Rebelion –

odhaľovanie antisociálneho správania na Webe

Pavol Návrat[1], Daniela Chudá[1], Mária Bieliková[1], Marián Šimko[1], Jakub Šimko[1], Ivan Srba[1], Michal Kompan[1], Róbert Móro[1], Irina Malkin Ondik[1], Peter Lacko[1], Alena Martonová[1], Kristína Machová[2], Ján Paralič[2], Peter Butka[2], Peter Bednár[2], Martin Sarnovský[2], Barbora Mesárošová[3], Radoslav Blaho[3], Lucia Sabová[3]

1 UISI FIIT STU BA, 2 KKUI FEI TU KE, 3 KP FiF UK BA

STU FIIT

SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ

APVV REBELION

Projekt sa zameriava na výskum nových modelov a metód pre automatizované rozpoznávanie anti-sociálneho správania v online komunitách.

STU FIIT
SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
FAKULTA INFORMATIKY A INFORMAČNÝCH TECHNOLÓGIÍ


WHAT IF I TOLD YOU
HALF THE NEWS ON FACEBOOK IS FAKE
imgflip.com

Motivácia a následky

- **Súčasný web** = voľný **priestor** pre vytváranie, šírenie a prijímanie informácií takmer bez omedzení

- Následky a rozsah
  - vytváranie osobných názorov a rozhodovaní počas politických volieb
  - štúdia Yelp - 16% recenzií je falošných.
  - degradácia úrovne diskusie, odrádzanie od zapájania sa do diskusií
  - 40% používateľov webu zažilo obťažovanie v online priestore
  - 18% európskej mládeže bolo niekedy v ich živote obeťou kyberšikany

Antisociálne správanie

- šírenie dezinformácií (napr. hoax alebo falošné správy)
- reakcie antisociálnych používateľov (napr. hejtovanie, manipulovanie diskusií alebo kyberšikana)

Rozvoj IT

- zosieťovanie, anonymita

+ prostriedok pre lepšie porozumenie a vysporiadanie

Celosvetový konsenzus:

na antisociálne správanie v online priestore nie je možné naďalej reagovať len ignorovaním

Súčasné riešenia

- **súčasný stav poznania neposkytuje efektívne riešenia**, ako regulovať a eliminovať anti-sociálne správanie
- **moderátor** online komunity
- čerpanie z davu
- pravidlá komunity pre odstraňovanie príspevkov
- CQA Yahoo! Answers
  - len 50% nahlásených používateľov bolo vylúčených,
  - až 40% z vylúčených nebolo nikdy nahlásených
- subjektívnosť, neškálovateľnosť, rezignácia...

Antisociálne správanie

kategorizácia výskumných článkov

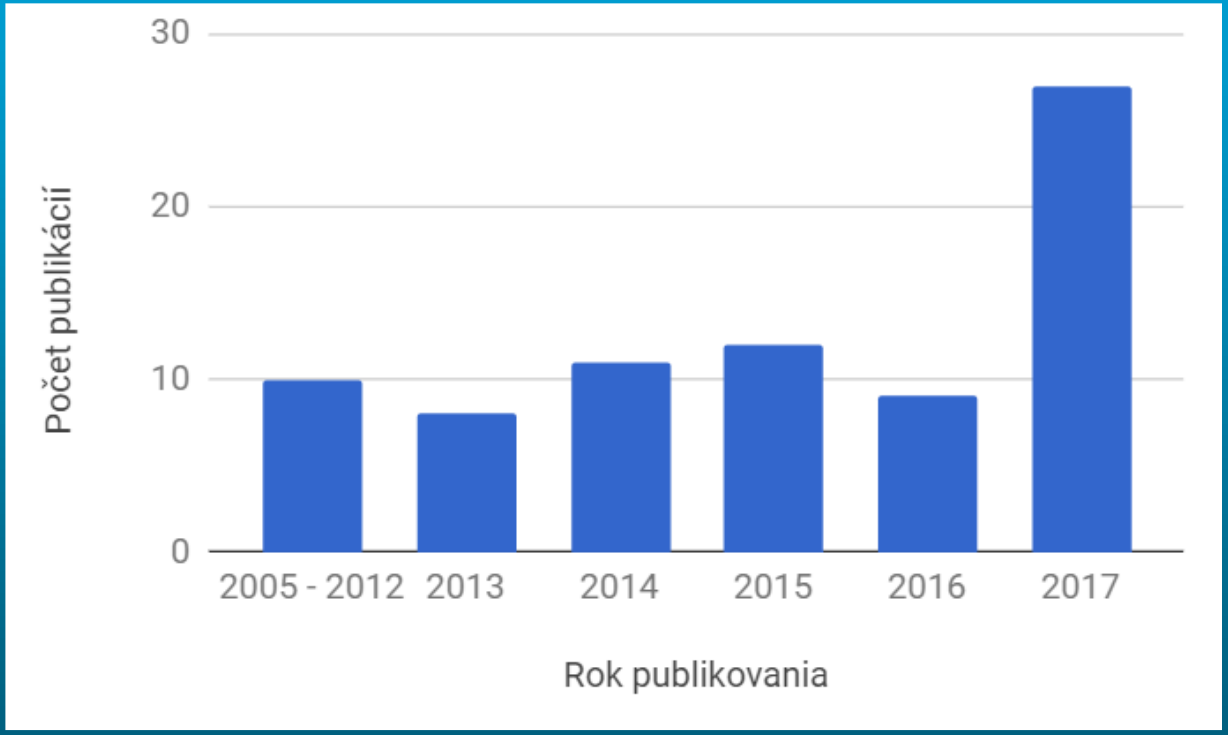
Šírenie dezinformácií:

