Kanditutkielma

Tietojenkäsittelytieteen kandiohjelma

# Tekstipohjaiset vastakkaishyökkäykset NLP-malleja vastaan

Akira Taguchi

26.4.2022

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-malleja vastaan

Ohjaajat — Handledare — Supervisors

Prof. Nikolaj Tatti

| Työn laji — Arbetets art — Level | Aika — Datum — Month and year | Sivumäärä — Sidoantal — Number of pages |
|---|---|---|
| Kanditutkielma | 26.4.2022 | 13 sivua |

Tiivistelmä — Referat — Abstract

**ACM Computing Classification System (CCS)**
Security and privacy
Computing methodologies → Artificial Intelligence → Natural language processing

Avainsanat — Nyckelord — Keywords

nlp, unicode, nlp attack, machine learning, Natural Language Processing, cyber security, adversarial example

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 käsittelyy. Luonnollisen kielen käsittely (eng. Natural Language Processing, NLP) on alati kasvavassa kuluttajakäytössä johtuen laskentatehon kasvusta, suurien tietomäärien saatavuudesta, onnistuneiden koneoppimismetodien kehittämisestä sekä laajemmasta ihmiskielen ymmärryksestä ja sen käytöstä eri konteksteissa (Hirschberg ja Manning, 2015).

Luonnollisen kielen käsittely on kohdennetun mainonnan keskiössä. Viesti ystävälle mainoskohdennetussa viestipalvelussa antaa työstettävän datan NLP-mallille: “Mikä elokuva meidän pitäisi katsoa viikonloppuna? “ NLP-mallin avulla automaattinen mainostaja ymmärtää mainostaa kyseiselle käyttäjälle miltei välittömästi sarjalippuja mainostavasta elokuvateatterista, suoratoistopalvelua tai mainostavaa aktiviteettikeskusta kyseiselle viikonlopulle. Tämän rahanarvoisen tarpeen löytäminen datasta automaation avulla 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 on alettu pilkkomaan suoraviivaisemmaksi dataksi (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. Koska laskentateho on kasvanut huomattavasti vuosien saatossa (Moore et al., 1965), meillä on myös puhdasta rautaa käsitellä suurta määrää dataa.

Tässä tutkielmassa tarkastellaan NLP-hyökkäysten käyttökohteita. Tähän kuuluu hyökkäystaksonomia, puolustusmetodit sekä NLP-mallien sekä niihin kohdistuvien hyökkäysten tulevaisuus. Hyökkäystaksonomiassa käymme läpi erilaisia tapoja hyökätä NLP malleja vastaan, hyökkäysten tarkoituksiin ja onnistumistodennäköisyyksiin. Puolustusmetodit ovat tärkeässä osassa, jotta haavoittuvuuteen kohdistuvat firmat saavat ohjeita vahingon mitigointiin ja ennaltaehkäisyyn. Koska NLP-mallit ovat eksponentiaalisessa nousussa kuluttajakäytössä, on tärkeää spekuloida mahdollisia kehityksiä koneoppimisessa sekä tästä syntyviä haavoittuvuuksia. Lopuksi käymme läpi mahdollisia luonnollisen kielen käyttö-

kohteita tulevaisuudesa sekä näistä aiheutuvia seurauksia eri osa-alueisiin akateemisella että kaupallisella puolella.

# 2 Hyökkäystaksonomia

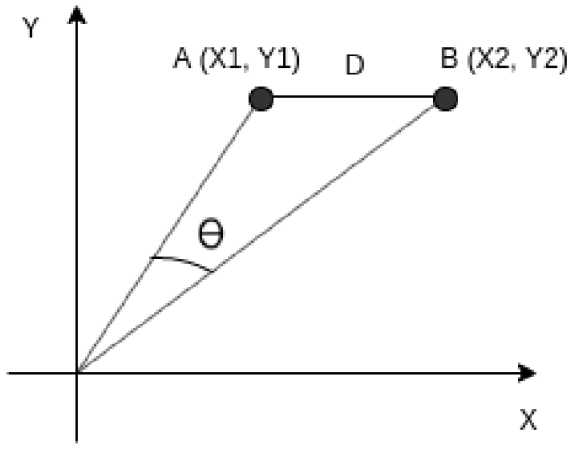
Käydään läpi hyökkästaksonomia, eli hyökkäysrajapinta, NLP-malleja vastaan.

