Fakultet organizacinih nauka
Univerziteta u Beogradu

Projekat iz predmeta primena veštačke inteligencije

Sentimenta analiza za klasifikaciju recenzija hotela sa sajta Booking.com

Beograd 2023.

Sadržaj rada

Tema rada

Ovaj rad istražuje primenu sentimentne analize za klasifikaciju hotelskih recenzija na Booking.com platformi. Cilj je poboljšanje korisničkog iskustva identifikacijom i razumevanjem pozitivnih i negativnih aspekata gostovanja u hotelima. Uz pomoć naprednih algoritama obrade prirodnog jezika, recenzije će biti automatski analizirane i klasifikovane kao pozitivne ili negativne, na osnovu komentara i ocena. Kroz ovaj proces, cilj je pružiti hotelskim preduzećima dublji uvid u zadovoljstvo gostiju i identifikovati oblasti za unapređenje usluge. Ovaj rad ima potencijal da podrži donošenje odluka u hotelijerstvu i doprinese optimizaciji korisničkog iskustva u hotelskom sektoru.

Sentimentna analiza

Sentimentna analiza (često nazvana i Opinion Meaning) je NLP (Natural Language Processing) pristup koji pomaže pri identifikaciji značenje podataka. Odnosno, analiza sentimenta proučava mišljenje, emocije ili stavove prema određenoj tematici, osobi ili entitetu. Ovi izrazi se mogu klasifikovati kao pozitivni, negativni ili neutralni. Na primer: „Stvarno mi se sviđa novi dizajn vaše veb stranice!“ → Pozitivno.

U poslednjih nekoliko godina, kako je omogućena dostupnost velikih skupova podataka i moćnih algoritama mašinskog učenja počinje sve veća primena ove analize sa efikasnim rezultatima. Glavni izazov u analizi sentimenta je suočavanje sa nijansama ljudskog jezika, poput sarkazma, ironije i figurativnih značenja jezika, što može otežati tačno određivanje znjačenja teksta. Uprkos ovim izazovima, analiza sentimenta ima potencijal da pruži dragocene uvide u javno mnjenje, ponašanje potrošača i druge važne oblasti istraživanja.

Kako funkcioniše sentimenta analiza?

Analiza osećanja koristi modele mašinskog učenja za analizu teksta ljudskog jezika. Korišćene metrike su dizajnirane da otkriju da li je ukupni sentiment dela teksta pozitivan, negativan ili neutralan.

Analiza se uglavnom zasniva na sledećim koracima:

1. **Prikupljanje podataka**. Proces priupljanja potrebnih podataka se uglavnom zasniva na scrapping bot-ovima.
2. **Čišćenje podataka**. Podaci koji se obrađuju moraju biti očišćeni od šumova. Takođe, u ovom delu se otklanjaju reči koje nemaju jasno značenje za dalju analizu teksta poput veznika, specijalnih karaktera, zamenica.

3. **Ekstrahovanje ključnih karakteristika (feature).** ML algoritmi automatski izdvajaju ključne reči u tekstu kako bi ih dalje grupisali u negativne ili pozitivne. Ovaj pristup uključuje poznatu tehniku pod nazivom “Bag of words” koja prati frekvenciju ponavljanja određenih reči u tekstu. Takođe, u ovom delu se može pomoću neuralnih mreža naći sličnost između pojmova u tekstu.
4. **Biranje ML modela**. Alat za analizu osećanja ocenjuje tekst koristeći automatski ili hibridni ML model zasnovan na pravilima. Sistemi zasnovani na pravilima vrše analizu osećanja po unapred definisanim pravilima zasnovanim na leksici i često se koriste u domenima kao što su pravo i medicina gde je potreban visok stepen preciznosti i ljudske kontrole. Automatski sistemi koriste ML i tehnike dubokog učenja za učenje iz skupova podataka. Hibridni model kombinuje oba pristupa.
5. **Treniranje modela.**
6. **Evaluacija modela.**

Tipovi sentimente analize

Sentimenta analiza se uglavnom fokusira na polarizovano značenje teksta (pozitivno, negativno ili neutralano značenje), međutim ova analiza može da se primenjuje i pri određivanju nešto kompeksnijih značenje podataka, kao što je prepoznavanje ironije, panike, zainteresovanosti itd.

