SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 000

Strojno učenje za analizu sentimenta u mikroblogovima

Ivan Križanić

Zagreb, svibanj 2020.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

TODO

SADRŽAJ

1. Uvod

Mikroblogovi su danas jedan od najčešće korištenih i najčešće proučavanih oblika komunikacije na internetu. Pronalazimo ih na iznimno popularnim društvenim mrežama, kao što su Twitter i Facebook, koji broje milijune korisnika diljem svijeta. Ljudi ih objavljuju u stvarnom vremenu, izražavajući svoje osjećaje, stavove i razmišljanja u svakodnevnom životu. Mnogi događaji i pojave u svijetu dobro su popraćeni reakcijama na drštvenim mrežama, stoga je korisno proučavati velike skupove objava kao izvor stajališta, preferenci, osjećaja i mnogih drugih svojstava koja se daju izvući iz značenja.

Ovaj se rad konkretno bazira na mikroblogovima društvene platforme Twitter. Takozvani Tweetovi, mikroblogovi platforme Twitter, kratke su poruke sačinjene od najviše 140 znakova. Prvi je objavljen 2005. godine, a dvije godine kasnije dnevno se objavljivalo 5000 mikroblogova. Po zadnjim poznatim podatcima taj broj iznosi preko 500 milijuna objava dnevno.(Salman Aslam, 2020) Radi se o iznimno velikom broju podataka koji kao skup mogu nositi korisne informacije, stoga ne čudi da postoje tvrtke koje u ponudi imaju analizu mikroblogova sa Twittera i drugih društvenih platformi (*brandmentions.com, mention.com*). Povratna informacija korisnika vrijedan je resurs kojim se tvrtke mogu obskrbiti, stoga analiza društvenih platformi ima velik ekonomski i društveni značaj. Obradi tako velikog broja podataka pristupa se tehnikama strojnog učenja, a konkretno područje koje se primjenjuje za ovakve zadatke naziva se obrada prirodnog jezika i još preciznije analiza sentimenta.

U radu sam se pozabavio problemom klasifikacije mikroblogova na one pozitivnog, neutralnog i negativnog sentimenta. Zadatak odgovara podzadatku A, četvrtog zadatka na natjecanju *Semeval 2017*, koji je u vrijeme održavanja privukao 39 timova iz cijelog svijeta. Ta je godina bila peta u nizu na kojoj se pojavio isti zadatak, što pokazuje interes zajednice za problem analize sentimenta. U sklopu zadatka napravio sam dva modela za klasifikaciju. Jedan pripada standardnom strojnom učenju i temelji se na SVM-modelu s linearnom jezgrom, a drugi pripada području dubokog učenja i temelji se na LSTM inačici modela .

2. Povezani radovi

Na temu analize sentimenta napisano je mnogo radova, a velik broj bavi se upravo mikroblogovima s društvenih mreža i to vrlo često upravo Twitterom. Uz to, u sklopu natjecanja *Semeval* neki natjecatelji objavljuju i rad u kojem se osvrću na svoju implementaciju rješenja. Stoga je dostupno puno informacija koje se mogu iskoristiti za vlastitu implementaciju, ali je istovremeno i otežano implementirati neviđeno rješenje. Najbolji rezultat moje implementacije ima točnost od 64%, što odgovara 14. mjestu na ljestvici predanih implementacija natjecanja *Semeval 2017*. (Sara Rosenthal, Noura Farra, Preslav Nakov, 2017) Prvo mjesto sa točnošću od 68.1% podijelila su dva tima: *DataStories* i *BB_twtr*. Upravo je tim *BB_twtr* zaslužan za aktualan *state-of-the-art* model u području analize sentimenta mikroblogova. Njihova trenutna implementacija hvali se da ostvaruje *F1-score* u iznosu od 68.5%.

U sljedećih nekoliko odlomaka osvrnut ću se na radove koji su mi služili kao izvor metoda i ostalih informacija koje sam koristio u izradi svoje implementacije.

