Univerzitet u Beogradu

Matematički fakultet

Sentiment Analysis with comparison of BERT and Naive Bayes

Autori: Zolotarev Igor, Anja Čolić

Profesor: dr Vladimir Filipović

Asistenti: Stefan Kapunac i Ivan Pop-Jovanov

Datum: Septembar 2024

Contents

1 Uvod

Analiza sentimenta se bavi klasifikovanjem emocionalnog stanja izraženog u tekstu, što može uključivati pozitivne ili negativne stavove. Ova vrsta postaje sve važnija kako za istraživanje tržišta, tako i za razumevanje mišljenja korisnika o proizvodima, uslugama i društvenim pitanjima.

Tradicionalni pristupi analizi sentimenta često koriste jednostavne modele poput Naivnog Bajesa, koji su brzi i efikasni, ali možda ne pružaju dovoljno duboko razumevanje složenih jezičkih obrazaca. S druge strane, savremeni modeli zasnovani na dubokom učenju, kao što je BERT (Bidirectional Encoder Representations from Transformers), omogućavaju obučavanje modela koji mogu bolje razumeti kontekst i nijanse u jeziku. BERT se oslanja na napredne tehnike, što mu omogućava da uči iz celokupnog konteksta rečenica, čime se poboljšava tačnost analize.

Cilj ovog rada je da uporedi performanse Naivnog Bajesa i BERT-a u analizi sentimenta koristeći IMDB recenzije kao dataset. Istražićemo kako se rezultati razlikuju izmeu ovih pristupa i razmotriti prednosti i nedostatke svakog od njih.

2 Sentiment analiza

Sentiment analiza, poznata i kao eng. *opinion mining*, predstavlja proces automatskog prepoznavanja i klasifikacije emocija ili stavova izraženih u tekstu. Ova metoda obrade prirodnog jezika (NLP) koristi se za identifikaciju polariteta teksta, tj. da li je sentiment pozitivan ili negativan. Glavni cilj sentiment analize je da interpretira i razume emocionalni ton teksta kako bi se izvukle korisne informacije o stavovima autora prema odredjenoj temi, proizvodu ili dogadjaju.

2.1 Koraci sprovodjenja sentiment analize

Sentiment analiza funkcioniše tako što koristi različite metode obrade teksta, uključujući leksičke pristupe, mašinsko učenje i neuronske mreže. Postupak sentiment analize može se podeliti u nekoliko ključnih koraka:

1. **Prikupljanje podataka:** Tekstualni podaci dolaze iz različitih izvora kao što su društvene mreže, recenzije, forumi, i vesti.

2. **Predobrada podataka:** Ovaj korak uključuje čišćenje teksta kako bi se uklonili šumovi (npr. specijalni znakovi, brojevi, bespotrebni razmaci). Takodje, uključuje tehnike poput *tokenizacije* (razdvajanje teksta na reči), *lemmatizacije* (redukovanje reči na njihov osnovni oblik) i uklanjanje zaustavnih reči (česte reči poput i“, ali“, da“).

3. **Izvlačenje karakteristika:** Karakteristike (eng. *features*) kao što su frekvencija reči, n-gram modeli, rečnički polaritet (pozitivne ili negativne reči) koriste se kako bi se tekst pretvorio u reprezentaciju pogodnu za modeliranje.

4. **Klasifikacija:** Nakon što je tekstualni skup podataka transformisan u odgovarajuće numeričke reprezentacije, koristi se model mašinskog učenja kako bi se klasifikovali stavovi izraženi u tekstu. Popularni modeli uključuju Naivni Bajesov klasifikator, podršku vektorskih mašina (SVM), BERT, itd.

5. **Evaluacija:** Klasifikacija se evaluira kroz metrike poput tačnosti, preciznosti, odziva i F1-score, kako bi se utvrdila uspešnost modela u prepoznavanju sentimenta.

2.2 Primena sentiment analize

Sentiment analiza se široko koristi u različitim oblastima, a neka od ključnih polja primene su:

- **Analiza društvenih medija:** Kompanije i organizacije koriste sentiment analizu da bi analizirale stavove i emocije korisnika na platformama kao što su Twitter, Facebook i Instagram.
- **Analiza recenzija proizvoda i usluga:** Kompanije koriste sentiment analizu da bi analizirale recenzije korisnika na sajtovima kao što su Amazon, Yelp ili TripAdvisor. Ova tehnika omogućava prepoznavanje pozitivnih i negativnih aspekata proizvoda ili usluga.
- **Politička analiza:** U političkim kampanjama, sentiment analiza se koristi za praćenje stavova birača na društvenim mrežama i medijima. Ona pomaže u proceni popularnosti kandidata i detekciji tema koje izazivaju emocije meu glasačima.
- **Finansijska tržišta:** Sentiment analiza se koristi u finansijskoj industriji za analizu vesti, članaka i društvenih komentara u vezi sa odreenim akcijama, valutama ili tržištima. Na osnovu toga se prave predikcije o promenama cena i trendovima.

