Kanditutkielma

Tietojenkäsittelytieteen kandiohjelma

Tekstipohjaiset vastakkaishyökkäykset NLP-luokittimia vastaan

Akira Taguchi

8.10.2022

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Ohjaaja(t)

Prof. Nikolaj Tatti

Yhteystiedot

PL 68 (Pietari Kalmin katu 5)
00014 Helsingin yliopisto

Sähkopostiosoite: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Study programme
Matemaattis-luonnontieteellinen tiedekunta	Tietojenkäsittelytieteen kandiohjelma

Tekijä — Författare — Author

Akira Taguchi

Työn nimi — Arbetets titel — Title

Tekstipohjaiset vastakkaishyökkäykset NLP-luokittimia vastaan

Ohjaajat — Handledare — Supervisors

Prof. Nikolaj Tatti

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kanditutkielma	8.10.2022	17 sivua

Tiivistelmä — Referat — Abstract

Tekstin automaattisella luokituksella on tärkeä rooli digiyhteiskunnassa. Tämä luokitus tapahtuu luonnollisen kielen käsittelyyn pohjautuvilla metodeilla. Metodeihin pohjautuvat malli ovat kuitenkin haavoittuvaisia, ja tässä tutkimuksessa käsitellään tekstipohjaisia vastakkaishyökkäyksiä NLP-luokittimia vastaan. Tutkielman alussa perehdytään automaattisen luokituksen käyttötarkoituksiin. Tämän jälkeen tutustutaan hyökkäystyyppeihin NLP-luokittimia vastaan. Lopuksi käsitellään puolustusmetodeita NLP-hyökkäyksiä vastaan.

ACM Computing Classification System (CCS)
Security and privacy → Human and societal aspects of security and privacy
Computing methodologies → Artificial Intelligence → Natural language processing

Avainsanat — Nyckelord — Keywords

Luonnolllisen kielen käsittely, vastakkaishyökkäys, koneoppiminen, tekoäly, ladonta, sensuuri

Säilytyspaikka — Förvaringsställe — Where deposited

Helsingin yliopiston kirjasto

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Koneoppimisen käyttötarkoitusten määrä kasvaa vuosi vuodelta suuremmaksi. Tätä teknologiaa voidaan hyödyntää muun muassa ihmisten puhuman kielen käsittelyyn. Luonnollisen kielen käsittely (eng. Natural Language Processing, NLP) on alati kasvavassa kuluttajakäytössä johtuen seuraavista syistä:

- laskentatehon kasvu
- suurien tietomäärien saatavuus
- onnistuneiden koneoppimismenetelmien kehitys
- sekä laajempi ihmiskielen ymmärrys ja sen käyttö eri konteksteissa (Hirschberg ja Manning, 2015).

Luonnollisen kielen käsittelyä voidaan hyödyntää kohdennetussa mainonnassa. Analysoimalla NLP-luokittimen avulla esimerkiksi käyttäjien lähettämiä viestejä toisilleen, voidaan saada selville tuote, jota kannattaa mainostaa yksilöllle. Viesti ystävälle viestipalvelussa antaa työstettävän datan NLP-luokittimelle: “Mikä elokuva meidän pitäisi katsoa viikonloppuna? “ NLP-luokittimen avulla automaattinen mainostaja ymmärtää mainostaa kyseiselle käyttäjälle esimerkiksi suoratoistopalvelua tai sarjalippuja mainostavaa elokuvateatteria. Tämän tiedon löytäminen suuresta määrästä dataa luonnollisen kielen käsittelyllä edellyttää kaikkia neljää aikaisemmin mainittua teknologista edistystä kultakin osa-alueelta.

Kaikkien neljän osa-alueen kehittyminen mahdollistaa luonnollisen kielen käsittelyn yleistymisen. Ihmiskielen ymmärtäminen tietokoneen tasolla on kehittynyt huomattavasti, kun ihmisen käyttämää kieltä, virkkeitä ja sanoja on alettu pilkkomaan helpommin ymmärrettäviksi paloiksi (Chowdhury, 2003). Jotta luonnollisen kielen käsittelyn malli olisi rakennettu älykkäästi, tarvitsemme edistyneitä koneoppimismetodeita. Tämä on tullut kehityksen saatossa mahdolliseksi (Jordan ja Mitchell, 2015). Koska datan määrä on kasvanut ja dataa on helpompaa hankkia (Gopalakrishnan, 2018), pystymme kouluttamaan mallin toimimaan mahdollisimman monessa eri tilanteessa. Laskentatehon huomattava kasvu vuosien mittaan (Moore et al., 1965) on alkanut mahdollistaa suurempien datamäärän käsittelyä kuin aikaisemmin.

Tässä tutkielmassa tarkastellaan NLP-hyökkäysten käyttökohteita. Tähän kuuluuvat hyökkäystyypit, puolustusmenetelmät sekä NLP-luokittimien sekä niihin kohdistuvien hyökkäysten tulevaisuus. Hyökkäystyypeissä käymme läpi erilaisia tapoja hyökätä NLP malleja vastaan, hyökkäysten tarkoituksiin ja onnistumisen todennäköisyyksiin. Puolustusmenetelmät ovat tärkeässä osassa, jotta haavoittuvuuteen kohdistuvat yritykset saavat ohjeita vahingon mitigointiin ja ennaltaehkäisyyn. On tärkeää myös spekuloida mahdollisia kehityksiä koneoppimisessa sekä tästä syntyviä haavoittuvuuksia. Lopuksi käymme läpi mahdollisia luonnollisen kielen käyttökohteita tulevaisuudessa sekä näistä aiheutuvia seurauksia eri osa-alueisiin sekä akateemisella että kaupallisella puolella.