Postoje različiti tipovi sentimentne analize koji se koriste u zavisnosti od cilja istraživanja. Neki od njih u gradirana sentimenta analiza, detekcija emocija, višejezičcna analiza.

Gradirana analiza osećanja

Ukoliko polarizovani odgovori nisu dovoljno precizni i jasni, može se koristiti prošireni model sentimentne analize gde se odgovori rangiraju po jačini:

- Veoma pozitivan
- Pozitivan
- Neutralan
- Negativan
- Veoma negativan

Ova skala je slična Liketovoj skali i može se predstaviti sa brojevima od jedan do pet, gde jedan predstavljama veoma negativan stav, a broj pet veoma pozitivan.

Detekcija emocija

Detekcija emocija sentimentne analize omogućava da se kroz ovaj model otkriju emocije korisnika poput frustracije, ironije i tuge.

Mnogi modeli ovog tipa koriste leksikone ili kompleksne algoritme mašinskog učenja. Jedna od negativnih strana korišćenja leksikona jeste što ljudi na različite načine izražavaju svoje mišljenje. Tako reči koje su u leksikonu kategorisane kao negativne mogu u nekim slučajevima imati pozitivno značenje. Na primer: "Vaš korisnički servis je užasan!", gde reč **užasan** ima negativno značenje, ali isto tako može predstavljati frazu kojom se izražava pozitivan stav: "Užasno mi se sviđa nova verzija sajta".

Tehnike detekcije emocija mogu se podeliti u dve glavne kategorije: metode zasnovane na pravilima i metode zasnovane na mašinskom učenju. Metode zasnovane na pravilima koriste unapred definisana pravila i heuristike za detekciju emocija, dok metode zasnovane na mašinskom učenju koriste statističke modele obučene na označenim podacima za prepoznavanje obrazaca i klasifikaciju emocija.

Višejezična sentimentna analiza

Višejezična analiza osećanja je vrlo često kompleksnija od ostalih tipova sentimentne analize. Ona uključuje mnogo pretprocesiranja i resursa. Većina ovih resursa je dostupna na mreži (npr. leksikoni osećanja), dok druge treba kreirati (npr. prevedeni korpusi ili algoritmi za detekciju buke).

Alternativno, ovaj model može da otkrije jezik pomoću jezičkog klasifikatora, a zatim da se koristi odgovarajući model analize sentimenta kako bi se podaci dalje obrađivali.

Svrha sentimentne analize

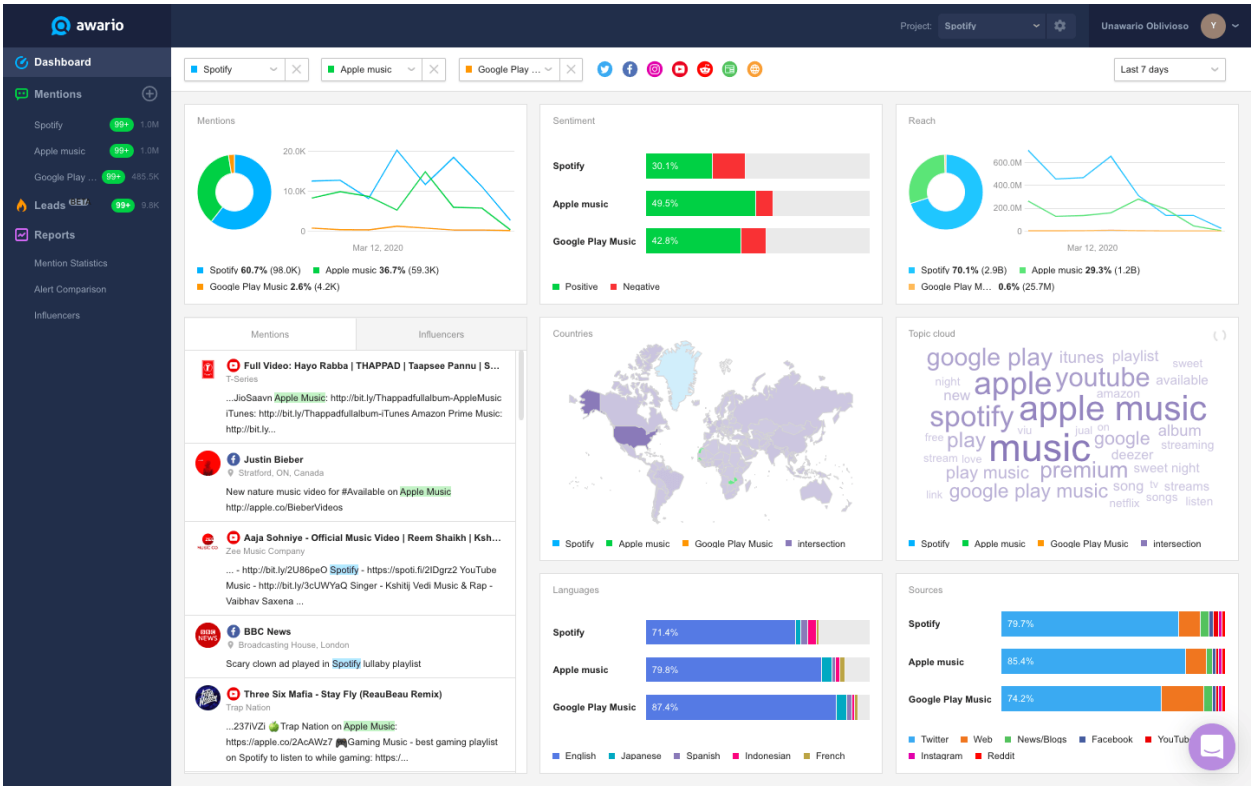
Sentimentna analiza je značajna za kompanije u razumevanju kako korisnici percipiraju njihov proizvod ili uslugu.

