. Vaikutus arvion tarkkuuteen lasketaan KNN:n avulla. Agentin koulutusdatassa on vain harmittomia sähköpostiviestejä, jotka on merkitty lähettäjän sähköpostiosoitteella. Koulutusdataa arvioimalla agentti oppii mitkä arvot esittävät tiettyä lähettäjää. Opittuaan lähettäjien profiilit agentti pystyy tunnistamaan uusista sähköposteista lähettäjän, ja mikäli lähettäjän sähköpostiosoite on eri kuin arvion tuottama osoite, sähköposti todetaan haitalliseksi.

Aryan ja Chamotran koneoppimisalgoritmit

Aryan ja Chamotran [14] ratkaisu oli muihin verrattuna selkeästi laajempi. He käyttivät sen komponenteissa useita koneoppimisalgoritmeja. Käytetyt koneoppimisalgoritmit, directed anomaly scoringia (DAS) lukuunottamatta, analysoivat eri ominaisuuksia sähköpostiviesteistä ja pisteyttävät ominaisuudet niiden epäilyttävyyden mukaan. Analyysien jälkeen DAS laskee sähköposteille kumulatiiviset pisteet, ja järjestää ne niiden epäilyttävyyden mukaan.

3.5 Artikkelien tutkimusmenetelmät ja -tulokset

Ratkaisujen toimivuuden tutkimisessa oli lieviä poikkeamia. Gascon ym. [11] ja Evans ym. [8] tutkivat toimivuutta kolmessa eri hyökkäysmallissa: sokea huijaus (blind spoofing), tunnettu ympäristö (known domain), ja tunnettu lähettäjä (known sender). Ding ym. [12] ja Ling ym. [13] keskittyivät tutkimaan eri ominaisuuksien sekä datan laajennusmenetelmän vaikutusta luokittelun tarkkuteen. Arya ja Chamotra [14] tutkivat toimivuutta reaalitilanteessa. Han ja Shen [4] suorittivat kolme koetta: spearphishing-viestien tunnistus, tuntemattoman kampanjan tunnistus, ja tunnetun kampanjan tunnistus. Han ja Shen huomioivat myös koulutuksessa käytettävän merkityn datan määrän vaikutuksen tarkkuuteen. Myös tutkimuksissa käytettyjen sähköpostien määrissä on merkittäviä eroja. Määrällä on merkittävä vaikutus ratkaisun toimintaan, sillä huonolla ja vähäisellä koulutusdatalla ei voi odottaa hyvää tulosta. Tutkimuksissa käytettyjen sähköpostien määrät on esitetty taulukossa 3.4.

Gascon ym. ja Evans ym.

Gascon ym. [11] ja Evans ym. [8] tutkivat ratkaisujensa toimivuutta kolmessa eri hyökkäysmallissa: sokea huijaus (blind spoofing), tunnettu ympäristö (known domain), ja tunnettu lähettäjä (known sender). **Sokeassa huijauksessa** hyökkääjällä

Taulukko 3.4: Koulutuksessa ja testauksessa käytettyjen viestien määrät

Artikkeli	**Spearphishing-viestit**	**Harmittomat viestit**	**Lähettäjiä**
Gascon ym.	380 301	380 302	17 381
Evans ym.	1201	7518	987
Ding ym.	417	13 916	-
Ling ym.	202	39 836	-
Arya ja Chamotra	-	68 000	-
Han ja Shen	1467	14 043	-

ei ole tietoa lähettäjästä jota yrittää jäljitellä. **Tunnetussa ympäristössä** hyökkääjä on saanut käsiinsä sähköposteja toiselta henkilöltä joka on samassa ympäristössä kuin jäljiteltävä kohde ja pystyy näiden perusteella väärentämään sähköpostin kuljetusominaisuuksia tarkasti. **Tunnetun lähettäjän** tilanteessa, hyökkääjällä on jäljitettävän kohteen sähköposteja joiden perusteella hän pystyy väärentämään useita otsaketietoja kalasteluviestiinsä.