2 Tekstin automaattinen luokitus

NLP-luokittimia käytetään analysoimaan tekstiä, joissa on tehokkaampaa korvata ihmisen manuaalisesti tekemä työ. Ensin käydään läpi neljä yleistä tapausta tekstin automaattisesta luokituksesta. Nämä neljä tapausta ovat roskapostin suodatus sähköposteista, vihapuheen sensurointi sosiaalisesta mediasta, valearvosteluiden tunnistus nettikauppojen arvosteluosioista sekä sentimenttianalyysi. Lopuksi käydään läpi tekstin automaattisen luokituksen edut verrattuna manuaaliseen, ihmisen tekemään luokitustyöhön.

2.1 Roskapostien suodatus

Sähköpostien automaattiseen luokitukseen joko roskaposteiksi tai asiaposteiksi käytetään NLP-luokittimia. Noin 70% liiketoiminnan sähköposteista on roskapostia. Näiden roskapostien tarkoitus voi muun muassa olla huijausta, ärsyttämistä tai loukkaamista (Garg ja Girdhar, 2021).

Roskapostin vaikutukset käyttäjästä riippuen ovat niin vakavia, että sähköpostipalvelun tarjoajan intresseissä on implementoida roskapostisuodatin. Käyttäjä pystyisi tarkastamaan vastaanotetusta sähköpostista, mikäli kyseinen sähköposti olisi esimerkiksi kalasteluroskapostia. Koska roskapostia lähetetään automaattisesti jokaiseen olemassa olevaan sähköpostiosoitteeseen päivittäin, menisi roskapostien tunnistamiseen ihmiseltä liian kauan aikaa päivittäin. Automaattisella roskapostin lähetyksellä tarkoitetaan tietokoneella ohjelmoitua sähköpostien lähettämistä eri sähköpostiosoitteisiin. Usein nämä sähköpostiosoitteetkin ovat hankittu tietokoneohjelmoinnin avulla, joten roskapostia lähetetään päivittäin paljon. Roskapostit saattavat sisältää viestin avaajaa järkyttävää tai provosoivaa mediaa. Roskaposti saattaa sisältää myös kalasteluyrityksiä. Kalasteluhyökkäyksessä tarkoituksena on huijata käyttäjää antamaan erilaisia tunnus-salasana-yhdistelmiä liittämällä roskapostiin esimerkiksi linkin viralliselta näyttävältä sivulle (Khonji et al., 2013). Sivulla käyttäjää kehotetaan kirjautumaan tunnuksillaan tuttuun palveluun, mutta oikeasti palvelu vain varastaa käyttäjän tunnukset. Roskaposti saattaa myös sisältää haittaohjelmia, joita käyttäjä voi saada koneelleen muun muassa lataamalla ja suorittamalla sähköpostin tiedostoja tai vierailemalla pahantahtoisella sivustolla. Tämä pahantahtoinen sivusto usein sisältää koodia, joka hyväksikäyttää usein jotain selaimen haavoittuvai-

suutta esimerkiksi asentaakseen tietokoneelle haittaohjelmia. Myös kiristysviestejä sekä sähköposteja eteenpäinlähettäviä haittaohjelmia kulkee roskapostien mukana, joita sähköpostipalvelun tarjoajat pyrkivät estämään roskapostisuodattimilla.

2.2 Vihapuheen sensurointi

Vihapuheen riittävään sensurointiin tarvitaan luonnollisen kielen käsittelyä. Suodattimen rakentaminen vihapuhetta vastaan pelkkien avainsanojen perusteella ei tuota toivottuja tuloksia. Katsotun vihapuheen sensuroinnille tarvitaan muun muassa meneillään olevan keskustelun suunta, tarkka ajanhetki, ajankohtaiset maailman tapahtumat, lähettäjän sekä vastaanottajan henkilöllisyys sekä kontekstuaaliset mediat, esimerkiksi kuvat, videot tai ääni (Schmidt ja Wiegand, 2017). Vihapuheen sensurointi manuaalisesti vaatii kontekstuaalista ymmärystä keskustelusta. Käytännössä tämä vaatisi yhdeltä tarkastajalta aiheen tutkimista sekä mahdollisiin uusiin vihapuhesanoihin tai vihapuhetta sisältäviin lauseisiin tutustumista. Työntekijöitä tarvittaisiin todennäköisesti paljon, mutta manuaalisella vihapuheen sensuroinnilla on myös toinen ongelma. Kaupallisen sisällön moderaattorit altistavat usein itsensä häiritsevälle sisällölle. Kaupallisella sisällöllä tarkoitetaan tässä tapauksessa esimerkiksi Facebookin, Googlen ja Twitterin sisältöä. Pienemmillä alustoilla moderointia harjoitetaan useammin vapaaehtoistyönä, jolloin häiritsevä sisältö jakaantuu usealle eri vapaaehtoiselle muutaman palkkatyöläisen sijasta. Kaupallisen sisällön moderoitava häiritsevä sisältö saattaa johtaa pitkäaikaiseen psykologiseen ja henkiseen kärsimykseen (Steiger et al., 2021).

