SVEUČILIŠTE U ZAGREBU
FAKULTET ORGANIZACIJE I INFORMATIKE
VARAŽDIN

Maja Matešić

TEHNIKE OBRADE PRIRODNOG JEZIKA ZA ANALIZU SENTIMENTA PODATAKA DRUTVENIH MEDIJA

RAD

UVOD U UMJETNU INTELIGENCIJU

Varaždin, 2024.

SVEUČILIŠTE U ZAGREBU

FAKULTET ORGANIZACIJE I INFORMATIKE

V A R A Ž D I N

Maja Matešić

Matični broj: 0016148202 (48521)

Studij: Informacijski i poslovni sustavi

TEHNIKE OBRADE PRIRODNOG JEZIKA ZA ANALIZU SENTIMENTA PODATAKA DRUTVENIH MEDIJA

RAD

Mentorica:

Izv. prof. dr. sc. Dijana Oreški

Varaždin, siječanj 2024.

Maja Matešić

Izjava o izvornosti

Izjavljujem da je ovaj RAD izvorni rezultat mojeg rada te da se u izradi istoga nisam koristio drugim izvorima osim onima koji su u njemu navedeni. Za izradu rada su korištene etički prikladne i prihvatljive metode i tehnike rada.

Autorica potvrdila prihvaćanjem odredbi u sustavu FOI Radovi

Sažetak

Projektni zadatak obuhvaća razvoj aplikacije koja primjenjuje tehnike obrade prirodnog jezika za analizu sentimenta podataka društvenih medija. Motivacija proizlazi iz potrebe za razumijevanjem stavova i osjećaja izraženih na društvenim platformama. Koristi se metoda analize sentimenta temeljena na strojnom učenju, posebice na modelima *Naive Bayes* i *Support Vector Machines*. U uvodu se ističe važnost analize sentimenta, dok se u glavnom dijelu detaljno opisuje formalizam umjetne inteligencije, teorija analize sentimenta, i konkretni algoritmi. Kritički osvrt razmatra izvedivost i primjenu, dok opis implementacije obuhvaća arhitekturu sustava i integraciju komponenti. Prikaz rada aplikacije ilustrira praktičnu primjenu na stvarnim podacima, a zaključak sažima postignute rezultate, identificira moguća poboljšanja te sugerira smjerove budućih istraživanja.

Ključne riječi: analiza sentimenta, *Naive Bayes*, metoda potpornih vektora, recenzije, društveni mediji, tehnike obrade prirodnog jezika, umjetna inteligencija, formalizam umjetne inteligencije, implementacija, prikaz rada aplikacije.

Sadržaj

1. Uvod 1

2. Metode i tehnike rada 2

3. Razrada teme 3
3.1. 3.1. Formalizam umjetne inteligencije za analizu sentimenta 3
3.2. Opis implementacije 4
3.2.1. Baza podataka 4
3.2.2. Biblioteke 6
3.2.3. Bayesov klasifikator 7
3.2.4. Metoda potpornih vektora 8
3.3. Prikaz rada aplikacije 10
3.3.1. Naivni Bayesov pristup 12
3.3.2. Metoda potpornih vektora 13
3.4. Kritički osvrt 14
3.4.1. Točnost modela 14
3.4.2. Nedostaci analize sentimenta 14

4. Zaključak 16

Popis literature 17

1. Uvod

Motivacija za odabir teme proizlazi iz sveprisutne uporabe društvenih medija u suvremenom društvu. Velika količina podataka generirana na tim platformama predstavlja izazov u analizi korisničkih stavova i osjećaja prema određenim temama. Analiza sentimenta pomoću tehnika obrade prirodnog jezika pruža mogućnost razumijevanja kako korisnici percipiraju određene događaje, proizvode ili trendove na društvenim medijima.

Strojno učenje se koristi u poslovnoj inteligenciji za analizu podataka. Algoritmi strojnog učenja mogu otkriti uzorke i trendove u velikim skupovima podataka, što je od pomoći u donošenju poslovnih odluka. U analizi sentimenta primjenjuje se strojno učenje i uz tu tehnologiju organizacije mogu lako identificirati probleme i poboljšati usluge ili proizvode. [1]

Sve povratne infromacije objavljene na internetu, poput komentara, oznaka „sviđa mi se“, recenzija, blogova, foruma prenose stavove, pobuđene emocije. Internet je mjesto gdje su mišljenja najčešće izražena. Diskutabilno je jesu li iste osobe koje pišu povratne infromacije upoznate sa subjektom komentiranja, također je upitno jesu li neanonimni, no mišljenja su prisutna. Zbog toga fenomena je interes za izrečenim stavovima na internetu sve veći.

