SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

ZAVRŠNI RAD br. 660

# Sustav za označavanje tekstualnog sadržaja navođeno aktivnim učenjem

Franjo Mindek

Zagreb, svibanj 2022.

**SVEUČILIŠTE U ZAGREBU**
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

Zagreb, 11. ožujka 2022.

# ZAVRŠNI ZADATAK br. 660

| | |
|---|---|
| Pristupnik: | **Franjo Mindek (0036523095)** |
| Studij: | Elektrotehnika i informacijska tehnologija i Računarstvo |
| Modul: | Računarstvo |
| Mentor: | doc. dr. sc. Mario Brčić |

Zadatak: **Sustav za označavanje tekstualnog sadržaja navođeno aktivnim učenjem**

Opis zadatka:

Strojno učenje je veoma uspješna, štoviše dominantna, paradigma za obradu teksta. No, efektivnost tih metoda znatno ovisi o kvaliteti ulaznih podataka. Za postojeća je rješenja uobičajeno da se utreniravaju nad podatcima koji su indiskriminativno skupljeni s interneta i labelirani za neku općenitu ili specifičnu svrhu. To rezultira potrebom za velikim skupovima da bi se ostvarila željena performansa. Za uspjeh samog model u određenoj zadaći je bitno imati kvalitetan skup podataka koji učinkovito pokriva prostor mogućnosti tako da model može pouzdano i precizno vršiti svoju namjenu. Aktivno učenje povećava uzorkovnu učinkovitost postupka utreniravanja modela jer se pronalaze čim informativniji uzorci. Time se smanjuju veličina, trošak i trajanje postupka označavanja skupa podataka. U ovom radu treba, koristeći dostupne alate, implementirati sustav za označavanje tekstualnog sadržaja koje je navođeno aktivnim učenjem. Označivačima podataka treba biti izloženo web sučelje koje ih vodi kroz postupak te im zadaje uzorke koji su visoko informativni za odabrani model.

Rok za predaju rada: 10. lipnja 2022.

# SADRŽAJ

# 1. Uvod

Zbog potrebe rješavanja UI[1]-potpunih problema sve se više susrećemo s raznim primjenama strojnog učenja. Ipak, zbog ulazno-izlazne povezanosti podataka i rezultata modela umjetne inteligencije nužno je imati kvalitetan skup podataka.

Većina problema umjetne inteligencije rješava se treniranjem modela nad indiskriminativno skupljenim podatcima koji su često označeni za preopćenitu ili nama krivo specifičnu svrhu. Rezultat toga je potreba za ogromnim skupovima podataka varirajuće kvalitete kako bi se približili željenoj efikasnošću modela. Uspjeh samog modela zasniva se na kvalitetnom skupu podataka koji efikasno razapinje prostor problema, što modelu omogućuje pouzdane i precizne odgovore.

Cilj aktivnog učenja je povećati uzorkovanu učinkovitost postupka utreniravanja modela minimizirajući broj podataka i maksimizirajući informacijsku dobit. Tim postupkom smanjujemo veličinu, trošak i vrijeme stvaranja skupa podataka te istovremeno povećavamo preciznost modela.