2.3 Valearviointien tunnistus

Ostosten tekemisten mahdollisuuden netissä sekä tuotteiden hyvän saatavuuden vuoksi kuluttajat joutuvat perustelemaan ostopäätöksensä yhä useammin tuotearvosteluihin. Oikeiden arvosteluiden lisäksi tuotesivulla saattaa olla valearvosteluja. Luonnollisen kielen käsittelyyn perustuvalla tekniikalla voidaan kyseiset valoarvostelut tunnistaa ja tuhota (Trupthi et al., 2019).

Valearvostelujen määrä ja kieliasu ovat pääsyyt NLP-luokittimien käyttöön edellä mainitussa käyttökohteessa. Valearvosteluita voidaan tuottaa eri syistä. esimerkiksi tuotteen näennäisen arvon laskeminen kilpailullisen tuotteen näennäisen arvon nostamiseksi. Valearvosteluita voidaan myös tehdä myös pelkästään pahantahtoisella tarkoituksella alentaa

tuotteen näennäisarvoa. Luokitin tunnistaa suuresta määrästä arvosteluja valearvostelut, vaikka kieliasu ei olisikaan formaali. Tässäkään tapauksessa pelkkä suodatin, joka perustuu avainsanoihin, ei riitä tunnistamaan valearvosteluita aidoista arvosteluista. Valearvostelujen tunnistus luonnollisen kielen avulla on markkinallisista syistä verkkokauppojen intresseissä. Valearviointien manuaalisessa tunnistuksessa on samanlaisia ongelmia kuin roskapostitunnistuksessa. Tämän lisäksi valearvostelut hukkuvat oikeiden ihmisten lähettämien arvostelujen sekaan, joita lähetetään todennäköisesti alemmalla kynnyksellä, kuin sähköposteja.

2.4 Sentimenttianalyysi

Tunnesävyn tunnistaminen omasta -tai kilpailijan tuotteesta tuottaa arvokasta tietoa tuotekehitykselle sekä markkinointi -ja asiakassuhteen ylläpidolle. Kuluttajien sentimentaalisuuden analysoiminen automaattisesti luonnollisen kielen käsittelyn tekniikoiden avulla on kustannustehokasta. Manuaalisesti tunnesävyjen selvittäminen jokaisesta olennaisesta netissä olevasta tekstistä vaatisi paljon resursseja eikä välttämättä tuottaisi yhtä paljon tai yhtä laadukkaita tuloksia, kuin automaattinen sentimentaalisuuden analysoiminen (Yi et al., 2003).

Tämä tarkastustyö voitaisiin tehdä manuaalisesti, mutta tarkastettavan sisällön määrän vuoksi tämä ei käytännössä ole kannattavaa ja useammissa tapauksissa onkin miltei mahdotonta. Yksittäisen keskustelun tai aiheen tutkiminen ei riittäisi analysoimaan tunnesävyjä, vaan tarkastajan täytyisi käydä läpi mahdollisimman monta netin tekstiä ja analysoida näistä tunnesävyt. Vaikka tekstin automaattisella luokituksella on paljon etuja verrattuna tekstin manuaaliseen luokitukseen, sisältää tekstin automaattinen luokitus kuitenkin tietoturvahiekkouksia.

3 Hyökkäystyypit

NLP-luokittimia vastaan, jotka ovat automaattisen tekstin luokituksen keskiössä, voidaan hyökätä. Tässä kappaleessa käydään läpi hyökkäystyypit NLP-luokittimia vastaan. Ensin käydään läpi roskapostisuodatuksen roskapostisuodatuksen ohitus, joka on NLP-hyökkäysten keskiössä. Sitten esitellään sensuuriohitus sekä ladontahyökkäykset. Ladontahyökkäyksistä käydään läpi näkymättömät merkit, homoglyfit, uudelleenjärjestelyt sekä poistatukset.

3.1 Roskapostin naamiointi asiapostiksi

Vastakkaishyökkäyksiä voidaan käyttää sähköposteissa roskapostisuodattimien ohitukseen. Roskapostisuodattimet toimivat koulutettujen NLP-luokittimien mukaan. Nämä luokittimet luokittelevat vastaanotetut sähköpostit joko hyväntahtoisiksi tai pahantahtoisiksi, eli roskaposteiksi (Kuchipudi et al., 2020).

Suodattimia vastaan toimii kolme vastakkaishyökkäystä. (1) Synonyymin korvaus. Synonyymin korvauksessa tarkoitus on korvata pahantahtoiset sanat hyväntahtoisiksi luokitelluilla synonyymeillä. Lauseiden samankaltaisuuksien vertailua demonstroidaan taulukossa 3.1. Ensimmäisessä sarakkeessa on viesti, jonka luokitin luokittelee joko roskapostiksi tai asiapostiksi toisessa sarakkeessa. Taulukon viimeinen rivi demonstroi synonyymien kovauksen läpäisevän roskapostisuodattimen. Pahantahtoisissa lauseissa pyritään nostattamaan samankaltaisuusastetta vaihtamalla sanoja synonyymeihin, kunnes NLP-luokitin tunnistaa viestin olevan asiapostia. (2) asiasanan injektointi. Asiasanan injektoinnissa asiasanoja lisätään sähköpostiin niin paljon, kunnes NLP-luokitin tunnistaa roskapostin olevan asiapostia. Sana ”asiaposti” tarkoittaa tässä yhteydessä tekstiä, jonka roskapostisuodatin on luokitellut hyväntahtoiseksi. Asiasanoja voidaan injektoida tietokannoista roskaposteihin muuttamatta viestin tarkoitusta rajusti. (3) roskapostisanojen väljennys. Roskapostisanojen väljennyksessä roskapostisanoihin sisällytetään välilyöntejä, jotta NLP-luokitin ei tunnistaisi näitä sanoja roskasanoiksi. Kun väljennystä on harjoitettu tarpeeksi, muuttuu roskaposti NLP-luokittimen näkökulmasta asiapostiksi. (Kuchipudi et al., 2020)

