**Procesiranje prirodnih jezika**

**Zadatak 5. Analiza sentimenta**

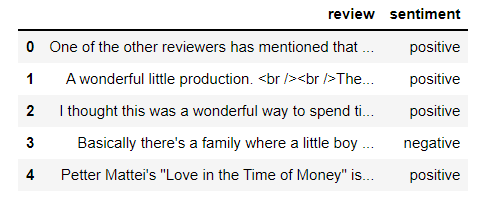
**Student Vesna Stojanović 1339**

**Niš, 2022.**

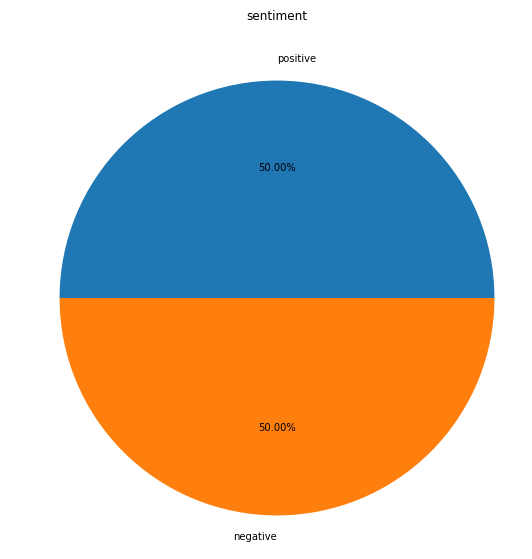
**Rešenje**

Dataset (Skup podataka) - Izabrani skup podataka nalazi se na sledećem linku: [**https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews**](https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews)**.**

U ovom zadatku korišćen je dataset koji sadrži listu IMDB review-a (utisaka filmova) u kojima su korisnici komentarisali odgledane filmove. Svaki utisak sadrži komentar i sentiment ('negative' ili 'positive'). U ovom zadatku korišćen je dataset koji sadrži listu IMDB review-a (utisaka filmova) u kojima su korisnici komentarisali odgledane filmove. Svaki utisak sadrži komentar (utisak) i sentiment ('negative' ili 'positive'). Ovo je IMDB skup podataka sa 50.000 recenzija filmova za obradu prirodnog jezika ili analitiku teksta.



Slika 1: Prikaz prvih 5 elementara data set-a

****

Slika 2: Zastupljenost sentimenta u dataset-u.

**Pretprocesiranje podataka**

Prvo se importuju biblioteke neophodne za rad kao što su nltk, pandas, matplotlib. Potom se uveze IMDB data set koji je prethodno download-ovani i takođe uradimo encoding jer poruke sadrže i emoji-eve i specijalne karaktere i reči koje nisu na engleskom pa je zbog toga potrebno setovati encoding. Prvi korak u rešavanju ovog zadatka je preprocesiranje. Za rešenje zadatka neophodno je znati tekst utiska i njegov sentiment. Promenljivama X i Y dodeljene su kolone koje sadrže utiske i sentimente. Nazivi sentimenata preimenovani su brojnim vrednostima (negative=0, positive=1).

**Tokenizacija, filtriranje, eliminacija stop reči, stemovanje**

Nad svakim utiskom iz skupa vrši se tokenizacija reči. Kreira se skup reči i interpunkcijskih znakova koji čine rečenicu. Nakon toga, iz svake reči (tokena) uklanjaju se brojevi i ostali nepotrebni karakteri i vrši se konverzija u mala slova. Nakon toga, vrši se eliminacija stop reči iz skupa. Korišćen je engleski skup stop reči iz NLTK alata jer su utisci na engleskom jeziku. Potom vrši se stemovanje dobijenih reči i vraćanje (join-ovanje) u rečenicu.

**Odredjivanje sentimenta utisaka korišćenjem Sentiment Analysis iz NLTK alata**

NLTK ima alat za određivanje sentimenta teksta (klasifikaciju) bez njegove prethodne obrade. Alat je pogodan za tekstove sa socijalnih mreža koje nemaju puno rečenica, tako da se ovo može upotrebiti nad data set-om IMDB utisaka filmova. Funkcija polarity\_scores se koristi u ovom zadatku. Njoj se prosledi tekst kao argument za koji je potrebno odrediti sentiment, a vraća 4 vrednosti: pos: Verovatnoća da je sentiment pozitivan (0-1) neu: Verovatnoća da je sentiment neutralan (0-1) neg: Verovatnoća da je sentiment negativan (0-1) compound: Normalizovana suma svih rejtinga. Ima vrednost izmedju -1 (ekstremno negativno) i 1 (ekstremno pozitivno) Tekst je pozitivan ako je vrenost compound >= 0.05, neutralan ako je compound između -0.05 i 0.05, i negativan ako je compound <= -0.05. Nakon određivanja sentimenta pomoću Sentiment Analysis alata i upoređivanja sa vrednostima iz dataset-a rezultat je 69.29%.

**Ekstrakcija atributa korišćenjem BOW metoda**

Za ekstrakciju atributa korišćen je bag of words pristup. Za prebrojavanje pojavljivanja reči korišćen je CountVectorizer iz biblioteke scikit-learn. Ovo je neophodno izvršiti radi kasnijeg izvršenja klasifikacije.

**Naivni Bajesov klasifikator iz NLTK**

Pošto za treniranje u NLTK alatu podaci u trening setu treba da budu oblika: (features, label), izvršeno je kreiranje dictionary-a funkcijom generate\_features, ključ je naziv atributa, a value vrednost za taj atribut. Set podataka deli se na trening i na test u standardnoj razmeri. Preciznost klasifikatora u ovom slučaju je 0.83.

**Decision Tree klasifikator iz NLTK**

Korišćen je i Decision Tree klasifikator koji nam pruža NLTK alat. Prosleđen je isti skup podataka koji smo koristili i za Naive Bajes klasifikator. Preciznost klasifikatora u ovom slučaju je 0.69.

**Korišćenje TF-IDF mere**

Bag of words pristup sa TF-IDF merom (Term Fruquency - Inverse Document Frequency). Ovo je pouzdaniji pristup. Vrednost za TF-IDF se dobija kao proizvod normalizovane TF i IDF (normalized\_term\_frequency \* idf). Glavni nedostatak je to što odbacuje redosled reči - ignoriše kontekst i značenje reči u dokumentu. Preciznost ovakve klasifikacije korišćenjem Naive Bajesovog klasifikatora je 0.65.

**Zaključak**

Najbolji rezultat definitivno daje Naivni Bajesov klasifikator 0.83. Decision Tree klasifikator daje slabije rezultate 0.69, dok Naivni Bajesov klasifikator korišćenjem TF-IDF mere daje najslabije rezultate 0.65. Ugrađeni Sentiment Analysis alat u okviru NLTK daje preciznost 0.69.