Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií

Oliver Leontiev
Kamilla Kisová

Špecifikácia inteligentného znalostného konateľa

Fact Checker

Predmet: Umelá inteligencia
Vyučujúci: Ing. Ivan Kapustík
5. Október 2020

1. Problémové prostredie

"Človek je spoločenská bytosť nadaná rozumovou schopnosťou, vedomím a článkovanou rečou" [5], uvádza filozofický opis človeka. Každodenne vidíme príklady sociálnych interakcií - buď v reálnom živote, alebo pomocou novodobých prostriedkov - telefonicky a na sociálnych sieťach. Regulovať obsah interakcií bol vždy veľkou filozofickou dilemou v demokratických prostrediach. S nástupom **sociálnych sietí** sa to stalo ešte väčším problémom. Hoci v roku 1996 internetové spoločnosti boli oslobodené od zodpovednosti obsahu vytváraného používateľmi, niektoré sociálne siete túto debatu znova otvorili [3]. Citlivé obsahy ako nahota, pornografia, vulgarita a násilie už dávno regulované sú, ale najnovší nápad - aj téma diskusie - je filtrovanie klamlivých správ.

1.1. Vlastnosti prostredia

Fact Checker má byť nezávislý softvér, ktorý má prístup do **sociálnej siete** - je kontrolujúca moc nad používateľmi.

Môžeme deliť prostredie na tri väčšie časti:

- platforma sociálnej siete
- databáza známych hoaxov a falošných faktov
- portály, domény overených článkov a dokumentov

Proces zistenia pravdivosti môžeme opísať v štyroch bodoch:

1. Všímanie postoja príspevku (stance detecting)
2. Hľadanie relevantných overených článkov a dokumentov
3. Overenie tvrdení na základe informácie získanej z relevantných dokumentov
4. Vyhodnotenie reputácie príspevku

Všímanie postoja - stance detecting - sa odohráva na platforme sociálnej siete. Keď používateľ inicializuje uverejnenie príspevku, Fact Checker začne konať v tomto prostredí. Môže vnímať text, dátum a autora príspevku.

Hľadanie relevantných článkov sa buď realizuje len v databáze hoaxov alebo aj na portáloch overených článkov.

Overenie tvrdení môžeme vnímať ako porovnanie príspevku a nájdených článkov, teda prostredím je databáza aj sociálna sieť.

Po **vyhodnotení reputácie** sa konateľ vráti na platformu sociálnej siete.

1.2. Prečo je potrebné inteligentné riešenie

Postupným sledovaním procesu sme si všimli, že individuálne kroky nie sú jednoznačné. Ak uvažujeme o "obyčajnom" programovaní, ako by sme mohli implementovať niečo, ako Fact Checker? Ako by sme mohli získať postoj z nejakého textu? Vyberieme najčastejšie slová alebo slovné spojenia? Ako určíme slovné spojenie? Ako začneme hľadať relevantné dokumenty, ak nevieme čo je zhrnutie textu? Aj keď sa nejako dostaneme k slovnému spojeniu, ako vyhľadáme články, ak neexistujú zhody s konkrétnym textom? Ako postupujeme pri overení tvrdení? Postupne porovnávame texty?

Všetky tieto otázky sú určite smiešne pre ľudskú inteligenciu. Táto obrovská časť znalostí, o ktoré sa potrebujeme oprieť, by dospela k neúspechu jednoduchých („neinteligentných") algoritmov. Fakt, že konateľ musí spracovať jazyk a aj interpretovať ho, značí, že musí byť v istej miere inteligentný. Musí sa naučiť ako zvládnuť ľudskú reč, trénovať sa a učiť sa. To je práve oblasť umelej inteligencie [4].

Ako sme už spomínali, regulovanie citlivých obsahov už je implementované na mnohých platformách. Tieto filtre fungujú na princípe machine learning a môžeme vidieť, že v niektorých prípadoch nie sú schopné identifikovať nevhodné časti textu. Preto aj výskumníci zaoberajúci sa detekovaním klamlivých správ začali experimentovať s **deep learningom** - o úroveň hlbšie v umelej inteligencii [2].