Asiasanan injektoinnille ja roskasanojen väljennykselle on olemassa erilaisia implementaatioita. Seuraavissa alaluvuissa tutustutaan ladontapohjaisiin vastakkaishyökkäyksiin.

Muokattu viesti	Ennustus
Ringtone Club: Get the UK singles chart on your mobile each week and choose any top quality ringtone! This message is free of charge.	roskapostia
Ringtone Club: **acquire** the UK single **graph** on your **Mobile_River** each **hebdomad** and **take** any **top_side caliber** ringtone! This **content** is **free_people** of charge.	roskapostia
Ringtone Club: **become** the UK **bingle graph** on your **nomadic** each **workweek** and **select** any **upper_side caliber** ringtone! This **subject_matter** is **liberate** of charge.	roskapostia
Ringtone Club: **go** the UK **one graph** on your **peregrine** each **calendar_week** and **pick_out** any **upside character** ringtone! This **substance** is **release** of charge.	asiapostia

Taulukko 3.1: Synonyymin korvaus. Vanhan viestin korvatut osat on lihavoitu. (Kuchipudi et al., 2020)

Muun muassa näitä hyökkäysmetodeita voidaan käyttää kahdessa aiemmin mainitussa roskapostisuodattimeen kohdistetussa hyökkäyksessä. Implementaatioita yhdistelemällä ja vaihtelemalla, saattaa NLP-luokittimen pahantahtoisuuden havaitseminen heikentyä entisestään, taaten hyökkääjälle varmemman onnistumisen.

3.2 Näkymättömät merkit

Näkymättömät merkit vaikuttavat tietokoneen NLP-luokittimen ymmärtämään sisältöön. Seuraavissa alikappaleissa käydään läpi ladontatason vastakkaishyökkäysmetodeita. Näitä yhdistelemällä, jopa alan johtava vihapuhefiltteri Google Perspective, on altis sensuurin ohitukselle vastakkaishyökkäyskoulutuksesta huolimatta (Gröndahl et al., 2018). Koulutus vastakkaishyökkäyksiä vastaan tässä kontekstissa tarkoittaa puolustavan NLP-luokittimen koulutusta syötteellä, joka yrittäisi hyökätä NLP-luokitinta vastaan. Kyseessä on siis NLP-luokitin, jonka luokitus lujittuu vastakkaishyökkäyksiä tuottavan NLP-luokittimen ulostulolla. Kyseinen hyökkäys perustuu Unicode-merkistöstandardiin, joka sisältää yksilöivät koodiarvot yli 100 000 kirjoitusmerkille, tähän kuuluvat myös aakkoset sekä erikoismerkit (Boucher et al., 2021).

Esimerkki tällaisesta erikoismerkistä on nollatilavuuden välilyönti -merkki, jonka Unicode merkintä on `U+200B`. Tällä merkillä voimme esimerkiksi vaikuttaa pelichattiin lähetettävän myrkyllissuodatettavaan merkkijonoon "olet huono"niin, että merkkijono menisi

NLP-luokittimen läpi chätistä. Merkkijono `olU+200Bet huU+200Bono` saattaisi mennä läpi chatin suodattimesta, mutta vastapuolelle viesti olisi edelleen `olet huono`.

Kontekstin poistamisen lisäksi näkymättömillä merkeillä voidaan myös tuoda ja syrjäyttää konteksteja toisilla. Esimerkiksi:
`Mikä pyhäinhäväistyksen rakennus!`
`Miten onnistuit tekemään tämän näin laiskasti?` -tekstin negatiivisuus voidaan syrjäyttää positiivisuudella syöttämällä NLP-luokittimelle sen sijaan teksti:
`Mikä pyU+200BhäinhävU+200BäistyU+200BksenU+200B rakennus!`
`Miten onnistuit tekemään tämän U+200BnäU+200Bin laU+200BiskasU+200Bti?`.

Taulukko 3.2 on esimerkki kontekstin syrjäyttämisestä näkymättömillä merkeillä. Esimerkissä tapahtuu käännös englannin kielestä ranskan kieleen. Vasemassa sarakkeessa on alkuperäinen, käännettävä viesti. Vasemmassa sarakkeessa taas on käännetty viesti ranskaksi. Alleviivattuihien kohtien väliin on upotettu nollatilavuuden välilyönti-merkki `U+200B`. Ihmiselle käännettävät viestit näyttävät samanlaisilta, mutta kääntäjälle jälkimmäisessä viestissä on kolme välilyöntiä enemmän. Viimeinen viesti pitäisi kääntyä viestiksi ”Teknologia on olemassa sitä varten”. Käännös on kuitenkin hyökkäyksen jälkeen ”Ympäristön kronologia on kronologia ympäristöstä ympäristölle”.