Gascon ym. [11] käyttivät selkeästi eniten sähköposteja. Koska he analysoivat vain sähköpostien otsakedataa, Gascon ym. muunsivat alkuperäisistä 760 603 sähköpostista 50%:n otsakedataa satunnaisesti jäljitelläkseen spearphishing-viestejä. Gasconin ym. käyttämät sähköpostit tulivat 17 381 lähettäjältä, siten että kaikki ovat lähettäneet vähintään kaksi sähköpostia. Sähköpostit ovat peräisin yrityksiltä sekä kaupallisilta sähköpostipalvemilta.

Evans ym. [8] keräsivät harmittomia sähköposteja viidestä julkisesti saatavilla olevasta lähteestä, yhteensä 987:lta eri lähettäjältä. Myös Evans ym. loivat itse testeissä käyttämänsä spearphishing-viestit imitoiden harmittomien viestien lähettäjiä. Luotujen sähköpostien lähettäjän osoitteen sekä verkkotunnuksen- ja lähettäjäkohtaiset piirteet he ottivat keräämistään harmittomista sähköposteista.

Molempien tutkimuksissa koulutus- ja testausdatajoukot koostuivat eriasteisesti hyökkäysmalleja mallintaen. Näin he pystyivät jäljentämään eri hyökkäysmallien esiintymistä ja mittaamaan ratkaisujen tehokkuutta eri hyökkäysmalleja vastaan. Taulukosta 3.5 nähdään Gasconin ym. [11] saavuttavan lähes vastaavat tulokset

kuin Evans ym. [8], Gasconin ym. säilyttäessä samalla matalan väärien positiivisten määrän.

Taulukko 3.5: Gascon ym. [11] ja Evans ym. [8] tulokset hyökkäysmalleittain

Ratkaisu	Sokea huijaus	Tunnettu ympäristö	Tunnettu lähettäjä
Gascon ym. KNN VP 0.01%	90.9%	72.7%	48.1%
Gascon ym. SVM VP 0.01%	92.4%	78.1%	30.1%
Gascon ym. KNN VP 10%	91.9%	78.4%	53.2%
Gascon ym. SVM VP 10%	92.9%	84.1%	33.9%
Evans ym.	90.6% VP 9.4%	92.0 % VP 8.0%	71.2% VP 28.8%

Tutkiakseen oman ratkaisunsa tarkkuutta paremmin, Evans ym. [8] vertailivat ratkaisujen eroja käyttäen omaa koulutus- ja testausdataansa. Vertailun tuloksena Evans ym. toteavat ratkaisujen välillä olevan vain pieniä eroja, merkittävimpänä havaintona Evansin ym. saanti (recall) oli Gasconia ym. [11] huomattavasti parempi (65.9% vs 17.7%) tunnetun lähettäjän tapauksessa.

Evans ym. [8] osoittivat että automaattisesti uuden analysoitavan ominaisuusjoukon generoimalla ratkaisun kohdatessa uuden hyökkäysmallin, se pystyy havaitsemaan sille tuntemattomia hyökkäysmalleja yhtä tarkasti kuin tiettyä hyökkäysmallia vastaan koulutettu versio. Ominaisuusjoukkojen generointi on siis toimiva ratkaisu ympäristössä, jossa analysoitavat hyökkäysmallit vaihtelevat. Haasteena Evansin ym. [8] ratkaisussa on ominaisuusjoukon valinta vahvistetun oppimisen avulla, heidän mittauksen mukaan vaihe kestää 1489.81 sekuntia. Evans ym. toteavatkin että reaalitilanteessa vahvistetun oppimisen vaihetta ei suoritettaisi usein.

Han ja Shen

Han ja Shen [4] toteuttivat ratkaisun tehokkuuden tutkimista varten kolme erillistä koetta, spearphishing-viestien tunnistaminen, tuntemattomien kampanjoiden tunnistaminen, sekä tunnettujen kampanjoiden tunnistaminen. Koulutuksessa ja testauksessa he käyttivät 1467 spearphishing-viestiä kahdeksasta hyökkäyskampanjas-

ta, sekä 14043 harmitonta viestiä. Viestit ovat peräisin Symanteciltä, joka on ne kerännyt asiakkailtaan.

