SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Analiza rada '*Your Sentiment Precedes You*: Using an author's historical tweets to predict sarcasm'

Mihael Nikić, bacc. ing.
Voditelj: *izv. prof. dr. sc. Jan Šnajder*

Zagreb, svibanj 2017.

SADRŽAJ

1. Uvod

Sarkazam (grč.sarkasmós; sarkazein = gristi usne od bijesa, sarx = meso) je zlobna, ljuta, zajedljiva, oštra i gorka poruga ("koja grize u meso"); pojačana ironija. To je oštra, pakosna poruga, jaka ironija, obično zasnovana na paradoksu. Polazište joj je negativan stav prema onomu kojem je upućena [8].

Ponekad razumijevanja sarkazma traži više od poznavanja samog konteksta rečenice, kao što je to primjer kod rečenice *Ja apsolutno obožavam ovaj restoran!* koja može biti sarkastična, ovisno o situacijskom kontekstu. Cilj rada, koji se analizira u sklopu ovog seminarskog rada, je prepoznati sarkazam kod korisnika na temelju njegovih prethodno objavljenih tvitova.[1] Za ostvarenje cilja korištene su dvije komponente: *kontrastni prediktor* (engl. *contrast-based predictor*), koji razotkriva postoji li sentiment kontrasta u tvitu koji promatramo te prediktor koji provjerava odgovara li sentiment izražen prema nekom entitetu u promatranom tvitu, onom sentimentu izraženom prema tom istom entitetu u prethodno objavljenim tvitovima (engl. *historical tweet-based predictor*). U nastavku teksta prethodno navedeni prediktor ćemo skraćeno nazivati *prošlosni prediktor*.

Ostatak seminarskog rada je strukturiran na sljedeći način. U nastavku prvog poglavlja dan je pregled sličnih radova. Drugo poglavlje sadrži pregled teorijskog dijela rada, u kojem se detaljnije opisuju postupci za detekciju sarkazma te primjena tih postupaka za gradnju modela koji u konačnici detektira sarkazam. Treće poglavlje daje kratak pregled strukture programskog rješenja i opis korištenog skupa podataka. U četvrtom poglavlju opisani su svi provedeni eksperimenti te analiza rezultata provedenih eksperimenata. Literatura je dana u šestom, a zaključak u sedmom poglavlju.

[1] tvit – naziv poruke koji se šalje preko društvene mreže *Twitter*

1.1. Slični radovi

Detekcija sarkazma se uglavnom oslanja na algoritme zasnovane na pravilima (engl. *rule-based algorithms*).

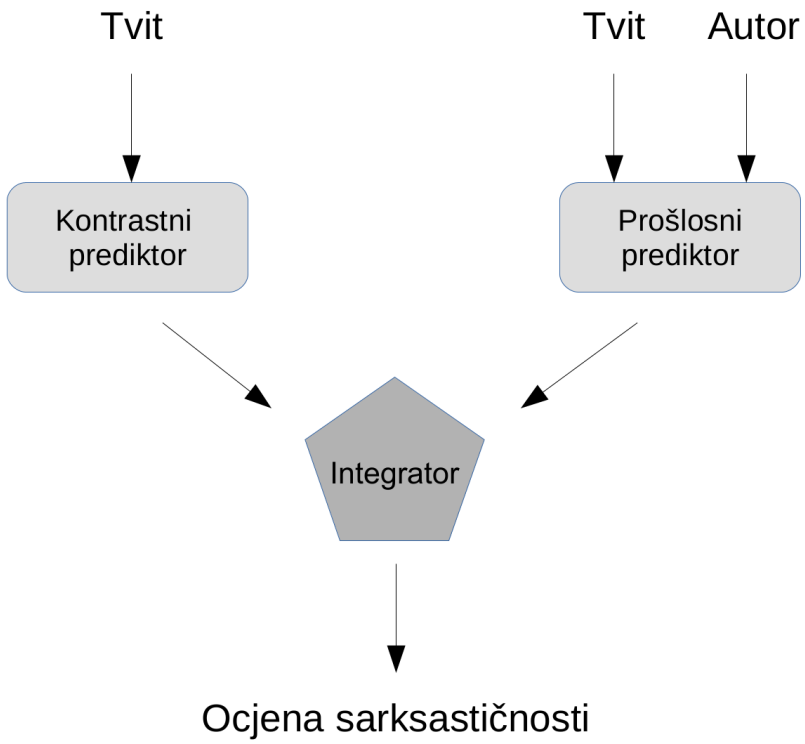
Npr. , Maynard and Greenwood (2014) zaključuju je li tvit sarkastičan na temelju pronalska suprotnog sentimenta u hashtagu[2] od onoga koji se nalazi u ostatku tvita. Slično, Riloff et al. (2013) ocjenjuje tvit sarkastičnim ako postoji kontrast između glagola i imenične fraze. Joshi et al. (2015) koristi lignvističku teoriju nazvanu *nepodudaranje konteksta* kao osnovu odabira značajki (engl. *feature design*) i opisa dvaju tipova značajki: implicitnog i eksplicitnog nepodudaranja.