Alkuperäinen viesti	Käännös
The technology is there to do it.	La technologie est là pour le faire.
The technology is there to do it.	La chnologie de l’environnement est la chnologie de l’environnement à l’environnement.

Taulukko 3.2: Hyökkäys näkymättömillä merkeillä (Boucher et al., 2021)

3.3 Homoglyfit

Homoglyfihyökkäykset NLP-luokittimia vastaan pohjautuvat siihen, että pahantahtoisten merkkien viralliset esitysmuodot näyttäytyvät hyväntahtoisilta merkkien virallisilta esityksiltä. Joissain kielissä tekstin merkitys muuttuu täysin yhden merkin vaihtuessa. Esimerkkinä homoglyfistä on `A` $\rightarrow$ `A`, missä viimeinen kirjain on todellisuudessa kyrillinen kirjain A. Taulkossa 3.3 homoglyfihyökkäys on muuntanut englanninkielisen tekstin
`I just can’t belive where she was` ranskankieliseen käännökseen
`I guess I can’t underestimate the location of the scribe and`.

Näkymättömien merkkien lailla homoglyfihyökkäyksen toteutus riippuu ympäristön fontista. (Boucher et al., 2021)

Alkuperäinen viesti	Käännös
I just can't believe where she was.	Je ne peux tout simplement psa croire où elle était.
I j̲ust can't beli̲eve where s̲he was.	Je crois que je ne peux pas sous-estimer l'endroit où se trouvait le scribe e.

Taulukko 3.3: Homoglyfihyökkäys (Boucher et al., 2021)

3.4 Uudelleenjärjestelyt

Uudelleenjärjestelyhyökkäys pohjautuu näennäisen tekstin uudelleenjärjestämiseen pahantahtoisesti. Pankkitilinumeron `1234567` pystyy esimerkiksi vaihtamaan kaksisuuntaisella algoritmilla tilinumeroksi `7654321` pankin palvelinpuolella maksajan huomaamatta mitään. Unicode-merkintä tälle suunnanvaihdolle on `U+200F`. Kaksisuuntaisella algoritmilla tarkoitetaan tässä Unicoden kaksisuuntaista algoritmia. Tämän algoritmin tarkoituksena asiallisissa tarkoituksissa kääntää kirjoitussuunta vasemmalle muun muassa arabiaa tai hepreaa kirjoittaessa. Uudelleenjärjestelyjä käytetään myös NLP-luokittimen sekoittamiseen, jolloin tulokset NLP-luokittimesta ovat käyttökelvottomia. Taulukossa 3.4 uudelleenjärjestelyhyökkäys merkeissä `la` aiheuttaa ranskankielisen käännöksen järjettömyyden. Tämänlaista hyökkäystä voisi käyttää digitaalista sanakirjaa tai kääntäjää vastaan. (Boucher et al., 2021) `U+200F` ladotaan näkymättömänä näkymättömien merkkien tapaan.

Alkuperäinen viesti	Käännös
A black box in your car?	Une boîte noire dans votre voiture ?
A b̲lack box in your car?	A b c h a c h a c h a c h a c h a c h a c h a c h e ?

Taulukko 3.4: Uudelleenjärjestelyhyökkäys (Boucher et al., 2021)

3.5 Poistatukset

Viimeisenä käydään läpi poistatushyökkäykset. Poistatushyökkäyksen tarkoituksena on poistaa käyttäjälle näkyvästä tekstistä ladontavaiheessa haluttu määrä tekstiä pois. Uhri esimerkiksi voisi olla myymässä asuntoaan, jolloin hän kopioi ja liittää sähköiseen sopimukseen hyökkääjän ehdottaman summan. Latomisvaiheessa käyttäjä kuitenkin unohtaa tarkistaa sopimuksen, jolloin poistatusmerkit saattavat poistaa myyntihinnasta esimerkiksi muutaman nollan.

Poistatushyökkäyksiä on vaikeampi toteuttaa aikaisempiin metodeihin verrattuna. Tämä johtuu useimpien käyttöjärjestelmien estosta kopioida poistatusta sisältävää tekstiä leikepöydälle suoraviivaisilla tavoilla, joilla uhri sen tekisi. Onnistuakseen poistatushyökkäyksessä, hyökkääjän tarvitsee yleisesti injektoida NLP-luokittimeen poistatus itse. Esimerkkejä poistatusmerkeistä ovat askelpalautin (BS, eng. backspace), delete (DEL) sekä vaunupalautus (CR, eng. carriage return). (Boucher et al., 2021) Kuva 3.4 havainnollistaa poistatushyökkäystä käytännössä. Esimerkissä näkymättömiä poistatusmerkkejä on pistetty sanojen väliin, muuttaen näin ladottujen lauseiden merkityksen. Viimeinen viesti pitäisi kääntyä viestiksi ”Tämä on tosiaan pakollinen valtiollemme.”. Käännös on kuitenkin hyökkäyksen jälkeen ”Tämä todellisuus on pakollinen rakkauden syntymälle”.

Alkuperäinen viesti	Käännös
This really is a must for our nation.	C’est vraiment une nécessité pour notre nation.
This rea_lly___ is a must for_ our na_tion.	Cette réalyya est un incontournable pour la naissance de l’amour.

