Predikcija kontroverznosti Reddit komentara

Dušan Blanuša
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
dusan.blanusa@uns.ac.rs

Tamara Lazarević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
tamaralazarevic@uns.ac.rs

***Apstrakt*—Povećanje popularnosti društvenih mreža i broja njihovih korisnika u poslednje dve decenije dovelo je do javne dostupnosti ogromne količine podataka. Ovi podaci mogu biti iskorišćeni za ispitivanje brojnih društvenih pojava. Mogućnost slobodnog javnog deljenja mišljenja na društvenim mrežama dovodi do polarizacije mišljenja, gde se neka od njih izdvajaju kao kontroverzna, tako što prouzrokuju veliku količinu i pozitivnih i negativnih reakcija. U radu izlažemo više pristupa za klasifikaciju kontroverznosti reddit komentara: kreiranje vektora osobina komentara upotrenom tf-idf i GloVe tehnike i klasifikovanje korišćenjem SVM klasifikatora, kao i klasifikovanje komentara upotrebom XLNET konvolutivne neuronske mreže. Svi navedeni pristupi su proizveli solidne rezultate, među kojima se kao najbolji pokazao pristup upotrebom konvolutivne neuronske mreže. Osim tekstualnih podataka, u srodnim istraživanjima je dodavanje netekstualnih karakteristika komentara dovelo do dodatnog povećanja performansi. Relevantne netekstualne karakteristike su ispitane u fazi analize podataka, a osobine za koje se pokazalo da su korisne za određivanje kontroverznosti komentara su uključene u tf-idf pristupu uz tekstualne osobine, što jeste dovelo do poboljšanja rezultata. Rano određivanje kontroverznosti potencijalno bi moglo omogućiti uspešnije održavanje reda u zajednicama. Budući da Reddit komentari predstavljaju kratke tekstualne sadržaje, napredak u klasifikaciji ovakvih sadržaja je značajan za oblast obrade prirodnog jezika.**

Ključne reči—kontroverznost; reddit; klasifikacija; komentar;

I. UVOD

Nagli rast drušvenih mreža u prethodnim godinama omogućio je korisnicima da javno dele svoja mišljenja i sadržaje. Ova otvorenost kreira platformu za formiranje polarizovanih mišljenja i kontroverznih diskusija.

Reddit je primer popularne onlajn platforme koja svojim korisnicima omogućava širenje sadržaja, znanja i mišljenja kroz objave koje se sastoje od tekstualnih i vizuelnih sadržaja. Omogućava korisnicima da izraze mišljenje o objavama i komentarima postavljanjem komentara i glasanjem za ili protiv. Glasovi su popularno nazvani *upvotes*, i predstavljaju pozitivne i *downvotes*, koji predstavljaju negativne glasove. Reddit korisnici mogu pristupiti zajednicama koje se nazivaju sabrediti (engl. *subreddits*) unutar kojih mogu učestvovati u diskusijama. Reddit komentari su markirani kao kontroverzni ako je ukupan broj glasova veći od propisane granice, i ukoliko se odnos glasova nalazi u okviru predefinisanih granica, što je prikazano na jednačini 1.

Jednačina 1 Kontroverznost komentara

$$k_i = \begin{cases} \textbf{kontroverzan}, & dg \leq \dfrac{\textbf{broj apvoutova}}{\textbf{broj glasova}} \leq gg \\ \textbf{nekontroverzan} & \end{cases}$$

Oznaka i predstavlja indeks komentara, dg i gg predstavljaju unapred definisane donje i gornje granice, u tom redosledu.

Kontroverzni sadržaj je sadržaj sa značajnim brojem i pozitivnih i negativnih ocena, i kao takav deli mišljenja i stavove zajednice u dve grupe Otkrivanje potencijalno kontroverznog sadržaja dovodi do efikasnijeg održavanja reda u zajednici, ili ukazuje na sadržaj koji nije zadovoljavajućeg kvaliteta. Osim toga, može predstavljati indikator problema u zajednicama kojima je potrebno posvetiti pažnju. Istraživanja pokazuju da se kontroverzni sadržaj na Redditu širi brže i dalje (1), pa je moguća primena pri kontrolisanju i širenju dometa korisnih informacija.

