KLASIFIKACIJA SENTIMENTA U TWITTER POSTOVIMA KORIŠĆENJEM UDALJENOG NADZORA
TWITTER SENTIMENT CLASSIFICATION USING DISTANT SUPERVISION

Nikola Jolić
Elektrotehnički fakultet
Univerziteta u Beogradu

Sadržaj – *U ovom radu predstavljen je novi način klasifikacije sentimenta u Twitter porukama, kao i opis jednog web-alata koji se ovom tematikom bavi. Ove poruke su klasifikovane kao pozitivne ili negativne u zavisnosti od upita pretrage. Predstavljeni su metodi mašinskog učenja za klasifikaciju i njihovi rezultati. Podaci za treniranje ovih algoritama su Twitter poruke zajedno sa emotikonima koji predstavljaju labele sa šumom. Pokazano je da algoritmi mašinskog učenja (Naive Bayes, Maximum Enthropy, SVM) imaju tačnost preko 80% kada su istrenirani sa podacima sa emotikonima. Takođe su opisani postupci preprocesiranja podataka u cilju postizanja visokih tačnosti.*

Abstract – *This document presents a novel approach in sentiment classification of Twitter messages, as well as it describes a web-application which deals in this subject. These message are classified as positive or negative with respect to a query term. It presents machine learning methods for classification as well as their results. Training data consists of Twitter messages with emoticons which represents noisy data. It is shown that machine learning algorithms (Naive Bayes, Maximum Enthropy, SVM) achieve accuracy over 80% when trained with data with emoticons. Also, data preprocessing methods are shown in terms of geting higher accuracies.*

UVOD

Twitter je popularna socijalna mreža (mikrobloging) gde korisnici ostavljaju kratke poruke („tvitove“). Ove poruke često pokazuju emocije korisnika o raznim temama. Ovaj rad predlaže način za automatsku ekstrakciju osećanja iz twitter poruka.

Ovo je veoma korisno, jer korisnici mogu koristiti analizu sentimenta da prouče više o proizvodima i uslugama pre nego što ih kupe. Takođe oglašivači mogu benefitirati tako što izuče mišljenje javnog mnjenja o njihovoj kompaniji i proizvodima. Organizacije mogu da dobiju važne povratne informacije o svojim novim proizvodima.

Većina istraživanja u ovoj tematici je fokusirana na klasifikaciji velikih tekstova, kao što su recenzije. Tvitovi se razlikuju od njih po tome što mogu sadržati maksimalno 140 karaktera i generalno nisu toliko pažljivo osmišljeni. Ali ipak, mogu dati dosta informacija o osećanju autora o određenoj temi.

Da bi se istrenirao klasifikator, obično su potrebni ručno obeleženi podaci. Sa velikim brojem različitih tema diskutovanih na Tviteru, bilo bi veoma teško prikupiti toliko veliki broj podataka. Predloženo rešenje je korišćenje „udaljenog nadzora“ gde se podaci sastoje od tvitova sa emotikonima. Npr. emotikon :) govori da se radi o tvitu sa pozitivnom emocijom, a :(govori da se radi o negativnoj emociji. Uz pomoć Tviter API-ja, lako je doći do podataka koji su nam potrebni.

DEFINICIJA SENTIMENTA

Za potrebe izrade ovog rada, difinisaćemo sentiment kao „lični pozitivni ili negativni osećaj“.
Često može biti nejasno da li tvit uopšte sadrži osećaj. Za taj slučaj se može postaviti pitanje da li se taj tvit može postaviti kao naslov novinskog članka. Ako može, tada se taj tvit može smatrati neutralnim. Takvi tvitovi nisu uzeti u razmatranje prilikom analize sentimenta. Uzeti su samo tvitovi koji izražavaju pozitivne ili negativne emocije.

PRISTUP PROBLEMU

Problem se rešava primenom različitih algoritama mašinskog učenja koji su Naive Bayes, Maximum Entropy i SVM. Izdvajanje svojstava se radi korišćenjem unigrama, bigrama, unigrama i bigrama i unigrama sa svojstvima iz govora. Prave se frejmvorci koji klasifikatore i izdvajače svojstava tretiraju kao dva entiteta. Oni nam omogućavaju da isprobamo različite kombinacije klasifikatora i izdvajača svojstava.

EMOTIKONI

Pošto emotikoni kao podaci za trenitanje algoritma predstavljaju podatke sa šumom, vrlo je važno prodiskutovati njihovu ulogu i upotrebu.

Za treniranje algoritma, emotikoni se izbacuju iz podataka za treniranje. Ukoliko ih ostavimo, javiće se negativan uticaj na tačnosti MaxEnt i SVM klasifikatora, a skoro nikakav na tačnost Naive Bayes. Razlika je u matematičkim modelima i težinskim selekcijama koje koriste ovi algoritmi.