Taulukko 3.5: Poistatushyökkäys (Boucher et al., 2021)

Vaikka NLP-luokittimia vastaan kohdistuvia hyökkäyksiä on monta ja osa hyökkäyksistä on vaikeasti havaittavia, näiltä pyritään silti suojautumaan eri menetelmillä.

4 Hyökkäyksiltä suojautuminen

Tässä kappaleessa käydään läpi erilaisia puolustusmenetelmiä NLP-hyökkäyksiä vastaan. NLP-hyökkäykset voidaan torjua korkean tason abstraktiolla, suurella yleisrasituksella, sekä alemman tason abstraktiolla, pienemmällä yleisrasituksella. Ensiksi esitellään korkean tason abstraktion OCR-puolustus, sitten kuvaillaan alemman tason abstraktion suorituskeskeinen puolustautuminen.

4.1 OCR-puolustus

OCR-metodia (eng. OCR, Optical character recognition), eli tekstintunnistusta on käytetty esimerkiksi printatun, skannatun tai käsinkirjoitetun tekstin muuntamiseen muokattavaksi tekstiksi. Joskus tekstin tunnistaminen on vaikeata johtuen esimerkiksi tekstin eri koista, tyyleistä, suuntauksesta tai monimutkaisesta tekstin taustasta. Esimerkiksi nollamerkki ”0” ja aakkonen ”o” voivat olla jollekkin OCR-työkalulle vaikeita erottaa toisistaan. Kyseistä virhelukua havainnollistetaan kuvassa 4.1 ja 4.2. Eri OCR-työkalut saavuttavat tekstintunnistuksen eri tavoilla, mutta esimerkiksi avoimen lähdekoodin OCR-työkalu Tesseractilla on seuraavat vaiheet tekstintunnistukseen kuvasta: (1) kuvan muuntaminen binäärikuviksi, (2) merkkien ääriviivojen tunnistamiset, (3) merkkien ääriviivojen tunnistaminen sanoiksi, (4) sanat tunnistetaan kahteen kertaan, jonka jälkeen teksti on tunnistettu kuvasta. Kuva 4.3 havainnollistaa edelläkuvattua Tesseractin tekstintunnistusprosessia (Patel et al., 2012).

OCR-puolustuksen läpikäynnin jälkeen voidaan huomata metodin eliminoivan huomattavan monta NLP-hyökkäystä. Näkymättömät merkit, poistatukset sekä uudelleenjärjestelymerkit poistuvat kokonaan tekstintunnistuksen jälkeen tuotetusta tekstistä, koska kyseiset merkit eivät ihmissilmällekkään näkyisi. Roskapostisuodatuksen -ja sensuurin ohituksen lisäksi OCR-puolustus alisuoriutuu eräällä toisellakin osa-alueella.

Tekstintunnituksen avulla epäselvyydet tekstin aidosta luonteesta voidaan hahmontaa uudelleen tulkitsemalla aineisto uudestaan visuaalisesti. Tämä menetelmä lisää yleisrasitusta huomattavasti riippuen käyttötarkoituksesta, mutta poistaa pahantahtoiset merkit ilman NLP-luokittimen uudelleenkoulutusta. (Boucher et al., 2021)

Kuva 4.1: Tekstitunnistettava kuva (Patel et al., 2012)

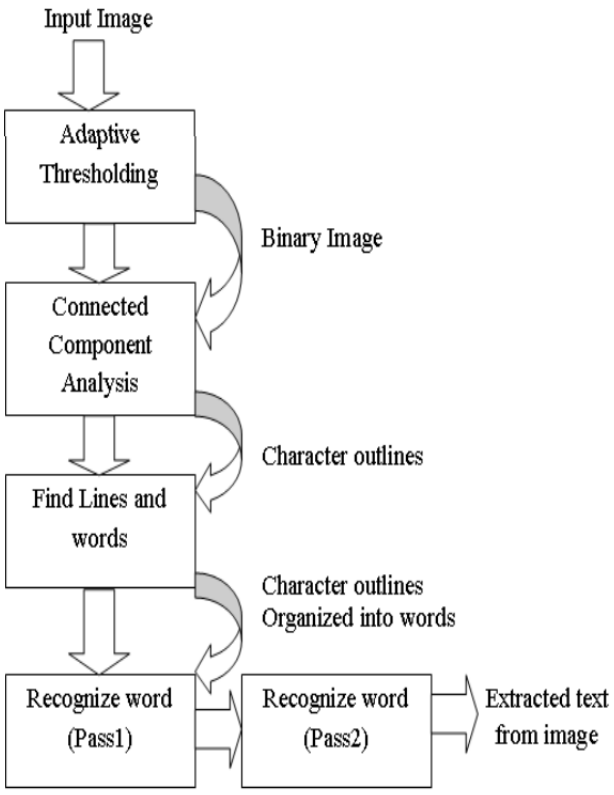
Alkuperäinen viesti	Tesseractin tunnistama teksti
MoonShine effect	MoonShihâ€□Ã

Taulukko 4.1: Tesseract epäonnistuu tunnistamaan ylätekstin ”Effect” kuvasta (Patel et al., 2012)

4.2 Suorituskykykeskeinen puolustus

Keskitymme seuraavaksi näkymättömiin merkkeihin, homoglyyfeihin, uudelleenjärjestelyihin ja poistatuksiin perustuvien hyökkyksien puolustamiseen. Jotkut suorituskykykeskeisistä puolustusmenetelmistä ovat kuitenkin laskennallisesti kalliita, eivätkä koneoppimismallin ulkoistaneet yritykset yleensä pysty taloudellisista syistä nojautumaan kyseisiin menetelmiin (Huang et al., 2019).

