**Izvještaj o projektu analize sentimenta**

1. **Uvod**

Analiza sentimenta jedno je od područja obrade prirodnog jezika, a koristi se pri određivanju jesu li pojedini podatci pozitivni, negativni ili neutralni (Kovačević i Kovačević, 2021). Ovakva se analiza ima razna područja primjene, poput analize tržišta i korisničke podrške. Cilj je analize sentimenta općenito, pa tako i ovog projekta, prepoznavanje emocionalnog tona izraženog u tekstu.

1. **Definiranje problema**

Projekt analize sentimenta predstavlja proces izrade modela koji što preciznije klasificira rečenice prema sentimentu na temelju korpusa sastavljenog od recenzija filmova na hrvatskom jeziku. Kako bi se taj cilj postigao, korišteno je nekoliko algoritama za strojno i duboko učenje, a to su SVM, KNN, LSTM, CNN te fino podešeni BERT model.

1. **Baza podataka**

Prvi je korak u projektu bio prikupljanje trideset filmova SF, horor i triler žanra koji su izašli u zadnjih nekoliko godina kako bismo razvili korpus recenzija na hrvatskom jeziku. Pomoću alata za otkrivanje granica rečenica razdvojili smo svaku rečenicu u svoj redak te ih tako umetnuli u tablicu za daljnju analizu.

Svakoj od 3262 sakupljenih rečenica dodali smo tri ID oznake. Prvo, MovieID u rasponu od 1 do 30 (npr. MovieID=**4**). Zatim ReviewID, koji je nastao dodavanjem brojeva na ID filma, ovisno o broju recenzija za pojedini film (npr. ReviewID=4**2**), i na kraju SentenceID – na prethodnu kombinaciju MovieID i ReviewID smo također dodali brojeve od jedan pa nadalje, ovaj put ovisno o tome koliko je pojedina recenzija imala rečenica (npr. SentenceID=42**17**).

Nakon prikupljanja i klasifikacije rečenica, učitali smo tablicu na GitHub organizaciju stvorenu za svrhe ovog projekta. Kada smo dovršili bazu recenzija, napravili smo prvotnu anotaciju podskupa podataka koji se sastojao od 150 rečenica u zasebnoj tablici. To nam je omogućilo da razjasnimo dvosmislenost pojedinih rečenica i tako stvorimo jasniju sliku o samom postupku anotiranja rečenica. U ovom dijelu projekta, svih pet članova je sudjelovalo u anotaciji, koristeći 0 za pozitivan, 1 za neutralan i 2 za negativan sentiment. Pokretanjem koda za određivanje slaganja između anotatora dobili smo stopu od 0,3182.

Potom je slijedila anotacija cijele baze od 3262 rečenice. U ovom su koraku sudjelovala tri anotatora čiji je zadatak bio rečenicama dodati ocjene od 0 do 2. Sukladno naknadnim uputama, oznake sentimenta su promijenjene s obzirom na prvotnu anotaciju te tako 0 označuje pozitivan sentiment, 1 označuje negativan sentiment, a 2 označuje neutralan sentiment. Kako bismo došli do zajedničkog dogovora oko konačne oznake te svakoj rečenici pridodali jednu ocjenu, zajedničkim smo dogovorom razriješili dvosmislene slučajeve. Zatim smo napisali novi kod koji nam je pomogao da lakše i učinkovitije izračunamo sporazum među anotatorima koristeći Fleiss' kappa koji nam je na kraju kao rezultat dao 0,7869.

Sljedeći korak bio je provesti istraživačku analizu podataka. Napravili smo novu CSV datoteku koja sadrži samo dva stupca našeg korpusa s oznakama „rečenica“ i „oznaka“ i upotrijebili kod kako bi tu datoteku podijelili na dva dijela, skup za obuku, koji uključuje 75 do 80% izvorne datoteke, i skup za testiranje, koji uključuje 25 do 30% izvornih CSV datoteka. Oba skupa spremljeni su kao dvije odvojene CSV datoteke.