Ljudi žele znati što drugi misle o različitim temama, od proizvoda i usluga do politike i aktualnih događaja. Ovaj projekt nije samo usmjeren na istraživanje teorijskih aspekata analize sentimenta već i na konkretnu implementaciju sustava koji koristi napredne tehnike umjetne inteligencije za donošenje relevantnih zaključaka iz podataka. Cilj je razviti jednostavni demonstrativnu aplikaciju, pridonoseći boljem razumijevanju obrade sentimenta na društvenim mrežama.

Struktura projekta obuhvaća pregled formalizma umjetne inteligencije za analizu sentimenta, teorijske osnove metode, kritički osvrt na praktičnu izvedivost, opis implementacije sustava, demonstraciju rada aplikacije te zaključak s prijedlozima za daljnja istraživanja.

Ovaj projektni rad fokusira se na Naive Bayes model strojnog učenja i Metodu potpornih vektora.

2. Metode i tehnike rada

Perspektiva će biti obogaćena relevantnom literaturom, znanstvenim člancima, radovima i teorijom s predavanja kako bi stvorili solidan temelj. Fokus će biti na konkretnim konceptima i algoritmima koje smo proučavali na kolegiju, prilagođenima analizi sentimenta na društvenim medijima. Kritički osvrt je očitovan u razmatranju prednosti i ograničenja odabranih metoda, posebno uzimajući u obzir konkretnu primjenu u analizi sentimenta na društvenim medijima.

Korištena je baza podataka koju je moguće preuzeti te su recenzije označene sentimentom (pozitivan, negativan, neutralan). Dodatno, podaci su pročišćeni od nepotrebnih riječi i znakova, kako bi se izbjegli netočni rezultati.

Metode obrade prirodnog jezika su proučene i primjenjene. Visual Studio Code i Python je odabrani set.

3. Razrada teme

Algoritmi strojnog učenja mogu otkriti uzorke i trendove u velikim skupovima podataka, što može pomoći u donošenju poslovnih odluka. Ova analiza sentimenta može identificirati ključne teme ili riječi koje se često pojavljuju među korisnicima, na primjer određivanje pozitivnog, negativnog ili neutralnog sentimenta teksta, što u krajnjem ishodu može pomoći organizacijama da brzo identificiraju problematične aspekte svojih proizvoda ili usluga i poduzmu korake za njihovo poboljšanje. [1] [2]

Trend dijeljenja iskustva i mišljenja će ostati pristuan jer korisnici se oslanjaju na povratne infromacije drugih korisnika, prijašnjih kupaca prije nego odaberu proizvod ili uslugu. Ovaj fenomen je posebno naglašen pojavom društvenih mreža i blogova, koji su omogućili ljudima da se međusobno povezuju i dijele svoje misli i osjećaje. Zbog toga je interes za izrečenim stavovima na internetu sve veći. Ljudi žele znati što drugi misle o različitim temama, od proizvoda i usluga do politike i aktualnih događaja. Organizacije od toga imaju korist ako se tim podacima znaju poslužiti. Analize sentimenta na internetu mogu se koristiti za istraživanje različitih tema. Na primjer, mogu se koristiti za istraživanje stavova o nekom proizvodu ili usluzi, praćenje promjena u stavovima opće populacije prema političkim kandidatima ili traženja savjeta i recenzija o raznim temama. Ova tehnika ima veliki potencijal za primjenu u različitim područjima, od marketinga i istraživanja tržišta do politike i društvenih znanosti. [3]

Integracija alata za poslovnu inteligenciju s alatima za strojno učenje omogućuje organizacijama da dobiju dublje uvide iz svojih podataka i donose informirane odluke na temelju tih analiza. To može biti ključno za poboljšanje korisničkog iskustva i povećanje konkurentnosti na tržištu. Dakle, internet je olakšao organizacijama da slušaju svoje kupce i prate trendove. [4]

3.1. 3.1. Formalizam umjetne inteligencije za analizu sentimenta

Transformirani i inteligentni pristupi rudarenju podataka sada omogućuju organizacijama da prikupljaju, kategoriziraju i alaliziraju korisničke recenzije, komentare i sl. s microblogging stranica koje su društveni mediji. Ova vrsta analize čini organizacije sposobnima da procjene što njihovi potrošaći žele, što odobravaju i koje se mjere mogu poduzeti kako bi održali i poboljšali povratne informacije.