Näkymättöimen merkkein tapauksessa, tietyt näkymättömät merkit voidaan poistaa suoraan syötteestä. Mikäli sovelluksessa näitä merkkejä ei voida poistaa, voidaan ne korvata *ei-*`<unk>` upotuksilla. Korvaus tapahtuu lähdekielisanakirjassa, jonne kuvataan tuntematon merkki ”ei-tuntemattomaksi tokeniksi”. Näin tuntemattomat merkit eivät voi häiritä ladontaa merkeillä, joiden asiallinen tarkoitus kyseisessä tekstissä ei ole latojalle yksiselitteistä. Homoglyfihyökkäysten torjuminen OCR-menetelmällä on ymmärrettävästi vaikeampaa verrattuina muihin merkkeihin. Paras keino torjua tällaisia hyökkäyksiä olisi kuvata osa homoglyfeistä niiden yleisemmin tunnettuihin vastineisiin. NLP-luokittimen yl-



Kuva 4.2: Tesseractin tekstintunnistuksen vaiheet (Patel et al., 2012)

läpitäjä joutuu tekemään tässä siis suurimman työn. Uudelleenjärjestelyhyökkäykset voidaan torjua poistamalla kaksisuuntais-ohjausmerkit
syötteestä, varoittamalla käyttäjää kaksisuuntais-ohjausmerkkien ilmestyessä syötteeseen tai käyttämällä kaksisuuntais-algoritmia halutun syötteen selvittämiseen. Puolustusmenetelmän valinta riippuu kontekstista, sillä esimerkiksi latinaa tai arabiaa kirjoittaessa ohjelma toimisi väärin pakottamalla käyttäjän syötteestä pois kaksisuuntais-ohjausmerkin `U+200F`. Poistatukset yleensä havaitaan NLP-luokittimien ulkopuolella syötteen annon alkuvaiheessa. NLP-luokittimen tasolla tähän tarvitsee harvemmin puuttua ja käyttäjälle voidaan pahimmassa tapauksessa lähettää varoitus poistatusmerkkien olemassaolosta syötteessä. On silti tärkeää tiedostaa poistatuksien puolustus, mikäli käyttöjärjestelmä unohtaa puuttua kyseiseen hyökkäystyyppiin. (Boucher et al., 2021)

Roskapostisuodattimeen -ja sensuurin ohitukseen kohdistuvat hyökkäykset voidaan yrittää torjua muun muassa DNE-metodilla (eng. Dirichlet Neighborhood Ensemble). Metodissa korvataan virtuaalisten lauseiden sanoja näiden synonyymeillä. Tämän jälkeen puolustava NLP-luokitin koulutetaan kyseisiä lauseita vastaan. Metodin tarkoituksena on siis puolustautua synonyymin korvausta vastaan, joka esiteltiin alikappaleessa 2.1, Roskapos-

tisuodatuksen ohitus (Zhou et al., 2020).

5 Yhteenveto

Kävimme läpi tässä tutkimuksessa tekijöitä luonnollisen kielen käsittelyn kehitykseen, joita ovat laskentateho, tietomäärä, koneoppiminen sekä ihmiskielen ymmärrys. Kävimme läpi hyökkäyspinta-alan ja puolustusmahdollisuudet NLP-luokittimista johtuvia tietoturvauhkia vastaan. Lopuksi käytiin myös läpi tekstipohjaisten vastakkaishyökkäysten tulevaisuutta NLP-luokittimia vastaan.

Kuten aikaisemmin mainittiin, neljä mahdollistajaa luonnollisen kielen käsittelyyn kuluttajakäytössä ovat laskentatehon kasvu, suurien tietomäärien saatavuus, onnistuneiden koneoppimismenetelmien kehittäminen sekä laajempi ihmiskielen ymmärrys ja käyttö eri konteksteissa. NLP-luokittimien mahdollistajien kehittyessä arvaamattomasti, on loogista tutkia myös NLP-hyökkäysten tulevaisuutta. Vastakkaishyökkäysten motiivit muovautuvat siis ajan myötä ja kasvattavat tahtomattaan näin hyökkäystyyppien määrää.

Hyökkäystyypit laajentuvat tulevaisuudessa eri formaatteihin. Koneoppimisen kukoistaessa voidaan NLP-luokittimia soveltaa tiedon ääni -tai videoformaatteihin. Tämä antaa puolestaan mahdollisuuden vastakkaishyökätä kyseiseen koneoppimismallia vastaan. Formaattien sisältäkin löytyy erinäisiä hyökkäystyyppejä. Esimerkiksi ääniformaateissa käytetään kuhunkin käyttötarkoitukseen sopivaa enkoodausta. Ei siis riitä, että hyökättävää ja puolustettavaa tulee uusien formaattien myötä, sillä formaattien sisälläkin tulee tapahtumaan jatkuvasti huomattavaa kehitystä.

Lisäksi haavoittuvuuksien löytö ruokkii itse itseään. Ensimmäisten vastakkaishyökkäysten kohdistuessa uuteen tietoformaattiin, syntyy tarve puolustukseen tätä vastaan. Toteutuksesta riippuen puolustusmenetelmän selvittäminen saattaa avata uusia ovia, jotka hyödyttävät hyökkääjiä. Usein haavoittuvuuden tarkastelu vastakkaishyökkäyksissä avaa enemmän mahdollisuuksia uusille hyökkäyksille kuin vanhojen hyökkäysten puolustuksille.