Nakon toga, nad testnim skupom izračunat je prosječan broj riječi po rečenici te određene najveća i najmanja rečenica. Uz to, izračunali smo broj rečenica po svakoj oznaci unutar testnog skupa. Za provedene izračune dobili smo sljedeće podatke: u testnom skupu nalazi se ukupno 653 testnih rečenica, od kojih 165 njih ima oznaku 0 (pozitivni sentiment), 58 rečenica ima oznaku 1 (negativni sentiment) i 430 rečenica ima oznaku 2 (neutralni sentiment). Kao prosječan broj riječi (duljina rečenice) dobili smo 21,33 – rečenice s najmanje broja riječi su one koje sadrže jednu riječ, dok se najduža rečenica sastoji od čak 95 riječi.

1. **Metodologija**

Nakon provođenja analize nad testnim skupom, slijedila je implementacija strojnog učenja nad našim podatcima. Napravili smo skup pod nazivom TRAIN koji se sastoji od tri skupa unutar kojih su test skupovi koje je svaka grupa kreirala zasebno (Test 1 je skup testova koje je izvorno kreirala naša grupa, Test 2 je kreirala druga grupa, itd.).

Kao metode strojnog učenja koristili smo SVM (engl. *Support Vector Machine*) i KNN (engl. *K-Nearest Neighbors*) algoritme koje smo implementirali u Pythonu te tako izračunali preciznost (engl. *precision*), odziv (engl. *recall*), točnost (engl. *accuracy*) i F1 rezultat (engl. *F1 score*) za navedene skupove. Koristili smo TF-IDF (engl. *Term Frequency – Inverse Document Frequency*) metodu za izdvajanje značajki.

Nakon strojnog učenja, krenuli smo s dubokim učenjem koje smo proveli nad podatcima koristeći LSTM (engl. *Long short-term memory*) i CNN (engl. *Convolutional Neural Networ*k) algoritme. Oba modela smo obučili na istom skupu podataka za obuku i testirali ih na testnim datotekama te koristili FastText srpsko-hrvatski model za ugradnju riječi (engl. *word embedding*).

Zadnji dio implementacije nad podatkovnim skupom bio je fino podešavanje našeg skupa podataka na BERT modelu (engl. *Bidirectional Encoder Representations from Transformers*), koristeći knjižnice s Hugging Face-a kako bi proces bio što jednostavniji. Rezultati evaluacije su nam pokazali da je fino podešeni BERT model napredovao sa svakim testom.

1. **Rezultati**

Kod strojnog učenja, za zasebno treniranje svakog testnog skupa, SVM i KNN su najbolje rezultate pokazali za Test 1 za kategorije preciznosti, odziva i točnosti, dok je Test 3 za SVM pokazao najbolji F1 rezultat.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algoritam** | **Train** | **Test 1: Group 1** | **Test 2: Group 2** | **Test 3: Group 3** |
| SVM | Train 1/2/3 | Preciznost: 0.6301 | Preciznost: 0.5402 | Preciznost: 0.5677 |
|  |  | Odziv: 0.6684 | Odziv: 0.6099 | Odziv: 0.5714 |
|  |  | F1: 0.5532 | F1: 0.4935 | F1: 0.5661 |
|  |  | Točnost: 0.6684 | Točnost: 0.6099 | Točnost: 0.5714 |

Tablica 1: Implementacija SVM algoritma nad zasebnim skupovima

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algoritam** | **Train** | **Test 1: Grupa 1** | **Test 2: Grupa 2** | **Test 3: Grupa 3** |
| KNN | Train 1/2/3 | Preciznost: 0.5941 | Preciznost: 0.4766 | Preciznost: 0.5125 |
|  |  | Odziv: 0.6684 | Odziv: 0.5964 | Odziv: 0.5126 |
|  |  | F1: 0.554 | F1: 0.4870 | F1: 0.5124 |
|  |  | Točnost: 0.6684 | Točnost: 0.5964 | Točnost: 0.5126 |