Ostali slični radovi se mogu pronaći u samom radu kojeg analiziramo [3].

[2]hashtag – bilo koja riječ koja ispred sebe ima oznaku '#' te na taj način predstavlja određeni pojam koji se može pretraživati, točnije na društvenim mrežama vodi do svih poruka koje u sebi sadrže taj hashtag

2. Postupci za izgradnju modela

2.1. Arhitektura



Slika 2.1: Arhitektura korištenog pristupa za detekciju sarkazma

Slika 2.1 prikazuje arhitekturu korištenog pristupa za detekciju sarkazma. Na ulazu model prima tekst koji predstavlja tvit i naziv autora tog tvita te na izlazu donosi predikciju sarkastičnosti danog tvita. Ovaj pristup predikcije sarkazma naziva se detekcija sarkazma na temelju pravila i sastoji se od tri modula:

- **Kontrastnog prediktora** (engl. *Contrast-based predictor*),
- **Prošlosnog prediktora** (engl. *Historical tweet-based predictor*),
- **Integratora** (engl. *Integrator*).

U sljedećim potpoglavljima sva tri modula su detaljnije objašnjena.

2.2. Kontrastni prediktor

Modul uzima u obzir samo promatrane tvitove (engl. *target tweets*). Kontrastni prediktor razotkriva postoji li sentiment kontrasta u tvitu, kojeg promatramo. Razlikujemo dvije vrste kontrasta:

- **Eksplicitni kontrast** – Tvit sadrži riječ jednog polariteta te još jednu drugog polariteta. Npr. rečenicu *Volim biti ignoriran* klasificirati ćemo kao sarkastičnu jer ima pozitivnu riječ *Volim* i negativnu riječ *ignoriran*[1].
- **Implicitni kontrast** – Tvit sadrži riječ jednog polariteta te jednu frazu drugog polariteta. Fraze koje uzimamo u obzir su izvučene iz skupa sarkastičnih tvitova. Izvlačenje fraza iz sarksastičnih tvitova je izvršeno na sljedeći način:

  1. Preuzet je skup od 5264[2] tvitova koji sadrže hashtag #sarcasm i pretpostavljeno je da su sarkastični;
  2. Iz njih su izvučeni svi n-grami, krenuvši od 3-grama pa sve do svih 10-grama (pri čemu 1-gram predstavlja jednu riječ);
  3. Od izvučenih fraza izbačene su sve koje se pojavljuju manje od 3 puta, a ostale su uzete u obzir prilikom detekcije sarkazma.

2.3. Prošlosni prediktor

Prošlosni prediktor za potrebe predikcije uzima u obzir neki tvit i naziv autora tog tvita. Cilj prošlosnog prediktora je identificirati odgovara li sentiment izražen u promatranom tvitu onom kojeg je autor u prošlosti iskazao. Koraci su sljedeći:

1. Ocijeniti sentiment promatranog tvita uz pomoć sustava za ocjenu sentimenta zasnovanog na pravilima (engl. *rule-based sentiment analysis sytem*),
2. Označiti vrste riječi (engl. *part of speech tagging*) u rečenici unutar promatranog tvita,

[1] Iako je primjer dan na hrvatskom, u stvarnoj implementaciji u obzir se uzimaju samo tvitovi koji su napisani na engleskom jeziku.