Jedna od karakteristika podataka poreklom sa društvenih mreža je njihova kratka priroda što predstavlja poseban izazov u oblasti obrade prirodnog jezika. U radu su upoređene različite metode formiranja osobina za klasifikaciju i različiti klasifikacioni modeli i upoređene njihove performanse za problem prepoznavanja kontroverznih komentara sa sajta reddit.com.

Relevantni radovi, koji su zasnovani na sličnim podacima poreklom sa društvenih mreža, bave se analizom podataka u različite svrhe. Među njima je otkrivanje postojanja eho komora koje grupišu korisnike koji dele mišljenja u zajednice (2), radovi koji se bave detekcijom sarkazma na osnovu tekstualnih sadržaja i drugih, netekstualnih osobina komentara i objava.

Predlažemo rešenje za detekciju kontroverznog sadržaja korišćenjem tekstualnih karakteristika Reddit komentara, unapređeno sa dodatnim netekstualnim karakteristikama koje utiču na kontroverznost komentara u Reddit zajednici.

U daljem radu biće istraženo više različitih pristupa za određivanje kontroverznosti komentara, korišćenjem više načina za kreiranje vektora osobina iz tekstualnog sadržaja, i poređenjem performansi različitih klasifikatora za te osobine. Osim tekstualnih, u obzir će biti uzete i netekstualne osobine komentara i opisana zapažanja o njihovom uticaju na određivanje kontroverznosti.

II. SRODNA ISTRAŽIVANJA

Prilikom odabira srodnih istraživanja, razmatrana su rešenja koja pripadaju istom domenu problema i koja su bazirana na sličnim obučavajućim skupovima podataka. U obzir su uzeta istraživanja koja se bave obradom tekstualnog sadržaja reddit komentara i objava u cilju klasifikacije kontroverznosti ili drugih ciljnih obeležja, rešenja koja se bave klasifikacijom kratkih tekstualnih sadržaja, kao i rešenja koja se bave analizom samog tekstualnog sadržaja reddit komentara. U radu koji se izdvojio kao jedan od najrelevantnijih (3), autori Hessel i Lee predstavljaju rešenje za klasifikaciju kontroverznosti reddit objava korišćenjem osobina dobijenih na osnovu tekstualnog sadržaja objava, kao i sadržaja i strukture ranih komentara koji iniciraju diskusiju. Obučavajući skup se bazira na 6 subreddita različitih tematika. Za klasifikaciju objava na osnovu tekstualnog sadržaja upoređuju više različitih modela, među kojima su najuspešniji modeli zasnovani na BERT osobinama, kao i modeli koji koriste ručno odabrane karakteristike u kombinaciji sa word2vec (4) embedding reprezentacijama.

U rešenju "Sarcasm Analysis using Conversation Context" (5) detekcija sarkazma izvršena je, među ostalim, i na obučavajućem skupu sačinjenom od reddit postova i komentara. Odabran je obučavajući skup od 50.000 instanci izbalansiranih među klasama. Najbolje rezultati ostvareni su korišćenjem LSTM modela. Primećeno je da recall malobrojnije klase znatno opada ukoliko se za obučavanje koriste neizbalansirani podaci.

Rad (6) predlaže hibridni pristup, u kojem se za modelovanje za detekciju sarkazma u diskusijama na onlajn društvenim mrežama koristi i sadržaj i kontekst. Iz toka diskusije se izvlače kontekstualne informacije, a za korisnike se kreiraju vektori koji opisuju njihov stil komunikacije i lične osobine. U kombinaciji sa konvolutivnim mrežama za izvlačenje tekstualnih karakteristika, postignut je značajan napredak u performansama klasifikacije na velikom Reddit korpusu.

Istraživanje "Sarcasm detection using context separators in online discourse" (7) za detekciju sarkazma koristi RoBERTa_large model za klasifikaciju sarkazma Twitter i Reddit obučavajućih skupova. Ispituju bitnost konteksta korišćenjem tri tipa ulaza: Odgovor, Kontekst+Odgovor i Kontekst+Odgovor (Razdvojeni).