Danas, ljudi sve više svoje mišljenje izražavaju na internetu u vidu komentara, online upitnika ili postova. Tako, sentimentna analiza pomaže da kompanije u realnom vremenu dobiju feedback svojih korisnika, čime se značajno ubrzava proces prikupljanja mišljenja korisnika.

Mišljenje korisnika je bitno za svaku kompaniju koja želi da ostane konkurenta na tržištu i postane profitabilna. Sentimentna analiza se može i proaktivno uključiti pri razvoju nekog proizvoda što dovodi do povećane lojalnosti i zadovoljstva korisnika.

Sentimentna analiza se može koristiti u mnogim industrijama, a neki od primera njene primene su:

- **Social Media Monitoring** - sentimentna analiza se koristi za praćenje osećanja i mišljenja korisnika na društvenim mrežama poput Instragama, Twitter-a i Facebook-a.
- **Monitoring brand awareness** - praćenje reputacije neke kompanije tokom vremena.
- **Analzing Customer experience** - analiza mišljenja korisnika onekom novom proizvodu.
- Procena uspešnosti **marketing kampanje**.
- **Određivanje ciljne grupe**, demografske karakteristike.



Neki od benefita korišćenja sentimentne analize su:

- Prikupljanje velike količine nestrukuiranih podataka sa različitih izvora.
- Real-life praćenje feedback-a kupaca, kao i mišljenja o samom brendu i proizvodu.
- Pruža feedback kako poboljšati proizvod ili uslugu.
- Identifikacija ključnih trendova u industriji.
- Može da skuplja mišljenje kupaca o konkurentskom proizvodu.

Projekat

Dataset

Za ovaj projekat korišten je dataset pod nazivom *515K Hotel Reviews Data in Europe.* Ovaj skup podataka sakupljen je sa sajta za rezervaciju smeštaja, Booking.com, uz pomoć Web scraping-a.

CSV fajl ima 17 polja i sadrži 515.000 recenzija korisnika o 1493 različita luksuzna hotela u Evropi.