## 2.1 Roskapostisuodatuksen ohitus

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

Suodattimia vastaan toimii kolme vastakkaishyökkäystä. Synonyymin korvaus, kelposanan injektointi sekä roskapostisanojen väljennys. Sana ”kelpo” tarkoittaa tässä yhteydessä tekstiä, jonka roskapostisuodatin on merkinnyt hyväntahtoiseksi. Synonyymin korvauksessa tarkoitus on korvata pahantahtoiset sanat hyväntahtoisiksi luokitelluilla synonyymeillä (taulukko 2.1). Sanojen synonyymisyys lasketaan taulukon 2.1 tapauksessa kuvan 2.1 kosini-samankaltaisuudella. Kelposanan injektoinnissa kelposanoja lisätään sähköpostiin niin paljon, kunnes NLP-malli tunnistaa roskapostin olevan kelpopostia. Kelposanoja voidaan injektoida tietokannoista roskaposteihin muuttamatta viestin tarkoitusta rajusti. Roskapostisanojen väljennyksessä roskapostisanoihin sisällytetään välilyöntejä, jotta NLP-malli ei tunnistaisi näitä sanoja roskasanoiksi. Kun väljennystä on harjoitettu tarpeeksi, muuttuu roskaposti NLP-mallin näkökulmasta kelpopostiksi. (Kuchipudi et al., 2020)

Kelposanan injektoinnille ja roskasanojen väljennykselle on olemassa erilaisia implementaatioita. Seuraavissa aliluvuissa tutustutaan ladontapohjaisiin vastakkaishyökkäyksiin. Muun muassa näitä hyökkäysmetodeita voidaan käyttää kahdessa aiemmin mainitussa roskapostisuodattimeen kohdistetussa hyökkäyksessä. Implementaatioita yhdistelemällä ja vaihtelemalla, saattaa NLP-mallin pahantahtoisuuden havaitseminen heikentyä entisestään, taaten hyökkääjälle varmemman onnistumisen.




**Kuva 2.1:** Sähköpostien samankaltaisuus voidaan laskea käyttäen kosini-samankaltaisuutta kahden sähköpostivektorin välillä. (Kuchipudi et al., 2020)

## 2.2 Neuroverkkohyökkäykset

Sanatason vastakkaishyökkäykset syvää oppimisverkkoa vastaan paljastavat näiden verkkojen heikkouksia. Vastakkaishyökkääminen näitä verkkoja vastaan on vaikeaa verrattuna kuvatason hyökkäyksiin. Tämä johtuu lauseiden diskreetistä luonteesta. Pienikin sana - tai merkkimuutos vaikuttaa lauseen viestiin sekä todennäköisemmin viestin luokitukseen, jonka NLP-malli päättää. (Zang et al., 2020)

## 2.3 Sensuurin ohitus

Kuvien sensuroimista internetissä voidaan soveltaa käyttäen syvää oppimisverkkoa, kuten Martin D. More, Douglas M. Souza, Jônatas Wehrmann, Rodrigo C Barros osoittavat julkaisuussaan *Seamless Nudity Censorship: an Image-to-Image Translation Approach based on Adversarial Training*. Syvää oppimisverkkoa hyödynnetään tunnistamaan intiimialueet ja tämän jälkeen vaatetuttamaan kuvien henkilöiden intiimialueet. (More et al., 2018)