Analiza sentimenta, poznata i kao „rudarenje mišljenja“, je disciplina računalne lingvistike koja se bavi određivanjem izraženog sentimenta u tekstu umjesto doslovnog značenja. Podaci su to koje je moguće sakupiti na društvenim mrežama, blogovima, forumimo i sl a korisnici ostavljaju povratne infromacije na sve entitete od proizvoda, usluga, organizacija, do pojedinaca, pitanja i slično. [5]

Glavne razine su:

1. Dokument – promatra se ukupan sentiment dokumenta, jedinstveni autor i jedan

entitet

2. Rečenica – detaljnija od analize dokumenta, preciznije proučava objekte

3. Aspekt – ukoliko postoji više autora ili entiteta, najdublje proučava mišljenje, fokusirana na suštinu, kompliciranija za implementaciju

Dijeli se na:

1. Utvrđivanje subjektivnosti: Određivanje je li tekst činjeničan (objektivan) ili je prisutan sentiment (subjektivan)

2. Utvrđivanje orijentacije: Identifikacija je li izraženi sentiment u tekstu pozitivan ili negativan

3. Utvrđivanje intenziteta: Ovaj korak uključuje procjenu je li izraženi sentiment u dokumentu slab, umjeren ili snažan.

Mnoge velike organizacije koriste sustave za analizu sentimenta kako bi dobile povratnu informaciju o svojim proizvodima ili uslugama. Na primjer, Google koristi Natural Language u sklopu Google Cloud-a, Amazon koristi Amazon Comprehend preko AWS-a (engl. Amazon web services, skraćeno AWS), IBM ima Watson Natural Language Understanding itd.

Ovi sustavi koriste tehnike obrade prirodnog jezika (NLP) za analizu teksta i identifikaciju sentimenta. Mogu se koristiti za analizu različitih vrsta podataka, kao što su recenzije proizvoda, komentari na društvenim mrežama ili razgovori s potrošačima.[6] [7]

Algoritme koji provode klasifikaciju nazivamo klasifikatorima. Postoji mnogo vrsta klasifikatora, dva klasifikatora koji se koriste pri demonstraciji su multinomijalni Naivni Bayesov klasifikator i linearna metoda potpornih vektora.

3.2. Opis implementacije

Moguće je implementirati linearnu regresiju, također se analiza sentimenta može postići metodom potpornih vektora, algoritma k-sredina, konvolucijskih neuronskih mreža, no odabrana je metoda Metode potpornih vektora Naive Bayes i (SVM).

3.2.1. Baza podataka

Najprije, baza „Hotel Reviews“ [8] koja je korištena preuzeta je data.world repozitorija bazi podataka, datasetova koje je moguće koristiti.

Sadrži informacije o recenzijama za hotele ostavljene na TripAdvisor-u, turističkom portalu koji nudi savjete za korisnike koji planiraju odmor.

Stupci su:

id, dateAdded, dateUpdated, ddress, categories, primaryCategories, city, country, keys, latitude, longitude, , postalCode, province, reviews.date, reviews.dateSeen, reviews.rating, reviews.sourceURLs, reviews.text, reviews.title, .userCity, reviews.userProvince, reviews.username, sourceURLs, websites

Primjer retka:
AVwc252WIN2L1WUfpqLP,
2016-10-30T21:42:42Z,
2018-09-10T21:06:27Z,
5921 Valencia Cir,
"Hotels,Hotels and motels,Hotel and motel reservations,Resorts,Resort,Hotel",
Accommodation and Food Services,
Rancho Santa Fe,
US,
us/ca/ranchosantafe/5921valenciacir/359754519, 32.990959,
-117.186136,
Rancho Valencia Resort
5,
https://www.hotels.com/hotel/125419/reviews
Our experience at Rancho Valencia was absolutely perfect from beginning to end!!!! We felt special and very happy during our stayed. I would come back in a heart beat!!!,Best romantic vacation ever!!!!,
Paula,
http://www.gayot.com/Hotels/Select-Your-City/United-States/California/San-Diego/Rancho-Valencia-Resort-Spa-Rancho-Santa-Fe-SDHOT02319-02",
http://www.ranchovalencia.com

Nakon rješavanja nepotrebnih stupaca, ostavila sam sljedeće stupce, kako bismo s njima krenuli u prvi korak implementacije, funkcije za pretprocesiranje teksta:

• **reviews.rating** – ocjena koju korisnik dodjeljuje hotelu, u rasponu od 1 do 5

• **reviews.sourceURLs** – URL izvor recenzije

• **reviews.text** – tekstualni sadržaj recenzije koju je korisnik napisao o hotelu, opis iskistva, dojmovi, pohvale, pritužbe

• **reviews.title** – naslov recenzije koji pruža sažetak ili ključnu informaciju o recenziji

• **reviews.userCity** – grad ili mjesto iz kojeg je korisnik ostavio recenziju

3.2.2. Biblioteke

Zatim, koristila sam nekoliko popularnih biblioteka za analizu podataka i obradu teksta.