III. OBUČAVAJUĆI SKUP

Sajt reddit.com je objavio skup podataka koji sadrži oko 1.7 milijardi javno dostupnih komentara, veličine preko jednog terabajta. Za izradu rešenja problema klasifikacije kontroverznih komentara preuzet je deo skupa koji sadrži komentare nastale u maju 2015. Godine sa sajta www.kaggle.com (8). Preuzeti skup se sastoji od 54504410 vrsta i 22 kolone, u vidu sqlite baze podataka, veličine 30 gigabajta. Skup sadrži velikom većinom nekontroverzne komentare, gde svega 3% ukupnog broja komentara čine kontroverzni i kao takav je nebalansiran.

Zbog obimnosti skupa, a i kako bi rad sa podacima bio jednostavniji, pet miliona vrsta za dalji rad preuzeto je iz baze podataka i sačuvano u obliku CSV (*coma separated values*) fajla.

Za različite pristupe korišćeni su podskupovi različitih veličina kreirani na osnovu navedenih 5 miliona kolona. Bez obzira na broj vrsta, svaki podskup, za svaki pristup, podeljen je na trening, validacioni i test podskup u odnosu 70/15/15.

IV. NETEKSTUALNE OSOBINE

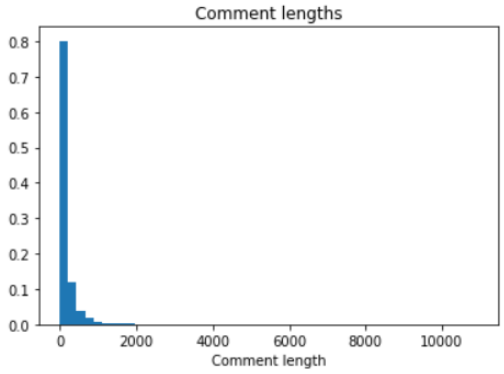
Osim osobina kreiranih na osnovu samog tekstualnog sadržaja komentara, mnoga radovi ostvaruju poboljšanje performansi uvođenjem kontekstnih osobina: vreme komentara (3), osobina korisnika (6), sadržaja na koji se komentar nastavlja (7) i slično. Obavljena je analiza netekstualnih obeležja obučavajućeg skupa u cilju otkrivanja pravilnosti i osobina koje bi mogle dovesti do poboljšanja performansi u kombinaciji sa tekstualnim osobinama.

Obučavajući skup sadrži ekstremno veliki broj podataka, pa je zbog hardverskih ograničenja odabran podskup originalnog skupa koji sadrži 5 miliona komentara. Popunjene su nedostajajuće vrednosti, a kolone sa tekstualnim sadržajem kovertovane su u numerički kako bi se moglo pristupiti obradi.

Razmatrani su uticaji sledećih netekstualnih osobina komentara na kontroverznost komentara: dužina komentara, autor, sabredit iz kojeg je komentar preuzet kao i *flair* autora.

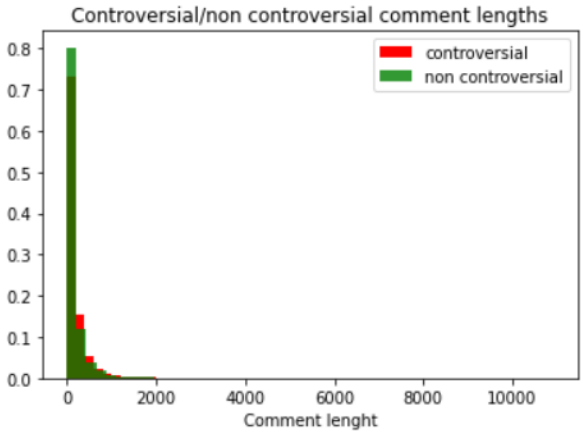
Flair sistem omogućava označavanje korisnika i objava. Svaki sabredit ima svoj skup flair-ova odabran od strane administratora. Neki sabrediti dozvoljavaju korisnicima da postave svoje, dok je na nekim samo moderatorima dozvoljeno da korisnicima dodele flair.