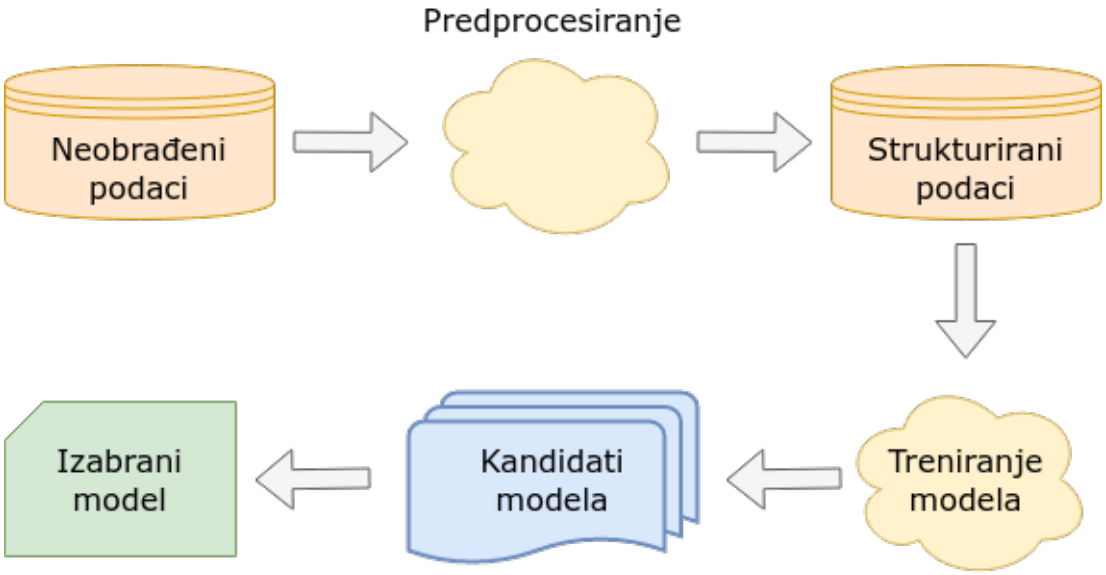
U ovom završnom radu koristeći javno dostupne alate, implementirati ćemo sustav za označavanje tekstualnog sadržaja navođeno aktivnim učenjem. Označivačima podataka bit će ponuđeno web sučelje koje će ih uvesti u postupak označavanje te će im zadavati uzorke visoke informacijske dobiti za odabrani model.

[1] UI - umjetna inteligencija

# 2. Označavanje podataka

Označavanje podataka je jedan od početnih koraka u procesu stvaranju modela strojnog učenja te ga svrstavamo u razinu predprocesiranja. Definiramo ga kao proces prepoznavanja neobrađenih podataka (slika, tekst, video...) koji se onda pomoću jedne ili više oznaka opisuju u kontekstu modela, što dozvoljava modelu strojnog učenja da može stvarati točna predviđanja.

Označavanje podataka često je najsporiji i najvažniji dio modeliranja modela strojnog učenja. U prosjeku više od 80% ukupnog vremena modeliranja UI potroši se na rad s podacima [4]. Vremenska neefikasnost označavanja podataka dovela je do potrebe razvoja modela UI koji pomažu u procesu označavanja podataka. Jedan primjer takve UI je aktivno učenje čiji ćemo utjecaj na označavanje podataka obraditi kasnije u radu.




**Slika 2.1:** Pojednostavljeni dijagram stvaranja modela strojnog učenja.

## 2.1. Strojno učenje i podatci

Strojno učenje je proces koji započinje obradom podataka, a završava stvaranjem modela strojnog učenja.

Samo strojno učenje definiramo kao programiranje računala na način da optimiziraju neki kriterij uspješnosti temeljem podatkovnih primjera ili prethodnog iskustva. Raspolažemo modelom koji je definiran do na neke parametre, a učenje se svodi na izvođenje algoritma koji optimizira parametre modela na temelju podataka ili prethodnog iskustva. [2]

Iz te definicije možemo uočiti povezanost između umjetne inteligencije i skupa podataka. Odgovor ili rješenje problema kojeg zahtjevamo od umjetne inteligencije je izlaz, ali taj izlaz možemo dobiti samo temeljem ulaza. Ako smo odabrali dobar skup podataka doći ćemo do traženog rješenja problema, no ako je naš skup podataka na bilo koji način iskrivljen ta iskrivljenost će se propagirati i na izlaze, što dovodi do krivih spoznaja.

Ukratko, ako naši podaci ne odgovaraju na pitanje problema onda neće ni model stojnog učenja.