1. pandas - za manipulaciju podacima i analizu podataka, u ovom slučaju za čitanje podataka iz .csv datoteke i stvaranje DataFramea

2. numpy - biblioteka za rad s višedimenzionalnim nizovima i matricama, u ovom kodu često korisna za numeričke operacije i rad s nizovima

3. nltk - Natural Language Toolkit (nltk) je biblioteka za obradu prirodnog jezika (NLP), za analizu sentimenta pomoću VADER analizatora sentimenta

4. matplotlib.pyplot - biblioteka za vizualizaciju podataka, za stvaranje grafova (npr. matplotlib inline za prikazivanje grafova unutar Jupyter bilježnice).

U ovom kodu koriste se različite biblioteke iz scikit-learn i nltk za obradu i analizu teksta, a sve to čini pripremu podataka i treniranje modela za klasifikaciju sentimenta (pozitivni, negativni, neutralni) na temelju recenzija.

```
from sklearn.model_selection import train_test_split \\
from sklearn.feature_extraction.text import TfidfVectorizer\\
from sklearn.svm import SVC\\
from sklearn.metrics import classification_report, accuracy_score\\
from nltk.tokenize import word_tokenize\\
from nltk.corpus import stopwords\\
from nltk.stem import WordNetLemmatizer \\
```

Slijedi uvod u korištene module i funkcije:

train_test_split iz `sklearn.model_selection`:

Funkcija se koristi za podjelu podataka na trening i test skup. Omogućuje stvaranje skupova za treniranje modela i za evaluaciju performansi modela.

TfidfVectorizer iz `sklearn.feature_extraction.text`:

TF-IDF (Term Frequency-Inverse Document Frequency) je tehnika vektorske reprezentacije teksta. Ova klasa pretvara kolekciju dokumenata u matricu značajki. Svaki dokument se predstavlja kao vektor koji odražava važnost riječi u odnosu na cijelu kolekciju dokumenata.

SVC iz `sklearn.svm`:

Ovaj modul pruža Support Vector Machine (SVM) klasifikator, a u ovom slučaju, koristi se za klasifikaciju teksta.

classification_report, accuracy_score iz `sklearn.metrics`:
`classification_report` pruža detaljan izvještaj o preciznosti, F1-score i drugim mjernim podacima za svaku klasu u modelu klasifikacije.
`accuracy_score` pruža ukupnu preciznost modela.

word_tokenize, stopwords, WordNetLemmatizer iz `nltk.tokenize, nltk.corpus, nltk.stem`:

`word_tokenize` se koristi za tokenizaciju (opojavničenje) teksta, odnosno dijeljenje teksta na pojedinačne riječi.

`stopwords` sadrži popis čestih riječi koje se često izostavljaju prilikom analize teksta jer obično ne nose semantički težak sadržaj („zaustavne“ riječi).

`WordNetLemmatizer` se koristi za lematizaciju, proces svođenja riječi na njihovu osnovu, pseudokorijen (npr. "running" postaje "run").

Kombinacija ovih alata i tehnika omogućuje izgradnju i treniranje modela koji može klasificirati recenzije hotela prema sentimentu. Prije treniranja modela, podaci se pripremaju tako da se iz teksta izdvoje značajke (TF-IDF vektorska reprezentacija), a zatim se koristi SVM klasifikator za predviđanje sentimenta.

3.2.3. Bayesov klasifikator

Tehnika za konstruiranje klasifikatora, radi se o probabilističkom modelu temeljenom na Bayesovom teoremu. Teorem opisuje vjerojatnost događaja na temelju prethodnih saznanja.

Teorem je izraza:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad P(B) \neq 0$$

A i B su događaji, $P(A|B)$ je vjerojatnost da se A dogodi, kada je B istinit, obrnuto $P(B|A)$ je vjerojatnost da će se dogoditi B ako je A istinit, $P(A)$ označava vjerojatnost da se dogodi A, a $P(B)$ da se dogodi B.

Promatranim instancama se dodaju oznake klasa iz nekog određenog skupa i one budu predstavnici vektori nekog modela. Na primjer, klasifikacija e-maila kao spam, filtrira e-poštu na temelju značajki poput riječi, frekvencije pojava određenih izraza, strukture poruka.

U analizi teksta za analizu sentimenta, vjerojatnost pojavljivanja riječi unutar teksta provjerava se kako bi se kreirao klasifikator.
Postupak uključuje:

- Treniranje klasifikatora na setu uzoraka s poznatim ishodima (negativan ili pozitivan

sentiment).

- Razbijanje svakog uzorka na vreću riječi.
- Pretprocesiranje uzoraka čišćenjem od neispravnih riječi.
- Dodjeljivanje vjerojatnosti pojave svake riječi ili fraze u uzorku.
- Naivna pretpostavka o nezavisnosti između riječi ili fraza.
- Pridjeljivanje vjerojatnosti vrećama riječi za davanje negativnog ili pozitivnog rezultata.
- Korištenje ovako prikupljenih podataka u treniranju modela.

