



Kandidatutkielma

Tietojenkäsittelytieteen kandiohjelma

# **Tekstipohjaiset vastakkaishyökkäykset NLP-malleja vastaan**

Akira Taguchi

19.4.2022

MATEMAATTIS-LUONNONTIETEELLINEN TIEDEKUNTA  
HELSINGIN YLIOPISTO

## Ohjaaja(t)

Prof. Nikolaj Tatti

## Yhteystiedot

PL 68 (Pietari Kalmin katu 5)  
00014 Helsingin yliopisto

Sähköpostiosoite: [info@cs.helsinki.fi](mailto: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	
Kandidutkielma	19.4.2022	11 sivua	
Tiivistelmä — Referat — Abstract			
<p><b>ACM Computing Classification System (CCS)</b></p> <p>Security and privacy</p> <p>Computing methodologies → Artificial Intelligence → Natural language processing</p>			
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

<b>1</b>	<b>Johdanto</b>	<b>1</b>
<b>2</b>	<b>Hyökkäystaksonomia</b>	<b>3</b>
2.1	Roskapostisuodatuksen ohitus . . . . .	3
2.2	Näkymättömät merkit . . . . .	3
2.3	Homoglyfit . . . . .	4
2.4	Uudelleenjärjestelyt . . . . .	5
2.5	Poistatukset . . . . .	6
<b>3</b>	<b>Puolustusmetodit</b>	<b>7</b>
3.1	OCR-puolustus . . . . .	7
3.2	Suorituskykykeskeinen puolustus . . . . .	7
<b>4</b>	<b>Skaalautuvuus</b>	<b>9</b>
4.1	Tulevaisuus NLP-hyökkäyksille . . . . .	9
<b>5</b>	<b>Yhteenveto</b>	<b>10</b>
	<b>Lähteet</b>	<b>11</b>



# 1 Johdanto

Koneoppimisen käyttötarkoitusten määrä kasvaa vuosi vuodelta suuremmaksi. Tätä teknologiaa voidaan hyödyntää muun muassa ihmisten puhuman kielen käsittely. 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 viikonloppulle. 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, puolustusmenetelmät sekä NLP-mallien sekä niihin kohdistuvien hyökkäysten tulevaisuus. Hyökkäystaksonomiassa käymme läpi erilaisia tapoja hyökätä NLP malleja vastaan, hyökkäysten tarkoituksiin ja onnistumistodennäköisyyksiin. Puolustusmenetelmät 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 hyökkästäksonomia, eli hyökkäysrajapinta, NLP-malleja vastaan.

### 2.1 Roskapostisuodatuksen ohitus

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

Suodattimia vastaan toimii kolme vastakkaishyökkäystä. Synonyymien korvaus, kelposan injektointi sekä roskapostisanojen väljennys. Sana ”kelpo” tarkoittaa tässä yhteydessä tekstiä, jonka roskapostisuodatin on merkinnyt hyväntahtoiseksi. Synonyymien korvauksessa tarkoitus on korvata pahantahtoiset sanat hyväntahtoisiksi luokitelluilla synonyymeillä (taulukko 2.1).

### 2.2 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.

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?.

Muokattu viesti	Kosini-saman-kaltaisuus	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: <b>acquire</b> the UK single <b>graph</b> on your <b>Mobile_River</b> each <b>hebdomad</b> and <b>take</b> any <b>top_side caliber</b> ringtone! This <b>content</b> is <b>free_people</b> of charge.	0,583	roskapostia
Ringtone Club: <b>become</b> the UK <b>bingle graph</b> on your <b>nomadic</b> each <b>workweek</b> and <b>select</b> any <b>upper_side caliber</b> ringtone! This <b>subject_matter</b> is <b>liberate</b> of charge.	0,583	roskapostia
Ringtone Club: <b>go</b> the UK <b>one graph</b> on your <b>peregrine</b> each <b>calendar_week</b> and <b>pick_out</b> any <b>upside character</b> ringtone! This <b>substance</b> is <b>release</b> of charge.	0,583	kelpopostia