Spearphishing-viestien tunnistamisessa Han ja Shen [4] vertasivat oman puolivalvotun menetelmänsä tuloksia, random forest algoritmilla saataviin tuloksiin. Lisäksi he suorittivat kokeen eri määrillä merkittyä koulutusdataa. Tuloksena he todistavat että heidän malli pystyy tehokkaasti tunnistamaan spearphishing-viestit vähäisellä merkityllä ja epätasaisella koulutusdatalla kuten käy ilmi taulukosta 3.6. Heidän menetelmänsä saavutti myös paremman tuloksen random forestiin verrattuna.

Taulukko 3.6: Han ja Shen [4], spearphishing-viestien tunnistus

Merkityn datan osuus	**Saannin KA (RF)**	**VPO KA (RF)**	**Saannin KA, Han ja Shen**	**VPO KA, Han ja Shen**
1%	0.7987	3.6247e-05	0.8700	4.3950e-05
2%	0.8000	4.3632e-05	0.8910	4.4910e-05
3%	0.8323	4.7608e-05	0.9106	5.1077e-05
4%	0.9251	6.3137e-05	0.9460	7.0652e-05
5%	0.9399	9.0109e-05	0.9607	9.0109e-05
6%	0.948	8.7266e-05	0.9707	8.7266e-05

Tuntemattomien kampanjoiden tunnistamisessa Han ja Shen [4] osoittavat että heidän ratkaisunsa pystyy hyvällä tarkkuudella tunnistamaan tuntemattomat kampanjat, kunhan koulutusdatassa on riittävästi tunnetuiksi kampanjoiksi merkittyjä sähköposteja. **Tunnettujen kampanjoiden tunnistamisessa** he osoittavat pystyvänsä hyvällä tarkkuudella luokittelemaan viestit kampanjoihin pienellä määrällä merkittyä dataa.

Ding ym. ja Ling ym.

Dingin ym. [12] ja Lingin ym. [13] ratkaisut olivat ainoat, jotka käyttivät välitys- ja maineominaisuuksia, sekä datan laajennusmenetelmää. Dingin ym. ja Lingin ym. tutkimukset painottuivat näiden ominaisuuksien vaikutukseen luokittelussa. Välitys-

ja maineominaisuuksilla oli positiivinen vaikutus, ja Dingin ym. kehittelemä, sekä Lingin ym. käyttämä KM-SMOTE datan laajennusmenetelmä vaikutti tulokseen merkittävästi, kuten taulukosta 3.7 käy ilmi.

Taulukko 3.7: Ding ym. [12] ja Ling ym, [13] ominaisuuksien vaikutus tulokseen

Ominaisuudet	**Saanti (%)**	**Precision(%)**	**F1-score(%)**
Ding Perus	84.78	81.25	82.98
Perus + VO	86.11	88.57	87.32
Perus + MO	84.21	91.43	87.67
Kaikki + KM-SMOTE	**95.56**	**98.85**	**97.16**
Ling Perus	83.61	82.26	82.93
Perus + VO	86.89	88.33	87.60
Perus + MO	83.61	91.07	87.18
Kaikki + KM-SMOTE	**95.08**	**93.55**	**94.31**

Ding ym. [12] ja Ling ym. [13] vertasivat myös eri koneoppimisalgoritmien saavuttamia tuloksia. Molemmat käyttivät seuraavia algoritmeja: random forest, decision tree, logistinen regressio (logistic regression), ja tukivektorikone (support vector machine). Molempien tulokset osoittivat random forestin suoriutuvan algoritmeista parhaiten.

Ding ym. [12] käyttivät tutkimuksessaan 417 ammatilaisten vahvistamaa spearphishing-viestiä, sekä 13916 sähköpostia joista 5562 on harmittomia, ja 8404 on tavallisia phishing-viestejä. Viestien alkuperästä on kerrottu niiden olevan peräisin tutkimusryhmän kanssa yhteistyötä tekeviltä yrityksiltä. Lingin ym. [13] tutkimusdata koostuu 202:sta spearphishing-viestistä, jotka he ovat itse keränneet valtion toimivaltaisten viranomaisten hyväksynnällä, 13916 satunnaisesti valittua sähköpostia joista osa on harmittomia ja osa tavallisia phishing-viestejä, sekä 1113 phishing-viestiä IWSPA kilpailussa käytetystä datasta.