Jedna od osnovnih osobina komentara je svakako njegova dužina. Ispitivan je uticaj dužine komentara na kontroverznost, sa hipotezom da su komentari koji su deo ozbiljne diskusije na neku temu duži. Dodato je novo obeležje, dužina, tako što je na tekstualni sadržaj komentara primenjen lambda izraz za izračunavanje dužine, a zatim je dužina svakog komentara grupisana u jedan od 50 binova. Rezultati postupka su prikazani na slici 1.



Slika 1 – Dužina komentara

Na slici 2 prikazan je grafik sa informacijama o dužini kontroverznih i nekontroverznih komentara. Dužine komentara svake od klasa podeljene su u 50 binova. Izvodi se zaključak da postoji manji broj najkraćih komentara, a veći broj dužih među kontroverznim komentarima, tako da su kontroverzni komentari nešto duži od nekontroverznih, što potvrđuje početnu hipotezu.



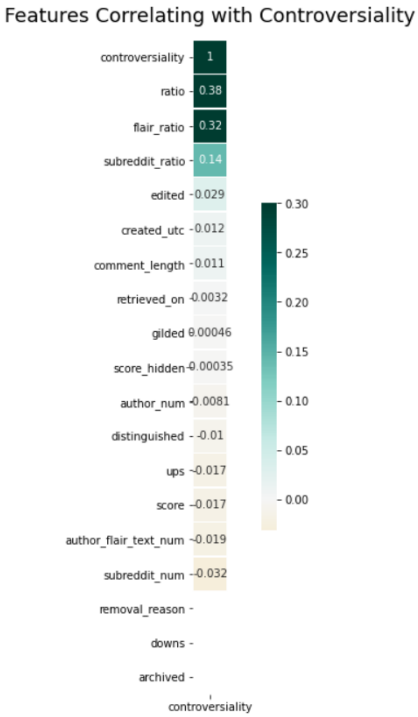
Slika 2 – Dužina kontroverznih i nekontroverznih komentara

Osim dužine, ispitivana je naklonost korisnika ka ostavljanju kontroverznih komentara, sa pretpostavkom da su neki autori naklonjeniji mišljenjima koje izazivaju kontroverzu. Podaci su transformisani tako što je dodato obeležje koje ukazuje na odnos broja kontroverznih komentara i svih komentara svakog od autora. Opadajućim sortiranjem novog obeležja utvrđeno je da postoji deo autora komentara, odnosno reddit korisnika, sa jako velikim odnosom kontroverznih komentara prema ukupnom broju ostavljenih komentara od strane tog korisnika, što ukazuje na moguću činjenicu da određeni korisnici imaju izražena mišljenja koja uzrokuju kontroverzu.

Isti princip primenjen je i na sabredite, gde se određeni sabrediti izdvajaju sa velikim odnosom kontroverznih komentara u odnosu na ukupan broj komentara. Neki od najkontroverznijih sabredita bave se fitnesom, politikom i problemima nejednakosti u društvu.

I za određene flair-ove uočena je tendencija da ostavljaju veći odnos kontroverznih komentara.

Proučena je i korelacija još nekih obeležja iz obučavajućeg skupa sa obeležjem kontroverznosti. Rezultati su prikazani na slici 3.



Slika 3 - Korelacija obeležja komentara i kontroverznosti

Neka od obeležja koja se izdvajaju su podatak da li je komentar izmenjen, vreme kreiranja komentara i podatak da li je komentar nagrađivan. Broj apvoutova i skor se ne mogu uzeti u obzir budući da se obeležje kontroverznosti direktno izračunava na osnovu ovih informacija.

Ispitivana je i zastupljenost određenih tema u kontroverznim komentarima. Tela kontroverznih komentara su tokenizovana, a zatim je kreiran skup imenica koje se u njima javljaju, sa pretpostavkom da imenice nose najveću količinu semantike. Wordcloud koji ilustruje učestalost reči prikazan je na slici 4.

