



Kandidatutkielma

Tietojenkäsittelytieteen kandiohjelma

Tekstipohjaiset vastakkaishyökkäykset NLP-malleja vastaan

Akira Taguchi

7.5.2022

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
URL: <http://www.cs.helsinki.fi/>

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	7.5.2022	14 sivua	
Tiivistelmä — Referat — Abstract			
<p>Tässä tutkimuksessa käsitellään tekstipohjaisia vastakkaishyökkäyksiä NLP-malleja vastaan. Tutustutaan hyökkäystaksonomiaan sekä puolustusmetodeihin yleisimmillä osa-alueilla. Tarkastellaan myös vastakkaishyökkäysten tulevaisuutta teknologian ja yhteiskunnallisten rakenteiden kehittyessä.</p> <p>Tullaan tunnistamaan vastakkaishyökkäysten rooli nykYTEknologiassa, sekä syyt merkittävyyteen koneoppimismallien käytössä. Huomaamaan myös hyökkäyspinta-alan kattavan laajeman alueen puolustuspinta-alan verrattuna vastakkaishyökkäyksissä. Huomataan syyn olevan jo-valjastettu, mutta reunatapauksissa hallitsematon tekoälyn voima.</p>			
<p>ACM Computing Classification System (CCS) Security and privacy → Human and societal aspects of security and privacy Computing methodologies → Artificial Intelligence → Natural language processing</p>			
Avainsanat — Nyckelord — Keywords			
Luonnollisen kielen käsittely, vastakkaishyökkäys, koneoppiminen, tekoäly, ladonta, sensuuri			
Säilytyspaikka — Förvaringsställe — Where deposited			
Helsingin yliopiston kirjasto			
Muita tietoja — övriga uppgifter — Additional information			

Sisältö

1	Johdanto	1
2	Hyökkäystaksonomia	3
2.1	Roskapostisuodatuksen ohitus	3
2.2	Neuroverkkohyökkäykset	4
2.3	Sensuurin ohitus	4
2.4	Näkymättömät merkit	5
2.5	Homoglyfit	6
2.6	Uudelleenjärjestelyt	7
3	Puolustusmenetelmät	8
3.1	OCR-puolustus	8
3.2	Suorituskykykeskeinen puolustus	8
3.3	Lujitus käyttäen lineaarisia luokittimia	9
4	Skaalautuvuus	10
5	Yhteenveto	12
	Lähteet	13

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 seuraavista syistä:

- laskentatehon kasvusta
- suurien tietomäärien saatavuudesta
- onnistuneiden koneoppimismenetelmien 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, 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 kuuluu hyökkäystaksonomia, puolustusmenetelmät sekä NLP-mallien sekä niihin kohdistuvien hyök-

käysten tulevaisuus. Hyökkäystaksonomiassa 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 Hyökkäystaksonomia

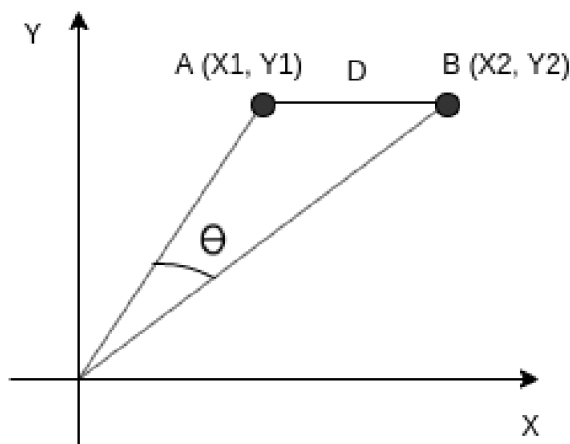
Käydään läpi hyökkästäksonomia, eli hyökkäysrajapinta, NLP-malleja vastaan.

2.1 Roskapostisuodatuksen ohitus

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

Suodattimia vastaan toimii kolme vastakkaishyökkäystä. Synonyymien korvaus, kelposanan 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). 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 entistään, 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 julkaisuissaan *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: Synonyymien korvaus. Vanhan viestin korvatut osat on lihavoitu. (Kuchipudi et al., 2020)

näkyä käyttäjälle silmintarkasteltavana (Gan et al., 2021). Puhtaan merkinnän uupues-
sa esimerkiksi tekstipohjaisesti vastakkaishyökkäyksestä myös helppo puolustaminen on
mahdollista (Pruthi et al., 2019).