Polja ovog dataset-a su:

- Hotel_Address: Address of hotel.
- Review_Date: Date when reviewer posted the corresponding review.
- Average_Score: Average Score of the hotel, calculated based on the latest comment in the last year.
- Hotel_Name: Name of Hotel
- Reviewer_Nationality: Nationality of Reviewer
- Negative_Review: Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
- Review_Total_Negative_Word_Counts: Total number of words in the negative review.
- Positive_Review: Positive Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Positive'
- Review_Total_Positive_Word_Counts: Total number of words in the positive review.
- Reviewer_Score: Score the reviewer has given to the hotel, based on his/her experience
- Total_Number_of_Reviews_Reviewer_Has_Given: Number of Reviews the reviewers has given in the past.
- Total_Number_of_Reviews: Total number of valid reviews the hotel has.
- Tags: Tags reviewer gave the hotel.
- days_since_review: Duration between the review date and scrape date.
- Additional_Number_of_Scoring: There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there.
- lat: Latitude of the hotel
- lng: longtitude of the hotel

Importovane biblioteke

Importing libraries

In [1]:
```
import tensorflow as tf
```

In [2]:
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

In [3]:
```
import pandas as pd
RANDOM_SEED = 42

np.random.seed(RANDOM_SEED)
tf.random.set_seed(RANDOM_SEED)
```

Tensorflow - TensorFlow dozvoljava programerima da kreiraju grafičke strukture podataka koji opisuju kako se podaci kreću kroz grafikon ili seriju procesnih čvorova. Olakšava proces rada sa neuralnim mrežama.

Numpy - Koristi se za rad sa matricama i kompleksnija izračunavanja.

Seaborn, matplot.lib - Biblioteke koje se koriste za vizuelizaciju.

Pandas - Koristi se za lakši rad sa skupom podataka (čišćenje, sređivanje)

Preprocessing

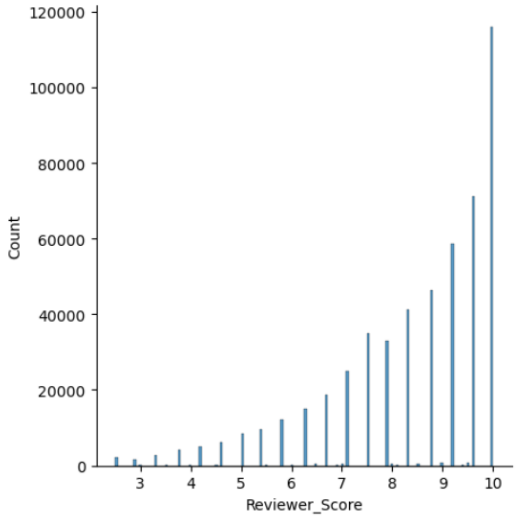
Pravi se nova varijabla ***review*** koja sadrži negativne i pozitivne komentare:

In [8]:
```
df['review'] = df['Negative_Review']+df['Positive_Review']
```

In [9]:
```
df['review'].head()
```

Out[9]:
```
0     I am so angry that i made this post available...
1    No Negative No real complaints the hotel was g...
2     Rooms are nice but for elderly a bit difficul...
3     My room was dirty and I was afraid to walk ba...
4     You When I booked with your company on line y...
Name: review, dtype: object
```

Potrebno je odrediti granicu po kojoj će se recenzije kategorisati u negativne ili pozitivne. Skala ocena je od jedan do deset. Kako je prikazano na sledećem grafu, većina korisnika je odgovorilo sa nekom višom ocenom, zato je nepravilno izbrati da polovina, odnosno ocena pet bude prag po kome će se recenzije raspoređivati. Zato je za treshold uzeta ocena 7. Ukoliko se pogleda na sajtu Booking, ovo je neki minimum koji mora da hotel ispuni za ocenu kako bi se kategorisao kao dobar.

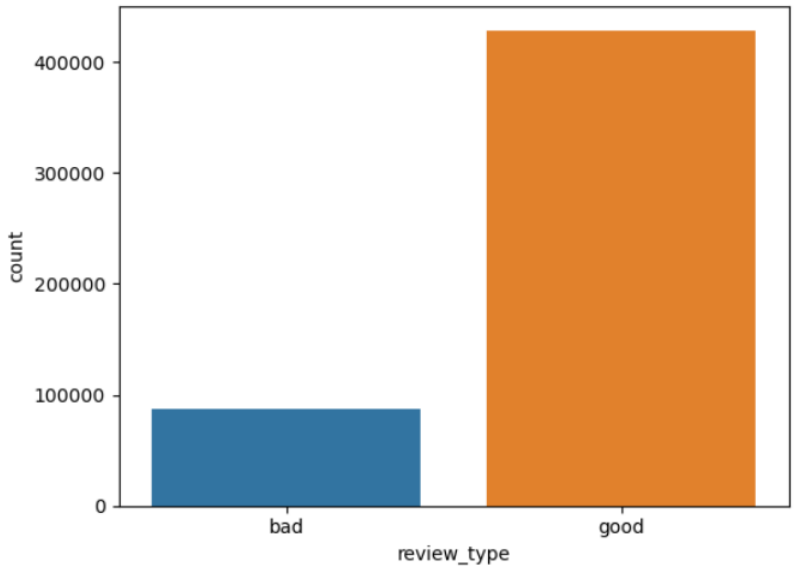



```
In [11]:  1 df['Reviewer_Score'].describe() #
Out[11]: count    515738.000000
         mean          8.395077
         std           1.637856
         min           2.500000
         25%           7.500000
         50%           8.800000
         75%           9.600000
         max          10.000000
         Name: Reviewer_Score, dtype: float64
```

