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`.

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-u sa linearnom jezgrom, a drugi pripada području dubokog učenja i temelji se na LSTM modelu.

2. Povezani radovi

Na temu analize sentimenta napisano je mnogo radova, a velik broj bavi se upravo mikroblogovima sa 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 62.9%, 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 osvrnuti ću se na radove koji su mi služili kao izvor metoda i ostalih informacija koje sam korsitio 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 mreža. Eksperimentirali su sa 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 mreže kao najbolje u području analize sentimenta, u svojoj sam implementaciju upotrijebio LSTM model kako bih se upoznao sa njegovim mogućnostima. Umjesto izrade vektora riječi iz velikog skupa mikroblogova, odlučio sam se koristiti gotove vektore iz biblioteke *Spacy* koji 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 sa 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 sa 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 podvrstu, *LSTM* mrežu. U svojoj su implementaciji koristili dvoslojni dvosmjerni LSTM model sa mehanizmom za pozornost koji pospješuje prepoznavanje korisnih težina. U *LSTM* sloju modela koristili su 150 neurona i trenirali sa 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 ofNostalgia 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 sa atributima svojestvenima njohovoj društevnoj ulozi, a za pojmove kojima je često pridjeljenja nekakva ocjena, npr. filmovi, igrice, glazba, 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. Naprednije i inovativne značajke koje čine ovaj rad posebnim nisam implementirao.

3. 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.

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`.

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`.

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.

Strojno učenje za analizu sentimenta u mikroblogovima

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.