Model računa sentiment uzorka temeljem vjerojatnosti pojave određene kombinacije riječi koje su prethodno istrenirane. Svaka riječ se zbraja kao negativna ili pozitivna vjerojatnost, čime se dobiva konačan rezultat sentimenta.

3.2.4. Metoda potpornih vektora

SVM (*Support Vector Machine*) smatra se posebno učinkovitom metodom jer je u stanju raditi s visokodimenzionalnim podacima i različitim vrstama značajki koje proizlaze iz teksta. Radi se o nadziranom pristupu, što znači da je to algoritam koji prođe treniranje s training data, zatim omogućuje rad s realnim podacima – test data. Tehnike koje se primjenjuju kod nadziranog učenja su metoda potpornih vektora, linearna regresija, logistička regresija i naivni Bayesovi klasifikatori [9].

Ono što se analizira jesu komentari, recenzije i poruke. Podaci se pripremaju na način da se tekst pretvori u numerički vektor. To ćemo postići tehnikom TF-IDF (*Term Frequency-Inverse Document Frequency*).

Ova implementacija koristi `scikit-learn` biblioteku za TF-IDF transformaciju. `fit_transform` metoda primjenjuje transformaciju na zadane tekstove, a rezultat je matrica koja predstavlja numeričke vektore. Svaka riječ u tekstu predstavljena je zasebnom značajkom u vektorima [2].

Slijedi isječak koda gdje prvi dio koda učitava podatke iz CSV datoteke "dataset.csv" i dijeli ih na značajke (`X`) i ciljnu varijablu (`y`). Zatim se tekst pretvara u vektore (TF-IDF) pomoću `TfidfVectorizer` klase. Podaci se razdvajaju na trening i test skupove pomoću `train_test_split` funkcije. Nakon toga, inicijalizira se model potpornih vektora pomoću `SVC` klase s linearnim kernelom. Model se trenira pomoću `fit` metode. Na kraju, model se evaluira pomoću `accuracy_score` i `classification_report` funkcija.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
```

```
from sklearn.metrics import accuracy_score, classification_report

# Učitavanje podataka iz CSV datoteke
df = pd.read_csv('dataset.csv')

# Podjela podataka na značajke (X) i ciljnu varijablu (y)
X = df['tekst']
y = df['sentiment']

# Pretvorba teksta u vektore (TF-IDF)
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(X)

# Razdvajanje podataka na trening i test skupove
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Inicijalizacija modela potpornih vektora
svm_model = SVC(kernel='linear')

# Treniranje modela
svm_model.fit(X_train, y_train)

# Predikcija na test skupu
predictions = svm_model.predict(X_test)

# Evaluacija modela
accuracy = accuracy_score(y_test, predictions)
report = classification_report(y_test, predictions)

print(f"Točnost modela: {accuracy}")
print("Izvještaj klasifikacije:\n", report)
```

Drugi dio koda koristi `nltk` biblioteku za klasifikaciju pozitivnih i negativnih recenzija. Funkcija `extract_features` pretvara recenzije u vektore značajki. Recenzije se dijele u pozitivne i negativne.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report

# Učitavanje podataka iz CSV datoteke
```

```
df = pd.read_csv('dataset.csv')

# Podijela podataka na značajke (X) i ciljnu varijablu (y)
X = df['review']
y = df['sentiment']

# Pretvorba teksta u vektore (TF-IDF)
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(X)

# Razdvajanje podataka na trening i test skupove
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_

# Inicijalizacija modela Naive Bayes
nb_model = MultinomialNB()

# Treniranje modela
nb_model.fit(X_train, y_train)

# Predikcija na test skupu
predictions = nb_model.predict(X_test)

# Evaluacija modela
accuracy = accuracy_score(y_test, predictions)
report = classification_report(y_test, predictions)

print(f"Točnost modela: {accuracy}")
print("Izvještaj klasifikacije:\n", report)
```

Podaci se učitavaju iz CSV datoteke `"dataset.csv"` i dijele se na značajke (`X`) i ciljnu varijablu (`y`). Tekst se pretvara u vektore (TF-IDF) pomoću `TfidfVectorizer` klase. Podaci se razdvajaju na trening i test skupove pomoću `train_test_split` funkcije. Na kraju, inicijalizira se model Naive Bayes pomoću `MultinomialNB` klase. Model se trenira pomoću `fit` metode. Na kraju, model se evaluira pomoću `accuracy_score` i `classification_report` funkcija.

3.3. Prikaz rada aplikacije

U ovom dijelu, prikazan je konkretan rad sustava za analizu sentimenta. Prikazat ću kako sam došla do željenih rezultata. Kompletna implementacija je u dokumentaciji.

