

Sistem za upravljanje korisnicima i objavama na forumu

Opis problema

Sistem za upravljanje korisnicima i objavama na forumima rešio bi problem mnogih socijalnih platformi koje u svom opisu imaju deljenje tekstualnog sadržaja. Glavni zadatak iz aspekta **objava** jeste da se uoče *trending* objave, ali i isprate one koje su potencijalno štetne za zajednicu i odreaguje pravilno. Sa druge strane, imamo **korisnike**, čije se ponašanje može ili nagraditi određenim bedževima/tokenima, ili kazniti restrikcijom aktivnosti (odnosno, suspenzijom naloga u krajnjem slučaju).

Ovakvu logiku možemo uočiti na dosta platformi. Neki od primera su:

- **Facebook** grupe sa labelama za najaktivnije članove,
- **StackOverflow** koji prati korisnike koji su najkvalitetnije doprineli zajednici korisnim odgovorima,
- i **Twitter** koji prati *Trending* teme i *tvitove*.

Naše rešenje bi objedinilo gore navedene pojave, uz dodatak sankcionisanja za toksično ponašanje. Predstavljao bi visoko-fleksibilnu implementaciju koja se lako može ugraditi u bilo koji sistem kome bi opisana evaluacija korisnika i njihove aktivnosti bila od koristi.

Metodologija

Naš sistem zasniva se na događajima koji pristižu u realnom vremenu. Bitno nam je vreme svake instance jer ovako možemo da imamo u uvid učestalost i šablone korisničkog ponašanja.

Događaji koje sistem prati su sledeći:

1. Postavljanje nove objave (sadržaj objave i vlasnik iste)
2. *Like post*
3. *Dislike post*
4. *Report post for harmful content*

Jedan od dva osnovna entiteta sistema je objava:

- Objava pored sadržaja ima podatak o vremenu kada je postavljena i korisniku koji ju je objavio
- Čuvaju se broj *lajkova*, *dislajkova*, i prijava objave
- Ukoliko je postavljena objava naglo izrazito popularnija u odnosu na ostale objave postavljene u sličnom vremenskom periodu, dobija **Trending** labelu. Ovakve objave ne smeju da imaju bilo kakvu negativnu oznaku. U sistemu je u svakom momentu samo ograničen broj objava *Trending*, i ovo se određuje srazmerno broju novopristiglih postova u određenom vremenskom periodu.

- Objava koja ima slab odnos *lajkova* i *dislajkova* dobija labelu **Poor Content**¹. U kompletnom sistemu bi ove objave trebale biti slabije prikazivane korisnicima i generalno manje promovisane.
- Objava koja je dobila veliki broj prijava labelirana je kao **Potentially Harmful**, i u tom slučaju ide na proveru. Provera može biti ručna, izvedena pomoću zasebnog API-ja, ili asistencijom neuronske mreže, međutim za obim ovog zadatka logika provere će biti apstrahovana.
- Ukoliko se proverom uspostavi da objava nije bila štetna, ona se označava kao **Poor Content**.
- Ukoliko se proverom uspostavi da objava jeste bila štetna po zajednicu (govor mržnje, spam, maliciozni linkovi etc.), dobija labelu **Harmful** i ima značajne konsekvence po korisnika autora. Idealno, u kompletnoj aplikaciji ove objave se ne prikazuju više drugim korisnicima ili se brišu iz javnog sistema.

Sledeći entitet značajan za sistem jeste korisnik:

- Svaka korisnička aktivnost se prati, i u zavisnosti od nje korisnik biva labeliran na adekvatan način. U okviru kompletnog sistema ove labele mogu pomoći i u realizaciji kompleksnijih funkcionalnosti sistema, ali i u vizualnom prenošenju važnih informacija drugim korisnicima.
- Ukoliko korisnik redovno objavljuje sadržaj koji dobija dobar odnos *lajkova* i *dislajkova* dobija labelu **Top User**. Ova labela može da se izgubi prestankom aktivnosti.
- Ukoliko korisnik redovno *lajkuje* generalno dobar sadržaj dobija labelu **Community Contributor**. Ona takođe može da se izgubi prestankom aktivnosti.
- Korisnik koji izvršava previše aktivnosti u jedinici vremena privremeno biva suspendovan i labeliran kao **Spammer**.
- Korisnik čije su objave (jedna ili više) označene kao *Harmful* dobija labelu **Harmful User**.
- Ukoliko korisnik rapidno *dislajkuje* i/ili prijavljuje objave biva labeliran kao **Potential Spammer**.
- Za *Potential Spammer*-a potrebno je izvršiti analizu prethodno urađenih akcija. Ako su su radnje opravdane, odnosno vidimo da su negativne reakcije i prijave bile na objavama koje su *Poor Content* ili *Harmful* ponašanje je opravdano i poništava se aktivna labela. Ukoliko je naizgled nasumično, korisnik dobija labelu **Spammer**.
- *Spammer* labela ističe posle određenog vremena, ali je *Harmful User* peramentna.

Poslednji pojam od značaja bile bi sankcije za korisnike:

- Korisnik koji je 3 puta dobio *Spammer* labelu je trajno suspendovan
- Korisnik koji trenutno poseduje *Spammer* labelu suspendovan je od svih aktivnosti privremeno
- *Harmful User* koji je napravio 3 *Harmful* objava je trajno suspendovan
- *Harmful User* koji dodatno dobije labelu *Spammer* je trajno suspendovan

¹ Podrazumevana labela koja bi se nalazila između *Trending* i *Poor Content* radi lakše realizacije mogla bi biti samo **Content**.

REALIZACIJA SISTEMA

Možemo izdvojiti nekoliko grupa pravila koja su logički odvojena:

1. Pravila za labeliranje objava
2. Pravila za labeliranje korisnika
3. Pravila za primenu kazni korisnika

Vremenske vrednosti svedene su sa realnih pojmova dana i meseci na znatno kraće – minute i sekunde. Odstupanje od realnog scenarija izvedeno je radi proverljivosti sistema u testnom okruženju.

PRAVILA ZA LABELIRANJE OBJAVA

Objava je postavljena u predhodnih 5min i
nema labelu *Poor Content* i
ima barem 10 lajkova i
spada u top 10% svih objava objavljenih u prethodnom 5min po broju lajkova
→ objava je označena kao *Trending*

Objava je duže od 5min imala labelu *Trending*
→ objava gubi labelu *Trending*

Objava ima status *Trending* i
objava ima status *Harmful* ili *Poor Content*
→ objava gubi labelu *Trending*

Objava ima odnos *like:dislike* jednak ili slabiji od 3:2
→ objava je označena kao *Poor Content*

Objava je dobila 10 ili više prijava
→ objava je labelirana kao *Potentially Harmful*

Objava je labelirana kao *Potentially Harmful*
→ pokreće se analiza objave

PRAVILA ZA UPRAVLJANJE KORISNICIMA

Korisnik je u prethodnih 3 min minimalno 5 puta minutno lajkovao objave označene kao *Content* ili *Trending* i
korisnik nije *Spammer*
→ korisnik postaje *Community Contributor*

Korisnik je *Community Contributor* i
u prethodnom minutu nije lajkovao minimalno 5 *Content* ili *Trending* objava
→ korisnik gubi status *Community Contributor*

Korisnik je *Community Contributor* i
korisnik je *Spammer*
→ korisnik gubi status *Community Contributor*

Korisnik je objavio minimalno jednu objavu svakih 30s u prethodna 2 minuta i
od gorepomenutih objava ni jedna nije *Poor Content* ili *Potentially Harmful* i
korisnik nije *Spammer* i
u prethodnih 24h korisnik nije imao *Harmful* objavu
→ korisnik postaje *Top User*

Korisnik je *Top User* i
objavio je *Poor Content*, *Potentially Harmful* ili *Harmful* objavu u prethodna 2 minuta
→ korisnik gubi status *Top User*

Korisnik je *Top User* i
korisnik ima labelu *Spammer*
→ korisnik gubi status *Top User*

Korisnik je *Top User* i
u perthodnih 30s nije postavio minimalno 1 objavu
→ korisnik gubi status *Top User*

Korisnik je izvršio više od 45 proizvoljnih akcija u minuti
→ korisnik postaje *Spammer*