2.4 Näkymättömät merkit

Näkymättömät merkit vaikuttavat tietokoneen NLP-mallin ymmärtämään sisältöön. Ky-
seinen hyökkäys perustuu Unicode-merkistöstandardiin, joka sisältää yksilöivät koodiarvot
kirjoitushetkellä yli 100 000 kirjoitusmerkille. Kuuluvat aakkoset sekä erikoismerkit.

Esimerkki tällaisesta erikoismerkistä on nollatilavuuden välilyönti -merkki, jonka Unico-
de merkintä on U+200B. Tällä merkillä voimme esimerkiksi vaikuttaa pelichattiin lähe-
tettävän myrkyllissuodatettavaan merkkijonoon "olet huono"niin, että merkkijono menisi
NLP-mallin läpi chätistä. Merkkijono olU+200Bet huU+200Bono saattaisi mennä läpi cha-
tin 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)

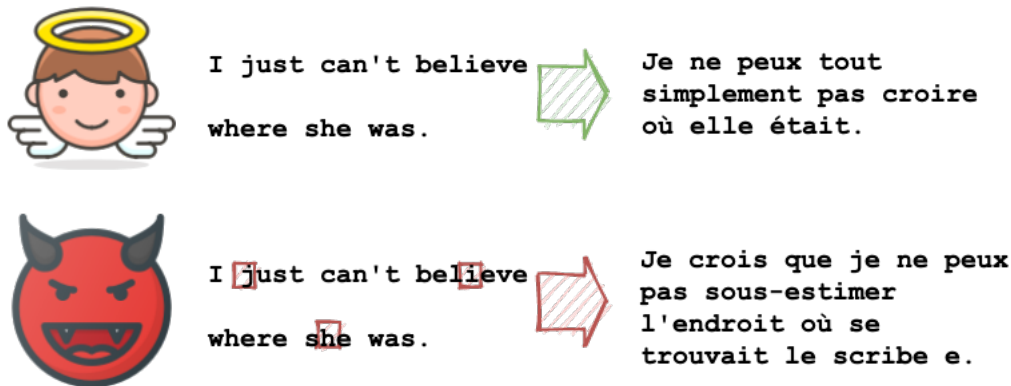
Poistatushyökkäykset kuuluvat näkymättömien merkkein kategoriaan, mutta onnistumistodennäköisyys poistatushyökkäyksille on alhainen. Poistatushyökkäyksiä on vaikeampi toteuttaa aikaisempiin metodeihin verrattuna. Tämä johtuu useimpien käyttöjärjestelmien estosta kopioida poistatusta sisältävää tekstiä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

2.5 Homoglyfit

Homoglyfihyökkäykset NLP-malleja vastaan pohjautuvat pahantahtoisten merkkien virallisten esitysmuotojen näyttävän hyväntahtoisten 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. Kuvassa 2.1 homoglyfihyökkäys on muuntanut englanninkielisen tekstin

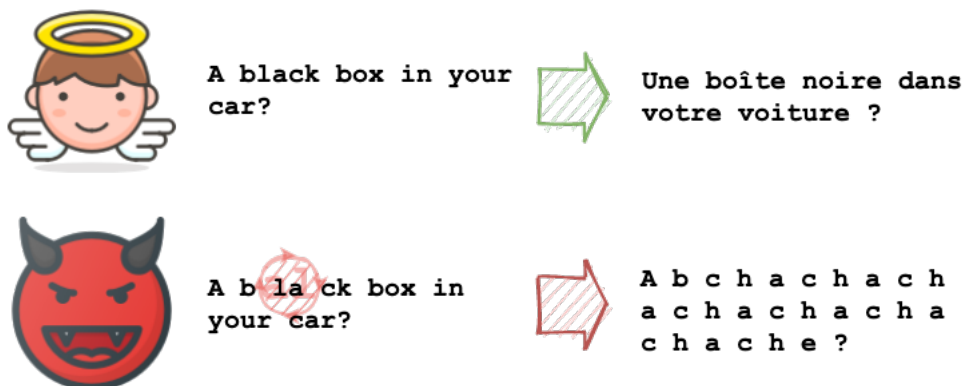
I just can't believe where she was ranskankieliseen käännökseen

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



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

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



Kuva 2.3: Homoglyfihyökkäys (Boucher et al., 2021)

2.6 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 sanakirjaa
tai kääntäjää vastaan. (Boucher et al., 2021) U+200F ladotaan näkymättömänä näkymät-
tömien merkkien tapaan.