NLP - Natural Language Processing - je vetva umelej inteligencie zaoberajúca sa spracovaním reči a jazyku. Vo vývoji je jeden NLP nástroj, **transformer**, ktorý funguje na architektúre neurálnej siete [2]. Kľúčovým prvkom v NLP je tzv. Attention, vďaka ktorému sa konateľ naučí, na ktoré slová a pojmy má dávať pozor [3]. Pomocou tejto funkcie je schopný vyťahať najdôležitejšie koncepty z textu a teda je vhodný na všímanie postoja pri našom inteligentom konateľovi. Aj výskumníci veria, že práve transformer bude ten prelom k dosiahnutiu Fact Checkera.

2. Špecifikácia znalostného konateľa

2.1. Ciele

Hlavnými cieľmi nástroja Fact Checker je **úspešne analyzovať súvislý text a odhaliť v ňom nepravdy a následne nedovoliť uverejnenie príspevku, ktorý nejakú obsahuje.** V širšom zmysle je cieľom **zabrániť šíreniu klamstiev po internete** a zjednodušiť prístup k pravdivým informáciám pre verejnosť tým, že nebude potrebné overovať si informácie u viacerých zdrojov. Zároveň odbremeniť prevádzkovateľov verejných internetových platforiem (napr. Twitter, Facebook) od zodpovednosti za zavádzajúce a klamlivé príspevky od užívateľov. Momentálne spoločnosti nie sú zodpovedné za obsah, ktorý je uverejňovaný na ich platformách, ale to sa môže čoskoro zmeniť [1], a taktiež pociťujú tlak zo strany investorov a reklamných spoločností.

Fact Checker je softvér, ktorý by fungoval ako prevencia a kontroloval by všetky príspevky, ktoré by sa niekto pokúsil zverejniť na stránku, ktorá by ho používala. Pokiaľ by ich vyhodnotil ako nepravdivé, zabránil by ich uverejneniu.

2.2. Vnemy

Fact Checker nevyužíva všetky vnemy v každom prostredí. Pri kontrolovaní príspevku, ktorý bol práve zaslaný na sociálnu sieť, ho väčšinou zaujíma iba obsah príspevku (text) a náhľad do databázy. Avšak pri kontrolovaní pravdivosti sa opiera aj o články na iných platformách, a vtedy využíva ostatné vnemy (meno autora, dátum uverejnenia a názov zdroja) na vyhodnotenie dôveryhodnosti. Preto sú pre Fact Checker aj tieto vnemy dôležité. Vnemy teda sú:

- Text príspevku
- Meno autora príspevku
- Dátum uverejnenia príspevku
- Názov domény alebo časopisu, kde je príspevok alebo článok uverejnený
- Náhľad do databázy známych hoaxov a falošných faktov

2.3. Typy akcií

- Blokovať zverejnenie príspevku
- Vyznačiť nepravdy v texte
- Navrhnúť pravdivý variant pre nepravdivé tvrdenie
- Zablokovaný text vrátiť autorovi (zobraziť mu ho)
- Pridať nepravdu do databázy známych hoaxov a falošných faktov

3. Druhy informácií a znalostí

3.1. Informácie

Údaje, ktoré získa konateľ vnímaním sú hneď organizované do informácií, t.j. naberú význam pre konateľa [4].

Údaje		Informácie
“Fero Mrkvička”	→	Fero je autor príspevku
2020-09-23 09:13	→	Toto je dátum uverejnenia
“Trump je nakazený”	→	Toto je text, s ktorým mám pracovať
`True`	→	Je reprezentácia pravdy
`False`	→	Je reprezentácia nepravdy
`Neutral`	→	Je reprezentácia neutrálnosti textu
“twitter.com”	→	Doména sociálnej siete
“hoax.db.com”	→	Doména databázy

3.2. Znalosti

Hranice medzi znalosťou a informáciou už nie sú také presné a jednoznačné. Znalosti môžeme vnímať ako implikáciu, odvodzovanie z informácií, alebo spracované informácie [4].

Informácie o autorovi môžu prispieť k znalostiam. Ak autor je overený používateľ, príspevok má tendenciu byť pravdivým; autor môže mať vyčíslenú hodnotu reputácie, čo tiež prispieva k vyhodnoteniu textu. Dátumy môžu implikovať znalosti, napr. pri hľadaní relevantných článkov. Ak sa nájdu dva dokumenty, prvý s dátumom, ktorý sa zhoduje s dátumom príspevku a druhý je už niekoľko mesiacov starý, tak prvý pravdepodobne má viac spoločné s príspevkom. `True`, `false` a `neutral` sú abstraktné definície pravdy, nepravdy, resp. neutrálnosti, a konateľ musí vedieť, že práve tieto hodnoty musí priradiť k príspevkom. Znalosťou je aj fungovanie jazyku, štruktúra textu a gramatika. Gramatika tiež môže prispieť k vyhodnoteniu dôveryhodnosti textu.