2.3 Izazovi u sentiment analizi

Sentiment analiza suočava se sa brojnim izazovima:

- **Sarkazam i ironija:** Detekcija sarkazma i ironije predstavlja veliki izazov, jer model može pogrešno interpretirati negativni ton kao pozitivan.
- **Kontekstualno razumevanje:** Sentiment analiza može imati teškoće u razumevanju konteksta i nijansi u tekstu. Na primer, ista reč može imati različito značenje u različitim kontekstima.
- **Višejezičnost:** Većina dostupnih modela trenirana je na tekstovima na engleskom jeziku, pa prelazak na druge jezike može zahtevati dodatne prilagodbe.

Uprkos izazovima, sentiment analiza je postala nezaobilazan alat u mnogim industrijama zbog svoje sposobnosti da izvuče korisne uvide iz velikih količina tekstualnih podataka.

3 BERT

BERT (Bidirectional Encoder Representations from Transformers) je napredni model za obradu prirodnog jezika (NLP) koji je predstavljen od strane Google-a u oktobru 2018. godine. BERT je postigao revolucionarne rezultate u mnogim zadacima NLP-a, uključujući analizu sentimenta, prepoznavanje entiteta, klasifikaciju teksta, prevodjenje, i odgovaranje na pitanja.

3.1 Osnovne karakteristike BERT-a

- **Bidirekcionalnost**: Tradicionalni jezički modeli obično obradjuju tekst ili sleva nadesno ili zdesna nalevo. BERT, s druge strane, koristi bidirekcionalni kontekst, što znači da simultano analizira reči u oba smera — uzimajući u obzir reči koje dolaze pre i posle date reči. Ovo omogućava BERT-u da bolje razume značenje reči u njihovom kontekstu.

- **Transformer arhitektura**: BERT je zasnovan na Transformer arhitekturi. Transformer koristi mehanizam self-attention da bi procenio važnost svake reči u rečenici u odnosu na sve druge reči. Na taj način BERT može paralelno analizirati sve reči u tekstu i efikasno procenjivati njihov meusobni odnos.

- **Pre-trening**: BERT se prvo trenira na ogromnim količinama nestruktuiranog teksta iz knjiga i Vikipedije, koristeći dva specifična zadatka: maskirano predvianje reči (Masked Language Model - MLM) i predvianje sledeće rečenice (Next Sentence Prediction - NSP). U MLM zadatku, odreeni procenat reči u rečenici se maskira, i BERT mora da predvidi te reči na osnovu njihovog konteksta. NSP zadatak podučava model da razume odnose izmedju rečenica.

3.2 Kako BERT funkcioniše?

1. **Unos teksta**

   Tekst koji se unosi u BERT se najpre razdvaja na tokene. BERT koristi WordPiece tokenizaciju koja deli reči na osnovne delove. Na primer, reč playing“ može biti podeljena na play“ i ing“.

2. **Maskirani jezički model (MLM)**

   Tokom pre-treninga, BERT nasumično maskira odredjeni procenat tokena u rečenici (oko 15%) i zatim pokušava da predvidi te maskirane tokene koristeći njihov kontekst. Na taj način model uči značenje reči unutar njihovog okruženja.

3. **Predikcija sledeće rečenice (NSP)**

   Pored MLM zadatka, BERT se trenira da predvidi da li je druga rečenica logički sledeća nakon prve. Ovaj zadatak omogućava BERT-u da razume odnose izmeu rečenica.

4. **Self-attention**

   Transformer arhitektura koristi mehanizam koji omogućava modelu da usmeri pažnju na relevantne reči u rečenici u odnosu na ostale. To znači da BERT može bolje obraditi složene zavisnosti izmedju reči, i to u oba smera.