[2] U originalnom radu korišten je skup od 8000 tvitova, ali nije objavljen pa je uz pomoć službenog *twitter API*-ja uspješno skinuto 5264 različitih tvitova koji sadrže hashtag #sarcasm

3. Izvući NNP[3] sekvencu kao ciljanu frazu (engl. *target phrase*) u promatranom tvitu,

4. Preuzeti kompletnu vremensku crtu (engl. *timeline*) od autora koristeći *Twitter API*[4],

5. Odabrati tvitove koje sadrže ciljanu frazu,

6. Za odabrane stare tvitove učiniti sljedeće:
   - Ocijeniti sentiment tvita odnosno rečenice uz pomoć sustava za ocjenu sentimenta zasnovanog na pravilima,
   - Glasanjem utvrditi koji sentiment je dominantan u odabranim tvitovima,

7. Proglasiti promatrani tvit sarkastičnim ako je sentiment promatranog tvita *različit* od onoga u starim tvitovima,

8. Ponavljati postupak za sve ostale NNP sekvence dok se nužni uvjet za proglašenje tvita sarkastičnim ne ispuni.

Mogući nedostatci ovog pristupa su:

- Ako prethodno objavljeni tvitovi sadrže sarkazam usmjeren prema ciljanoj frazi, dok promatrani tvit ne, prediktor će neispravno klasificirati promatrani tvit sarkastičnim.
- Ako i prethodno objavljeni tvitovi i promatrani tvit sadrže sarkazam usmjeren prema ciljanoj frazi, prediktor će neispravno klasificirati promatrani tvit nesarkastičnim.
- Ako se entitet spomenut u promatranom tvitu nikad nije pojavio u prethodno objavljenim autorovim tvitovima, onda se izlaz prošlosnog prediktora nemože uzeti u obzir.

2.4. Integrator

Ovaj modul kombinira predikcije prošlosnog i kontrastnog prediktora. Postoje četiri vrste ovog modula:

[3] NNP (*proper noun, sing.*) – vlastita imenica, u jednini

[4] `https://dev.twitter.com/overview/api`

1. **Isključivo prošlosni** (engl. *Only historical tweet-based*) – Integrator koji uzima u obzir samo izlaze prošlosnog prediktora. U slučaju da se ciljna fraza nije spominjala u starim tvitovima, tvit će biti proglašen ne-sarkastičnim,

2. **ILI** (engl. *OR*) – Integrator koji proglašava tvit sarkastičnim ako to učini **bilo koji** od prediktora,

3. **I** (engl. *AND*) – Integrator koji proglašava tvit sarkastičnim ako to učine isključivo **oba** prediktora,

4. **Opušteni-I** (engl. *Relaxed-AND*) – Sličan prethodnom, s tim da za razliku od prethodnog ne uzima u obzir izlaze prošlosnog prediktora, ako navedeni nije pronašao niti jedan stari tvit u kojem se pojavljuje ciljani sentiment.

3. Implementacija

3.1. Skup podataka

Za implementaciju programskog rješenja korišteni su sljedeći skupovi podataka:

- 5264 tvitova koji sadrže hashtag #sarcasm – kao što je i navedeno, koristimo ih za vađenje implicitnih sentimenata,
- Ručno anotiran korpus načinjen od 2278 tvitova – služi kao testni korpus koji ćemo koristiti u eksperimentima.[1]

Oba prediktora koja implementiramo oslanjaju se na leksikone sentimenta: Kontrastni prediktor zbog detekcije kontrasta, dok prošlosni prediktor zbog identifikacije sentimenta u tvitu. Za potrebe eksperimenta koriste se dva leksikona:

- *VADER*[2] leksikon koji je izgrađen u sklopu *VADER-Sentiment-Analysis* modula *NLTK* biblioteke[3]
- Leksikon 2 (**L2**) – lista sa pozitivnim i negativnim riječima od Mohammad i Turney (2013)

3.2. Implementacija programskog rješenja

Programsko rješenje je implementirano je u programskom jeziku *Python*. Osim programskog jezika *Python* korištene su i gotove *Python* biblioteke:

- *NLTK (Natural Language Tool Kit)* – jedna od najpopularnijih biblioteka za obradu prirodnog jezika. Sadrži gotove funkcije za pred obradu teksta, provjeru ispravnosti riječi, analizu sentimenta i slično.

[1] 80% tvitova iz originalnog korpusa nije moguće preuzeti zbog postavki privatnosti ili nedostupnosti.