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

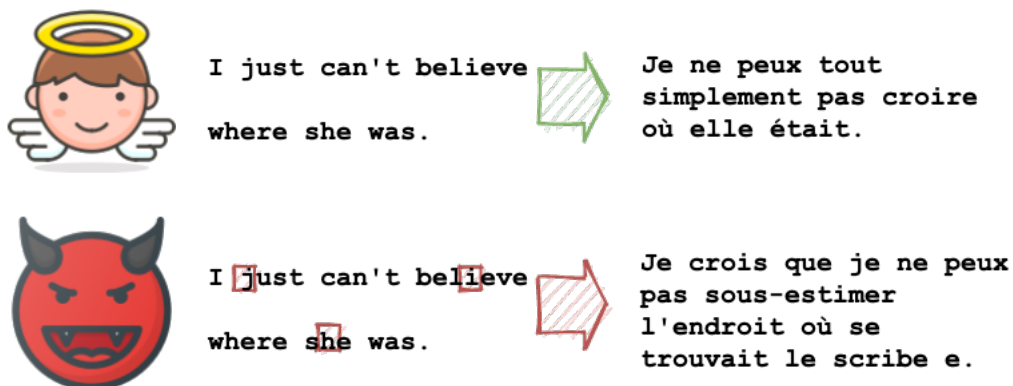
## 2.3 Homoglyfit

Homoglyfyihyökkäykset 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  $A \rightarrow A$ , missä viimeinen kirjain on todellisuudessa kyrillinen kirjain A. Kuvassa 2.1 homoglyfyihyökkäys on muuntanut englanninkielisen tekstin

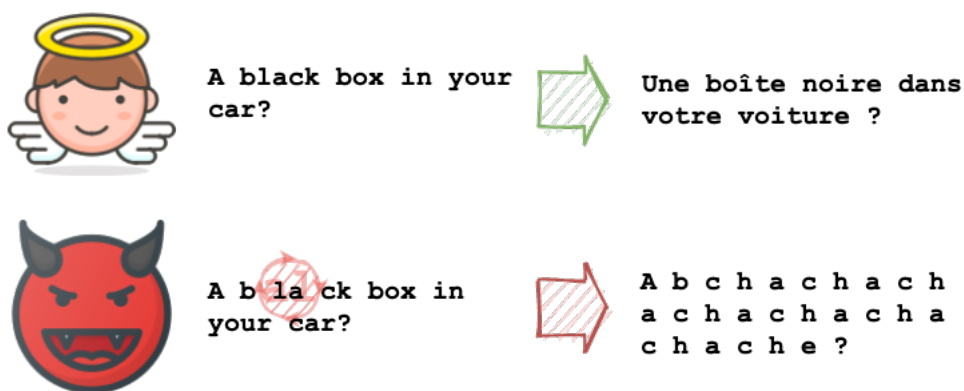
I just can't belive where she was ranskankieliseen käännökseen

I guess I can't underestimate the location of the scribe and.

Näkymättömien merkkien lailla homoglyfyihyökkäyksen toteutus riippuu ympäristön fontista.



Kuva 2.1: Homoglyfihyökkäys (Boucher et al., 2021)



Kuva 2.2: Homoglyfihyökkäys (Boucher et al., 2021)

## 2.4 Uudelleenjärjestelyt

Uudelleenjärjestelyhyökkäys pohjautuu näennäisen tekstin uudelleenjärjestämiseen pahan-  
tahtoisesti. Pankkitilinumeron 1234567 pystyy esimerkiksi vaihtamaan kaksisuuntaisella-  
algoritmillä tilinumeroksi 7654321 pankin palvelinpuolella. maksajan huomaamatta mi-  
tään. Unicode-merkintä tälle suunnanvaihdolle on U+200F. Uudelleenjärjestelyjä käyte-  
tään myös NLP-mallin sekoittamiseen, jolloin tulokset NLP-mallista ovat käyttökeltotomia. Kuvassa 2.2 uudelleenjärjestelyhyökkäys merkeissä la aiheuttaa ranskankielisen  
käännöksen järjettömyyden. Tämänlaista hyökkäystä voisi käyttää digitaalista sanakir-  
jaa tai kääntäjää vastaan. U+200F ladotaan näkymättömänä näkymättömien merkkien  
tapaan.