3 Puolustusmenetelmät

Käydään läpi puolustusmenetelmät NLP-hyökkäyksiä vastaan.

3.1 OCR-puolustus

NLP-hyökkäykset voidaan torjua korkean tason abstraktiolla korkealla yleisrasituksella sekä alemman tason abstraktiolla alemmalla yleisrasituksella. Näytöltäluvun (eng. OCR, On-Screen-Reading) 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-mallin uudelleenkoulutusta. (Boucher et al., 2021)

3.2 Suorituskykykeskeinen puolustus

Keskitymme seuraavaksi näkymättömiin merkkeihin, -homoglyfeihin, uudelleenjärjestelyihin -ja poistatuksiin perustuvien hyökkäysten puolustamiseen. Suorituskykykeskeiset puolustusmenetelmät ovat kuitenkin laskennallisesti kalliita, eivätkä koneoppimismallin ulkoistaneet yritykset pysty kustantamaan kyseisiä metodeita (Huang et al., 2019).

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, joista ladontamoottori ei ole aivan varma. (Boucher et al., 2021)

Homoglyfihyökkäysten torjuminen OCR-menetelmällä on ymmärrettävästi vaikeampaa verrattuina muihin merkkeihin. Paras keino torjua tällaisia hyökkäyksiä olisi kuvata osa homoglyfeistä niiden yleisemmin tunnettuihin vastineisiin. NLP-mallin ylläpitäjä joutuu tekemään tässä siis suurimman työn. (Boucher et al., 2021)

Uudelleenjärjestelyhyökkäykset voidaan torjua riisumalla kaksisuuntais-ohjausmerkit syötteestä, varoittamalla käyttäjää kaksisuuntais-ohjausmerkkien ilmestyessä syötteeseen tai käyttämällä kaksisuuntais-algoritmia halutun syötteen selvittämiseen. Puolustusmene-

telmän valinta riippuu kontekstista, sillä esimerkiksi latinaa tai arabiaa kirjoittaessa ohjelma toimisi väärin pakottamalla käyttäjän syötteestä pois kaksisuuntais-ohjausmerkin U+200F. (Boucher et al., 2021)

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. (Boucher et al., 2021)

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 tulkimuksissaan *Understanding and Quantifying Adversarial Examples Existence in Linear Classification*. Tutkielma pohjautuu kahteen olennaiseen määritelmään:

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

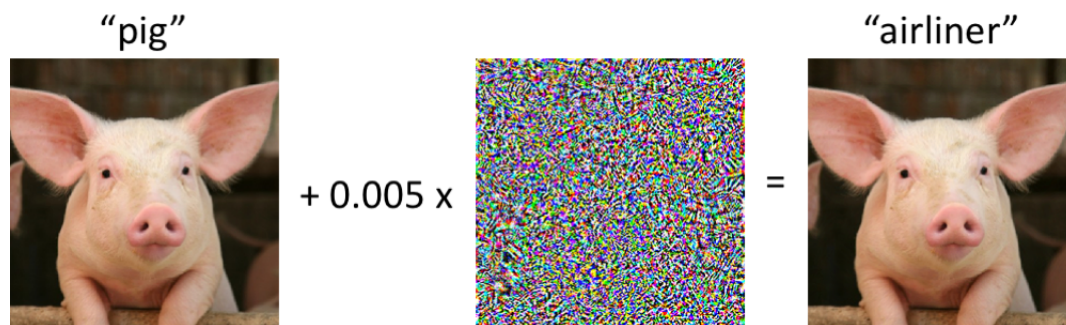
Määritelmä 2. Olkoon luokitin C . Datavektorin x vahva vastakkaisesimerkki (ε, δ) on toinen datavektori x' niin, että $\|x - x'\| \leq \varepsilon$ ja $|(x - x') \cdot \mu| \leq \delta$, mutta $C(x) \neq C(x')$.