[2] VADER – Valence Aware Dictionary and sEntiment Reasoner

[3] U originalnom radu korišten je Leksikon 1 (**L1**) od Pang i Lee (2004), ali zbog nedostupnosti navedenog korišten je *VADER* leksikon iz *NLTK* biblioteke

- Konkretno za analizu sentimenta korišten je *VADER-Sentiment-Analysis* modul koji je posebno prilagođen za analizu sentimenta izraženih na društvenim mrežama,
- Sustav za analizu sentimenta riječ ili rečenicu ocjenjuje realnom vrijednošću na temelju kojeg sentiment klasificiramo pozitivnim ako je vrijednost veća od 0, negativno ako je vrijednost manja od 0 te neutralno ako je jednaka 0. Ako je sentiment klasificiran kao neutralan, onda se promatrana riječ ili rečenica ne uzima u obzir za predikciju sarkazma bilo kojim postupkom.

– *scikit-learn* – biblioteka koja pruža jednostavne i učinkovite alate za dubinsku analizu podataka (engl. *data mining*) i analizu podataka (engl. *data analysis*). U sklopu rada, iz biblioteke su korištene metode za izračun evaluacijskih mjera poput preciznosti, odziva i f1-mjere.

– *twython* – biblioteka koja služi kao omotač (engl. *wrapper*) službenom *Twitter API*-ju s ciljem preuzimanja tvitova nekog korisnika korištenjem samog programskog jezika *Python*.

Implementacija programskog rješenja načinjena je od nekoliko nezavisnih komponenti koje obavljaju određene zadaće:

1. Implementacija apstraktnoga razreda prediktora iz kojeg su izvedeni preostala dva prediktora nalazi se u:

   `predictor.predictor.Predictor`

2. Implementacija prošlosnog prediktora nalazi se u razredu:

   `historical_predictor.h_predictor.HistoricPredictor;`

3. Implementacija kontrastnog prediktora nalazi se u razredu:

   `contrast_predictor.c_predictor.ContrastPredictor;`

4. Implementacije sva četiri oblika integratora (i njihovog baznog razreda) nalaze se u razredima koji započinju sa `integrators.*` i završavaju sa:

   `integrator.Integrator;`
   `and_integrator.ANDIntegrator;`
   `or_integrator.ORIntegrator;`
   `only_hist_integrator.OnlyHistoricalTweetIntegrator;`
   `relaxed_and_integrator.RelaxedANDIntegrator;`

4. Eksperimenti

4.1. Provedba eksperimenata

Za potrebe eksperimenata koristimo dva prethodno navedena rječnika (VADER i L2) te radimo dvije različite provjere i uspoređujemo rezultate s onima iz [3]:

- **Detekcija sarkazma sa VADER leksikonom (VD1)** - U ovoj provjeri se koristi *VADER* leksikon iz *NLTK* biblioteke,
- **Detekcija sarkazma sa L2 (SD2)** - U ovoj provjeri se koristi leksikon L2.

Dobiveni rezultati ilustrirani su tablicama 4.1 i 4.2 te prikazuju preciznost (P), odziv (R) i f-score (F) za VD1 i SD2.

	P	**R**	**F**
Isključivo prošlošni prediktor	0.159	0.069	0.096
ILI	0.266	**0.337**	0.297
I	0.200	0.040	0.066
Opušteni-I	**0.298**	0.307	**0.302**

Tablica 4.1: Srednja preciznost, odziv i F-score VD1 pristupa za sve četiri konfiguracije integratora

Tablica 4.3 prikazuje dobivene rezultate za SD2 provjeru u radu kojeg analiziramo [3]. Za VD1 provjeru rezultati nisu dostupni budući da *VADER* leksikon nije korišten u originalnom radu, već leksikon od Pang i Lee (2004) koji je nedostupan za preuzimanje, kao što je prethodno i rečeno.

	P	**R**	**F**
Isključivo prošlošni prediktor	0.188	0.088	0.120
ILI	0.195	**0.216**	**0.205**
I	**0.375**	0.029	0.054
Opušteni-I	0.219	0.157	0.183

Tablica 4.2: Srednja preciznost, odziv i F-score SD2 pristupa za sve četiri konfiguracije integratora

	P	**R**	**F**
Isključivo prošlošni prediktor	0.496	0.499	0.497
ILI	0.842	**0.927**	**0.882**
I	0.779	0.524	0.627
Opušteni-I	**0.880**	0.884	**0.882**

Tablica 4.3: Srednja preciznost, odziv i F-score SD2 pristupa za sve četiri konfiguracije integratora rada [3]

4.2. Analiza rezultata

Iz dobivenih rezultata vidljiva je nepodudarnost rezultata dobivenih reimplementacijom rada i stvarnih rezultata dobivenih u radu [3].