Arya ja Chamotra

Arya ja Chamotra [14] tutkivat ratkaisunsa toimivuutta reaalitilanteessa Centre for Development of Advanced computing (CDAC) verkossa. Puolen vuoden aikana se

analysoi 68000 sähköpostia. Arya ja Chamotra ottivat 472 sähköpostin osajoukon kolmen päivän ajalta tulleista sähköposteista, tutkivat ne manuaalisesti ja vertasivat ratkaisun luokitteluihin. Ratkaisu luokitteli sähköposteista 20 spearphishing-viesteiksi, joista viisi (5) oli vääriä positiivisia. Lopuista 452 sähköpostista se luokitteli kaksi (2) väärin negatiivisiksi. Näiden laskelmien perusteella tulosindikaattorit olivat: saanti 0.882, tarkkuus 0.75, ja F1-tulos 0.8105.

3.6 Vertailu

Ratkaisut ovat selkeästi toisistaan eroavia toteutukseltaan, sekä koulutus- ja testausdataltaan, joten tulosten suora vertailu ei kerro parasta ratkaisua. Taulukosta 3.8 huomataan että Ding ym. [12] saavuttivat parhaan tuloksen tarkkuudessa ja F1-pisteissä, mutta paras saanti oli Gasconilla ym. [11]. Artikkeleissa ei kerrottu ratkaisujen vaatimien resurssien määrää tai suoritusnopeuksia, joten niiden vertailu reaalitilanteissa ei ole mahdollista. Ominaisuuksien kannalta Evansin ym. [8] sekä Hanin ja Shenin [4] ratkaisut pystyvät reagoimaan muuttuviin hyökkäysmalleihin, mikä tekee niistä toimivampia reaalitilanteissa.

Taulukko 3.8: Tulokset ja arvioitavien ominaisuuksien määrät

Artikkeli	**Saanti (%)**	**Tarkkuus(%)**	**F1-score(%)**	**Arvioitavia ominaisuuksia**
Gascon ym.	99.3	87.9	-	46
Evans ym.	98.1	86.2	-	-
Ding ym.	95.56	98.85	97.16	27
Ling ym.	95.08	93.55	94.31	38
Arya ja Chamotra	88.2	75	81.05	>46
Han ja Shen	97.07	-	93.26	∼33

Vertailussa on myös huomioitava Gasconilla ym. ja Evansilla ym. [8] olleen muista poikkeava lähestymistapa spearphishing-viestien tunnistamiseen, heidän ratkaisuissaan lähettäjistä luotiin profiilit ja verrattiin viestin oletettua lähettäjää olemassa olevaan profiiliin. Muut keskittyvät spearphishing-viestien piirteiden tunnistami-

seen viesteistä. Lähettäjäprofiileihin painottuva ratkaisu toimii vain organisaatioissa, joissa voidaan olettaa lähettäjiltä löytyvän tarpeeksi sähköpostiviestejä profiilien muodostamista varten.

Arvioitavien ominaisuuksien määrä

Arvioitavien ominaisuuksien määrällä vaikuttaa olevan yhteys tulokseen. Arya ja Chamotra [14], jotka käyttivät eniten eri ominaisuuksia saivat selkeästi huonoimman tuloksen, kun taas Ding ym. [12], sekä Han ja Shen [4] saavuttivat parhaimpia tuloksia pienellä määrällä arvioitavia ominaisuuksia. Lisäksi Dingin ym. ja Lingin ym. [13] käyttämä data sekä ominaisuudet olivat erittäin lähellä toisiaan, mutta Ding ym. saavuttivat paremman tuloksen pienemmällä määrällä ominaisuuksia. Ominaisuuden määrän kasvulla lienee siis negatiivinen vaikutus tulokseen.

Evansin ym. [8] dynaaminen ominaisuusosajoukkojen valinta saavutti erinomaisen saannin sekä hyvän tarkkuuden. Toteutukseltaan ratkaisu on samankaltainen kuin Gasconin ym. [11], joka saavutti hieman Evansia ym. paremmat tulokset. Gasconin ym. ja Evansin ym. käyttämä data oli myös samankaltaista, molemmat käyttivät itse luotuja spearphishing-viestejä koulutus- ja testausdatana. Tulokset ovat siis vertailukelpoisia. Evansin ym. saavuttama tulos dynaamisella ominaisuuksien valinnalla on merkittävä, sillä dynaaminen valinta mahdollistaa reagoinnin muuttuvaan ympäristöön, ja vähentää manuaalista työtä.