Izbacivanje emotikona podstiče klasifikatore da uče iz drugih svojstava (unigrama i bigrama) koji se nalaze u tvitovima. Pomoću njih se određuje polaritet.

Emotikoni se tretiraju kao podaci sa šumom, jer ne mogu tačno odrediti sentiment poruke.

IZBACIVANJE SVOJSTAVA

Jezički model Tvitera sadrži mnoštvo različitih svojstava. Sledeća svojstva možemo iskoristiti tako da smanjimo prostor svojstava.

Korisnička imena Korisnici često u svoje tvitove uključuju korisnička imena (usernames) drugih korisnica koristeći ih kao referencu ka njima. Standard je takav da se koriste uz prefiks @.

Korišćenje linkova Korisnici često i ubacuju u svoje tvitove linkove. Tada se za ceo link može koristiti klasa ekvivalencije sa ključem „URL“.

Ponovljena slova U tvitovima se neretko koristi „kežual“ model jezika. Tako će neko umesto „gladan“ napisati „glaaaadaaaaan“, time hiperbolisati svoje osećanje. Da bi se smanjio prostor svojstava, svaki nailazak na slovo koje se pojavljuje više od drugi put se ignoriše i ceo token se pretvara u „glaadaan“.

METODI MAŠINSKOG UČENJA

U testiranju se koriste sledeći klasifikatori: ključne reči, Naive Bayes, Maximum Entropy (MaxEnt) i Support Vector Machines (SVM).

OSNOVNI

Vebsajt Twittratr je sajt koji radi analizu sentimenta Tviter postova. Kao osnovni pristup, koristi se njihova lista reči, koja je javno dostupna i besplatna. Za svaki tvit, prebroje se pozitivne i negativne i na osnovu toga tvit se polarizuje kao pozitivan ili negativan.

NAIVE BAYES

Naive Bayes je jednostavan model koji dobro radi na kategorizaciji teksta. Koristi se multinomijalni model. Klasa *c* je dodeljena tvitu *d*, gde je

$$c* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c)\sum_{i=1}^{m} P(f|c)^{n_i(d)})}{P(d)}$$

MAXIMUM ENTROPY

Ideja u Maximum Entropy modelima je se uzmu što uniformniji modeli da bi se zadovoljila zadata ograničenja. Za razliku od Naive Bayes pristupa, MaxEnt, ne pretpostavlja nikakavu nezavisnost između svojstava. To znači da možemo dodati svojstva kao što su bigrami i fraze u MaxEnt, ne brinući da li će se preklapati. Model je predstavljem sledećom formulom:

$$P_{ME}(c|d,\lambda) = \frac{\exp[\Sigma_i \lambda_i f_i(c,d)]}{\Sigma_{c'} \exp[\Sigma_i \lambda_i f_i(c,d)]}$$

Težinski vektor (lambda) odlučuju o važnosti svojstva u klasifikaciji. Veća težina označava da to svojstvo više utiče o odluci o polaritetu tvita.

SUPPORT VECTOR MACHINES

Koristi se SVM sa linearnim jezgrom. Ulazni podaci su vektori veličine *m*. Svaki ulaz u vektoru odgovara prisutnosti svojstva. Npr. kod unigrama, svako svojstvo predstavlja reč pronađenu u tvitu. Ukoliko je pristuna, vrednost je 1, u suprotnom 0. Koristi se metod prisutnosti, za razliku od broja, tako da se ulazni podaci ne moraju skalirati i time se ubrzava samo vreme procesuiranja.

EVALUACIJA

EKSPERIMENT

Za data-set se koristi Tviterov API pomoću kojeg se pomoću ulaznog upita dobija skup tvitova koji odgovaraju traženom upitu. Tviterov API sadrži parametar naosnovu koga biramo na kojem ćemo jeziku da nam vrati tvitove.

Postoje razni emotikoni koji mogu prestavljati pozivna ili negativna osećanja. Npr. :), kao i :-) predstavljaju oba pozitivno osećanje. U Tviterovom API-ju, ako kao upit prosledimo :), dobijećemo rezultate svih tvitova sa

pozitivnim emotikonima. Isti je slučaj i sa negativnim za upit :(.