2.1. Rad tima *BB_twtr* - najuspješniji model današnjice

Prvi u nizu radova na koje se želim osvrnuti jest rad pobjednika natjecanja *Semeval 2017*, a ujedno i aktulani *state-of-the-art* model u području analize sentimenta mikroblogova. Radi se o radu *BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs* (Cliche, 2017). Poblemu su pristupili tehnikama dubokog učenja. Prva faza rada bavi se izradom vektora riječi koji su dalje korišteni u treniranju CNN i LSTM modela mreža. Eksperimentirali su s tri različite tehnike izrade vektora riječi (*Word2Vec, FastText, GloVe*). U drugoj su fazi nenadziranim učenjem razdijelili sentiment na negativan i pozitivan, jer je prije toga sentiment polariteta u vektorima bio vrlo slab. U trećoj su fazi provodili nazdirano učenje koristeći podatke sa natjecanja i model izgrađen od 10 CNN i 10 LSTM mreža koje koriste različit broj epoha za treniranje i različite vektore riječi. U podzadatku A postigli su točnost od 68.1%, a model su koristili i u ostala 4 podzadatka natjecanje te su u svim zadatcima ostvarili

najbolji rezultat.

Budući da navode CNN i LSTM modele mreža kao najbolje u području analize sentimenta, u svojoj sam implementaciju upotrijebio LSTM model mreže kako bih se upoznao s njegovim mogućnostima. Umjesto izrade vektora riječi iz velikog skupa mikroblogova, odlučio sam se koristiti gotove vektore iz biblioteke *Spacy* koja koristi vektore izrađene metodom *Word2Vec*. Pri tome gubim prednosti posebnih značajki koje su karakteristične za jezik mikroblogova, a koje bi se mogle pokazati u vektorima nastalim na temelju mikroblogova, ali pristup je jednostavniji i štedi znatnu količinu računalne obrade koja bi bila potrebna za izradu vlastitih vektora.

2.2. Rad tima *DataStories*

U podzadatku A natjecanja *Semeval 2017*, zadatka 4, prvo mjesto dijelila su dva tima, ali tim *DataStories* imao je niži *F1-score*. Svoj su pristup opisali u radu *DataStories at SemEval-2017 Task 4: Deep LSTM with Attention forMessage-level and Topic-based Sentiment Analysis* (Baziotis et al., 2017). S obzirom na to da su prethodnih godina ostvarili slabije rezultate, dok su timovi koji su koristili pristup dubokog učenja pretežno zauzeli pozicije na vrhu, *DataStories* tim odlučio je skrenuti pažnju s klasičnog strojnog učenja na duboko učenje. Rad su podijelili na dva osnovna koraka: obradu teksta i traniranje modela. Za obradu teksta implementirali su vlastite funkcije koje su primjenjive u općoj upotrebi, ali su usmjerene na obradu mikroblogova s Twittera. Za izradu vektora riječi koristili su 330 milijuna neoznačenih mikroblogova na engleskom jeziku. Na vokabularu od 660 tisuća riječi koristili su *GloVe* metodu izrade vektora. U obradi teksta koristili su vlastiti tokenizator koji je prilagođen Twitteru i posjeduje mogućnost izvlačenja raznih elemenata poput datuma, valuta, emotikona i sličnih sadržaja. Za razliku od njih, u svojoj implementaciji koristim implementaciju tokenizatora iz biblioteke *SpaCy* jer je pristupačna i široko korištena. U daljnoj su obradi primjenili standardne postupke pročišćavanja teksta koji se koriste u obradi prirodnog jezika.

Osvrnuli su se na CNN i naglasili problematiku gubitka informacije o poretku riječi prilikom uporabe istih. Iz tog su razloga preferirali RNN, konkretnije napredniju izvedenicu koja primjenjuje ćelije s dugoročnom memorijom odnosno LSTM. U svojoj su implementaciji koristili dvoslojni dvosmjerni model s mehanizmom za pozornost koji pospješuje prepoznavanje korisnih težina. U LSTM sloju modela koristili su 150 neurona i trenirali s podskupovima od 128 podataka. U testiranju su naveli kako mehanizam pozornosti doprinosi rezultatu za 0.04% te ga stoga nisam implementirao u svoj model.

2.3. Rad tima *TakeLab* - pristup klasičnim strojnim učenjem

Za razliku od velikog broja ekipa na natjecanju, tim *TakeLab* odlučio se za pristup klasičnim metodama strojnog učenja. Koristili su skup ručno izrađenih značajki i trenirali na SVM-modelu s linearnom jezgrom. Kao značajke koriste *Tf-Idf* i gotove vektore riječi, ali i neke specifične značajke poput leksikona pozitivnih i negativnih riječi te posebnu značajku po kojoj je rad dobio ime: "*Nedavne smrti i moć nostalgije*", odnosno originalni engleski naziv *Recent Deaths and the Power of Nostalgia in Sentiment Analysis in Twitter* (Lozić et al., 2017). Značajka se temelji na činjenici da je sentiment mikroblogova koji spominju nedavno preminute ljude pretežno pozitivan jer se ljudi obično prisjećaju pozitivnih stvari vezanih za pokojnika. Također su iskoristili svojstva nostalgije koja upućuju na pretežno pozitivan sentiment prilikom spominjanja pojmova i pojava iz prijašnjih vremena. Za značajne ljude kreirali su značajke koje opisuju osobe s atributima svojestvenima njohovoj društevnoj ulozi, a za pojmove kojima je često pridjeljenja nekakva ocjena, npr. filmovi, igrice, glazba i slično, napravili su značajke koje donose informacije o uspješnosti i popularnosti pojma. Svojim SVM-modelom ostvarili su solidan plasman u nekoliko zadataka, a u zadatku kojime se bavi ovaj rad ostvarili su 16. mjesto.