Ovaj dataset sadrži popis od 1000 hotela i njihovih recenzija koje je osigurao *Datafiniti Business Database. Dataset* uključuje lokaciju hotela, naziv, ocjenu, podatke o recenzijama,

naslov, korisničko ime i još mnogo toga. Podaci o recenzijama omogućuju povezivanje ključnih riječi u tekstu recenzije s ocjenama. [11]

Podaci su pretprocesirani. Pretprocesiranje je obilježeno uklanjanjem URL-a, uklanjanjem zaustavih riječi, normaliziranjem teksta, te je razlika bila znatna kada se radi o recenzijama.

Podaci su učitani iz .csv datoteke, zatim je inicijaliziran VADER analizator aentimenta, VADER (*Valence Aware Dictionary and sEntiment Reasoner*) je alat za analizu sentimenta. Inicijalizira se `SentimentIntensityAnalyzer` objekt iz biblioteke `nltk`.

Definiraju se funkcije za klasifikaciju sentimenta i na temelju srednjeg sentimenta se određuju kategorije: *'positive', 'negative' ili 'neutral'.*

Također, osim klasifikacije za `reviews.text`, odnosno recenzije, odrađena je i klasifikacija za naslove receczija. Rezultati se ispisuju.

Rezultati za `reviews.text`:

reviews.text	**sentiment**
currently bed writing past hr dog barking sque...	neutral
live md aloft home away home stayed 1 night st...	positive
stayed family daughter wedding accommodating s...	positive
stayed visiting maryland live cute hotel great...	positive
travel lot job constantly staying hotel arrive...	positive
...	...
staying 4 year visit son attends boise state a...	positive
hard review oceanfront hotel go ocean necessar...	positive
live close needed stay somewhere night due ren...	positive
rolled laid head woke continental breakfast ro...	positive
filthy outdated noisy neighbour worst nearly e...	negative

Rezultati za `reviews.title`:

reviews.title	sentiment_title
Never again...beware, if you want sleep.	positive
ALWAYS GREAT STAY...	positive
Wonderful stay	positive
Great Hotel Experiece!	positive
Short stay for business.	neutral
...	...
Stay here every time I visit Boise	neutral
Picture Window Ocean View!	neutral
Clean, comfortable and quiet	positive
Passing through	neutral
Polde	neutral

3.3.1. Naivni Bayesov pristup

Izvješće daje informacije o performansama *Naivnog Bayes* klasifikatora za svaku klasu. *Breakdown* ključnih metrika:

Preciznost: Omjer točno predviđenih pozitivnih promatranja prema ukupno predviđenim (točno + netočno / pravi + lažni) pozitivima.

Odziv: Omjer točno predviđenih pozitivnih promatranja prema svim stvarnim pozitivima. Izraz glasi pravi pozitivi dijeljeno s pravi pozitivi + lažni pozitivi.

F1-ocjena: Težinska srednja vrijednost preciznosti i odziva. Dobar način prikaza da klasifikator ima dobru vrijednost i za lažno pozitivne i lažno negativne rezultate.
Točnost modela iznosi: Naive Bayes točnost modela: 0.5281954887218046

Tumačenje:

- Za klasu 1.0, preciznost, odziv i F1-ocjena modela su sve nula, što ukazuje da model nije ispravno predvidio nijedan primjerak ove klase. Isti uzorak primjećuje se za klase 2.0 i 3.0.
- Za klasu 4.0, preciznost i odziv su niski, što sugerira da je model imao problema s ispravnim prepoznavanjem primjeraka ove klase. F1-ocjena također je niska.
- Za klasu 5.0, preciznost, odziv i F1-ocjena su relativno viši, što ukazuje na bolje performanse za ovu klasu.
- Ukupna točnost iznosi 0,53 (ili 53%), što znači da je model ispravno predvidio kategoriju ocjene za 53% instanci u testnom skupu.

Rezultati ukazuju na to da model ima poteškoća u postizanju dobrih rezultata za sve klase, te da bi moglo biti prostora za poboljšanje, možda prilagodbom hiperparametara, kori-

štenjem drugog modela ili dobivanjem više podataka. Također, neravnoteža u broju instanci pojedinih klasa također bi mogla utjecati na performanse modela.

Naivni Bayesov klasifikator procjenjuje sentiment uzorka temeljem vjerojatnosti pojavljivanja određenih kombinacija riječi, koje su prethodno istrenirane. Svaka riječ se tretira kao nezavisna značajka, pridonoseći pozitivnoj ili negativnoj vjerojatnosti, te se ti doprinosi zbrajaju kako bi se dobio konačan rezultat.