Podaci za trenitanje su postprocesuirani sa sledećim filterima:

- Emotikoni su izvađeni. Ovo je bitno za potrebe treniranja algoritama. Ukoliko ne bi bili izvađeni iz skupa podataka, algoritmi kao što su MaxEnt i SVM bi stavljali velike težine emotikonima, pa bi zbog relativne neodređenosti značenja emotikona, moglo uticati na tačnost.
- Svaki tvit koji sadrži i pozitivno i negativno osećanje je uklonjen. Ovo se radi iz razloga što se ne želi da pozitivni tvitovi budu kategorisani kao negativni i suprotno.
- „Retweetovi“ su uklonjeni. Retweetovanje je proces kopiranja nečijeg tvita i postavljanja preko drugog naloga. Ukoliko bi se i takvi tvitovi uzeli u razmatranje, tada bi jedan isti tvit bio ubrojan više puta, a to nije cilj analize.
- Tvitovi sa emotikonom „:P“ su uklonjeni. Ukoliko se u Tviterov API kao upit unese uvakav emotikon, rezultat su tvitovi sa negativnim osećanjem, međutim, u praksi se takav emotikon često koristi i za izražavanje pozitivnog osećanja, te je i tog razloga uklonjen.
- Ponovljeni tvitovi su uklonjeni. Povremeno, Tviterov API kao rezutat može vratiti jedan isti tvit dvaput. softver proverava poslednjih 100 tvitova koji su vraćeni i ukoliko ima ponovljenih, brišu se. Takođe je isti postupak i sa „retweetovima“.

REZULTATI I DISKUSIJA

Istražuje se upotreba unigrama, bigrama, unigrama i bigrama i unigrama sa svojstvima iz govora.

Unigrami Unigramski izdvajač svojstava je najjednostavniji način za izvlačenje svojstva iz tvitova. Algoritmi mašinskog učenja rade bolje nego osnovni pristup pomoću ključne reči. Daju tačnosti od 81%, 81,4% i 82,9% za Naive Bayes, MaxEnt i SVM. Ovo je veoma slično rezultatima alata Sentiment140 koji su 81,3%, 80,5% i 82,2%.

Bigrami Bigrami se koriste kod svojstava koji sadrže uz sebe i negaciju, npr. „nije dobro“ ili „nije loše“. U eksperimentima, negacija uz unigrame ne poboljšava tačnost.

Bigrami su veoma retki i zbog toga tačnost može da opada za slučajeve MaxEnt i SVM. Ovaj problem se može rešiti tako što se koriste i bigrami i unigrami

Unigrami i Bigrami U poređenju sa unigramima, tačnost se povećala za slučaj Naive Bayes (81,3% na 82,7%) i za MaxEnt (80,5% na 82,7%), ali ne i za SVM.

Delovi iz govora Delovi iz govora (Parts of speech – POS) se koriste kao svojstva zato što jedna reč može imati više značenja, tako da se mora uklopiti u kontekst da bi se mogla odrediti konotacija u kojoj se nalazi.

BUDUĆI RAD

Tehnike mašinskog učenja se dobro pokauju u klasifikaciji osećanja u tvitovima, međutim, veruje se da se i ti dobri rezultati mogu još unaprediti.

Neki od predloga za poboljšanje rada su sledeći:

- **Semantika** Predloženi algoritmi klasifikuju u celosti osećanje iz tvita, ali polaritet te klasifikacije može zavisiti od perspektive iz koje se interpretira. U tom slučaju, semantika može pomoći. Semantika se može izvući iz raznih faktora. Jedan od njih može biti jezik na kom tvit napisan ili država u kojoj se korisnik nalazi.
- **Tvitovi specifičnog domena** Klasifikacija tvitova radi sa prilično velikom tačnošću (reda veličine 80%). Ova tačno se postiže čak i sa veoma velikim vokabularom. Ukoliko bi se vokabular smanjio, tako da za određeni upit pretrage koristi odgovarajući domen vokabulara, tada bi se ista ili veća tačnost postizala mnogo brže.
- **Podrška za neutralne tvitove** U stvarnom svetu, neutralne izjave se ne mogu zanemariti, pa se tako i odgovarajuće osećanje može pripisati i neutralnim tvitovima.
- **Internaciolaziacija** Pretraga Tviterovima API-jem se vrši samo na jednom jeziku. Ukoliko bi se ta pretraga proširila na više jezika, tada bi rezultati bili raznovrsniji i uzeti sa većeg uzorka ispitanika i te bi se mogli uzeti kao merodavniji.

ZAKLJUČAK

Pokazano je da je uz korišćenje emotikona kao labela sa šumom postignut efikasan način za sprovođenje učenja sa udaljenim nadgledanjem. Algoritmi mašinskog učenja mogu dostići visok nivo tačnosti kada se koristi ovaj metod. Iako postovi sa Tvitera imaju drugačiju karakteristiku u poređenju sa drugim tekstovima, ovi algoritmi su pokazali da je moguće ekstrahovati osećanje sa podjednakim performansama kao i iz drugih.

LITERATURA

[1] YUINFO šablon za autore radova
http://www.yuinfo.org/YUINFO_Template.docx
[2] Go A, Bhyani R, Huang L, „Twitter Sentiment Classification using Distant Supervision“, pages 1-6