Korisnik je objavio više od 10 objava u minuti
→ korisnik postaje *Spammer*

Korisnik ima barem jednu *Harmful* objavu
→ korisnik postaje *Harmful User*

Dislike/report aktivnost je sačinjala barem ½ svih korisnikovih radnji u protekla 24h i
korisnik je imao više od 30 aktivnosti u prethodna 24h
→ korisnik je označen kao *Potential Spammer*

Korisnik je *Potential Spammer* i

80% *dislike* aktivnosti u prethodna 24h su izvršene nad *Poor Content* ili *Harmful* objavama i

80% *report* aktivnosti u prethodna 24h su izvršene nad *Poor Content* ili *Harmful* objavama

→ korisnik gubi status *Potential Spammer*

Korisnik je *Potential Spammer* i

80% *dislike* aktivnosti u prethodna 24h su izvršene nad *Content* ili Trending objavama i

→ korisnik postaje *Spammer*

Korisnik je *Potential Spammer* i

80% *report* aktivnosti u prethodna 24h su izvršene nad *Content* ili Trending objavama i

→ korisnik postaje *Spammer*

Korisnikova „najsvežija“ labela *Spammer* je starija od 1min

→ korisnik gubi labelu *Spammer*

PRAVILA ZA PRIMENU KAZNI KORISNIKA

Korisnik je *Spammer* i

korisnik nije prethodno trajno suspendovan

→ korisnik je suspendovan 2min

Korisnik je 3 puta dobio *Spammer* labelu i

korisnik nije prethodno trajno suspendovan

→ korisnik je trajno suspendovan

Korisnik je napravio 3 ili više *Harmful* objava i

korisnik nije prethodno trajno suspendovan

→ korisnik je trajno suspendovan

Korisnik je *Harmful User* i

korisnik je dobio labelu *Spammer* nakon labele *Harmful User* i

korisnik nije prethodno trajno suspendovan

→ korisnik je trajno suspendovan

Primeri rezonovanja

Korisnik je objavio štetnu objavu (*forward chaining example*)

Korisnik objavljuje štetnu objavu koja potom biva viđena od strane drugih korisnika. Objava dobija preko 10 prijava nakon čega se kao potencijalno štetna analizira. *Third-party* analiza objave zaključila je da jeste u pitanju štetna objava, i time ažurirala podatak o istoj. Korisnik (ukoliko je ovo prvi put) dobija oznaku *Harmful user*. Korisnik gubi bilo koje druge pozitivne labela koje je možda u međuvremenu stekao.

Korisnik je rapidno prijavljivao sve objave (*CEP example*)

Korisnik, u ovom slučaju spammer, je rapidno prijavljivao sve objave prikazane na korisničkoj tabli. Napravio je preko 30 aktivnosti u poslednja 24h i polovina ili više su bile negativne interakcije. Korisnik postaje potencijalni spammer. Analizom sadržaja koji je korisnik prijavio utvrđuje se da 80% ili više prijavljenih objava nije zaista bilo lošeg sadržaja (po procenama drugih korisnika). Korisnik postaje spammer, gubi druge pozitivne labela, dobija suspenziju.

Određivanje uslova za *Trending* objavu (*accumulate example*)

Accumulate neće vršiti masovne transformacije ali hoće računati potreban uslov po broju pozitivnih interakcija kako bi novopristigla objava dobila status *Trending*.

Određivanje uslova za *Trending* objavu 2 (*globals example*)

Računanje gorepomenutog praga moglo bi da se svede na računanje istog svakih sat vremena, a ne prilikom svake interakcije korisnika. Tada se novodostignuti broj pozitivnih interakcija samo poredi sa prethodno izračunatom vrednosti, čime ne gubimo previše na tačnosti ali dobijamo na performansama sistema.

Izmena ponašanja sistema (*globals example 2*)

Većina značajnih vrednosti trebala bi da bude u vidu konstanti koje vlasnici aplikacije potom mogu lako da menjaju prema svojim potrebama.

Suspenzija korisnika (*međusobno isključivanje pravila*)

Ukoliko se pravilo kojim korisnik dobija trajnu suspenziju izvrši, nema potrebe izvršavati bilo koje drugo pravilo iz ove kategorije.