### 2.1.1. Strukturiranost skupa podataka

Sama definicija strojnog učenja osim podataka spomenula je i postojanost prethodnog iskustva. Općenito strojno učenje možemo podijeliti u 3 paradigme koje se između ostaloga razlikuju po izvoru informacija nad kojima se treniraju. Ovisno o izabranoj paradigmi i samoj namjeni modela drukčije ćemo pristupati procesu strukturiranja podataka.

Paradigme strojnog učenja su:

1. Nadzirano učenje - definirano je korištenjem označenih skupova podataka. Ti skupovi podataka treniraju i „nadziru“ algoritam u procesu klasifikacije[1] ili regresije[2]. To jest, na temelju poznatih podataka predviđamo izlaz/rješenje problema. Koristeći označene podatke model može mjeriti svoju preciznost .

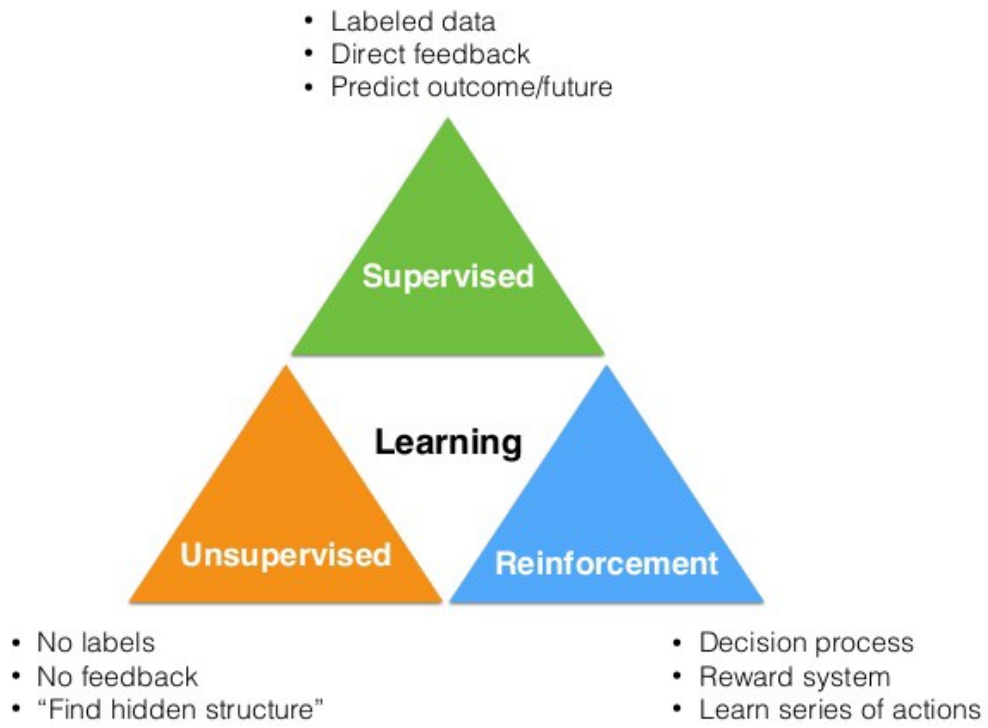
2. Nenadzirano učenje - alalizira i grupira neoznačene skupove podataka. Cilj je da algoritam otkrije skrivene uzorke ponavljanja u podacima bez potrebe ljudskog nadzora. Ima upotrebu u grupiranju, otkrivanju stršećih/novih vrijednosti, otkrivanju povezanosti i smanjenju dimenzionalnosti.

[1] Klasifikacija - problem pridjeljivanja kategorije podacima

[2] Regresija - problem otkrivanja veza zavisnih i nezavisnih varijabli

3. Podržano/ojačano učenje - učenje optimalne strategije na temelju pokušaja s odgođenom nagradom. Definirano je agentom, okolinom i akcijama koje agent može raditi.