Eri ominaisuudet

Merkittävin ero ominaisuuksissa on Gasconin ym. [11] ja Evansin ym. [8] valinta käyttää vain otsakedataa ominaisuuksien muodostamiseksi, kun taas muut käyttivät myös viestin sisältöä. Gasconin ym. ja Evansin ym. ratkaisu vaatii lähettäjäprofiilien muodostamiseksi lähettäjien lähettämiä sähköposteja, mutta koen että koska hyökkääjien ideaalit kohteet imitoitaviksi ovat usein yritysten hierarkiassa korkealla

tasolla, löytyy myös näiltä kohteilta riittävästi historiallista dataa profiilin muodostamiseksi. Ainoastaan otsakedatan käyttäminen pienentää myös analysoitavan datan määrää, sillä tekstisisällön analysointi ja aiheiden tunnistaminen vaativat todennäköisesti enemmän resursseja. On kuitenkin todettava että tekstinsisältöä analysoivat tutkimukset ovat tarkkuudeltaan päässeet parempiin tuloksiin kuin Gascon ym. ja Evans ym.

Ding ym. [12] sekä Ling ym. [13] puolestaan osoittivat maineominaisuuksilla olevan positiivinen vaikutus tarkkuuteen ja F1-tulokseen. Kun taas Gasconilla ym. [11] ja Evansilla ym. [8] oli saanti erittäin korkealla, mutta tarkkuus huomattavasti matalampi. Mainetietojen lisäys Gasconin ym. määrittelemien otsakeominaisuuksien joukkoon johtaisi potentiaalisesti erittäin hyvään tulokseen.

Jatkokäsittelyt

Tutkimuksissa hyödynnetyistä jatkokäsittelykeinoista Dingin ym. [12] esittämä KM-SMOTE datan laajennusmenetelmä sekä Hanin ja Shenin [4] kampanjoiden tunnistus, ovat toimivia ratkaisuja. Datan laajennus auttaa vähentämään epätasapainoisen datan vaikutusta, ja kampanjoiden tunnistus helpottaa uusien viestien luokittelua ja tekee ratkaisusta kestävämmän muuttuvassa ympäristössä. Aryan ja Chamotran [14] hämäys-hiekkalaatikoinnin vaikutuksesta ei esitetty tuloksia, mutta oletettavasti se vaikutti positiivisesti, kuitenkin se on resurssivaatimuksista johtuen hankala toteuttaa isolla mittakaavalla.

Koneoppimisalgoritmit

Taulukosta 3.3 huomataan KNN:n, sekä random forestin olevan tutkimuksissa käytetyimmät algoritmit. Vertailuja algoritmien välillä suorittivat Ding ym. [12], Ling ym. [13], sekä Han ja Shen [4]. Ding ym. ja Ling ym. päätyivät siihen että random forest oli heidän vertailussaan paras. Han ja Shen taas vertasivat omaa KNN:ää hyö-

dyntävää algoritmiaan random forestiin, ja tulokset osoittavat heidän algoritminsa suoriutuvan random forestia paremmin. Näiden vertailujen perusteella voidaan todeta KNN:n soveltuvan parhaiten spearphishing-viestien luokitteluun.

Kokonaisuudet

Gasconin ym. [11] ja Evansin ym. [8] ratkaisu toimii kunhan hyökkääjä pyrkii imitoimaan jotain lähettäjää, josta on luotu lähettäjäprofiili. Muiden ratkaisuissa taas tekstinsisällöstä ja muista ominaisuuksista pyritään tunnistamaan spearphishing-viesteille ominaisia piirteitä, mikä on vankempi vieraita lähettäjiä sisältävässä ympäristössä. Gascon ym. tuovat kuitenkin ilmi että hyökkääjän on helppo imitoida lähettäjän tekstuaalisia ominaisuuksia, jos hyökkääjällä on hallussa imitoinnin kohteen lähettämiä sähköposteja. Mutta tällöin otsakeominaisuuksiin perustuva lähettäjäprofiili pystyy erottamaan imitoijan aidosta lähettäjästä.