Slika 4 - Wordcloud sa učestalostima reči u kontroverznim komentarima

Može se uočiti da se ističu teme kao što su: međuljudski odnosi, gejming, politika i problemi u društvu, reddit i svakodnevni život. U cilju grupisanja imenica u smislene celine, za 100 najčešćih reči su izdvojeni njihovi GloVe embedinzi, a zatim je izvršeno klasterovanje u 15 klastera. Za svaki od klastera izdvojen je centar, a zatim pronađen GloVe embedding sa najmanjim kosinusnim restojanjem od datog centra. Kao rezultat ovog procesa dobijeno je sledećih 15 tema:

1. fact
2. game
3. sh*t
4. time
5. thing
6. gun
7. reddit
8. child
9. state
10. fight
11. month
12. way
13. sub
14. war
15. edit

Zaključak je da se glavni motivi kontroverznih komentara poklapaju sa prethodno urađenim ručnim klasterovanjem tema, i da neke od njih zaista odgovaraju temama najkontroverznijih sabredita.

V. IZRADA REŠENJA

Ideja i pristup rešavanju problema i pretprocesiranju ulaznih podataka su bili isti za sve pristupe u izradi rešenja; od ulaznih podataka napraviti vektore, te ih kasnije klasifikovati u jednu od dve kategorije, prema tome da li su komentari označeni kao kontroverzni ili nekotroverzni. Embedinzi su kreirani u dva pristupa; manuelno, upotrebom TF-IDF tehnike uz dodatnu lematizaciju, i formiranjem *document-level* embedinga upotrebom GloVe (9) vektora. Kao klasifikatori korišćeni su SVM klasifikator i konvoluciona neuronska mreža XLNet (10) na bazi LSTM-a.

A. *SVM i TF-IDF*

Problemu binarne klasifikacije ulaznih podataka pristupilo se prvo implementacijom SVM (Support Vector Machine) klasifikatora. SVM predstavlja model nadgledanog učenja koji analizira ulazne podatke zarad njihove klasifikacije i regresije. Kako je ulazni skup podataka posedovao binarno labelirane ciljne klase a problem je klasifikacione prirode, SVM je bio idealan početni korak.

Prvobitno preuzeti skup podataka je sadržao 97% komentara obeleženih kao nekontroverzni, što je dovelo do velikog disbalansa i nemogućnosti obučavanja modela nad takvim skupom. Zbog toga je bilo neophodno isti i balansirati. Iz početnog skupa odabrano je jedanaest hiljada komentara označenih kao kontroverzni i isto toliko negativno označenih čime je set izbalansiran. Navedeni postupak poznat je kao *downsampling*. Kako se SVM usporava sa povećanjem broja ulaznih vektrora, ukupan broj od 22000 komentara je optimalne veličine za njegovo obučavanje. Radi smanjenja broja reči u korpusu i fokusiranja na reči koje nose značenje, čime se postižu bolje performanse modela, iz teksta komentara su izbačeni svi linkovi i karakteri koji nisu slova i brojevi.

Priprema ulaznih podataka predstavlja prvi korak klasifikacije podataka. U slučaju rešenja primenom SVM-a, ovu fazu čine tokenizacija i lematizacija nakon koje sledi TF-IDF vektorizacija. Lematizacija je, kao postupak dobijanja morfološki logičkih korena reči, bila idealna za smanjenje broja reči u korpusu, uz zadržavanje značenja rečenica. Korišćen je *WordNetLemmatizer* za Pajton programski jezik za engleski jezik. Steming (eng. *Stemming*) je odbačen u korist lematizacije, i pored svoje lakše obrade, upravo zbog nezadržavanja logike rečenica kreiranjem korena reči jednostavnim otklanjanjem prefiksa i sufiksa. Cilj ovog koraka je svođenje sličnih reči na zajednički oblik, te tako „stop“, „stopped“ i „stoping“, postaju „stop“ i broj različitih reči se znatno smanjuje. Prvi

veliki korak u pripremi podataka za klasifikator je lematizacija celog korpusa tekstualnih tela komentara. Nad svakom reči u obučavajućem skupu je odrađena lematizacija, nakon čega je sačuvan rečnik jedinstvenih reči u korpusu.