Tablica 2: Implementacija KNN algoritma nad zasebnim skupovima

Nadalje, u kategoriji TRAIN, Test-1 pokazuje najbolje rezultate u kategorijama odziv i točnost, a Train-2 pokazuje najbolje rezultate u kategoriji preciznost te za F1 rezultat. KNN je postigao dosljedno niže rezultate u usporedbi s SVM-om u svim kategorijama.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Algoritam** | **Train** | **Test 1: Grupa 1** | **Test 2: Grupa 2** | **Test 3: Grupa 3** |
| SVM | TRAIN | Preciznost: 0.6119 | Preciznost: 0.5699 | Preciznost: 0.6070 |
|  |  | Odziv: 0.6782 | Odziv: 0.6222 | Odziv: 0.6073 |
|  |  | F1: 0.6182 | F1: 0.5624 | F1: 0.6055 |
|  |  | Točnost: 0.6782 | Točnost: 0.6222 | Točnost: 0.6073 |
| KNN | TRAIN | Preciznost: 0.5066 | Preciznost: 0.5641 | Preciznost: 0.5460 |
|  |  | Odziv: 0.6398 | Odziv: 0.6117 | Odziv: 0.5454 |
|  |  | F1: 0.5233 | F1: 0.5479 | F1: 0.5451 |
|  |  | Točnost: 0.6398 | Točnost: 0.6117 | Točnost: 0.5454 |

Tablica 3: Implementacija SVM i KNN u TRAIN kategoriji

Kod dubokog učenja, LSTM je najbolje rezultate postigao kod Testa 3, dok je kod Testa 1 i Testa 2 bio znatno slabiji. CNN model je najbolje ocjene također postigao kod Testa 3 za svaku od mjerenih kategorija te je u usporedbi s LSTM algoritmom pokazao znatno bolje rezultate kod Testa 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algoritam** | **Test 1** | **Test 2** | **Test 3** |
| LSTM | Preciznost: 0.4114 | Preciznost: 0.5091 | Preciznost: 0.7101 |
|  | Odziv: 0.4006 | Odziv: 0.4911 | Odziv: 0.6905 |
|  | F1: 0.3621 | F1: 0.4757 | F1: 0.6899 |
|  | Točnost: 0.3982 | Točnost: 0.5479 | Točnost: 0.6898 |

Tablica 4: Implementacija LSTM algoritma

|  |  |  |  |
| --- | --- | --- | --- |
| **Algoritam** | **Test 1** | **Test 2** | **Test 3** |
| CNN model | Preciznost: 0.4023 | Preciznost: 0.5364 | Preciznost: 0.7824 |
|  | Odziv: 0.3577 | Odziv: 0.5324 | Odziv: 0.7776 |
|  | F1: 0.3388 | F1: 0.5165 | F1: 0.7738 |
|  | Točnost: 0.3675 | Točnost: 0.6046 | Točnost: 0.7768 |

Tablica 5: Implementacija CNN algoritma

Pri usporedbi rezultata sa strojnim učenjem, SVM TRAIN je ponudio najbolju F1 ocjenu i točnost za Test 1 i Test 2, pokazujući najdosljedniju izvedbu u cijeloj testnoj skupini. KNN algoritam je pokazao umjerene rezultate tijekom testova, ali nikada nije postigao najbolje rezultate. Test 3 je pokazao najbolje rezultate i za LSTM i za CNN algoritme unutar svih kategorija.

Kod fino podešenog BERT modela, Test 3 je pokazao najbolju ukupnu izvedbu, postigavši ​​najbolje rezultate s preciznošću od 0,8264, odzivom od 0,8274, F1 rezultatom od 0,8263 i točnošću od 0,8272. Nasuprot tome, Test 1 i Test 2 pokazali su umjerenu izvedbu, pri čemu je Test 1 postigao niže rezultate u svim kategorijama, dok je Test 2 pokazao je poboljšani odziv, ali sveukupno manje uravnotežene rezultate. Ovi rezultati su nam sugerirali da se zbog razlika u distribuciji podataka model bolje generalizira pod uvjetima predstavljenim u Testu 3.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algoritam** | **Test 1** | **Test 2** | **Test 3** |
| BERT | Preciznost: 0.4445 | Preciznost: 0.5841 | Preciznost: 0.8264 |
|  | Odziv: 0.4440 | Odziv: 0.5999 | Odziv: 0.8274 |
|  | F1: 0.4442 | F1: 0.5412 | F1: 0.8263 |
|  | Točnost: 0.6141 | Točnost: 0.5655 | Točnost: 0.8272 |

