# Analiza sentimenta tvitova na ruskom jeziku
# Upotrebom neuronskih mreža

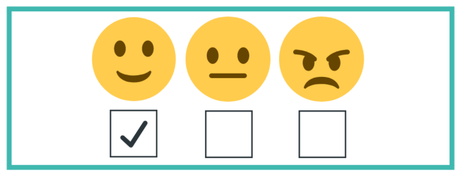
Marieta Rakoš, kontakt: marieta.rakos@gmail.com
Univerzitet u Novom Sadu, Fakultet tehničkih nauka

## Apstrakt

Ideja ovog rada je proučavanje algoritama za analizu sentimenta kratkog teksta.
Poznat algoritam korišćen za ovu temu je Naive Bayes, ali se u ovom radu fokusiram na neuronske mreže.
Implementirana je nauronska mreža koju smo koristili na vežbama kao i neuronske mreže koje nudi Tensor Flow.

## Uvod

Analiza sentimenta predstavlja određivanje da li je tekst pozitivan ili negativan (ili neutralan)

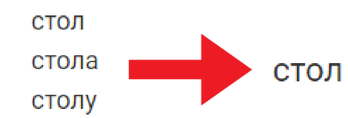


Svodi se na nekoliko koraka:
- Učitavanje podataka iz csv fajla
- Predprocesiranje
- Treniranje
- Testiranje
- Real life testiranje

## Postupak

### Predprocesiranje

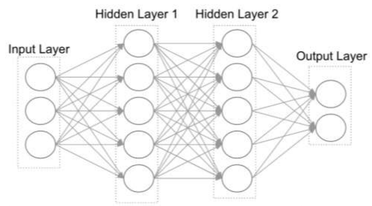
Svaki učitani tvit potrebno je obraditi kako bi se mogli propustiti kroz neuronsku mrežu. Najpre se tvit deli na reči, a zatim se izbacuju svi znakovi koji ne pripadaju ruskoj ćirilici. Svaka reč podvrgava se stemingu tj. svođenju na osnovni oblik:




Pronađene reči se smeštaju u vokabular gde im se pridružuje broj koliko puta se pojavljuju u našem skupu tvitova. Izbacuju se neutralne reči – stop words, jer one ne utiču na sentiment. Uzimam 5000 najčešćih reči, jer reči koje se pojavljuju retko nisu od koristi.

Implementirana su dva načina pripremanja podataka za obučavanje:
•Tvit se pretvara u vektor nula i jedinica, gde nula označava da se odgovarajuća reč ne nalazi u tvitu, a jedinica suprotno. Svi vektori su dužine 5000
"Я люблю мороженое" postaje [0, 0, 0, 1, 0, 0 …, 1, 0, 0]
•Tvit se pretvara u vektor gde se svaka reč zamenjuje brojem koji je pridružen toj reči u vokabularu. Vektori su dužine broja reči u tvitu, pa je potrebno sve tvitove svesti na istu dužinu - dodati padding u vidu nula ili oduzeti višak reči
"Я люблю мороженое" postaje [0, 23, 2111] (bez paddinga)

Prva ideja se propušta kroz neuronsku mrežu sa vežbi i kroz tensorflow neuronsku mrežu sa fully connected slojevima.
Ulaz u mrežu je tvit pretvoren u vektor, prvi sloj ima 125, drugi 25 a izlaz 2 neurona.




Druga ideja se propušta kroz RNN mrežu, koja je konstruisana na osnovu tutorijala priloženom na zvaničnom sajtu TensorFlow https://github.com/Hvass-Labs/TensorFlow-Tutorials/blob/master/20_Natural_Language_Processing.ipynb

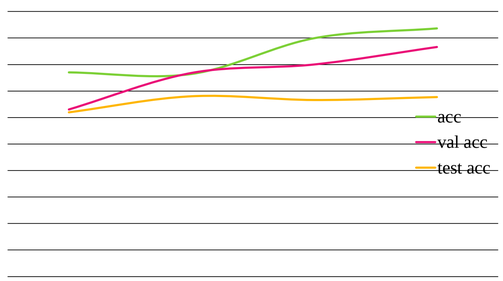
## Rezultati

### Neuronska mreža sa vežbi

Ovaj model se nije dobro pokazao pre svega zato što se tvitovi iz trening skupa obrađuju jedan po jedan, ne postoji obrađivanje manjeg skupa istovremeno, kao što je to implementirano kod tensorflow neuronske mreže. Iako su ova dva modela istih razmera, prvi nije moguće ni kreirati, a ako se veličina smanji i do tri puta, fitovanje traje i preko 3 sata, nakon čega dolazi do terminacije sesije na Google Colab-u.

### TensorFlow neuronska mreža

Ovaj model je znatno uspešniji u odnosu na prvi. Korišćen je batch veličine 128. Nakon tri epohe postiže se accuracy od 74,9% na trenirajućem, a 69,3% na validacionom skupu, a 68% na test skupu. Na svake 3 sledeće epohe, povećava se razlika između accuracy na trenirajućem i test skupu, što nije poželjno.




### Rekurentna neuronska mreža RNN

Ovaj model je veoma interesantan slučaj. Sa istim parametrima kao u tutorijalu, postiže se accuracy oko 69%. Ukoliko se izmeni vrednost bilo kog parametra vrednošću koja se koristi u prethodnom modelu, accuracy može da bude i skoro 0% ili skoro 100% ili čak i 100%. Konkretno, aktivacija na poslednjem sloju se promeni iz sigmoid u softmax, ili learning rate umesto 0,001 u 0,75, itd.

## Zaključak

Uz detaljnija istraživanja i rad, školska neuronska mreža bi se mogla poboljšati. Nema smisla smanjiti broj reči u rečniku i tako smanjiti input, jer se time povećava šansa da sve reči nepoznatog tvita budu nepoznate ili vrlo malo poznate.Ali i tako poboljšanu, teško je poreditи sa Tensorflow neuronskom mrežom.

### TensorFlow neuronska mreža

Iako kroz epohe postaje sve veća razlika između train accuracy i test accuracy, ona je i dalje oko 70%, što nije loše, obzirom na to da se koristi i upola manje tvitova nego što postoji u fajlu. Čak je i posle samo tri epohe, rezultati su zadovoljavajući. Dalje istraživanje bi pomoglo da se postigne još veća tačnost.

Što se tiče RNN modela, zaključuje se da je vrlo nestabilan, jer mala promena parametra dovodi do drastičnih promena performansi. Čak i kada se fituje više puta sa istim parametrima, rezultati su veoma različiti.

## Dalja istraživanja

- Analiza sentimenta pomoću CNN
- Zavisnost uspešnosti klasifikovanja od nasumično odabranog skupa tvitova za treniranje
- Dodatno nameštanje hiper parametara