## 2.5 Poistatukset

Poistatushyökkäyksiä on vaikeampi toteuttaa aikaisempiin metodeihin verrattuna. Tämä johtuu useimpien käyttöjärjestelmien estosta kopioida poistatusta sisältävää tekstiä leikkipöydälle.

# 3 Puolustusmetodit

## 3.1 OCR-puolustus

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

## 3.2 Suorituskykykeskeinen puolustus

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

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

Homoglyyfihyökkäysten torjuminen OCR-metodilla on ymmärrettävästi vaikeampaa verrattuna 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 syötteenannon alkuvaiheessa. NLP-mallin tasolla tähän tarvitsee harvemmin puuttua ja käyttäjälle voidaan pahimmassa tapauksessa lähettää varoitus poistatusmerkkien olemassaolosta syötteessä. On silti

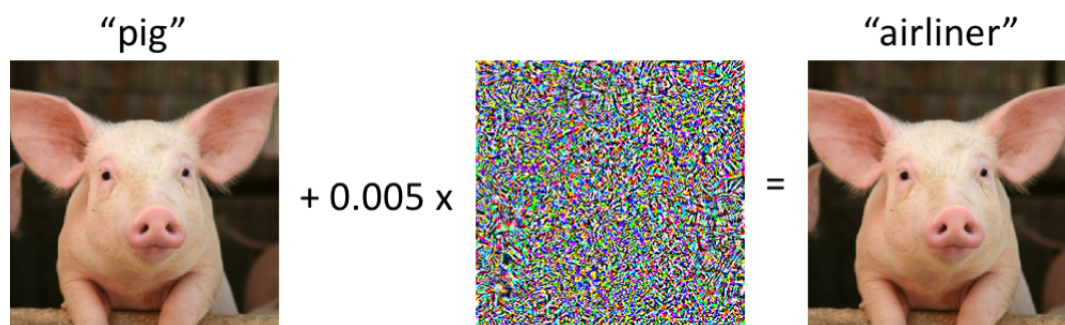
tärkeää tiedostaa poistatuksien puolustus, mikäli käyttöjärjestelmä unohtaa puuttua kyseiseen hyökkäysrajapintaan.

## 4 Skaalautuvuus

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

### 4.1 Tulevaisuus NLP-hyökkäyksille

Koska NLP-mallit ovat jo nyt raskaassa kuluttajakäytössä, kohdistuvat haavoittuvuudet myös tulevaisuudessa kuluttajapuolen NLP-malleihin. Koska osa NLP-hyökkäyksistä on lähestulkoon huomaamattomia (Gan et al., 2021), voidaan hyökkäyksiä suorittaa yhä enemmän osapuolista ja intresseistä riippumatta. Ympäristöaktivistit saattavat haluta manipuloida Googlen kuvahaun NLP-Mallin liittämään possuihin liittyneen ympäristöinsidentin lentokoneyhtiöön x (kuva 4.1).



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

# 5 Yhteenveto

Luonnollisen 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](https://arxiv.org/abs/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](https://arxiv.org/abs/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](https://doi.org/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](https://arxiv.org/abs/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](https://doi.org/10.1145/3407023.3407079). URL: <https://doi.org/10.1145/3407023.3407079>.
- Mađry, A. ja Schmidt, L. (2018). "A Brief Introduction to Adversarial Examples". DOI: [10.1126/science.aaa8685](https://arxiv.org/abs/1802.03740). URL: [https://gradientscience.org/intro\\_adversarial/](https://gradientscience.org/intro_adversarial/).
- Moore, G. E. et al. (1965). *Cramming more components onto integrated circuits*.