Potaknut njihovim radom, a i mnogim drugima koji koriste ručno izrađene značajke, u svojoj implementaciji iskoristio sam leksikon pozitivnih i negativnih riječi, te sam iskoristio SVM-model. Naprednije i inovativne značajke koje čine ovaj rad posebnim nisam implementirao.

3. Model

3.1. O zadatku 4 natjecanja *Semeval 2017* i analizi sentimenta u mikroblogovima

Verzija zadatka s kojim sam se bavio u ovom radu peta je u nizu na najtecanju *Semeval*. Kao i svih prijašnjih godina, zadatak je bio poprilično popularan i privukao 48 timova koji su sudjelovali u različitim podzadatcima. U zadatku se kroz godine pojavilo nekoliko podzadatka kao što su ocjena pripadnosti sentimenta mikrobloga određenoj temi i skaliranje pripadnosti na skali od 1 do 5. Osnovna verzija zadatka bavi se klasifikacijom mikroblogova u tri razine polariteta, preciznije u mirkoblogove pozitivnog, neutralnog i negativnog sentimenta.

Analiza sentimenta u tekstovima kao što su mikroblogovi s društvenih mreža donosi razne poteškoće s kojima se ne moramo nositi kada je riječ o standardnijim oblicima teksta. Problematične karakteristike mikroblogova su niska ograničenost broja znakova koja uzrokuje sažet izraz, ali i povećava uporabu kolovijalnih izraza, skraćenica i raznih suvremenih novotvorenica koje bismo mogli okarakterizirati kao *slang*. Prisutni su i razni elementi koji ne pripadaju prirodnom jeziku kao što su emotikoni, hiperlinkovi i razne oznake kao npr. oznaka korisničkog imena koja ima oblik *"@user"*. Hiperlinkovi u takvim kratkim tekstovima često nose velik teret značenja, odnosno često tek uz informaciju o sadržaju na koji hiperlink pokazuje možemo pravilno ocijeniti sentiment same poruke. Problem je i u pravopisu, korisnici često mijenjaju riječi radi postizanje vizualnog ili nekog drugog efekta, pa tako možemo naići na tvorevine poput: "ŁoŁ", "ca$h", "©ool", i slične koje bi bilo poželjno prepoznati i pretvoriti u smislene riječi ili kratice. Također je pristuno nizanje istog slova u riječima poput "*cool*" koje možemo pronaći u obliku kao što je "*cooool*" ili negaciji *"no"* kojoj se često nadodaje zadnje slovo *"o"*. Takvim je riječima također poželjno ukloniti suvišne znakove kako bi se pronašče u rječniku, ali treba imati na umu da takvo ponavljanje znakova nosi značenje u sebi, a koje bismo klasičnim ispravljanjem pravopisa izgubili.

Korisno bi bilo prepoznati i skrivene riječi kao što su *"F**k"*. *"S**t"*, *"N***a"* jer su to često riječi koje mogu znatno utjecati na sentiment objave, no to nije tako jednostavan zadatak zbog raznih metoda kojima se takve, često proste riječi, pokušavaju ukomponirati u tekstove.

Kada se odmaknemo od početne obrade teksta nailazimo na nove poteškoće kao što su korištenje sarkazma i učestalost ciničnog tona koji u potpunisti mijenjaju polaritet sentimenta, a koje je vrlo teško prepoznati iz perspektive modela. Ograničenost duljine poruke posljedično donosi manjkavost izraza koji se često bolje razumiju ukoliko se posjeduje znanje o svijetu i vremenu u kojem su napisani, a ne samo o jeziku i značenju istog, što je još jedno svojstvo koje je vrlo teško ostvarivo modelima strojnog učenja.

