Kanditutkielma

Tietojenkäsittelytieteen kandiohjelma

NLP-hyökkäysten käyttökohteet

Akira Taguchi

23.2.2022

Matemaattis-luonnontieteellinen tiedekunta

Helsingin yliopisto

Ohjaaja(t)

FT 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

NLP-hyökkäysten käyttökohteet

Ohjaajat — Handledare — Supervisors

FT Nikolaj Tatti

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Kanditutkielma	23.2.2022	7 sivua

Tiivistelmä — Referat — Abstract

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

Luonnollisen kielen prosessointi on kuitenkin sellaisenaan haavoittuvainen erillaisille tekstipohjaisille hyökkäyksille.

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

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

Avainsanat — Nyckelord — Keywords

Natural Language Processing, cyber security

Säilytyspaikka — Förvaringsställe — Where deposited

Helsingin yliopiston kirjasto

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Sisällys

Ohjelmistojen hyökkäysrajapinta-ala kasvaa jatkuvasti. Osa haavoittuvaisuuksista korjataan heti havainnoinnin jälkeen, osa mitigoidaan ja osan vaikutusalue on manifestoituu vasta tulevaisuudessa . Luonnollisen kielen prosessointi (eng. Natural Language Processing, NLP) on osoittautunut hyväksi hyökkäysrajapinnaksi tätä teknologiaa hyödyntäviä osapuolia vastaan (Boucher et al., 2021). NLP-järjestelmät on tehty tulkitsemaan ihmisen luonnollista kieltä. Tämän kielen konekääntäminen aloitettiin jo vuonna 1949.

Tässä tutkielmassa tarkastellaan NLP-hyökkäysten käyttökohteita. Tähän kuuluu oleellisen historian esittely, hyökkäystaksonomia sekä puolustusmetodit. On tärkeää ymmärtää luonnollisen kielen prosessoinnin tarkoitus, jotta voidaan syventyä hyökkäysiä mahdollistaviin ongelmiin sekä näiden ratkaisemiseen.

2 Taustaa

Ehdotukset kielten välisten sanojen välittämisestä koodeilla esitettiin 1700-luvulla Leibnizin ja Descartesin johdolla. Vuonna 1957 Georgetown-IBM-kokeen tekijät väittivät 3-5 vuoden jälkeen konekääntämisen olevan ratkaistu ongelma.

Tarve konekääntämiselle kumpuaa tietokoneen vajeesta ymmärtää ihmisen puhumaa kieltä. Ohjelmoinnissa tämän käännöksen tekee ihminen kutsuessaan esimerkiksi python-tulkilla "print("Hello world")". Käännöksen tapahtuminen tietokonepuolella tuottaa mielekkäämpiä ongelmia. Komento "Tulosta syntymäpäiväni"saattaa koneoppimismallista riippuen tulostaa näytölle merkkijonon "1.1.1970"tai fyysisen kuvan syntymästäsi lähikirjastosi tulostimeen.

NLP-hyökkäyksissä keskiössä on juuri tämän tulkitsemisen vaikeuden hyväksikäyttäminen pahansuopiin tarkoituksiin.

3 Hyökkäystaksonomia

Käymme seuraavaksi läpi neljä erilaista häiriöhuomaamatonta hyökkäysmetodia. Nämä hyökkäykset eivät siis näy visuaalisesti ihmiskäyttäjälle näyttöpäätteellä erilaisina verrattuna viattomaan tekstiin. Tarkemmin keskitymme Unicoden ja muiden enkoodausmenetelmien hyväksikäyttämiseen NLP-malleja vastaan.

3.1 Näkymättömät merkit

Näkymättömät merkit vaikuttavat tietokoneen NLP-mallin ymmärtämään kontekstiin. Esimerkki tällaisesta on zero-width space -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+200Bksen` `U+200Brakennus! Miten onnistuit tekemään tämän U+200BnäU+200Bin laU+200Biskas` `U+200Bti?`.

3.2 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` $\rightarrow$ `A`, missä viimeinen kirjain on todellisuudessa kyrillinen kirjain A. Näkymättömien merkkien lailla homoglyyfihyökkäyksen toteutus riippuu ympäristön fontista.

3.3 Uudelleenjärjestelyt

Uudelleenjärjestelyhyökkäys pohjautuu näennäisen tekstin uudelleenjärjestämiseen pahantahtoisesti. Pankkitilinumeron `1234567` pystyy esimerkiksi vaihtamaan bidi-algoritmilla tilinumeroksi `7654321` maksajan huomaamatta mitään.

3.4 Poistatukset

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

4 Puolustusmetodit

NLP-hyökkäykset voidaan estää alhaisemmalla tasolla overheadilla sekä korkeammalla tasolla edistyneen teknologian turvin.

5 Yhteenveto

Lähteet

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