Ono što se prvo može uočiti je da prošlosni prediktor ostvaruje puno lošije rezultate od očekivanih, tijekom evaluacije. Razlog je vjerojatno u neuspješnom pronalasku dovoljnog broja NNP sekvenci u promatranom tvitu, a i u prethodno objavljenim tvitovima te pojava velikog broja neutralno ocjenjenih sentimenata koji se zanemaruju.

Također jedan od razloga može bit taj što se prilikom implementacije koristio sustav za analizu sentimenta koji nije identičan onom koji se koristi u analiziranom radu. U radu je spomenuto da se za potrebe analize sentimenta koristi sustav temeljen na pravilima (engl. *rule-based sentiment analysis system*) i spomenuta je preciznost koju navedeni sustav postiže na *Sentiment140*[1] korpusu (preciznost iznosi 58.49%, što je manje od preciznosti iznosa 65.38% koju postiže sustav korišten za potrebe reimplementacije

rada, na tom istom korpusu), ali nije spomenut i sam postupak implementacije takvog sustava.

Pronalazak eksplicitnog kontrasta u tvitovima je opisan kao postupak traženja riječi suprotnog polariteta u rečenici po uzoru na postupak iz Joshi et al. (2015), ali nije eksplicitno navedeno je li sličnost samo u tome što se pronalaze riječi suprotnog polariteta ili su neke druge značajke iz navedenog rada također posmatrane. Slično vrijedi i za pronalazak implicitnog kontrasta.

[1] `http://help.sentiment140.com/for-students`

5. Zaključak

Na temelju dobivenih rezultata vidljivo je da reimplementacija nije u potpunosti ispravno izvedena budući da su rezultati dobiveni eksperimentima znatno lošiji od onih koji su dobiveni u samom radu [3].

Razlozi tome su mnogi, krenuvši od mogućih grešaka prilikom same reimplementacije pa sve do problema ponekih preapstraktno definiranih rješenja određnih problema u samom radu.

Sljedeći korak bi svakako bila izrada prošlosnog prediktora koji postiže bolje rezultate prilikom evaluacije. Također, osim prošlosnog prediktora poželjno bi bilo unaprijediti i sustav pronalaženja eksplictnog odnosno implicitnog kontrasta s tim da je prioritet na unaprijeđenju prošlosnog prediktora budući da je i cilj samog rada proizvesti predikciju sarkazma na temelju autorovih prethodno obajvljenih tvitova.

6. Literatura

[1] Dmitry Davidov, Oren Tsur, i Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. U *Proceedings of the fourteenth conference on computational natural language learning*, stranice 107–116. Association for Computational Linguistics, 2010.

[2] Aditya Joshi, Vinita Sharma, i Pushpak Bhattacharyya. Harnessing context incongruity for sarcasm detection. U *ACL (2)*, stranice 757–762, 2015.

[3] Anupam Khattri, Aditya Joshi, Pushpak Bhattacharyya, i Mark James Carman. Your sentiment precedes you: Using an author's historical tweets to predict sarcasm. U *6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, stranica 25, 2015.

[4] Diana Maynard i Mark A Greenwood. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. U *LREC*, stranice 4238–4243, 2014.

[5] Saif M Mohammad i Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[6] Bo Pang i Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. U *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, stranica 271. Association for Computational Linguistics, 2004.

[7] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, i Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. U *EMNLP*, svezak 13, stranice 704–714, 2013.

[8] Wikipedia. Sarkazam — Wikipedia, slobodna enciklopedija. `http://hr.wikipedia.org/w/index.php?title=Sarkazam`, 2017. [Online; accessed 09-April-2017].

7. Sažetak

Sarkazam je zlobna, ljuta, zajedljiva, oštra i gorka poruga; pojačana ironija. Ponekad razumijevanje samog sarkazma zahtjeva više od poznavanja situacijskog konteksta. U ovom seminarskom radu analiziramo jedan postojeći rad koji na temelju prethodno objavljenih tvitova nekog autora pokušava odrediti da li je trenutno promatrani tvit sarkastičan ili ne.

Predikcija se radi pomoću modela koji na ulazu prima tekst koji prestavlja tvit i naziv autora tog tvita te na izlazu radi predikciju sarkastičnosti danog tvita. Pristup se sastoji od tri modula: Kontrasnog prediktora, prošlosnog prediktora i integratora.

Na kraju rada radimo eksperimente, čije rezultate uspoređujemo s onima iz analiziranog rada [3] te donosimo zaključak.