3.2. Odabir metoda i pristupa

U rješavanju problema koristio sam dva različita pristupa kako bih se osvijestio o prednostima i manama jednog i drugog. Prvi pristup pripada klasičnim metodama strojnog učenja i temelji se na vlastaručnoj izradi značajki i upotrebi SVM-modela. Drugi pristup pripada grani strojnog učenja koja se naziva duboko učenje i temelji se na značajkama nastalima od vektora riječi i LSTM inačici modela povratne neuronske mreže.

3.2.1. Klasično strojno učenje - SVM model

3.2.2. Duboko učenje - LSTM model

3.3. Značajke

S obzirom na to da paralelno koristim dva pristupa izrade i treninga modela, izrada i korištenje značajki također je podijeljena u dva smjera. Izrada značajki za klasično strojno učenje znatno je opsežniji i kreativniji proces nego izrada istih za pristup dubokim učenjem. Moglo bi se reći da je srž klasičnog pristupa upravo u izradi značajki jer se modeli sami po sebi ne mogu značajno konfigurirati, pa rezultat najviše ovisi o onome što mu na ulazu pružimo. Kod dubokog učenja postoje jednostavni standardni pristupi koji su ponekad gotovo mandatorni.

3.3.1. Značajke u klasičnom pristupu

Značajke u ovom pristupu čine glavnu okosnicu uspjeha modela, pa je stoga izradi posvećen znatan udio vremena. Značajke se mogu grupirati u tri kategorije:

– brojanje riječi i vektori riječi

– polaritet i sentiment riječi

– brojanje prisutnosti elemenata

Brojanje riječi i vektori riječi

U ovoj se kategoriji nalaze dvije vrste značajki koje se međusobno isključuju, odnosno ne koriste se istovremeno. Prva značajka je uobičajena kao početni uzorak značajki koji se koristi za treniranje osnovnog modela (engl. *baseline model*) kao referenca za daljnje eksperimente. Radi se se o metodi vreće riječi (engl. *Bag-of-Words*), odnosno preciznije o primjeni mjere učestalosti riječi *TF-IDF* (engl. *term frequency-inverse document frequency*). Mjera se definira na sljedeći način: potrebno je definirati dvije zasebne statističke mjere – mjeru učestalosti izraza (*tf*) i inverznu učestalost u dokumentima (*idf*). Prvu mjeru koja označava učestalost pojave riječi računamo na sljedeći način:

$$\mathrm{tf}(t, d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

gdje t predstavlja izraz, d predstavlja dokument, odnosno skup svih izraza u promatranom tekstu, $f_{t,d}$ predstavlja broj pojavljivanja izraza u tekstu, a brojnik predstavlja neveći broj pojavljivanja nekog izraza u tekstu. Ova se normalizirana verzija učestalosti pojave izraza koristi radi sprječavanja pristranosti prema velikim tekstovima. Druga mjera koju određujemo je inverzna učestalost izraza u dokumentu i računa se na sljedeći način:

$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

gdje D predstavlja skup svih tekstova na kojima računamo učestalost, a nazicnik razlomka predstavlja broj pojavljivanja izraza u tekstu, dok je N ukupan broj dokumenata u skupu D. Konačna mjera jednaka je umnošku dviju prethodno izračunatih mjera, odnosno:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

U implementaciji rješenja koristim gotovu metodu iz knjižnice *scikit-learn*. Korištenjem nastalih značajki u SVM-modelu dobio sam točnost od 39.5%, što je korektan osnovni model s obzirom na točnost nasumičnog odabira koja iznosi 33.3%.

Druga značajka koju sam uveo, a koja pripada ovoj kategoriji, temelji se na vektorima riječi iz knjižnice *Spacy*. Radi se o vektorima s 300 dimenzija nastalima primjenom *Word2vec* (Mikolov et al., 2015) metode. *Word2vec* je model plitke neuronske mreže s dva sloja koji treniranjem pokušava rekonstruirati lingvističko značenje riječi. Kao ulaz koristi vrlo velik skup tekstualnih podataka koji mogu biti raznog porijekla, kao npr. članci *Wkipedie*, objave na društvenim mrežama, primjerci elektroničke pošte itd. Kao rezultat nastaju višedimenzionalni vektori čije se dimenzije obično kreću između 100 i 1000 dimenzija. Vektori su u prostoru smješteni na način da su vektori riječi bliskog znanja prostorno bliski jedan drugome.
Koristeći prethodno izrađene informacije proizašle iz metode *tf-idf* kodirao sam ulazne informacije vektorima riječi na način da sam izračunao prosječnu vrijednost vektora svih riječi koje se pojavljuju u mikroblogu, a za broj značajki u tako nastalom vektoru odabrao sam vrijednost broja riječi u najduljem mikroblogu. Poboljšanje nastalo korištenjem ovih značajki značajno je povečalo uspješnost modela koji je nakon treniranja imao točnost od 60.9%

