Utjecaj AI tehnologija na sigurnost djece na Internetu

Zlatan Morić
Head of Cyber Security Department @ Algebra University College

AI (umjetna inteligencija)

- skup tehnika i alata koji omogućuju računalima da rješavaju probleme i donose odluke slične onima koje bi donijeli ljudi.
- Primjeri korištenja AI tehnologija u svakodnevnom životu:
  - Virtualni asistenti
    - razvoj virtualnih asistenata, poput Siri i Alexe
  - Personalizirana reklama
    - personaliziranje reklama, omogućujući oglašivačima da šalju relevantne reklame korisnicima na temelju povijesti pretraživanja
  - Automatizacija posla
    - automatizacija rutinskih poslova, poput procjene rizika
  - Sigurnost
    - prepoznavanje lica i analiza ponašanja

Važnost sigurnosti djece na Internetu

- Online prijetnje
  - Internet može biti opasno mjesto za djecu, s obzirom na to da postoje mnogi online predatori, cyber-nasilnici i drugi koji bi mogli pokušati izložiti djecu opasnostima
- Neprimjereni sadržaj
  - Djeca su često izložena neprimjerenom sadržaju na Internetu, uključujući pornografiju, nasilje i druge štetne stvari
- Utjecaj na zdravlje i razvoj
  - Pretjerano korištenje interneta može imati negativan utjecaj na zdravlje djeteta, uključujući smanjenje tjelesne aktivnosti, povećanje stresa i manjak sna
- Utjecaj na odnose
  - Djeca koja se aktivno uključuju u društvene mreže i online zajednice također su izložena riziku od loših odnosa i utjecaju negativnih komentara

Identifikacija i sprečavanje online prijetnji

- Analiza teksta
  - identifikacija potencijalno štetne poruke i sadržaj, uključujući one koji sadrže seksualne prijetnje ili vrijeđanje
- Analiza slika i video zapisa
  - identifikacija potencijalno štetnih sadržaje, kao što su dječje pornografije i nasilne slike
- Učenje iz primjera
  - AI alati mogu biti naučeni iz primjera prijetnji i štetnog sadržaja kako bi razvili sposobnost prepoznavanja sličnog sadržaja u budućnosti
- Automatizirano upozoravanje
  - Kada se prepozna potencijalno štetan sadržaj, mogu automatski upozoriti odgovorne osobe ili čak zabraniti pristup

AI tehnologije još uvijek imaju ograničenja
i ne može zamijeniti ljude
u borbi protiv online prijetnji

AI filtri za cenzuru pomažu

- Blokiranje sadržaja
  - automatski blokiraju sadržaj koji se smatra neprimjerenim za djecu, kao što su nasilne slike, pornografija i vrijeđanje.
- Analiza ključnih riječi
  - identifikacija potencijalno štetnog sadržaja
- Učenje iz primjera
  - uče iz primjera sadržaja koji je blokiran kako bi razvili sposobnost prepoznavanja sličnog sadržaja u budućnosti
- Personalizirana kontrola
  - omogućavaju roditeljima ili korisnicima da personaliziraju stupanj kontrole sadržaja koji se može prikazivati

Tehnologije

- Klasifikacija teksta
  - učenje iz primjera označenih tekstova za identificira i kategorizira sadržaje koji uključuju cyberbullying, vrijeđanje, seksualno eksplicitne poruke i sl.
- Prepoznavanje slika
  - detekciju i blokiranje nasilnih, seksualno eksplicitnih ili drugih neprimjerenih slika na Internetu
- Prepoznavanje govora
  - prepoznavanje i sprečavanje cyberbullyinga i drugih neprimjerenih sadržaja u audio obliku
- Predikcija
  - predikcija cyberbullyinga i drugih vrsta online prijetnji, koristeći podatke o ponašanju korisnika na društvenim mrežama.

Izazovi

- Transparentnost
  - zabrinutosti o tome kako su napravljeni AI algoritmi i koji su njihovi kriteriji i preciznost za prepoznavanje potencijalnih prijetnji
- Sloboda govora
  - AI tehnologije za cenzuru sadržaja mogu značajno utjecati na slobodu govora
- Diskriminacija
  - pristranost AI tehnologije za zaštitu što može dovesti do diskriminacije
- Implementacija
  - osigurati implementaciju na ispravan način, kako bi se osigurala njihova učinkovitost u zaštiti djece, bez stvaranja dodatnih prijetnji ili ugrožavanja privatnosti
- Izmjene
  - mogućnost izmjene i iskorištavanja AI tehnologije za zloupotrebu, uključujući stvaranje neprimjerenog sadržaja za djecu ili kršenje privatnosti

Zloupotreba AI tehnologije

- Automatsko generiranje uvredljivog sadržaja
  - AI tehnologije poput Generative Adversarial Networks (GANs) i Natural Language Processing (NLP) se koriste za automatsko generiranje neprimjerenog sadržaja koji se koristi za cyberbullying
- Povećana anonimnost
  - AI tehnologije, kao što su virtualno maskiranje i avatari, omogućavaju napadačima da ostaju anonimni dok šire svoj neprimjereni sadržaj
- Brzo širenje
  - AI tehnologije za automatizaciju marketinških aktivnosti omogućavaju napadačima da brzo šire svoj neprimjereni sadržaj na svim društvenim mrežama

Deepfake

- Lažne fotografije i video snimke
  - za stvaranje lažnih fotografija i video snimaka djece koji se šire po Internetu u svrhu seksualnog iskorištavanja i/ili cyberbullyinga.
- Lažni online profile
  - za stvaranje lažnih online profila djece kako bi se prikupljali podaci i širila mreža prijatelja
- Lažne poruke
  - za stvaranje lažnih poruka koje se šalju djeci s ciljem stvaranja prijateljstva, prikupljanja informacija ili zastrašivanja

OpenAI i ostali

- Javno dostupni AI alati
- Jednostavni za korištenje
- Teško detektirati da nije riječ o osobi (pogotovo djeci)

AI i sigurnost djece na Internetu

- AI filtri za cenzuru i softveri za detekciju i prevenciju su primjer korištenja AI-a koji pomažu u smanjenju dostupnosti neželjenog sadržaja za djecu, poput pornografije i nasilja.
- AI može pomoći u identificiranju i prevenciji online prijetnji, poput cyberbullyinga i seksualnog zlostavljanja djece.
- AI može također biti iskorišten za cyberbullying i druge oblike ugroza, što zahtijeva pažljivo praćenje i regulaciju. Postoje i etička pitanja u vezi s upotrebom AI tehnologije, kao što su privatnost i diskriminacija.

Mjere zaštite

- Edukacija
  - Obrazovanje djece, roditelja i nastavnika o opasnostima na internetu i kako ih se može spriječiti
- Osnaživanje tehnologije
  - Razvijanje AI tehnologija za što precizniju identifikaciju i blokiranje neprimjerenog sadržaja za djecu
- Zajednička odgovornost
  - Stvaranje zajedničke platforme za rješavanje online prijetnji (tvrtke, država i roditelji)

Mjere zaštite

- Zakonodavno okruženje
  - Usvajanje i provedba strogih zakona i propisa koji će regulirati korištenje interneta i AI tehnologije
- Prilagođavanje tehnologije
  - Prilagođavanje tehnologije tako da se uključuje sigurnost djece kao ključni faktor prilikom dizajniranja i razvoja proizvoda
- Svijest
  - Podizanje svijesti o važnosti zaštite djece na Internetu i potrebi za aktivnim pristupom u borbi protiv online prijetnji

Thank you for your attention!

?