Nadzirano učenje je jedina paradigma za čije je modeliranje nužan označeni skup podataka. Zbog tog svojstva paradigme, spominjanje strojnog učenja u tekstu većinski će se odnositi na nadzirano učenje.




**Slika 2.2:** Vizualna reprezentacija 3 paradigme strojnog učenja [3]

## 2.2. Karakteristike dobrog skupa podataka

Karakteristike modela strojnog učenja, a time i skupova podataka nad kojima se trenira uvelike ovise o namjeni umjetne inteligencije. Svejedno možemo opisati određena svojstva koja su uvijek poželjna unutar skupa podataka neovisno o njihovoj primjeni.

1. Potpunost - pretpostavka je da u skupu podataka ne postoje prazne „ćelije". Svaki uzorak unutar skupa podataka mora sadržavati informacije o svim traženim oznakama. Neispravne uzorke treba nadopuniti ili izbrisati.

2. Objedinjenost - potrebno je pokriti cijelo područje problema koje se rješava. Primjer, model koji predviđa emocije mora preko slika lica osobe mora u svom skupu podataka sadržavati slike lica ljudi iz raznih kultura i različitih fizičkih karakteristika. Kako različite kulture drukčije iskazuju emocije, a i postoje fizičke razlike u strukturama lica, bez dovoljne objedinjenosti model ne bi zadovoljavao svojstvo generalizacije.

3. Relevantnost - podatci ne bi trebali pokrivati područje izvan problema. Kod skupova podataka sakupljenih s vanjskih izvora potrebno je filtrirati samo nama korisne informacije. Često je posljedica nerazumijevanja svojstva objedinjenosti podataka.

4. Ortogonalnost - prostor problema koji opisujemo trebamo opisati sa što manje ponavljanja između uzoraka. [1]

5. Konzistentnost - kriterij opisivanja uzoraka skupa podataka mora ostati konzistentan. S potrebom ogromnih skupova podataka od više desetaka tisuća do nekoliko milijuna uzoraka rijetko će sve uzorke opisati jedna osoba. U takvim slučajevima s više označivača podatci svejedno moraju ostati konzistentno opisani.

6. Ispravnost - po principu „*garbage in, garbage out*“, neispravni podatci skupa podataka dovode do neispravnih rezultata modela.

7. Suvremenost - podatci moraju opisivati trenutno stanje problema koji se opisuje. Primjer, model koji predviđa rezultat sportskog natjecanja treba biti upoznat s trenutnim stanjem suparničkih timova. U protivnom, on ne ispunjava svoju svrhu.

Upravo zbog potreba zadovoljavanja što više korisnih svojstava skupova podataka dovelo je do razvoja mnogih internih i javno dostupnih sustava označavanja podataka te razvoja UI sa svrhom ubrzanja i povećanja preciznosti tog procesa. [4]

## 2.3. Alati za označavanje podataka

### 2.3.1. Label Studio

### 2.3.2. Doccano

# 3. Aktivno Učenje

# 4. Problem

# 5. Zaključak

Zaključak.

# LITERATURA

[1] Salem Alelyani, Huan Liu, i Lei Wang. The effect of the characteristics of the dataset on the selection stability. U *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, stranice 970–977. IEEE, 2011.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[3] Dhairya Parikh. Learning paradigms in machine learning. URL `https://medium.datadriveninvestor.com/learning-paradigms-in-machine-learning-146ebf8b5943`.

[4] Cognilytica Research. Data Engineering, Preparation, and Labeling for AI 2019. Technical report, Cognilytica Research, 01 2019.

**Sustav za označavanje tekstualnog sadržaja navođeno aktivnim učenjem**

**Sažetak**

Sažetak na hrvatskom jeziku.

**Ključne riječi:** Ključne riječi, odvojene zarezima.

**Title**

**Abstract**

Abstract.

**Keywords:** Keywords.