Polaritet i sentiment riječi

S obzirom na to da je zadatak klasifikacija s obzirom na polaritet mikrobloga, dotakao sam se polariteta i sentimenta samih riječi. Za to sam iskoristio dva leksikona. Prvi od njih je leksikon ocjena riječi temeljen koji sadrži ocjene po atributima zadovoljstva, uzbuđenosti i dominantnosti po imenu *Affective Norms for English Words* (Bradley i Lang, 1999). Prva verzija sastojala se od nešto više od tisući riječi, no 2013. godine proširena je na 14000 (Warriner et al., 2013). Pri implementaciji rješenja ugledao sam se na pristup prisutan u repozitoriju *dwzhou/SentimentAnalysis* (Doris Zhou). Drugi leksikon koji sam koristio je zapravo običan popis pozitivnih i negativnih riječi. Dodavanjem značajki dobivenih korištenjem leksikona podigao sam točnost modela za 1%, odnosno postigao točnost od 61.9%.

Brojanje prisutnosti elemenata

U ovoj se kategoriji nalaze značajke nastale brojanjem ili promatranjem prisutnosti raznih elemenata u mikroblogovima. Elementi čiju sam prisutnost naznačavao zastavicama 0 ili 1 su: *e-mail* adrese, hiperlinkovi, znakovi valute, datumi, telefonski brojevi itd. Za neke sam elemente bilježio točan broj pojavljivanja u mikroblogu, a izrazi za koje sam bilježio tu informaciju su: izrazi sa nizanjem znakova (npr. *"cooool"*), izrazi napisani velikim slovima, cenzurirani izrazi (npr. *"F****), broj ponovljenih riječi, po-

jave takozvanih *hashtagova* i broj uskličnika. Kao dodatnu značajku nadodao sam i ukupan broj riječi u rečenici. Dodavanjem ovih značajki ostvario sam porast točnosti od 0.4%, odnosno postigao sam konačnu točnost od 62.32%

3.3.2. Značajke u dubokom učenju

4. Zaključak

Zaključak.

LITERATURA

Christos Baziotis, Nikos Pelekis, i Christos Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. U *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, stranice 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

Margaret M Bradley i Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.

Mathieu Cliche. BB_twtr at SemEval-2017 task 4: Twitter sentiment analysis with CNNs and LSTMs. U *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, stranice 573–580, Vancouver, Canada, Kolovoz 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2094. URL `https://www.aclweb.org/anthology/S17-2094`.

Doris Zhou. Implementations of various sentiment analysis methods in python. URL `https://github.com/dwzhou/SentimentAnalysis`. [Online; accessed 13-April-2020].

David Lozić, Doria Šarić, Ivan Tokić, Zoran Medić, i Jan Šnajder. TakeLab at SemEval-2017 task 4: Recent deaths and the power of nostalgia in sentiment analysis in twitter. U *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, stranice 784–789, Vancouver, Canada, Kolovoz 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2132. URL `https://www.aclweb.org/anthology/S17-2132`.

Tomas Mikolov, Kai Chen, Gregory S Corrado, i Jeffrey A Dean. Computing numeric representations of words in a high-dimensional space, Svibanj 19 2015. US Patent 9,037,464.

Salman Aslam. Twitter by the numbers: Stats, demographics and fun facts, 2020. URL https://www.omnicoreagency.com/twitter-statistics/. [Online; accessed 13-April-2020].

Sara Rosenthal, Noura Farra, Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. 2017.

Amy Beth Warriner, Victor Kuperman, i Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

POJMOVNIK

CNN konvolucijska neuronska mreža (engl. *Convolutional neural network*). 3

LSTM ćelija s dugoročnom memorijom (engl. *Long short-term memory*). 1, 3, 6

RNN povratna neuronska mreža (engl. *Reccurent neural network*). 3

SVM stroj potpornih vektora (engl. *Support-vector machine*). iv, 1, 4, 6

Strojno učenje za analizu sentimenta u mikroblogovima

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: strojno učenje, duboko učenje, obrada prirodnog jezika, analiza sentimenta, analiza mikroblogova, Semeval

Machine Learning for Sentiment Analysis in Microblogs

Abstract

Abstract.

Keywords: Machine learning, Deep learning, Natural language processing, Sentiment analysis, Microblogs analysis, Semeval.