Samozrejme znalosti, teda rôzne odvodzovanie z textu, názvov a z dátumov, sa môžu zmeniť. Závisí to hlavne od trénovania konateľa, vstupy pre Fact Checker v tomto štádiu prispievajú k jeho fungovaniu.

4. Zhodnotenie správania znalostného konateľa

Fact Checker by fungoval ako brána medzi užívateľom, ktorý sa chystá uverejniť príspevok a konkrétnou sociálnou sieťou alebo inou platformou. Z pohľadu užívateľa by sa udialo nasledovné:

1. **Ak jeho príspevok neobsahuje falošné fakty** alebo iný nežiaduci obsah, tak sa úspešne uverejní a užívateľ nie je nijako vyrušený. Alternatívne sa mu môže na obrazovke zobraziť malá značka, ktorá oznamuje, že Fact Checker vyhodnotil jeho príspevok ako bezproblémový.

2. **Ak príspevok obsahuje aspoň jednu nežiadúcu informáciu** (klamlivú, nebezpečnú), tak sa neuverejní a užívateľovi sa zobrazí text jeho príspevku s vyznačenými časťami (tvrdeniami), ktoré Fact Checker považuje za falošné fakty. K týmto tvrdeniam by boli napísané ich pravdivé varianty a/alebo odkaz na zdroj, ktorý ich vyvracia.

Fact Checker najprv text príspevku zanalyzuje a objaví v ňom tvrdenia (napr. „Nosenie rúška znižuje riziko nákazy.", „Zem je plochá.", „Dnes ráno som vypil kávu."), následne začne eliminovať tvrdenia, ktoré sú neutrálne („Dnes ráno som vypil kávu"), a to všetko pomocou rôznych algoritmov (založených na princípe deep learning) [2]. Následne začne kontrolovať pravdivosť zvyšných tvrdení. V tom mu pomôže jeho databáza a prístup na iné portály a platformy.

Prvým krokom je skúsiť nájsť v databáze vyhodnocované tvrdenie alebo jemu podobné. Databáza by mala byť prispôsobená potrebám Fact Checkera, aby v nej vedel čo najrýchlejšie hladať a pridávať do nej nové klamstvá, ktoré objavil. Táto databáza by zároveň bola open-source, vďaka čomu by do nej mohli pridávať aj iné programy a zároveň aj iný užívatelia [3].

Ak Fact Checker nenájde dostatočnú zhodu v databáze, tak by kontroloval pravdivosť tvrdení pomocou porovnávania s inými príspevkami a článkami na internete, ku ktorým by priradil hodnotu dôveryhodnosti na základe štruktúri a spisovnosti textu, mena autora, dátumu vydania a pravdivosti iných tvrdení v týchto príspevkoch.

5. Poďakovanie

Kamille Kisovej za sekcie 1. a 3. (50%) a Oliverovi Leontievovi za sekcie 2. a 4. (50%).

6. Zdroje

[1] ROMM, Tony - DWOSKIN, Elizabeth. 2020. *Trump signs order that could punish social media companies for how they police content, drawing criticism and doubts of legality.* Washington Post. https://www.washingtonpost.com/technology/2020/05/28/trump-social-media-executive-order/

[2] DICKSON, Ben. 2020. *This stance-detecting AI will help us fact-check fake news.* The Next Web. https://thenextweb.com/neural/2020/03/14/this-stance-detecting-ai-will-help-us-fact-check-fake-news-syndication/

[3] LAMBERT, Nathan. 2020. *AI & Arbitration of Truth.* Towards Data Science. https://towardsdatascience.com/ai-arbitration-of-truth-808b57a93a97

[4] NÁVRAT, P. a kol. *Umelá inteligencia.* Bratislava: STU, 2015. ISBN 978-80-227-4344-0.

[5] Wikipédia. *Človek (filozofia)* https://sk.wikipedia.org/wiki/%C4%8Clovek_(filozofia)