SpremljAI: SISTEM ZA ZAZNAVO LAŽNIH PROFILOV NA SPLETU

Gregor Gabrovšek[1]
[1] Univerza v Ljubljani, Fakuklteta za računalništvo in informatiko, 1000 Ljubljana
gg4792@student.uni-lj.si

Povzetek

Sistem SpremljAI omogoča samodejno zaznavanje lažnih profilov na spletnih platformah. Z uporabo jezikovnih modelov določi avtorja posameznega besedila, četudi se ta skriva za večimi uporabniškimi imeni. Sistem samodejno zbira nove objave uporabnikov, sproti zaznava potencialne *spammerje* in o dogodkih v realnem času obvešča upravitelje spletnih portalov.

Ključne besede: umetna inteligenca; jezikovni modeli; zaznavanje lažnih profilov; obdelava naravnih jezikov

1 MOTIVACIJA

Diskusije na spletnih platformah, kot so socialna omrežja in forumi, so postale pomemben del našega vsakdanjika. Zaradi svoje anonimnosti pa pogosto postanejo gojišča osebnih napadov in žaljivk, prepogosto usmerjenih proti najšibkejšim skupinam v družbi.

V Sloveniji se soočamo s porastom sovražnega govora, ki ima kot posledico čedalje večji razkol med državljani ter nezmožnost sodelovanja pri grajenju naše prihodnosti.

Administratorji spletnih platform, ki omogočajo komentiranje uporabnikov (npr. portal RTV SLO) imajo polne roke dela z uporabniki, ki objavljajo sovražno vsebino. Težava je tudi v tem, da se odstranjeni uporabniki lahko enostavno registrirajo z novim uporabniškim imenom in še naprej pišejo enake komentarje kot prej.

Položaj kliče po sistemu, ki bi upraviteljem takih platform olajšal delo in jih samodejno opozoril na uporabnike, ki ustvarjajo lažne profile za širjenje sovražnega govora.

2 ZASNOVA SISTEMA

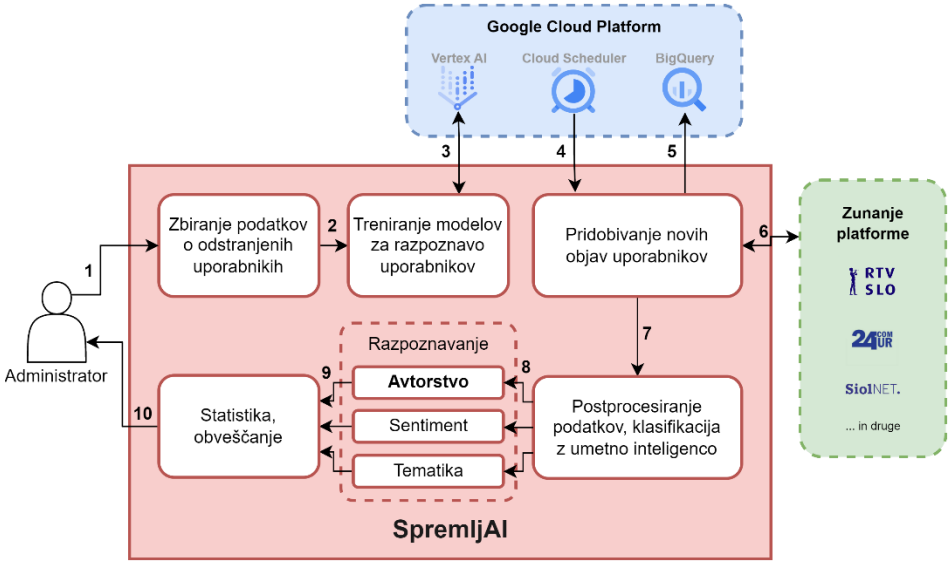
Da lahko sistem doseže svoj cilj, mora v grobem opravljati sledeče naloge: zbiranje novih objav uporabnikov, učenje jezikovnih modelov ter klasifikacija novih komentarjev s temi modeli, grupiranje potencialnih lažnih profilov z že poznanim uporabnikom ter obveščanje administratorja o ugotovitvah. Diagram sistema se nahaja na Sliki 1.

2.1 Zbiranje objav uporabnikov na spletnih platformah

Glavni vir podatkov so seveda platforme, kjer naključni uporabniki lahko objavljajo svoja besedila. Primeri takih platform so RTV SLO, 24ur, SiolNET in podobne. Sistem mnogokrat dnevno zažene komponento za zbiranje besedil uporabnikov (v podanih primerih so ta besedila v obliki komentarjev na objavljeno novico) in jih shrani v podatkovno bazo. Zbrani podatki vključujejo tudi čas objave,

naziv/ID uporabnika ter potencialno ostali metapodatki. Besedilo je tudi postprocesirano za lažjo kasnejšo obdelavo.