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.

NLP-mallien haavoittuvuudet tulevat olemaan tulevaisuudessa näkyvämpiä kuluttajapuolella. 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).



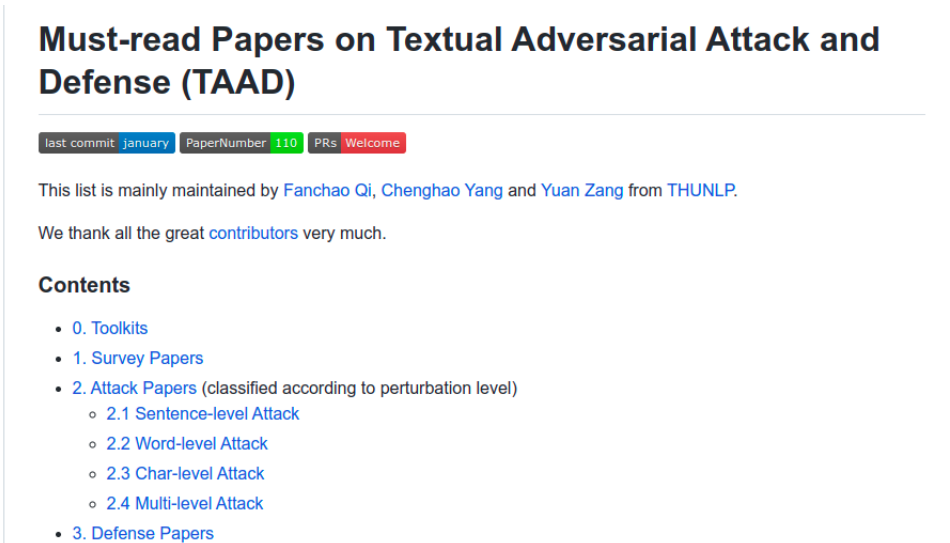
Kuva 4.1: Hyökkäys lentoyhtiötä kohtaan. (Mądry ja Schmidt, 2018)

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 löytö ruokkii itse itseään. Ensimmäisten vastakkaishyökkäysten kohdistuessa uuteen tietformaattiin, 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. Tämä nä-

kyy muun muassa GitHub-repositoriossa *Must-read Papers on Textual Adversarial Attack and Defense (TAAD)*, jossa hyökkäystutkimusten määrä suhteessa puolustustutkimusten määrään on 75 : 23 (kuva 4.2).



Must-read Papers on Textual Adversarial Attack and Defense (TAAD)

last commit January PaperNumber 110 PRs Welcome

This list is mainly maintained by [Fanchao Qi](#), [Chenghao Yang](#) and [Yuan Zang](#) from [THUNLP](#).

We thank all the great [contributors](#) very much.

Contents

- [0. Toolkits](#)
- [1. Survey Papers](#)
- [2. Attack Papers](#) (classified according to perturbation level)
 - [2.1 Sentence-level Attack](#)
 - [2.2 Word-level Attack](#)
 - [2.3 Char-level Attack](#)
 - [2.4 Multi-level Attack](#)
- [3. Defense Papers](#)

Kuva 4.2: Hyökkäystutkimusten määrä verrattuna puolustustutkimuksiin.

5 Yhteenveto

Luonnollisen 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. Tarve ihmiskielen koneelliseen ymmärrykseen on tuonut mukanaan kiinnostuksen lisäksi tietoturvatietoisuutta aiheesta.

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

Koneoppiminen on todennäköisesti vasta kehitysvaiheen alkupuolella. Jo nyt näkemämme vastakkaishyökkäykset osoittavat useampia haavoittuvaisuuksia, kuin mitä vastaan pystymme puolustautumaan. Selvästi suurin syy tälle on tekoälyn valjastettu voima, jota emme vielä pysty hallitsemaan kaikissa reunatapauksissa.

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/1802.03748). 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](https://doi.org/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](https://doi.org/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](https://doi.org/10.18653/v1/2020.acl-main.540). URL: <https://doi.org/10.18653/v1/2020.acl-main.540>.