Luonnolisen kielen käsittely on muovautunut tärkeäksi osaksi tietokoneteollisuutta. Koneoppimisen avulla kuluttajan käyttämästä ihmiskielestä saadaan käyttöön rahanarvoista mainontatietoa, jota yritys pystyy käyttämään joko itse tai myymään sen eniten tarjoavalle taholle. Rahanarvoisen hyödyn lisäksi luonnollisen kielen käsittely tarjoaa myös yleishyödyllisiä ratkaisuja, kuten vihapuheen esto ja roskapostisuodattimet. Tarve ihmiskielen koneelliseen ymmärrykseen ja haluun valjastaa kestävästi sen hyödyt ovat tuoneet mukanaan kiinnostuksen luonnollisen kielen käsittelyn tietoturvaan.

Lähteet

Boucher, N., Shumailov, I., Anderson, R. ja Papernot, N. (2021). *Bad Characters: Imperceptible NLP Attacks*. arXiv: 2106.09898 [cs.CL].

Chowdhury, G. G. (2003). "Natural language processing". *Annual Review of Information Science and Technology* 37.1, s. 51–89. DOI: https://doi.org/10.1002/aris.1440370103. eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370103. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103.

Garg, P. ja Girdhar, N. (2021). "A Systematic Review on Spam Filtering Techniques based on Natural Language Processing Framework". Teoksessa: *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, s. 30–35. DOI: 10.1109/Confluence51648.2021.9377042.

Gopalakrishnan, K. (2018). "Deep learning in data-driven pavement image analysis and automated distress detection: A review". *Data* 3.3, s. 28.

Gröndahl, T., Pajola, L., Juuti, M., Conti, M. ja Asokan, N. (2018). "All You Need is "Love": Evading Hate Speech Detection". Teoksessa: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. AISec '18. Toronto, Canada: Association for Computing Machinery, s. 2–12. ISBN: 9781450360043. DOI: 10.1145/3270101.3270103. URL: https://doi.org/10.1145/3270101.3270103.

Hirschberg, J. ja Manning, C. D. (2015). "Advances in natural language processing". *Science* 349.6245, s. 261–266. DOI: 10.1126/science.aaa8685. eprint: https://www.science.org/doi/pdf/10.1126/science.aaa8685. URL: https://www.science.org/doi/abs/10.1126/science.aaa8685.

Huang, X., Alzantot, M. ja Srivastava, M. (2019). *NeuronInspect: Detecting Backdoors in Neural Networks via Output Explanations*. DOI: 10.48550/ARXIV.1911.07399. URL: https://arxiv.org/abs/1911.07399.

Jordan, M. I. ja Mitchell, T. M. (2015). "Machine learning: Trends, perspectives, and prospects". *Science* 349.6245, s. 255–260.

Khonji, M., Iraqi, Y. ja Jones, A. (2013). "Phishing Detection: A Literature Survey". *IEEE Communications Surveys & Tutorials* 15.4, s. 2091–2121. DOI: 10.1109/SURV.2013.032213.00009.

Kuchipudi, B., Nannapaneni, R. T. ja Liao, Q. (2020). "Adversarial Machine Learning for Spam Filters". Teoksessa: *Proceedings of the 15th International Conference on Availability, Reliability and Security*. ARES '20. Virtual Event, Ireland: Association for Computing Machinery. ISBN: 9781450388337. DOI: 10.1145/3407023.3407079. URL: https://doi.org/10.1145/3407023.3407079.

Moore, G. E. et al. (1965). *Cramming more components onto integrated circuits*.

Patel, C., Patel, A. ja Patel, D. (2012). "Optical character recognition by open source OCR tool tesseract: A case study". *International Journal of Computer Applications* 55.10, s. 50–56.

Schmidt, A. ja Wiegand, M. (huhtikuu 2017). "A Survey on Hate Speech Detection using Natural Language Processing". Teoksessa: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, s. 1–10. DOI: 10.18653/v1/W17-1101. URL: https://aclanthology.org/W17-1101.

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J. ja Lease, M. (2021). "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support". Teoksessa: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery. ISBN: 9781450380966. DOI: 10.1145/3411764.3445092. URL: https://doi.org/10.1145/3411764.3445092.

Trupthi, M., Pabboju, S. ja Gugulotu, N. (2019). "Deep Sentiments Extraction for Consumer Products Using NLP-Based Technique". Teoksessa: *Soft Computing and Signal Processing*. Toim. J. Wang, G. R. M. Reddy, V. K. Prasad ja V. S. Reddy. Singapore: Springer Singapore, s. 191–201. ISBN: 978-981-13-3393-4.

Yi, J., Nasukawa, T., Bunescu, R. ja Niblack, W. (2003). "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques". Teoksessa: *Third IEEE International Conference on Data Mining*, s. 427–434. DOI: 10.1109/ICDM.2003.1250949.

Zhou, Y., Zheng, X., Hsieh, C.-J., Chang, K.-w. ja Huang, X. (2020). *Defense against Adversarial Attacks in NLP via Dirichlet Neighborhood Ensemble*. DOI: 10.48550/ARXIV.2006.11627. URL: https://arxiv.org/abs/2006.11627.