3.3 BERT-ova primena

BERT se široko koristi u različitim NLP zadacima zbog svoje sposobnosti da efikasno razume kontekst u kojem se nalaze reči. Neke od ključnih primena su:

- **Analiza sentimenta:** Identifikacija tonova i emocija u tekstu.
- **Odgovaranje na pitanja (Question Answering):** Automatsko odgovaranje na pitanja na osnovu unetog teksta.
- **Prepoznavanje entiteta (Named Entity Recognition - NER):** Identifikacija imena, mesta, organizacija u tekstu.
- **Prevoenje teksta:** Korišćenje BERT-a za poboljšanje tačnosti prevodjenja.

4 Naive Bayes

Naivni Bajes (Naive Bayes) je jednostavan, ali veoma efikasan algoritam za klasifikaciju koji se oslanja na Bajesovu teoremu, a koristi se za predvidjanje verovatnoće pripadnosti klasi na osnovu skupa podataka. Iako se naziv "naivni" koristi zbog pretpostavke o nezavisnosti izmeu atributa, Naivni Bajes često daje izvanredne rezultate u praksi, posebno u zadacima klasifikacije teksta.

4.1 Osnovne karakteristike Naivnog Bajesa

1. **Bajesova teorema:** Naivni Bajes koristi Bayesovu teoremu koja izračunava verovatnoću dogadjaja na osnovu prethodnih informacija. Teorema:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

   Gde:

   - $P(C|X)$ je verovatnoća klase C, pod uslovom da vazi atribut X.
   - $P(X|C)$ je verovatnoća atributa X, pod uslovom da vazi C.
   - $P(C)$ je verovatnoća klase C.
   - $P(X)$ je ukupna verovatnoća podataka (atributa X).

2. **Naivna pretpostavka nezavisnosti:** Pretpostavlja se da su svi atributi u skupu podataka medjusobno nezavisni, što znači da se verovatnoća pojavljivanja jednog atributa ne menja zbog prisustva drugog atributa. Iako ova pretpostavka retko važi, Naivni Bajes i dalje daje dobre rezultate jer smanjuje složenost računanja verovatnoća.

3. **Vrste Naivnog Bajesa:**

   - **Gaussov Naivni Bajes:** Koristi se za kontinuirane atribute i pretpostavlja da su vrednosti atributa normalno (Gaussovski) distribuirane.
   - **Multinomialni Naivni Bajes:** Koristi se za diskretne atribute, posebno za zadatke klasifikacije teksta, gde su atributi reči ili tokene u dokumentu.
   - **Bernulijev Naivni Bajes:** Takodje se koristi za diskretne podatke, ali je posebno efikasan kada su atributi binarni (0 ili 1).

4. **Treniranje i klasifikacija:**

   - **Treniranje:** Naivni Bajes se trenira izračunavanjem verovatnoće svake klase $P(C)$ na osnovu učestalosti pojavljivanja svake klase u skupu za obuku. Takoe se izračunava $P(X|C)$ za svaki atribut X u odnosu na svaku klasu C.
   - **Klasifikacija:** Da bi se klasifikovala nova instanca X, Naivni Bajes izračunava verovatnoću $P(C|X)$ za svaku moguću klasu C i bira klasu sa najvećom verovatnoćom.

4.2 Kako Naivni Bajes funkcioniše?

1. **Sakupljanje podataka:**

   Naivni Bajes se obučava na skupu podataka gde je svaka instanca obeležena klasom. Na primer, za zadatak klasifikacije e-mail poruka kao spam“ ili nije spam“, e-mailovi su označeni odgovarajućim etiketama, a atributi su reči unutar poruka.

2. **Izračunavanje verovatnoća:**

   Nakon što je model obučen, za novu instancu podataka X (na primer, novi e-mail), model koristi Bayesovu teoremu kako bi izračunao verovatnoću da X pripada svakoj mogućoj klasi (npr. spam ili nije spam).

3. **Predikcija:**

   Klasa sa najvećom verovatnoćom $P(C|X)$ biće predikcija modela za novu instancu. Na primer, ako verovatnoća da je e-mail spam $P(spam|X)$ veća od verovatnoće da nije spam $P(nijespam|X)$, model će ga označiti kao spam.

4.3 Prednosti Naivnog Bajesa

- **Brzina**: Naivni Bajes je veoma brz i može obraditi veliki broj podataka u realnom vremenu.
- **Jednostavnost**: Implementacija Naivnog Bajesa je relativno jednostavna, što ga čini pristupačnim za mnoge primene u mašinskom učenju.
- **Efikasan za male skupove podataka:** Iako su drugi algoritmi skloni pretreniravanju na malim datasetovima, Naivni Bajes može raditi dobro čak i sa ograničenom količinom podataka. Nizak zahtev za memorijom: Zbog jednostavne strukture, Naivni Bajes ne zahteva mnogo memorijskih resursa.