Koska sensuuria voidaan soveltaa hyödyntäen koneoppimismalleja, voidaan sensuuri myös ohittaa hyödyntäen koneoppimismallin heikkouksia. Vastakkaishyökkäys voisi tunnistaa sensurointia aiheuttavia pikseliyhdistelmiä, ja tässä tutkimuksessa esiteltyjä hyökkäystapoja käyttäen sensuurin laukaiseminen voidaan estää. Tällöin kyseessä ei kuitenkaan enää ole puhdas merkintä (eng. clean label), sillä vastakkaishyökkäyksen todellinen tarkoitus

| Muokattu viesti | Kosini-samankaltaisuus | 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. | 1 | 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. | 0, 583 | 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. | 0, 583 | 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. | 0, 583 | kelpopostia |

**Taulukko 2.1:** Synonyymin korvaus. Vanhan viestin korvatut osat on lihavoitu. (Kuchipudi et al., 2020)

näkyy käyttäjälle silmintarkasteltavana (Gan et al., 2021). Puhtaan merkinnän uupuessa esimerkiksi tekstipohjaisesti vastakkaishyökkäyksestä myös helppo puolustaminen on mahdollista (Pruthi et al., 2019).

## 2.4 Näkymättömät merkit

Näkymättömät merkit vaikuttavat tietokoneen NLP-mallin ymmärtämään kontekstiin. Esimerkki tällaisesta on nollatilavuuden välilyönti -merkki, jonka Unicode merkintä on `U+200B`. Tällä merkillä voimme esimerkiksi vaikuttaa pelichattiin lähetettävän toksissuodatettavan merkkijonoon "olet huono"niin, että merkkijono menisi NLP-mallin läpi chätistä. Merkkijono `olU+200Bet huU+200Bono` saattaisi mennä läpi chatin suodattimesta, mutta vastapuolelle viesti olisi edelleen `olet huono`. (Boucher et al., 2021)

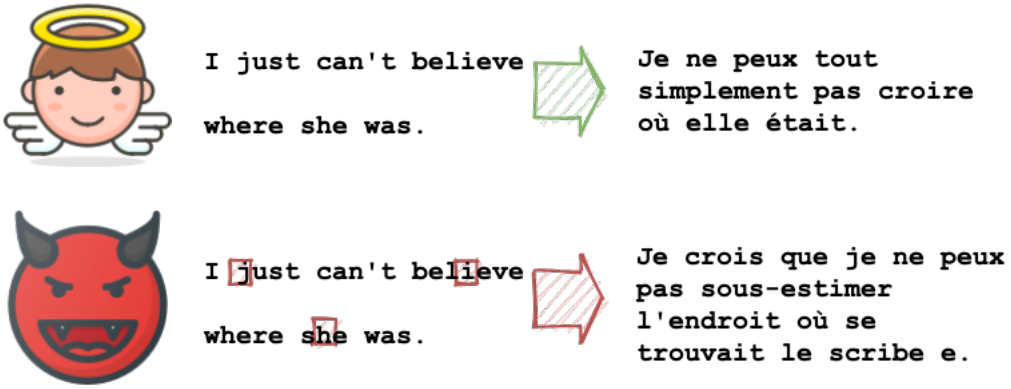
Kontekstin poistamisen lisäksi näkymättömillä merkeillä voidaan myös tuoda ja syrjäyttää konteksteja toisilla.
`Mikä pyhäinhäväistyksen rakennus!`
`Miten onnistuit tekemään tämän näin laiskasti?` -tekstin negatiivisuus voidaan

syrjäyttää positiivisuudella syöttämällä NLP-mallille sen sijaan teksti
`Mikä pyU+200Bhäinhävu+200Bäistyu+200Bksenu+200B rakennus!`
`Miten onnistuit tekemään tämän U+200Bnäu+200Bin lau+200Biskasu+200Bti?`. (Boucher et al., 2021)