Medijana je 8.8, ali je to dosta visok kriterijum kako bi se hotel kategorisao kao dobar.

Sledeća funkcija omogućava da se recencije kategorišu u dve grupe (good, bad) na osnovu Review Score-a koje imaju. Pravi se nova kolona pod nazivom **review_score:**

```
In [12]:  1 df['review_type'] = df['Reviewer_Score'].apply(lambda x: "bad" if x <7 else "good")
```



Potrebno je izbalansirati dataset posto je procenat dobrih recenzija veći od onih koje su negativno ocenjene.
Kako dataset sadrži dovoljan broj podataka, iskorištena je metoda ***under-sampling***, tako da se broj podataka u dominantnoj klasi smanjio na broj podataka koji se nalazi u klasi bad. Uzorak je uzet nasumično.

```
good_df = good_review.sample(n = len(bad_review), random_state=101)
bad_df = bad_review
```

Napravljen je novi dataframe koji ima 86851 podataka.

```
In [24]: print(good_df.shape, bad_review.shape)
         (86851, 19) (86851, 19)

In [25]: review_df = good_df.append(bad_df).reset_index(drop=True)
```

Za dalju analizu, izbačene su sve kolone osim review i review type. Review predstavlja recenziju korisnika u obliku teksa, dok review_type je kategorija recenzije(good, bad).

```
In [27]: review_df = review_df[['review', 'review_type']]
```

Embedding

Kako bi se podaci koristili dalje za sentimentnu analizu potrebno je pretvoriti tekst u oblik koji je razumljiv kompjuteru. Ovo se može uraditi putem embedding-a, odnosno, pretvaranja reči u vektore. Svrha ovoga je pronalaženje veze između semantičkog značenja i sintakse reči.

Prvo, izlazna varijabla - *review_type* je u tekstualnom obliku. Za ovu kategoričku varijablu može se koristiti One-hot encoding tehnika da bi se predstavila binarnim vektorom. Ovom tehnikom, sve varijable se prebacuju u vekorski oblik, gde su svi elementi nula, dok je element koji odgovara datoj kategoriji predstavljen jedinicom.

```
type_one_hot = OneHotEncoder(sparse = False).fit_transform(
    review_df.review_type.to_numpy().reshape(-1,1)
)
```

OneHotEncoder je deo sklearn.preprocessing biblioteke.

Sparse = False, znači da rezultat treba da bude u 2D matrici a ne u obliku Sparse matrice.
reshape(-1,1), preoblikuje 1-dimenzionalni niz u 2-dimenzionalni niz sa jednom kolonom. Ovo je neophodno jer metoda fit_transform očekuje 2D ulaz.

Sada kada je sređena izlazna varijabla, potrebno je napraviti test i train set.

In [35]:
```
from sklearn.model_selection import train_test_split
train_reviews, test_reviews, y_train, y_test =\
  train_test_split(
    review_df.review,
    type_one_hot,
    test_size=.1,
    random_state=RANDOM_SEED
  )
```

Poslednji korak pre pravljenja neuralne mreže potrebno je tekst pretovoriti u vektore. Kako bi se to desilo, korišten je **Universal Sentence Encoder.** Univerzalni dekoder rečenica (USE) je unapred obučeni model dubokog učenja koji je razvio Google i koji pretvara rečenice ili kratke tekstove u vektorske reprezentacije fiksne dužine. Dizajniran je da uporedi semantičko značenje i kontekstualne informacije rečenica, omogućavajući efikasno poređenje, pronalaženje i razumevanje tekstualnih podataka.