4.4 Ograničenja Naivnog Bajesa

- **Pretpostavka nezavisnosti:** Glavno ograničenje Naivnog Bajesa je njegova osnovna pretpostavka o nezavisnosti atributa. U stvarnim podacima atributi često nisu nezavisni, što može negativno uticati na performanse modela.
- **Osetljivost na podatke sa nultim vrednostima:** Ako se tokom obučavanja ne pojavi odredjeni atribut za neku klasu, Naivni Bajes može izračunati nultu verovatnoću za taj atribut u budućim predikcijama, što može dovesti do problema. Ovaj problem se može rešiti tehnikom Laplasove korekcije (dodavanje male pozitivne vrednosti svim učestalostima).

4.5 Primene Naivnog Bajesa

- **Klasifikacija teksta:** Naivni Bajes je jedan od najčešće korišćenih algoritama za zadatke klasifikacije teksta, uključujući:
  1. **Analiza sentimenta:** Razvrstavanje teksta prema njegovom tonu, npr. pozitivan ili negativan sentiment.
  2. **Filtriranje spam poruka:** Klasifikacija e-mailova kao spam“ ili nije spam“ na osnovu njihovog sadržaja.
  3. **Klasifikacija dokumenata:** Razvrstavanje dokumenata po temama na osnovu reči koje sadrže.
- **Filtriranje sadržaja:** Naivni Bajes se koristi za prepoznavanje relevantnog sadržaja u preporučivačkim sistemima.
- **Prepoznavanje entiteta:** Korišćenje Naivnog Bajesa za identifikaciju ključnih reči i entiteta u tekstu, kao što su imena, mesta ili organizacije.

5 Dataset

IMDB dataset recenzija

Za potrebe sentiment analize u ovom radu korišćen je IMDB dataset, koji sadrži 50,000 filmskih recenzija izvučenih sa sajta Internet Movie Database (IMDB).

5.1 Struktura dataset-a

Dataset je podeljen na dva atributa:

- **Tekst recenzije:** Ovaj atribut sadrži stvarne recenzije filmova napisane od strane korisnika IMDB-a. Recenzije variraju u dužini i mogu sadržati različite oblike izražavanja, od formalnog jezika do neformalnog slenga.
- **Sentiment:** Ovaj atribut sadrži oznaku koja pokazuje polaritet sentimenta. Postoje dve moguće vrednosti: *positive* (pozitivno) ili *negative* (negativno), što predstavlja stav autora prema filmu.

Recenzije su jasno obeležene kao pozitivne ili negativne, što omogućava jednostavnu binarnu klasifikaciju sentimenta. Svaka recenzija se sastoji od jednog teksta i pridruženog sentimenta. Primer podataka može se videti u tabeli 1.

ID	Tekst recenzije	Sentiment
0	One of the other reviewers has mentioned that this is ...	Positive
1	A wonderful little production. ¡br /¿¡br /¿The...	Positive
2	I thought this was a wonderful way to spend time on a rainy afternoon.	Positive
3	Basically there's a family where a little boy...	Negative
4	Petter Mattei's "Love in the Time of Money" is...	Positive
...	...	...
49995	I thought this movie did a down right good job...	Positive
49996	Bad plot, bad dialogue, bad acting, idiotic dialogue...	Negative
49997	I am a Catholic taught in parochial elementary...	Negative
49998	I'm going to have to disagree with the previous review...	Negative
49999	No one expects the Star Trek movies to be high art...	Negative

Table 1: Primeri recenzija iz IMDB dataset-a

5.2 Obim i podela podataka

IMDB dataset se sastoji od ukupno 50,000 recenzija, koje su ravnomerno podeljene na pozitivne i negativne. To znači da postoji 25,000 pozitivnih i 25,000 negativnih recenzija, što dataset čini izbalansiranim i pogodnim za treniranje i evaluaciju modela.

6 Rezultati

Kroz eksperimentalnu evaluaciju, testirani su modeli BERT i Naivni Bajes koristeći isti dataset za sentiment analizu IMDB recenzija filmova. Rezultati prikazani putem matrica konfuzije i ROC krivih pružaju uvid u preciznost i tačnost svakog modela.

1. **Matrica konfuzije**:

   BERT model je ostvario tačnost od 83.83%. Ukupno je tačno klasifikovano 4136 negativnih i 4177 pozitivnih recenzija, dok su pogrešno klasifikovane 803 negativne i 801 pozitivna recenzija.