- Falošné a skreslené správy (angl. fake and biased news, 32)
- Falošné recenzie (angl. fake reviews, 7)
- Hoax (5)

Reakcie používateľov:

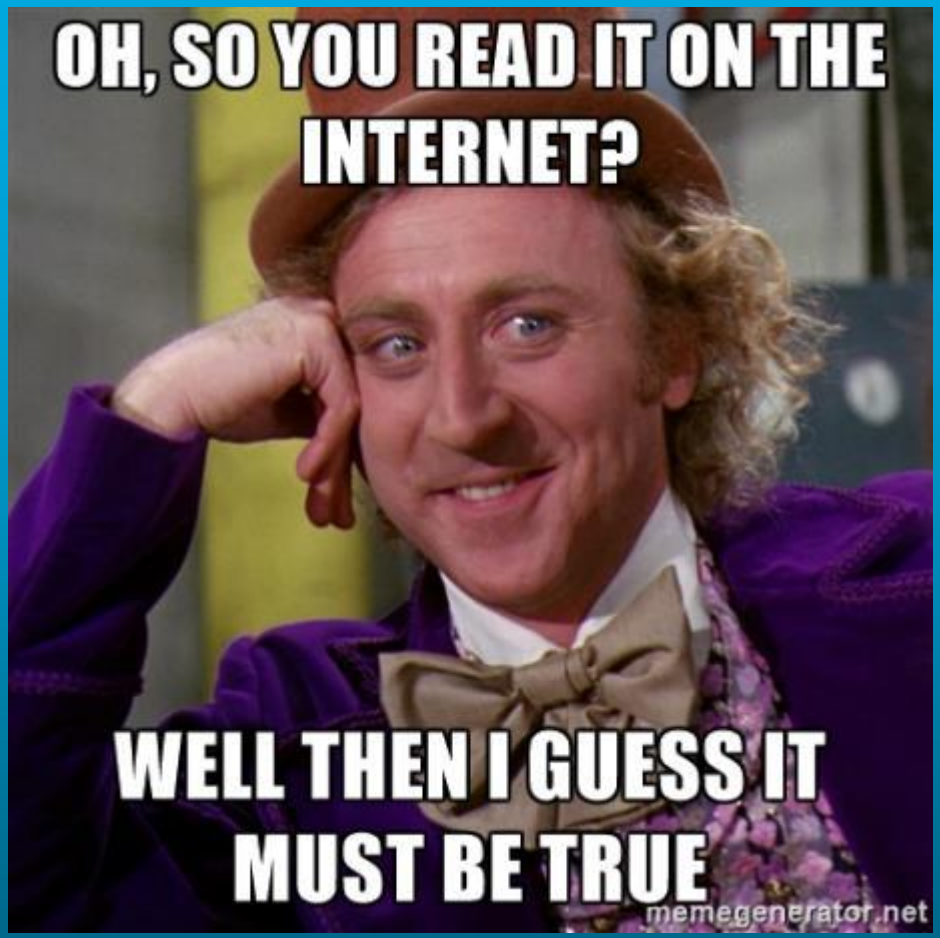
- Trolovanie (angl. trolling, 13)
- Manipulovanie diskusií (angl. sockpuppeting, 6)
- Zneužívanie komunity (angl. community abuse, 5)
- Vandalizmus (angl. vandalism, 4)
- Záškodnícke správanie hráčov online hier (angl. griefing, 4)
- Kyberšikana (angl. cyberbulling, 4)
- Hejtovanie (angl. hating, 2)
- Spamovanie (angl. spamming, 1)