Tablica 6: Rezultati fino podešenog BERT modela

Analizom rezultata algoritama strojnog i dubokog učenja te fino podešenog BERT modela na tri testna skupa podataka pokazalo se da su SVM i KNN najbolje ocjene ostvarili nad prvim testnim skupom, dok su algoritmi dubokog učenja i fino podešeni BERT model ostvarili najbolje rezultate nad trećim testnim skupom.

1. **Zaključak**

Pri provedbi procesa analize sentimenta nad filmskim recenzijama na hrvatskom jeziku u svrhu ovog projekta uspješno su primijenjene različite metode strojnog i dubokog učenja.

Usporedba modela pokazala je da u kontekstu strojnog učenja SVM daje stabilnije rezultate od KNN-a, iako je za oba modela najuspješniji bio skup Test 1. Kada je riječ o algoritmima dubokog učenja, oba su modela, LSTM i CNN, značajno bolje rezultate postigli na Testu 3. Razlog tome može biti bolja distribucija oznaka sentimenata te distribucija sličnija podatcima u skupu za treniranje ili pak složeniji jezični obrasci kojima je sentiment izražen, što je također slučaj u kojemu duboke neuronske mreže mogu dati bolje rezultate. Za kraj, najbolje rezultate postigao je fino podešeni BERT model, osobito na skupu Test 3, što pokazuje kako suvremeni modeli zaista jesu izrazito učinkoviti pri analizi sentimenta.

1. **Nedostatci i ograničenja**

Najveće je ograničenje pri provedbi ovog projekta ograničen broj resursa na hrvatskom jeziku. Od samih recenzija filmova, do FastText modela za ugradnju riječi, rad na podatcima na hrvatskom jeziku primjetno je teži od onoga na, primjerice, podatcima na engleskom jeziku.

Osim toga, poneke je poteškoće stvarala neravnomjerna raspoređenost oznaka sentimenta (0, 1 i 2) u samim skupovima podataka, posebice zato što je neutralnih rečenica bilo znatno više od ostalih. Uz to, povezane su i poteškoće uzrokovane neujednačenostima između skupova podataka među različitim grupama. Neuravnoteženost u ova dva segmenta tako je utjecala na neke od rezultata.

Za kraj, važno je napomenuti kako se korpus na kojemu se temelji ovaj projekt sastavljen od tek nešto više od 3000 rečenica. Iako relativno velik broj, može biti premalen za treniranje složenijih modela poput BERT-a. Manji broj primjera općenito otežava generalizaciju modela i može dovesti do poteškoća u kasnijem prepoznavanju sentimenta.

1. **Reference**

Kovačević, A. i Kovačević, Ž. (2021). ALATI ZA ANALIZU SENTIMENTA. *Polytechnic and design, 9* (3), 167-174. <https://doi.org/10.19279/TVZ.PD.2021-9-3-02>. Pristupljeno 5. lipnja 2025.

1. **Poveznice**

Demo: <https://huggingface.co/spaces/HighFive-OPJ/demo>

Kod:

1. Završna anotacija: <https://huggingface.co/datasets/HighFive-OPJ/OPJ-Corpus/blob/main/Final%20Annotation.py>
2. Istraživačka analiza podataka: <https://huggingface.co/datasets/HighFive-OPJ/Exploratory_data_analysis/blob/main/testing.py>

Baza podataka: <https://huggingface.co/datasets/HighFive-OPJ/OPJ-Corpus/blob/main/Movies%20-%20Movies_Sentences.csv>

Modeli:

1. Strojno učenje: <https://huggingface.co/datasets/HighFive-OPJ/Implementation_1-Machine_learning/tree/main>
2. Duboko učenje: <https://huggingface.co/datasets/HighFive-OPJ/Deep_Learning/tree/main>
3. LLM: <https://huggingface.co/datasets/HighFive-OPJ/Deep_Learning/tree/main>