Tako obrađene komentare je bilo potrebno pretvoriti u vektore razumljive klasifikatoru. Za tu potrebu je odabran TF-IDF *vectoriser* koji dodeljuje vrednost rečima na osnovu frekvencije pojavljivanja. TF komponenta se odnosi na učestanost reči u dokumentu – što se reč više pojavljuje u jednom dokumentu, to je važnija, relevantnija za taj dokument. IDF komponenta odnosi se na učestanost reči ostalim dokumentima – što je reč češća u svim ostalim dokumentima, to je manje relevantna za taj konkretni dokument. Svaka reč u rečenici se pretvara u njoj odgovarajući broj, koji predstavlja njenu važnost u toj rečenici, te se svaka rečenica pretvara u onoliko dimenzioni vektor koliko ima jedinstvenih reči u celom korpusu. Reči koje se pojavljuju u rečenici dobijaju odgovarajuću vrednost, dok se ostale zamenjuju nulom. Na taj način se postiže jedinstvena dimenzionalnost svih vektora.

Vektori dobijeni kao proizvod primenjivanja postupaka navedenih u prethodnim korakacima su spremni za fitovanje klasifikatora. Odabran je SVM model iz *Sklearn Python* biblioteke, sa linearnim kernelom. Klasifikator je prvo fitovan podacima proizašlim isključivo iz tekstualnog sadržaja komentara, odnosno TF-IDF vektorima. Takav model je predstavljao osnovu za dalje razvijanje i menjanje ulaznih vektora u cilju povećavanja tačnosti klasifikacije. Pored tekstualnog sadržaja komentara, skup podataka je obuhvatao i sabredit kojem je pripadao komentar, autora, da li je sadržaj editovan ili nije, kao i dodatno izračunatu dužinu komentara. Nakon obučavanja prvobitnog klasifikatora, ulazni podaci su prošireni dodatnim, netekstualnim informacijama.

Inicijalni model, fitovan samo TF-IDF vektorima, dao je tačnost od 59% na test podacima, dok je model fitovan proširenim vektorima bio precizan u 66% slučajeva (tabela 1).

TF-IDF	59%
TF-IDF i subredit	64.33%
TF-IDF i da li je editovan	63%
TF-IDF i dužina komentara	64,6%
TF-IDF i svi prethodni	66%

Tabela 1 - Tačnost SVM pristupa

B. XLNet

XLNet je pretrenirani, generalizovani, autoregresivni model za razumevanje prirodnog jezika i predstavlja unapređenu verziju BERT pretreniranog modela specijalizovanu za klasifikaciju teksta. XLNet koristi Transformer XL (11) za ekstrakciju osobina iz tekstualnih ulaza, čime dobija dublji smisao o kontekstu jezika. Korišćenje XLNet pretreniranog modela se svodi na dotreniravanje modela na individualnim ulaznim podacima, kao što su u našem slučaju Reddit komentari.

Kreiranje rešenja korišćenjem XLNet pretreniranog modela se sastoji iz četiri koraka.

1. Učitavanje podataka
2. Kreiranje trening embedinga
3. Treniranje modela
4. Evaluacija performansi

Za XLNet klasifikator je korišćen isti skup podataka kao i za ostale pristupe, samo što je ovaj put učitano pet miliona komentara iz baze. Takav skup je bilo neophodno balansirati, nakon čega je brojao 123480 kontroverznih i isto toliko nekotroverznih komentara.

Nakon učitavanja podaka, kreirani su embedinzi. Za tu potrebu je preuzet XLNetTokenizer, koji se bavi razlaganjem složenih podataka na manje celine, odnosno razdvaja rečenice tako da svaka reč postane jedan token. Od tako tokenizovanih podataka se kreiraju vektori, menjanjem svakog tokena za XLNet indeks, uz povećavanje dimenzije svakog dobijenog vektora kako bi odgovarao najdužem vektoru. Kako bi model bio upotrebljiv i na rečenicama koje su veće od najvećih u trening skupu, vektorima je naknadno dodeljenja dužina od 128 indeksa. Za svaki vektor je zatim kreirana *attention* maska, koja je imala jedinicu na poljima gde je postojao indeks, a nulu tamo gde ga nije bilo, odnosno na mestima koja su dodata povećanjem dužine vektora. Ulazni vektori, labele kontroverznosti i maske su zatim podeljene na trening, test i validacioni skup u odnosu 70:15:15. Poslednji korak u pripremi podataka za obučavanje modela bilo je kreiranje *torch* tenzora od prethodno obrađenih vektora.