U ovom projektu koristi se **Multilingual Universal Sentence Encoder**.

In [38]:
```
use = hub.load("https://tfhub.dev/google/universal-sentence-encoder-multilingual-large/3")
```

In [39]:
```
from tqdm import tqdm
X_train = []
for r in tqdm(train_reviews):
    emb = use(r)
    review_emb = tf.reshape(emb, [-1]).numpy()
    X_train.append(review_emb)
```

```
100%|██████████| 156331/156331 [5:09:17<00:00,  8.42it/s]
```

In [40]:
```
X_train = np.array(X_train)
```

In [41]:
```
X_test = []
for r in tqdm(test_reviews):
    emb = use(r)
    review_emb = tf.reshape(emb, [-1]).numpy()
    X_test.append(review_emb)

X_test = np.array(X_test)
```

```
100%|██████████| 17371/17371 [38:15<00:00,  7.57it/s]
```

Objašnjenje koda:

- **tqdm** - koristi se za vizuelizaciju progresa petlje koja se izvršava.
- **emb = use(r)** - zove Universal Sentence Encoder i vrši vektorizaciju reči.
- **review_emb = tf.reshape(emb, [-1]).numpy()** - vektorizovan tekst se prebacuje u jednodimenzionalni niz.

Pravljenje modela

Za model se koristi sekvencijalni model Keras biblioteke. Ovaj model je linearan oblik slojeva i omogućava da se slojevi (layers) dodaju jedan za drugim. Svaki sloj vrši određenu operaciju ili transformaciju na ulaznim podacima.

Model nije kompleksan i sadrži tri nivoa neuroana. Prvi nivo ima 256 jedinica ili neurona, kao ulazni parametar uzima X_train. Aktivaciona funkcija je relu.

```
model.add(
    keras.layers.Dense(
        units=256,
        input_shape=(X_train.shape[1], ),
        activation='relu'
    )
)

model.add(keras.layers.Dropout(rate=0.5))

model.add(
    keras.layers.Dense(
        units=128,
        activation='relu'
    )
)

model.add(keras.layers.Dropout(rate=0.5))

model.add(keras.layers.Dense(2, activation = 'softmax'))
```

Sledeći nivo ima 128 neurona, dok poslednji, izlazni nivo sadrži samo dve jedinice koje predstavljaju klase.

Treniranje modela

Treniranje modela se vrši kroz 10 epoha, odnosno iteracija. Batch_size pokazuje koji broj neurona se propušta kroz mrežu odjendom. Validation_split je 0.1, što znači da će se 10% train dataset-a korititi za validaciju tačnosti modela tokom treniranja. Verbose funkcija prikazuje liniju koja predstavlja procenat dokle je kod stigao sa izvršavanjem. Na kraju, uz pomoć funkcije shuffle, skup podataka je izmešan što sprečava da dodje do bias-a.

```
history = model.fit(
    X_train, y_train,
    epochs = 10,
    batch_size = 16,
    validation_split = 0.1,
    verbose = 1,
    shuffle = True
)
```