Unatoč nazivu *"naivni"*, ovaj klasifikator pretpostavlja nezavisnost između riječi koje utječu na sentiment. Iako ova pretpostavka nije uvijek točna u stvarnim podacima gdje riječi mogu biti međusobno povezane, naivni Bayesov klasifikator i dalje pruža dobre rezultate. Jedna od prednosti ovog pristupa leži u tome što zahtijeva relativno malen skup podataka za treniranje u usporedbi s drugim metodama analize sentimenta.

3.3.2. Metoda potpornih vektora

Drugi naziv je strojevi za podršku vektorima ili metoda jezgrene funkcije. Radi se o modelu predviđanja, u okrilju nadziranog učenja za analizu regresije i klasifikaciju, konstruira se kako bi klasificirao dvije grupe podataka na temelju označenog skupa za treniranje, zatim kreira model koji je zadužen za dodjeljivanje novih istanci u jednu od dviju klasa. Koristi se i za klasificiranje više klasa i kada radi na princip jedna klasa vs. druge klase.

Ovo nije probabilistički model. Korišteni skup podataka, odnosno stupac koji je odabran za obradu sadrži samo tekst, sve se značajke odnose na učestalost pojavnosti riječi i podaci su vizualizirani kao n-dimenzionalni vektori u prostoru. Radi se podjeli između dviju kategorija, koju osigurava vektor, granica. Moguće je koristiti se s više vektora, teži se naći onaj vektor koji je maksimalno udaljen od točaka jednog skupa i drugog.

Ključne karakteristike su:

- Margina – udaljenost između točaka koje se smatraju „kritičnima“ (one koje su najbliže plohi razdvajanja)

- Potporni vektor – najbliži plohi razdvajanja

- Kernel trik – omogućuje pretvaranje neseparabilnih problema u separabilne

Koraci korištenja SVM klasifikatora su sljedeći:

1. Podjela skupa podataka:
    - Skup podataka podijeljen je na trening i test skupove pomoću `train_test_split` funkcije. Trening skup se koristi za treniranje SVM modela, dok se testni skup koristi za evaluaciju performansi modela.
2. TF-IDF vektorizacija:

- Tekstualni podaci su vektorizirani pomoću TF-IDF vektorizacije pomoću `TfidfVectorizer`. Ova tehnika pretvara tekst u numerički oblik, pri čemu su riječi reprezentirane njihovim TF-IDF vrijednostima.

3. Treniranje SVM klasifikatora:

   - SVM model je instanciran pomoću `SVC()` (Support Vector Classification) i zatim treniran na trening skupu koji je prethodno vektoriziran.

4. Predviđanja i evaluacija modela:

   - Na temelju treniranog modela, napravljena su predviđanja na testnom skupu. Zatim su izračunate metrike performansi, poput točnosti i izvješća o klasifikaciji (precision, recall, F1-score).
   - Na kraju, ispisuje se točnost i izvješće o klasifikaciji za SVM model na testnom skupu.

Evidentno je da je točnost nešto viša nego s Naivnim Bayesovim pristupom:

Točnost Metode potpornih vektora 0.5620300751879699

3.4. Kritički osvrt

3.4.1. Točnost modela

Komentar za točnost od oko 0.53 nam govori da bi preciznost bila još veća da se koristio neki leksikon ili slično, što bi koristilo više resursa, no preporučljivo je za organizacije koje bi pratile svoj napredak na društvenim medijima. Također, u tom bi se slučaju program izvodio na serverima, što je efektivnije od lokalnog izvođenja na osobnom računalu.

3.4.2. Nedostaci analize sentimenta

Analiza sentimenta je moćna tehnika koja može biti korisna u raznim područjima. Međutim, kao i svaka druga tehnika, ima i svoje nedostatke.

Jedan od glavnih nedostataka analize sentimenta je osjetljivost na ironiju i sarkazam. Ljudi često koriste ironiju i sarkazam za izražavanje negativnih emocija, ali ti izrazi često mogu biti nejasni za računala. Računala mogu pogrešno shvatiti ironiju kao pozitivan sentiment, što može dovesti do pogrešnih zaključaka.

Negacije i dvostruke negacije također mogu predstavljati problem za analizu sentimenta. Negacije se često koriste za izražavanje suprotnog sentimenta, ali računala ne mogu uvijek pravilno razumjeti kontekst u kojem se negacije koriste. Na primjer, rečenica "Nisam zadovoljan ovim proizvodom" može se shvatiti kao negativna, ali računalo može pogrešno zaključiti da je pozitivna jer riječ "zadovoljan" ima pozitivnu konotaciju. Višeznačnost riječi također

može predstavljati problem za analizu sentimenta. Neke riječi mogu imati više značenja, a računala ne mogu uvijek pravilno razumjeti koje značenje je relevantno u određenom kontekstu.