Preuzet je XLNetForSequenceClassification model i dotreniravanje je pokrenuto na NVIDIATesla T4 grafičkoj kartici. Testiranjem je utvrđeno da je model davao optimalne rezultate nakon obučavanja u dve epohe, sa *forward i backward* prolazima.

U toku obučavanja, model je validiran posle svake epohe (tabela 2.).

1. epoha	65.746%
2. epoha	67.737%

Tabela 2 – Validaciona tačnost XLNet pristupa

Nakon kraja obučavanja, model je testiran na test podacima, gde je preciznost klasifikacije bila 67.888%.

C. GloVe embedinzi + Klasifikator

Kao pristup kreiranju embedinga komentara na osnovu teksta, koji će biti ulaz u klasifikator primenjen je i pristup korišćenjem GloVe (9) embedinga. Kolekcija pretreniranih embedinga je javno dostupna za preuzimanje. GloVe embedinzi formiraju vektorski prostor reči sa strukturom koja ekodira značenje, i ova metoda pokazuje *state-of-of-the-art* rezultate na brojnim problemima sličnosti reči. Zasniva se na statistici zajedničkog ponavljanja reči u korpusu.

Korišćen je korpus od 6 milijardi tokena sa embedinzima od 300 dimenzija. Veći broj dimenzija omogućava bolje enkodovanje semantike i u teoriji se odražava u boljim performansama pri upotrebi.

Zbog ograničenja klasifikatora, iz masivnog obučavajućeg skupa izdvojen je podskup od 500.000 zapisa. Od tog broja samo su 2.2% komentara kontroverzni.

Problem jake neizbalansiranosti rešen je primenom *downsampling* metode, čime je obučavajući skup sveden na približno 11.000 kontroverznih, i isto toliko nekontroverznih komentara.

Komentari su pripremljeni za fazu klasifikacije tako što su iz njih uklonjeni znakovi interpunkcije za koje se pokazalo da najviše utiču na nepronalaženje reči komentara u GloVe korpusu. Nakon toga, tekstualni sadržaj komentara podeljen je na tokene. Svaki token je potražen u GloVe korpusu i odgovarajući embeding je dodat u skup vektora za komentar koji se trenutno obrađuje. Nepronađeni tokeni se najčešće odnose na izraze koji su karakteristični za redit i sleng. Primeri su: 'downvoted', 'subreddit', 'lmao', 'rekt'.

Na kraju procesuiranja komentara, od svih embedinga tokena za taj komentar, formira se vektor komentara na osnovu srednje vrednosti embedinga skupa embedinga.

Obučavajući skup je podeljen na trening, test i validacioni u odnosu 70/15/15.

Korišćenjem validacionog dela skupa, pristupa se traženju optimalnog klasifikatora za dati problem. Isprobano je 11 različitih tipova klasifikatora sklearn biblioteke. Dobijeni su sledeći f1 skorovi iz tabele 3.

'gnb'	0.617
'svm1'	0.633
'svm2	0.635
'svm3’	0.520
'mlp1'	0.568
'mlp2'	0.581
'ada'	0.605
'dtc'	0.555
'rfc'	0.598
'gbc'	0.627
'lr'	0.636

Tabela 3 - F1 skorovi klasifikatora

Na osnovu podataka iz pretrage, odabran je SVM (Support Vector Machine) klasifikator kao jedan od najboljih, i pristupljeno je Grid Search pretrazi za najboljim parametrima za ovaj model. Isprobano je 12 kombinacija za parametre C gama ovog klasifikatora, sa 5 foldova za svaku kombinaciju.

Kao optimalne vrednosti parametara pronađene su:

- 'C': 1
- 'gamma': 1

Klasifikovanjem na test skupu sa ovim parametrima dobijeni su rezultati iz tabele 4.

	Precision	Recall	F1-score	support
0	0.66	0.63	0.65	1670
1	0.62	0.65	0.63	1553
Accuracy			0.64	3323
Macro avg	0.64	0.64	0.64	3223
Weighted avg	0.64	0.64	0.64	3323

Tabela 4 – Performanse SVM klasifikatora

VI. ZAKLJUČAK

U ovom radu su prikazani različiti pristupi problemu binarne klasifikacije kontroverznosti tekstualnog sadržaja Reddit komentara, praćenog dodatnim, netekstualnim podacima. Predstavljeni su obučavajući skup, njegovo balansiranje, pretprocesiranje, kreiranje embedinga, obučavanje modela, validacija i testiranje.

