Detekcija sarkazma pomoću Naivnog Bajesa i konvolucionih neuronskih mreža

Jovana Jevtić, Dragana Filipović
Softversko inženjerstvo i informacione tehnologije, Fakultet tehničkih nauka – Novi Sad
Mentori: Aleksandar Lukić, Branislav Anđelić

UVOD

Sa porastom upotrebe društvenih mreža, kao i online načina komunikacije, ljudi u današnje vreme putem pisane forme iskazuju svoja mišljenja, i tada nastaje problem prepoznavanja sarkazma, koji se inače izražava dikcijom i intonacijom tona govornika, kao i samom mimikom i gestikulacijom. Odlučile smo se da upotrebom Naivnog Bajesa i konvolucionih neuronskih mreža izgradimo rešenje koje će doprineti prepoznavanja sarkazma u tekstu.

SKUP PODATAKA

Skup podataka koji je u obliku json formata, sadrži sledeće informacije:

- article_link
- headline
- is_sarcastic

U ovom projektu su samo headline i is_sarcastic korišćeni.
Skup sadrži 26709 podataka.
Kod konvolucione neuronske mreže skup podataka je deljen na deo za treniranje (70%), deo za validaciju (20%) i deo za testiranje (10%).
Kod Naivnog Bajesa skup je podeljen na deo za treniranje (70%) i deo za testiranje (30%).

PRETPROCESIRANJE

Prva stvar koja je rađena u oba pristupa je pretprocesiranje teksta, koje je rađeno tako što su izbačeni znakovi interpunkcije, brojevi i engleski članovi, jer je tekst na engleskom jeziku. Nakon toga je rađena stemizacija teksta.

METODOLOGIJE

1. Konvolucione neuronske mreže

Kod neuronske mreže posle pretprocesiranja, sledeći korak je bilo pretvaranje samih reči u oblik koji računar može da obradi, a to su brojevi. Koristili smo Tokenizer iz Keras biblioteke i na taj način dobili reči reprezentovane vektorima brojeva. Sledeći korak kod neuronskih mreža je kreiranje modela same mreže. Mreža se sastoji iz više slojeva, konkretno tri konvoluciona, max-pooling, dropout i potpuno povezanih slojeva. Kao loss funkcija korišćena je binary crossentropy, a kao optimizaciona je korišćena Adam, sa 0.0001 learning rate-om.

2. Naivni Bajes

Kod Naivnog Bajesa smo brojali reči, odnosno kreirali smo više rečnika kod kojih jedan sadrži sve reči i broj pojavljivanja u svakom naslovu, drugi sadrži broj reči koje se pojavljuju u sarkastičnim i nesarkastičnim naslovima, dok su u trećem rečniku brojani sarkastični i nesarkastični naslovi. Korišćena je sledeća formula

$$P(s_i|T) = e^{\sum_{t \in T} ln(\frac{P(t|s_i)}{P(t)}) + ln(P(s_i))}$$

gde je:
P(S) = P(naslov je sarkastičan ili ne), P(s1) za sarkastične, P(s2) za nesarkastične
P(T) = P(napisan je niz reči koje predstavljaju naslov)
P(T|S) = P(određen je niz reči koji predstavlja naslov određenog tipa)
P(S|T) = ovo je ono što računamo-> za dati niz reči koji predstavlja naslov izračunati verovatnoću da je taj naslov određenog tipa

TESTIRANJE I REZULTATI

Kod treniranja neuronske mreže, nakon eksperimentisanja, batch size je postavljen 256, a broj epoha na 32, jer nakon tog broja dolazi do overfittinga. Treniranje je trajalo 10 minuta i dobijeni su rezultati prikazani u tabeli.
Kod Naivnog Bajesa, treniranje je trajalo 4 sekunde i dobijeni su sledeći rezultati, koji su iskazani pomoću metrike accuracy:

	Sarkastični naslovi	Nesarkastični naslovi
Naivni Bajes	85.345	87.867
Konvoluciona mreža	83.261	78.571

ZAKLJUČAK

Ponekad je i ljudima u usmenoj komunikaciji teško prepoznati sarkazam, pa s toga se može zaključiti da je problem prepoznavanja sarkazma za računar još kompleksniji, iz razloga što računar nema dodatne informacije koje govornik iskazuje putem mimike, gestikulacije i intonacije. Ipak rezultati koji su prikazani u tabeli nisu loši, iako postoje načini na koji bi se oni mogli poboljšati, pogotovo kod konvolucionih neuronskih mreža. Prvenstveno, skup podataka nad kojima se mreža trenira je potrebno unaprediti, a nakon toga isprobati neke od vrsta word embeddinga, zatim poboljšati sam model mreže i na kraju isprobati različite vrednosti parametara kao što su learning rate, batch size i broj epoha.

REFERENCE

- https://www.kaggle.com/wflazuardy/sarcasm-detection-with-keras-preprocessing
- https://iq.opengenus.org/text-classification-using-cnn/
- Vežbe i predavanja