Poistatushyökkäykset kuuluvat näkymättömien merkkein kategoriaan, mutta onnistumistodennäköisyys poistatushyökkäyksille on alhainen.Poistatushyökkäyksiä on vaikeampi toteuttaa aikaisempiin metodeihin verrattuna. Tämä johtuu useimpien käyttöjärjestelmien estosta kopioida poistatusta sisältävää tekstiä leikepöydälle. (Boucher et al., 2021)

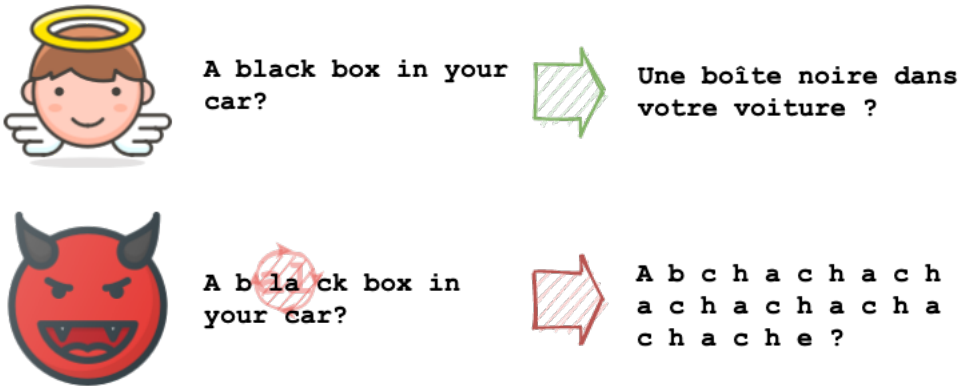
## 2.5 Homoglyfit

Homoglyyfihyökkäykset NLP-malleja vastaan pohjautuvat pahantahtoisten merkkien virallisten esitysmuotojen näyttävän hyväntahtoisten merkkien virallisilta esityksiltä.Joissain kielissä tekstin merkitys muuttuu täysin yhden merkin vaihtuessa. Esimerkkinä homoglyyfistä on `A` → `A`, missä viimeinen kirjain on todellisuudessa kyrillinen kirjain A. Kuvassa 2.1 homoglyyfihyökkäys 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`.




**Kuva 2.2:** Homoglyyfihyökkäys (Boucher et al., 2021)

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




**Kuva 2.3:** Homoglyyfihyökkäys (Boucher et al., 2021)

## 2.6 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`. Uudelleenjärjestelyjä käytetään myös NLP-mallin sekoittamiseen, jolloin tulokset NLP-mallista ovat käyttökelvottomia. Kuvassa 2.2 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.

# 3 Puolustusmetodit

## 3.1 OCR-puolustus

NLP-hyökkäykset voidaan estää alhaisemmalla tasolla korkealla yleisrasituksella sekä korkeammalla tasolla edistyneen teknologian turvin (Boucher et al., 2021). Näytöltäluvun (eng. OCR, On-Screen-Reading) avulla epäselvyydet tekstin aidosta luonteesta voidaan uudelleenrenderöidä tulkitsemalla aineisto uudestaan visuaalisesti. Tämä metodi lisää yleisrasitusta huomattavasti riippuen käyttötarkoituksesta, mutta poistaa pahantahtoiset merkit ilman NLP-mallin uudelleenkoulutusta.

## 3.2 Suorituskykykeskeinen puolustus

Keskitymme seuraavaksi näkymättömiin merkkeihin, -homoglyyfeihin, uudelleenjärjestelyihin -ja poistatuksiin perustuvien hyökkyksien puolustamiseen. Suorituskykykeskeiset puolustusmetodit ovat kuitenkin laskennallisesti kalliita, eivätkä koneoppimismallin ulkoistaneet firmat pysty kustantamaan kyseisiä metodeita (Huang et al., 2019).