Načini za poboljšanje analize sentimenta Postoji nekoliko načina za poboljšanje točnosti analize sentimenta. Jedan od načina je uključivanje konteksta u analizu. Računala mogu biti bolja u prepoznavanju ironije i sarkazma ako imaju više informacija o kontekstu u kojem se izrazi koriste.

Korištenje strojnog učenja također može pomoći u poboljšanju točnosti analize sentimenta. Strojno učenje se može koristiti za treniranje računala na skupu podataka teksta koji je označen po sentimentu. Računalo može naučiti identificirati obrasce u tekstu koji su povezani s određenim sentimentom. Istraživači također rade na razvoju novih metoda analize sentimenta koje su otpornije na ironiju, negacije i višeznačnost.

4. Zaključak

Komentari korisnika na web portalima, društvenim mrežama i blogovima predstavljaju dragocjen izvor informacija o stavovima i povratnim informacijama o aktualnim događajima i trendovima. Automatizacija prikupljanja i analize tih podataka postaje ključna, a jedan od učinkovitih pristupa je analiza sentimenta. Kroz implementaciju različitih algoritama za analizu sentimenta pomoću programskog jezika Python i strojnog učenja, projekt je pružio uvid u mogućnosti analize sentimenta u različitim područjima.

Dataset "Recenzije hotela" sadrži popis od 1000 hotela i njihovih recenzija koje je osigurao Datafiniti Business Database. Dataset uključuje lokaciju hotela, naziv, ocjenu, podatke o recenzijama, naslov, korisničko ime i još mnogo toga. Možete koristiti ovaj skup podataka za usporedbu recenzija hotela po državama, eksperimentiranje sa skoringom sentimenta i drugim tehnikama obrade prirodnog jezika. Podaci o recenzijama omogućuju vam povezivanje ključnih riječi u tekstu recenzije s ocjenama.

Za analizu sentimenta možete koristiti metode Naive Bayes i SVM. Naive Bayes je nadzirani algoritam strojnog učenja koji se koristi za klasifikacijske probleme. Izgrađen je na Bayesovom teoremu i omogućuje izradu jednostavnih klasifikatora na temelju njega. SVM, s druge strane, vrlo je popularan model koji primjenjuje geometrijsku interpretaciju podataka. Mapira točke podataka u prostor kako bi se maksimizirala udaljenost između dvije kategorije. SVM-ovi mogu učinkovito izvršiti nelinearnu klasifikaciju pomoću takozvanog kernel trika. Kernel trik sastoji se od upotrebe specifičnih jezgrenih funkcija koje pojednostavljuju mapiranje između izvornog prostora u prostor više dimenzija.

Upotrebom ovih metoda možete klasificirati recenzije kao pozitivne, negativne ili neutralne. Aplikacija se može koristiti za analizu sentimenta u recenzijama hotela i može se primijeniti u turističkoj industriji za poboljšanje kvalitete usluge.

Ova implementacija, bazirana na algoritmima strojnog učenja i programskom jeziku Python, pružila je korisne uvide u praktičnu primjenu analize sentimenta na različitim područjima, uključujući marketing. Automatizirana analiza sentimenta pokazuje se kao snažan alat za donošenje informiranih poslovnih odluka temeljenih na stavovima i osjećajima korisnika.

Popis literature

[1] S. J. Russell i P. Norvig, ur., *Artificial Intelligence: A Modern Approach* (Pearson Series in Artificial Intelligence), 4. izdanje, u sur. s M. Chang, J. Devlin, A. Dragan i dr. Harlow, UK: Pearson Education Limited, 2022., 1166 str., ISBN: 978-1-292-40113-3.

[2] M. Ahmad, S. Aftab, M. Bashir i N. Hameed, „Sentiment Analysis using SVM: A Systematic Literature Review," *International Journal of Advanced Computer Science and Applications*, sv. 9, 2018.

[3] B. Liu, „Sentiment Analysis and Subjectivity," *Handbook of Natural Language Processing*, Chapman i Hall, 2010.

[4] B. Pang i L. Lee, *Opinion Mining and Sentiment Analysis*. Foundations i Trends in Information Retrieval, 2008.

[5] „Approaches of NLP and Sentiment Classification." (2020.), adresa: `https://kaggle.com/subhamoybhaduri/approaches-of-nlp-and-sentiment-classification`.

[6] V. Singh i S. K. Dubey, „Opinion Mining and Analysis: A Literature Review," 2014.

[7] M. Schatten, J. Ševa i B. Okreša Đurić, „Synesketch: An Introduction to Social Semantic Web Mining & Big Data Analytics for Political Attitudes and Mentalities Research," 2015.

[8] *Hotel Reviews Dataset*, `https://data.world/datafiniti/hotel-reviews`.

[9] S. Patel. „Chapter 2: SVM (Support Vector Machine)." (2017.), adresa: `https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72`.