Aryan ja Chamotran ratkaisua [14] on vaikea verrata muihin vähäisten tulosten perusteella. He tutkivat ratkaisun toimintaa vain reaalitilanteessa, mistä johtuen heidän tuloksensa perustuu huomattavasti pienempään aineistoon. Heidän ratkaisunsa olettaisi pystyvän tunnistamaan spearphishing-viestit laajan ominaisuusjoukon analysoinnin ansiosta, vaikkakin resurssi- ja aikavaatimukset ovat varmasti muita korkeammat. Ding ym. [12] ja Ling ym.[13] saivat kaikki tulosmittarit huomioiden erinomaiset tulokset. He eivät kuitenkaan paljastaneet kaiken käyttämänsä koulutus- ja testausdatan lähteitä tai kuvailleet käyttyjen spearphishing-viestien rakennetta ja hienostuneisuutta, joten tulosten vertailukelpoisuudesta ei ole takuita. Näistä huolimatta he kuitenkin osoittivat välitys- ja mainetietojen, sekä datan laajennuksen käytöllä olevan positiivinen vaikutus tulokseen. Hanin ja Shenin [4] ratkaisu tuotti myös hyvän tuloksen saannin osalta, sekä hyvän F1-tuloksen kampanjoiden tunnistuksesta. Tarkempia tuloksia he eivät ilmoittaneet, mikä vaikeuttaa vertailua muihin tuloksiin.

4 Yhteenveto

Spearphishing-viestien tunnistamiseen koneoppimisen avulla on useita erilaisia ratkaisuja ja lähestysmistapoja. Tässä tutkielmassa vertailin kuutta artikkelia ja niissä esitettyjä ratkaisuja viestien tunnistamiseen tutkimuskysymysten näkökulmasta. Tutkielman tutkimuskysymykset ovat: TK1. *Mitkä ominaisuudet sähköpostiviestissä soveltuvat tunnistamiseen?*, TK2. *Mitä koneoppimisalgoritmeja hyödynnetään tunnistamiseen?*, ja TK3. *Mitä muita ominaisuuksia hyvällä tunnistamisratkaisulla on?*. Vertailun pohjalta ei voida todeta yhtä ratkaisua selkeästi parhaaksi. Jokainen artikkeli tuo tunnistamiseen uusia ja toimivia menetelmiä. Paras ratkaisu lienee näitä hyviä menetelmiä hyödyntävä kokonaisuus.

TK1. Ominaisuudet

Arvioitavien ominaisuuksien määrällä ja valinnalla on selkeä vaikutus tulokseen. Ratkaisut jotka arvioivat vähemmän ominaisuuksia saavuttivat paremmat tulokset. Ominaisuuksien määrän kasvu vaikuttaa myös koneoppimisalgoritmien suoriutumiseen, määrän kasvaessa vaadittavat resurssit kasvavat ja luokittelusta tulee monimutkaisempaa. Ominaisuuksien määrä tulisi siis pitää pienenä, mikä vaatii ominaisuuksien täsmällistä valintaa. Ratkaisu ominaisuuksien valinnalle on Evansin ym. [8] esittämä ominaisuusosajoukkojen valinta vahvistetun oppimisen avulla. Evans ym. todistivat ratkaisunsa pystyvän reagoimaan muuttuviin hyökkäysmalleihin, sekä saavuttamaan hyvän tuloksen dynaamisella ominaisuuksien valinnalla. Suoritusajan

ja resurssien tarpeen minimointia varten tarvitaan kuitenkin vielä jatkotutkimuksia, nykyisellään ominaisuuksien valinnassa kestää liian kauan. Arvioitavien ominaisuuksien valinta dynaamisesti vähentää myös manuaalista vaadittavaa työtä ominaisuuksien päivittämiseen ja hienosäätämiseen uusien hyökkäysmuotojen ilmetessä.