Tietyt näkymättömät merkit voidaan poistaa suoraa syötteestä. Mikäli applikaatiossa näitä merkkejä ei voida poistaa, voidaan ne korvata non-<unk> upotuksilla.

Homoglyyfihyökkäysten torjuminen OCR-metodilla on ymmärrettävästi vaikeampaa verrattuina muihin merkkeihin. Paras keino torjua tällaisia hyökkäyksiä olisi mapata osa homoglyyfeistä niiden yleisemmin tunnettuihin vastineisiin. NLP-mallin ylläpitäjä joutuu tekemään tässä siis suurimman jalkatyön.

Uudelleenjärjestelyhyökkäykset voidaan torjua riisumalla kaksisuuntais-ohjausmerkit syötteestä, varoittamalla käyttäjää kaksisuuntais-ohjausmerkkien ilmestyessä syötteeseen tai käyttämällä kaksisuuntais-algoritmia halutun syötteen selvittämiseen. Puolustusmetodin valinta riippuu kontekstista, sillä esimerkiksi latinaa tai arabiaa kirjoittaessa ohjelma toimisi väärin pakottamalla käyttäjän syötteestä pois kaksisuuntais-ohjausmerkin `U+200F`.

Poistatukset yleensä havaitaan NLP-mallien ulkopuolella syötteen annon alkuvaiheessa. NLP-mallin 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 hyökkäysrajapintaan.

## 3.3 Lujitus käyttäen lineaarisia luokittimia

NLP-malleja voidaan lujittaa käyttämällä tukivektorikonetta, mikä perustuu lineaarisiin luokittimiin. Xupeng Shi ja A. Adam Ding käsittelevät tukivektorikoneen lujitusmahdollisuuksia tuktimuksessaan *Understanding and Quantifying Adversarial Examples Existence in Linear Classification*. Tutkielma pohjautuu kahteen olennaiseen määritelmään:

**Määritelmä 1.** Olkoon luokitin $C$. Datavektorin $x$ vastakkaisesimerkki $\varepsilon$ on toinen datavektori $x'$ niin, että $||x - x'|| \leq \varepsilon$, mutta $C(x) \neq C(x')$.

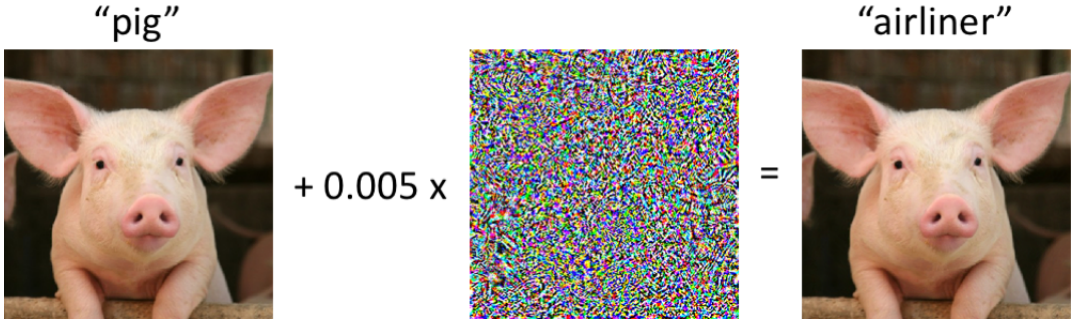
**Määritelmä 2.** Olkoon luokitin $C$. Datavektorin $x$ vahva vastakkaisesimerkki $(\varepsilon, \delta)$ on toinen datavektori $x'$ niin, että $||x - x'|| \leq \varepsilon$ ja $|(x - x') \cdot \mu| \leq \delta$, mutta $C(x) \neq C(x')$.

Tutkielmassa ehdotetaan puolustusmetodeita perustuen ylläoleviin määritelmiin sekä niiden sovelluksiin käyttäen lineaarisia luokittimia.