Za kreiranje embedinga su korišćena dva pristupa, menuelni upotrebom TF-IDF tehnike uz dodatnu lematizaciju i formiranjem *document-level* embedinga upotrebom GloVe vektora. Korišćeni su SVM klasifikator i konvoluciona neuronska mreža XLNet na bazi LSTM-a.

GloVe embeding je u kombinaciji sa SVM klasifikatorom bio precizniji od manuelnog pristupa (tabela 5).

TF-IDF i SVM	58%
GloVe i SVM	64%

Tabela 5 – Poređenje GLoVe i TF-IDF pristupa

Isti slučaj je primetan i u odnosu SVM i XLNet klasifikatora, XLNet je bio u proseku 4% tačniji kada je klasifikovan samo tekstualni sadržaj (tabela 6).

SVM (GLoVe)	64%
XLNet	67.888%

Tabela 6 – Poređenje GLoVe + SVM I XLNet pristupa

Potvrđen je i obrazac dobijen obradom netekstualnog sadržaja skupa podataka. Na kontroverznost komentara nisu uticale samo reči njegovog teksta, nego i dužina komentara, da li je on editovan, kao i kom sabreditu pripada. Tako je dokazano da je najpreciznij bio upravo pristup sa kombinacijom vektora nastalih na osnovu tekstualnog sadržaja i netekstualnih informacija (tabela 7).

TF-IDF	58%
TF-IDF i netekstualni podaci	66%

Tabela 7 – Poređenje performansi sa i bez netekstualnih osobina

Ovako kreirano rešenje omogućava otkrivanje potencijalno kontroverznog sadržaja što dalje dovodi do efikasnijeg očuvanja reda i mira u onlajn zajednicama.

Planovi za dalji razvoj rešenja problema obuhvataju:

- Kombinovanje XLNet klasifikatora sa netekstualnim sadržajem
- Korišćenje većeg korpusa embedinga za GloVe, treniranog na podacima sa društvenih mreža, zbog većeg prisustva slenga
- Proširivanje GloVe pristupa netekstualnim sadržajem
- Proširivanje TF-IDF pristupa N-gramima radi boljeg razumevanja suštine rečenice
- Uključivanje strukture komentara kao parametra
- Uključivanje prethodnih komentara unutar komentara u cilju modelovanja konteksta
- Uključivanje klastera dobijenih analizom podataka kao klasifikacionih obeležja

BIBLIOGRAFIJA

1. *Controversial information spreads faster and further in Reddit.* **Mantzaris, Jasser Jasser and Ivan Garibay and Steve Scheinert and Alexander V.** 2020.
2. **Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini.** *Echo chambers on social media: A comparative analysis.* 2020.
3. **Hessel, Jack, and Lillian Lee.** *Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features.* 2019.
4. *Distributed representations of words and phrases and their compositionality.* **Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean.** 2013. NeurIPS.
5. *Sarcasm Analysis using Conversation Context.* **Debanjan Ghosh, Alexander R. Fabbri, Smaranda Muresan.** 2018.
6. *CASCADE: Contextual Sarcasm Detection in Online Discussion Forums.* **Devamanyu Hazarika, Soujanya Poria, Sruthi Gorantla, Erik Cambria, Roger Zimmermann, Rada Mihalcea.** 2018.
7. *Sarcasm Detection using Context Separators in Online Discourse.* **Dadu, Kartikey Pant and Tanvi.** 2020.
8. [Online] https://www.kaggle.com/reddit/reddit-comments-may-2015.
9. *GloVe: Global Vectors for Word Representation.* **Jeffrey Pennington, Richard Socher, and Christopher D. Manning.** 2014.
10. **Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le.** XLNet: Generalized Autoregressive Pretraining for Language Understanding. 2019.
11. **Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov.** Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. 2019.
12. *How did the discussion go: Discourse act classification in social media conversations.* **Subhabrata Dutta, Tanmoy Chakraborty and Dipankar Das.** 2018.