Vývoj počtu publikácií podľa rokov publikovania zahrnutých v prehľadovej štúdii



Otvorené problémy

1. Metódy strojového učenia sú aplikované bez hlbšieho porozumenia skúmaného antisociálneho správania.
2. Nevyužívanie celého spektra dostupných informácií o obsahu a používateľoch.
3. Absencia multiplatformovej detekcie antisociálneho správania.
4. Potreba skorej detekcie/predikcie.


OH, SO YOU READ IT ON THE INTERNET?
WELL THEN I GUESS IT MUST BE TRUE
memegenerator.net

Záver – ciele projektu

Automatizovaná detekcia a predikcia antisociálneho správania sa človeka v online priestore:

- analýza dát (kvantitatívnymi a kvalitatívnymi prístupmi informatických vied a psychológie),
- technologická stránka, kde sa do popredia dostávajú nové prístupy a technológie v umelej inteligencii.

Záver – ciele projektu

1. s využitím analýzy dát a kvalitatívnych prístupov vyskúmať do hĺbky fenomény antisociálneho správania ľudí v online komunitách a priniesť nové poznatky
2. s využitím dolovania v dátach navrhnúť a verifikovať nové modely, ktoré reprezentujú dátový pohľad na antisociálne správanie ľudí v online priestore
3. s využitím strojového učenia navrhnúť a verifikovať nové metódy umelej inteligencie pre automatizovanú detekciu a predikciu rôznych typov antisociálneho správania

Ďakujeme za pozornosť

- Pre položenie silných základov predkladaného projektu sme vykonali rozsiahlu prehľadovú štúdiu, v rámci ktorej sme vyhľadali a systematicky zakategorizovali 77 výskumných článkov zaoberajúcich sa antisociálnym správaním. Prehľadová štúdia ukázala enormný nárast záujmu výskumníkov o túto ob-lasť v aktuálnom roku 2017, v ktorom sa počet publikácií oproti roku 2016 zvýšil trojnásobne (obr. 1).
- Najväčšia časť analyzovaných príspevkov sa zaoberá návrhom metód (35 publikácií) a prípadovými štúdiami (19 publikácií). V menšej miere je zastúpený aj návrh a realizácia inovatívnych aplikácií (5 publikácií), ktoré slúžia používateľom najmä na automatické rozlíšenie falošných správ a hoaxov.
- Najčastejšie riešenou úlohou v analyzovaných príspevkoch je detekcia antisociálneho správania (55 publikácií) a realizácia exploratívnych analýz (22 publikácií), ktorých cieľom je lepšie pochopenie anti-sociálneho správania. Naproti tomu, výskum predikcie (t.j. predpovede že antisociálne správanie v budúcnosti nastane) je evidentne len v začiatkoch (identifikovali sme len 3 publikácie).

Výskumné výzvy

1. Nejednoznačná hranica medzi prosociálnym a antisociálnym správaním.
2. Charakter dát - neanotované datesety, dátové sady sú prirodzene nevyvážené.
3. Potreba dodržať technologickú neutrálnosť.

Metóda, postupy

1. Získanie dát.
2. Integrácia dát.
3. Predspracovanie a čistenie dát.
4. Exploratívna analýza dát.
5. Dolovanie v dátach.

Metóda, postupy

1. Konštrukcia čŕt
   - Extrakcia čŕt
   - Transformácia čŕt
   - Selekcia čŕt
2. Trénovanie modelov.
3. Vyhodnocovanie modelov.