   Naivni Bajes klasifikator postigao je tačnost od 85.86%, što je nešto bolje od BERT-a. Ispravno je klasifikovano 4324 negativne i 4191 pozitivne recenzije, dok su pogrešno klasifikovane 615 negativnih i 787 pozitivnih.

   Ovi podaci pokazuju da, iako Naivni Bajes klasifikator ima nešto višu ukupnu tačnost, BERT model pokazuje bolju ravnotežu u klasifikaciji pozitivnih i negativnih primera.

2. **ROC Kriva**:

   Površina ispod ROC krive (AUC) za BERT model iznosi 0.92, što ukazuje na visoku sposobnost modela da razlikuje izmeu pozitivnih i negativnih klasa.

   S druge strane, AUC za Naivni Bajes iznosi 0.86, što je solidan rezultat, ali i dalje niži u poreenju sa BERT-om.

Iako Naivni Bajes model postiže bolju ukupnu tačnost, BERT model pokazuje veću efikasnost u klasifikaciji kada su klase uravnotežene, što se potvruje njegovom višom AUC vrednošću.

Na osnovu ovih rezultata, može se zaključiti da je BERT model superiorniji kada je reč o balansu izmeu klasa i sposobnosti razlikovanja pozitivnih i negativnih recenzija. Ipak, Naivni Bajes postiže nešto veću tačnost i lakši je za implementaciju, posebno na manjim dataset-ima ili u situacijama sa ograničenim resursima. Stoga, izbor modela zavisi od specifičnih potreba sistema – BERT je bolji za složenije zadatke sa težim balansom klasa, dok je Naivni Bajes koristan za jednostavnije primene.

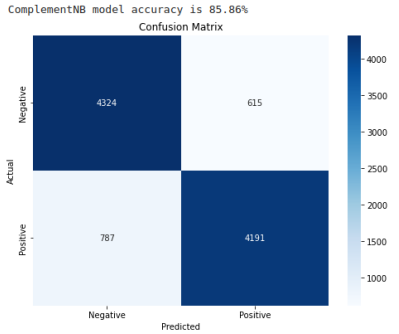

Figure 1: Matrica konfuzije - Naivni Bajes

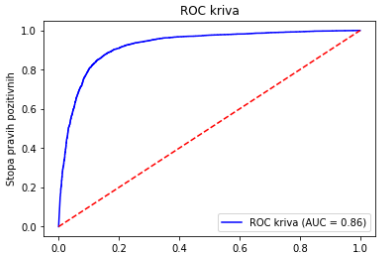

Figure 2: ROC kriva - Naivni Bajes

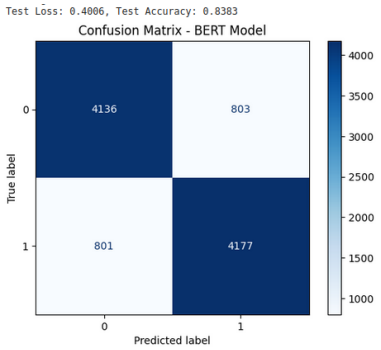

Figure 3: Matrica konfuzije - Bert

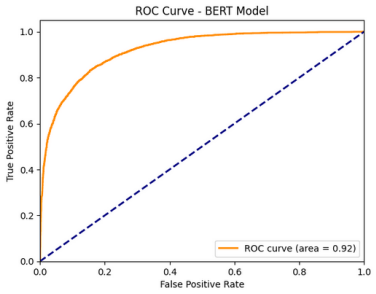

Figure 4: ROC kriva - Bert

7 Zaključak

Ovaj rad je produbio znanje autora o temi sentiment analize i pokazao moc navedenih modela. Za dalji napredak ovog projekta, bio bi značajan veći i bolji skup podataka za trening, koji je izbegnut zbog manjka resursa.

8 Literatura

1. https://www.kaggle.com/code/myr9988/sentiment-analysis-using-bert
2. https://www.upgrad.com/blog/bayes-theorem-in-machine-learning/
3. https://github.com/MATF-RI/Materijali-sa-vezbi/tree/master
4. https://www.linkedin.com/pulse/what-bert-how-trained-high-level-overview-suraj-yadav/
5. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
6. McCallum, A., Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. AAAI-98 Workshop on Learning for Text Categorization.
7. Zhang, H. (2004). The Optimality of Naive Bayes. AAAI Conference on Artificial Intelligence.