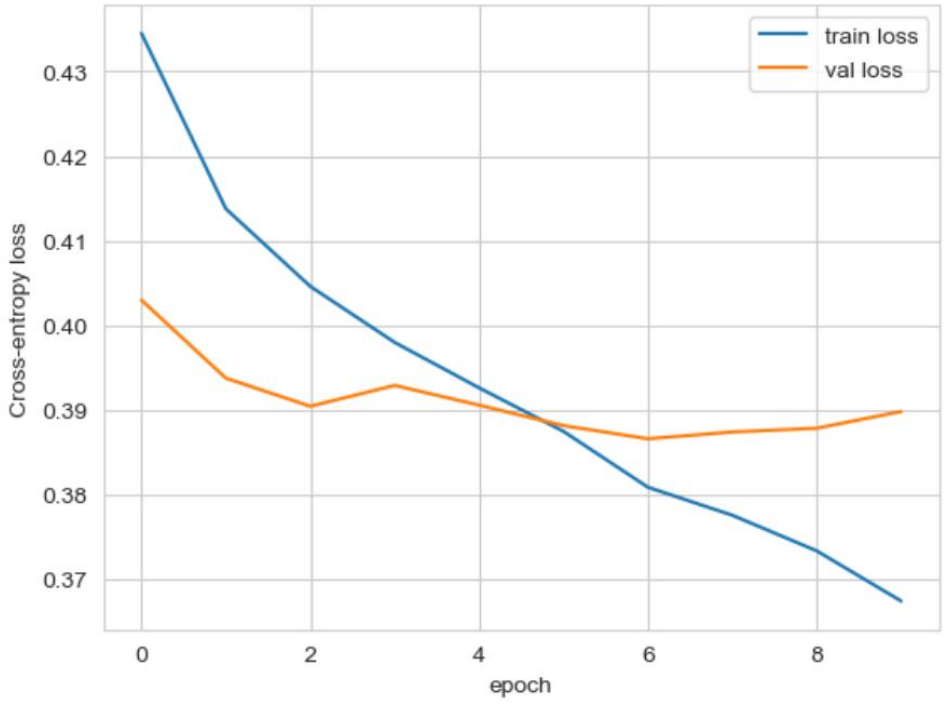
Model.fit započinje proces učenja, gde model uči iz dataseta. Tokom iteracija, on namešta težine tako da funkcija gubitka (loss function) bude što manja.

Model.compile funkcija se koristi pre samog testiranja modela. Loss funkcija je definisana kao categorical_crossentropy što se često koristi kod problema klasifikacije, gde je za izlaznu varijablu primenjen One-hot encoding princip. Optimizer određuje algoritam kojim se menjaju težine (weights) tokom treninga modela. Pomoću dela metrics, biramo koja će se metrika koristiti za definisanje rada modela koji se trenira.

```
model.compile(
    loss = 'categorical_crossentropy',
    optimizer = 'adam',
    metrics = ['accuracy']
)
```

Evaluacija modela

Na sledećem grafu prikazana je loss i validation funkcija kroz epohe tokom sentimentne analize. Ove dve funkcije se uglavnom prikazuju zajedno na grafu kako bi se videlo koja od ove dve funkcije treba korigovati. Ovaj tip prikazivanja pokazuje da li je došlo do overfitting-a ili underfitting-a.



Na kraju, model ima tačnost predikcije podataka (accuracy) od 83,2.

Zaključak

Sentimentna analiza omogućava lakši i brži uvid u iskustvo korisnika, naglašava tačke poslovanja koje je neophodno unaprediti i može poslužiti kao pomoć u poboljšanju rada firme. U ovom projektu, na jednostavnom primeru, pokazana je implementacija sentimentne analize za poslovanje u hotelijerstvu. Upotrebljen je Univerzalni dekoder rečenica (USE) kako bi se reči iz teksta vektoriozovale i dalje koristile kao ulazni parametar za model analize osećanja. Model je linearan i se sastoji iz tri nivoa. Prvi nivo ima 256 neurona, dok poslednji ima dva, kao dva moguća izlaza ovog modela. Nakon 10 epoha treniranja dobija se accuracy modela od 83.2. Vizuelizacijom loss funcije potvrđeno je da nije došlo do overfitting-a ili underfitting-a.

Literatura

baeldung. (2022, February 19). Training and Validation Loss in Deep Learning | Baeldung on Computer Science. Www.baeldung.com. https://www.baeldung.com/cs/training-validation-loss-deep-learning

Wikipedia. (2019, March 14). Sentiment analysis. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Sentiment_analysis

Gupta S. 2018 Jan 7. Sentiment Analysis: Concept, Analysis and Applications. Towards Data Science. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17.

What is sentiment analysis (opinion mining)? Definition from SearchBusinessAnalytics. (n.d.). SearchBusinessAnalytics. https://www.techtarget.com/searchbusinessanalytics/definition/opinion-mining-sentiment-mining

Monkeylearn. 2018 Jun 20. Sentiment Analysis: Nearly Everything You Need to Know | MonkeyLearn. MonkeyLearn. https://monkeylearn.com/sentiment-analysis/.

What is Sentiment Analysis? - Sentiment Analysis Explained - AWS. Amazon Web Services, Inc. https://aws.amazon.com/what-is/sentiment-analysis/#:~:text=Sentiment%20analysis%20is%20the%20process.