Käytettäviä ominaisuusryhmiä tulisi myös tutkia tarkemmin. Evans ym. [8] käyttivät ominaisuusosajoukkojen valinnassa vain otsakedataa. Ding ym. [12] sekä Ling ym. [13] todistivat maine- ja välitysominaisuuksien vaikuttavan positiivisesti tulokseen. Voidaan olettaa että Evansin ym. ratkaisu saavuttaisi nykyistä paremman tuloksen jos se käyttäisi otsakedatasta saatujen ominaisuuksien lisäksi tekstisisällöstä saatavia ominaisuuksia.

Tutkielman pohjalta kaikkien artikkeleiden käsittelemien sähköpostiviestien ominaisuuksien, voidaan todeta soveltuvan spearphishing-viestien tunnistamiseen. Artikkeleissa ei kuitenkaan tuotu ilmi analyysien kestoa, jolla on merkittävä vaikutus ratkaisun soveltumiseen reaalitilanteissa. Monessa ratkaisussa analysoidaan tekstinsisältöä, mikä pidemmissä viesteissä johtaa keston kasvamiseen. Tämä huomioon ottaen Gasconin ym. [11] sekä Evansin ym. [8] valinta analysoida ainoastaan otsakedataa lienee reaalitilanteisiin soveltuvin.

TK2. Koneoppimisalgoritmit

Koneoppimisalgoritmeillä on tärkeä rooli spearphishing-viestien tunnistamisessa, ja katsauksen pohjalta KNN on tunnistamiseen parhaiten soveltuva algoritmi. Pääasiallinen tehtävä koneoppimisalgoritmeillä on viestien luokittelu, mutta niitä hyödynnettiin myös muissa tehtävissä. Evans ym. [8] käyttivät KNN:ää vahvistetun oppimisen agentissaan, Han ja Shen [4] hyökkäyskampanjoiden tunnistamisessa. Random forest on toinen koneoppimisalgoritmi, jota ratkaisuissa käytettiin. Hanin ja Shenin suorittamaan vertailuun tukeutuen voidaan kuitenkin todeta että KNN on random forestia paremmin soveltuva spearphishing-viestien luokitteluun.

TK3. Muut ominaisuudet

Ding ym. [12], ja Ling ym. [13] osoittivat KM-SMOTE datan laajennusmenetelmän vaikutuksen luokittelun tulokseen olevan positiivinen. Datan laajennuksen hyödyntäminen jatkotutkimuksissa ja kaupallisissa ratkaisussa toisi merkittävän edun. Sen avulla vähäisellä spearphishing-viestien määrällä ei olisi hankaloittavaa vaikutusta koneoppimismallien kouluttamiseen. Spearphishing-hyökkäyskampanjoiden tunnistaminen saapuvista viesteistä [4] on toinen toimintatapa joka olisi hyvä saada käyttöön. Kampanjoiden tunnistaminen suoritetaan luokittelun jälkeen, joten se ei vaikuta suoritusaikaan ja uusien viestien luokittelun pitäisi helpottua. Kampanjatiedosta voitaisiin hyötyä myös jakamalla sitä tietoturvayrityksille ja organisaatioille, edistäen kampanjoiden torjumista korkeammalla tasolla. Aryan ja Chamotran [14] hämäys-hiekkalaatikointi komponentti vaikuttaa myös kelvolliselta toimintatavalta, sitä voisi hyödyntää pienemmässä roolissa väärien positiivisten karsimiseen. Hyvässä spearphishing-viestien tunnistamisratkaisussa tulisi siis olla käytössä vähintäänkin datan laajentamismetodi sekä kampanjoiden tunnistuskomponentti.

Yhteenveto

Tutkielman perusteella tehokkaan ratkaisun kehittäminen spearphishing-viestien tunnistamiseen vaatii vielä jatkotutkimuksia. Erityisesti tulisi tutkia eri menetelmien vaatimia resursseja ja suoritusaikoja, liian pitkä analyysi ei sovellu reaalitilanteisiin. On myös kehitettävä vahvistetun oppimisen hyödyntämistä ominaisuuksien valinnassa, sekä määriteltävä analyysille optimaaliset ominaisuusryhmät. Tutkielmassa saatiin kuitenkin vastaukset tutkimuskysymyksiin, ja niiden perusteella voidaan paremmin määritellä millainen on hyvä tunnistamisratkaisu.