# 4 Skaalautuvuus

Neljän aikaisemmin mainitun NLP-mallin mahdollistajien kehittyessä arvaamattomasti, on loogista tutkia NLP-hyökkäysten tulevaisuutta.

Koska NLP-mallit ovat jo nyt raskaassa kuluttajakäytössä, kohdistuvat haavoittuvuudet myös tulevaisuudessa kuluttajapuolen NLP-malleihin. Koska osa NLP-hyökkäyksistä on lähestulkoon huomaamattomia (**DBLP:journals/corr/abs-2111-07970**), voidaan hyökkäyksiä suorittaa yhä enemmän osapuolista ja intresseistä riippumatta. Ympäristöaktivistit saattavat haluta manipuoloida Googlen kuvahaun NLP-Mallin liittämään possuihin liittyneen ympäristöinsidentin lentokoneyhtiöön x (kuva 4.1).




**Kuva 4.1:** Hyökkäys lentoyhtiötä kohtaan. (Mądry ja Schmidt, 2018)

Vastakkaishyökkäysten motiivit muovautuvat siis ajan myötä ja kasvattavat tahtomattaan näin hyökkäystaksonomiaa.

Hyökkäystaksonomia laajentuu tulevaisuudessa eri formaatteihin. Koneoppimisen kukoistaessa voidaan NLP-malleja soveltaa tiedon ääni -tai videoformaatteihin. Tämä antaa puolestaan mahdollisuuden vastakkaishyökätä kyseiseen koneoppimismallia vastaan. Formaattien sisältäkin löytyy erinäisiä hyökkäysrajapintoja. 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.

Haavoittuvuuksien löyty ruokkii itse itseään. Ensimmäisten vastakkaishyökkäysten kohdistuessa uuteen tietoformaattiin, syntyy tarve puolustukseen tätä vastaan. Toteutuksesta riippuen puolustusmetodin selvittäminen saattaa avata uusia ovia, jotka hyödyttävät hyökkääjiä.

# 5 Yhteenveto

Luonnolisen kielen käsittely on kätevä työkalu käsittelemään ihmisten puhumaa kieltä tietokonemaailmassa.

Luonnollisen kielen käsittely on kuitenkin sellaisenaan haavoittuvainen erillaisille tekstipohjaisille hyökkäyksille.

Tässä tutkielmassa tutustuimme näiden luonnollisen kielen prosessoinnin historiaan, aikaisemmin mainittujen hyökkäysten mahdollistajiin sekä näiden torjuntametodeihin.

# 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.

Gan, L., Li, J., Zhang, T., Li, X., Meng, Y., Wu, F., Guo, S. ja Fan, C. (2021). "Triggerless Backdoor Attack for NLP Tasks with Clean Labels". *CoRR* abs/2111.07970. arXiv: 2111.07970. URL: https://arxiv.org/abs/2111.07970.

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

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.

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.

Mądry, A. ja Schmidt, L. (2018). "A Brief Introduction to Adversarial Examples". DOI: 10.1126/science.aaa8685. URL: https://gradientscience.org/intro_adversarial/.

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

More, M. D., Souza, D. M., Wehrmann, J. ja Barros, R. C. (2018). "Seamless Nudity Censorship: an Image-to-Image Translation Approach based on Adversarial Training".

Teoksessa: *2018 International Joint Conference on Neural Networks (IJCNN)*, s. 1–8. DOI: 10.1109/IJCNN.2018.8489407.

Pruthi, D., Dhingra, B. ja Lipton, Z. C. (heinäkuu 2019). "Combating Adversarial Misspellings with Robust Word Recognition". Teoksessa: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, s. 5582–5591. DOI: 10.18653/v1/P19-1561. URL: https://aclanthology.org/P19-1561.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q. ja Sun, M. (2020). "Word-level Textual Adversarial Attacking as Combinatorial Optimization". Teoksessa: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.540. URL: https://doi.org/10.18653%2Fv1%2F2020.acl-main.540.