Komponenta za zbiranje besedil je tudi enostavno razširljiva, torej lahko nova stranka (npr. upravitelj nekega foruma) zlahka omogoči podporo tudi za svojo spletno stran.



Slika 1: Shema sistema

2.2 Treniranje jezikovnih modelov

Ključni del sistema SpremljAI so jezikovni modeli, torej globoke nevronske mreže, katerih naloga je v našem primeru klasifikacijske narave – za dano besedilo morajo ugotoviti če oz. kateremu avtorju pripada to besedilo. Za inicializacijo sistema mora skrbnik spletne strani sistemu podati informacijo o odstranjenih uporabnikih, za katere obstaja verjetnost, da se bodo znova registrirali pod drugim uporabniškim imenom.

Iz podatkovne baze nato sistem pridobi komentarje teh odstranjenih uporabnikov in s pomočjo tehnike *fine-tuninga* obstoječ jezikovni model prilagodi za nalogo dodeljevanja besedila enemu od teh avtorjev.

Strategija treniranja je precej odvisna od števila uporabnikov. Če je skupno število vseh uporabnikov na platformi majhno (<100), lahko en sam model celo naučimo na vseh uporabnikih, ki nam kasneje lahko celo omogoči odkrivanje, če ena oseba upravlja več profilov. Če pa je uporabnikov zelo veliko, naučimo modele le na besedilih odstranjenih uporabnikov.

Ključna lasnost te komponente sistema je konfigurabilnost – en sam jezikovni model lahko sicer dobro razlikuje tudi med do 100 uporabniki, točnost pa se hitro viša z zmanjšanjem števila uporabnikov v učni množici. Če pa želimo višjo točnost, moramo ustvariti več modelov, ki razpoznajo le del množice uporabnikov. Uporabnik sistema SpremljAI mora torej sprejeti odločitev, kako pomembna je točnost glede na čas in denar, ki ga imamo na voljo (treniranje jezikovnih modelov je seveda precejšen časovni in finančni zalogaj).

2.3 Razpoznavanje

Ko že imamo ustvarjene jezikovne modele in sveže zbrane podatke iz neke dotične spletne platforme, lahko na podatkih opravimo analizo ujemanja avtorstva. To pomeni, da z uporabo jezikovnih modelov klasificiramo besedila novih uporabnikov in ugotovimo, ali se katero besedilo z veliko verjetnostjo ujema z enim od uporabnikov, ki ga jezikovni model lahko razpozna.

Ker točnost modelov ni stoodstotna, je možna nastavitev, da mora sistem k istemu uporabniku klasificirati več besedil nekega novega uporabnika, da lahko z gotovostjo rečemo, da je prišlo do ujemanja.

3 PREDNOSTI SISTEMA

Sistem SpremljAI je uporaben za spletne platforme, na katerih je omogočeno javno in anonimno objavljanje besedil, na primer komentarjev pod novicami. Na večjih portalih, kot je RTV SLO, uporabniki lahko skupno napišejo na tisoče komentarjev dnevno, kar je veliko breme za moderatorje, ki morajo vsebino tudi pregledovati. Sistem SpremljAI lahko moderatorsko ekipo opolnomoči, da nekatere uporabnike, pri katerih obstaja visoko tveganje, da bodo objavljali vsebino s sovražnim govorom, hitro zazna (nato pa jih bodisi natančneje nadzoruje, bodisi jih nemudoma znova odstrani). Tako preventivno ravnanje torej zniža kasnejše potrebe po moderiranju, kar lahko pomeni nižji strošek moderiranja.

4 ZAKLJUČEK IN POTENCIALNE IZBOLJŠAVE

Z našim sistemom je torej mogoč učinkovit nadzor nad uporabniškimi komentarji na spletnih platformah za javni diskurz, kar ima potencial za znižanje količine lažnih informacij in sovražnega govora. Vir takega govora je običajno majhno število vztrajnih uporabnikov, ki smo jim s pomočjo umetne inteligence lahko lažje kos.

V trenutni implementaciji sistema je že podprto učenje modelov, zbiranje besedil iz spletnih portalov in določanje avtorstva, šele v razvojni fazi pa je prototip nadzorne plošče, kjer bi imel administrator/moderator spletne strani pregled nad novimi opozorili glede lažnih profilov.

Logičen naslednji korak je tudi vključitev ostalih klasifikacijskih modelov v sistem SpremljAI, kar je tudi že prikazano na Sliki 1. Možni primeri so detekcija sovražnega govora s pomočjo analize sentimenta (naš sistem bi lahko še pred moderatorjem zaznal, če besedilo vsebuje sovražni govor) in klasifikacija stališč uporabnika (glede na uporabnikova besedila lahko npr. sklepamo o politični pripadnosti – nenadna povišana aktivnost velikega števila uporabnikov z enako pripadnostjo bi torej lahko nakazovala na poskus vpliva na javno mnenje), če naštejemo le nekaj možnosti.