Lähdeluettelo

[1] M. H. U. Sharif ja M. A. Mohammed, "A literature review of financial losses statistics for cyber security and future trend", *World Journal of Advanced Research and Reviews*, 2022. DOI: 10.30574/wjarr.2022.15.1.0573.

[2] GOV.UK, *Cyber Security Breaches Survey 2023*, https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2023/cyber-security-breaches-survey-2023, viitattu: 9.4.2024, 2023.

[3] J. I. Hong, "The state of phishing attacks", *Communications of the ACM*, vol. 55, nro 1, s. 74–81, 2012. DOI: 10.1145/2063176.2063197.

[4] Y. Han ja Y. Shen, "Accurate spear phishing campaign attribution and early detection", teoksessa *SAC*, ACM, 2016, s. 2079–2086. DOI: 10.1145/2851613.2851801.

[5] M. Jakobsson ja S. Myers, *Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft*. 605 Third Avenue, New York, NY, United States: Wiley-Interscience, 2006, ISBN: 978-0-471-78245-2.

[6] C. Griffiths, "The Latest 2024 Phishing Statistics", tekninen raportti, 2024, viitattu: 9.4.2024. url: https://aag-it.com/the-latest-phishing-statistics/.

[7] K. Krombholz, H. Hobel, M. Huber ja E. Weippl, "Advanced social engineering attacks", *Journal of Information Security and Applications*, vol. 22, s. 113–122,

2015, Special Issue on Security of Information and Networks, ISSN: 2214-2126. DOI: https://doi.org/10.1016/j.jisa.2014.09.005.

[8] K. Evans, A. Abuadbba, T. Wu et al., "RAIDER: Reinforcement-Aided Spear Phishing Detector", *International Conference on Network and System Security*, vol. 13787, 2022.

[9] Barracuda Networks, "2023 spear-phishing trends", tekninen raportti, 2023, viitattu: 15.4.2024. url: https://www.barracuda.com/reports/spear-phishing-trends-2023.

[10] Trend Micro, "Pawn Storm in 2019", tekninen raportti, 2020, viitattu: 9.4.2024. url: https://documents.trendmicro.com/assets/white_papers/wp-pawn-storm-in-2019.pdf.

[11] H. Gascon, S. Ullrich, B. Stritter ja K. Rieck, "Reading Between the Lines: Content-Agnostic Detection of Spear-Phishing Emails", teoksessa *Research in Attacks, Intrusions, and Defenses*, sarja Lecture Notes in Computer Science, vol. 11050, Springer, 2018, s. 69–91. DOI: 10.1007/978-3-030-00470-5_4.

[12] X. Ding, B. Liu, Z. Jiang, Q. Wang ja L. Xin, "Spear Phishing Emails Detection Based on Machine Learning", teoksessa *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, IEEE, 2021. DOI: 10.1109/CSCWD49262.2021.9437758.

[13] Z. Ling, H. Feng, X. Ding, X. Wang, C. Gao ja P. Yang, "Spear Phishing Email Detection with Multiple Reputation Features and Sample Enhancement", teoksessa *Science of Cyber Security*, sarja Lecture Notes in Computer Science, vol. 13580, Springer, 2022, s. 522–538. DOI: 10.1007/978-3-031-17551-0_34.

[14] S. Arya ja S. Chamotra, ”Multi Layer Detection Framework for Spear-Phishing Attacks”, teoksessa *Lecture Notes in Computer Science*, vol. 13146, Springer, 2021, s. 38–56. DOI: 10.1007/978-3-030-92571-0_3.

[15] S. Duman, K. Kalkan-Cakmakci, M. Egele, W. Robertson ja E. Kirda, ”Email-Profiler: Spearphishing Filtering with Header and Stylometric Features of Emails”, teoksessa *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, kesäkuu 2016. DOI: 10.1109/compsac.2016.105.

[16] G. Stringhini ja O. Thonnard, ”That Ain’t You: Blocking Spearphishing Through Behavioral Modelling”, teoksessa *Lecture Notes in Computer Science*. Springer International Publishing, 2015, s. 78–97, ISBN: 9783319205502. DOI: 10.1007/978-3-319-20550-2_5.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall ja W. P. Kegelmeyer, ”SMOTE: synthetic minority over-sampling technique”, vol. 16, nro 1, s. 321–357, 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953.