SVEUČILIŠTE U RIJECI
FAKULTET INFORMATIKE I DIGITALNIH TEHNOLOGIJA
Radmile Matejčić 2, Rijeka

Akademska godina 2023./2024.

OSNOVNI PODACI O PREDMETU		
Naziv predmeta	**Metode obrade prirodnog jezika**	
Studijski program	Sveučilišni diplomski studij Informatika	
Status predmeta	Izboran za modul IIS	
Semestar	3.	
Bodovna vrijednost i nastavno opterećenje	ECTS koeficijent opterećenosti studenata	6
	Broj sati (P+V+S)	30+30+0
Nositelj predmeta	**Prof. dr. sc. Sanda Martinčić-Ipšić**	
E-mail	smarti [at] uniri.hr	
Ured	O-409	
Vrijeme konzultacija	Utorkom 12-13 uz prethodni dogovor	
Asistent	**Karlo Babić, mag. inf.**	
E-mail	karlo.babic [at] inf.uniri.hr	
Ured	O-419	
Vrijeme konzultacija	Petkom 15:30-16:30 uz prethodni dogovor	

DETALJNI OPIS PREDMETA

Ciljevi predmeta

Cilj predmeta je primijeniti postupke strojnog i dubokog učenja za nestrukturirane tekstualne podatke, te riješiti standardne zadatke računalne analize prirodnog jezika poput: klasifikacije tekstova, pretraživanje informacija u nestrukturiranim podacima, automatskog sažimanja dokumenta, ekstrakcije informacija (npr. entiteta i ključnih riječi), izlučivanje tema iz tekstova, razvoj sustava za praćenje mišljenja u komentarima, otkrivanje toksičnog diskursa ili emocija iz korisničkih komentara, otkrivanje lažnih vijesti, razvoj dijaloških sustava, generiranja tekstova, analiza semantike, parafraziranja i razumijevanja prirodnog jezika te drugih zadataka.

Uvjeti za upis predmeta

Oslušan predmet Strojno i duboko učenje

Očekivani ishodi učenja za predmet

Očekuje se da će nakon uspješno ispunjenih obveza na predmetu student moći:

I1. Vrednovati i kritički procijeniti principe, metode i algoritme računalne obrade tekstova za rješavanje standardnih problema (zadataka) računalne analize prirodnog jezika.

I2. Dizajnirati i razviti odgovarajući model strojnog i/ili dubokog učenju u kombinaciji s klasičnim simboličkim pristupima za zadani zadatak iz područja obrade prirodnog jezika.

I3. Vrednovati metode strojnog i dubokog učenja za postavljeni zadatak (problem) iz područja obrade prirodnog jezika.

I4. Procijeniti primjenjivost elemente arhitekture duboke mreže ili druge duboke strukture na postavljeni problem iz područja obrade prirodnog jezika s obzirom na dostupne podatke, postavljene arhitekture te procesorske kapacitete.

I5. Procijeniti razumljivost dobivenog modela s obzirom na provedenu evaluaciju problema oskudnosti i neuravnoteženosti podataka.

I6. Implementirati sustav za obradu prirodnog jezika za specifični problem (zadatak) .

I7. Osmisliti, planirati i pripremiti tekstualni skup podataka iz vanjskih nestrukturiranih izvora pa i društvenih mreža za specifični problem (zadatak) u praktičnoj primjeni uz uvažavanje pravnih i etičkih aspekata.

Sadržaj predmeta

Na predmetu se obrađuju sljedeći sadržaji:

- Problemi obrade prirodnog jezika i teksta uključujući potrebne statističke, lingvističke i računalne osnove za razvoj metoda računalne analize prirodnog jezika. I1
- Korpusi, prethodna obrada teksta: korjenovanje, lematizacija, zaustavne riječi, tokenizacija. Jezični resursi.I7
- Uvod u duboko učenje za tekstualne podatke. Logistička regresija. Funkcije gubitka. I4
- Reprezentacije teksta: model rijetke vektorske reprezentacije (TF-IDF), model neuređene vreće riječi (BOW), modeli gustih reprezentacija s vektorima niske dimenzionalnosti (embedding). Neprekidna vreća riječi (Continuous bag-of-words) i Skip-gram. I2, I4, I7
- Statistički jezični modeli. Neuralni jezični modeli. Veliki jezični modeli. I2, I5
- Pretraživanja informacija, Modeli sličnosti, dohvaćanje i rangiranje dokumenata. Semantička reprezentacija riječi, rečenica i tekstova. Semantička sličnost. Metode evaluacije. I1, I2, I3
- Metode dubinske analize teksta. Klasifikacija teksta. Grupiranje teksta. Principi evaluacije. I2
- Zadaci klasifikacije teksta: otkrivanje mišljenja, stavova, emocija, toksičnih komentara, lažnih vijesti i drugih. Problemi klasifikacije s većim brojem klasa (multiclass) i labela (multilabel). Interpretacija dobivenih modela. Rad s neuravnoteženim klasama. I1, I2, I3, I6
- Modeli za duboko učenje: Duboka unaprijedna mreža (Deep feed-forward network). Povratne neuronske mreže (Rekurentne neuronske mreže). Dvosmjerne povratne mreže. Ćelija s dugoročnom memorijom (LSTM), Upravljačka rekurentna jedinica (GRU).I2, I4
- Modeliranje dugih sljedova. Označavanje vrste riječi i imenovanje entiteta. I1, I2, I4, I6
- Mehanizmi pažnje (attention). Transformeri. Učenje principima transfera zadataka (transfer learning), principi učenja s jednim (one-shot learning) ili nekoliko primjera (few-shoots learning). I2, I4
- Primjeri problema/zadataka: Ekstrakcija informacija. Ekstrakcija ključnih riječi. Ekstrakcija relacija. Principi evaluacije ekstrakcije. Ekstraktivno i apstraktivno sažimanje teksta, generiranje teksta. Principi evaluacije generiranog teksta. Principi evaluacije. I1, I3, I4, I6
- Automatsko otkrivanje tema u tekstu. Latentne reprezentacije teksta. Principi evaluacije latentnih modela. I1, I3, I6
- Koherentnost teksta, razrješavanje koreferenciranja, parafraziranje. Određivanje i provjeravanje točnosti činjenica. I1, I3, I6
- Semantika i razumijevanje jezika. I1, I6
- Trendovi u računalnoj analizi prirodnog jezika i veliki (fundamentalni) jezični modeli (foundation models). Pravni i etički aspekti. I7

Način izvođenja nastave	☒ predavanja	☒ samostalni zadaci
	☒ seminari i radionice	☐ multimedija i mreža

	☒ vježbe	☐ laboratorij
	☒ obrazovanje na daljinu	☒ mentorski rad
	☒ terenska nastava	☐ ostalo _____
Komentari	Nastava će se izvoditi kombinirajući rad u učionici (predavanja i vježbe), samostalni rad izvan učionice, uz povremene seminare i radionice povezane s industrijom uz korištenje sustava za e-učenje.	

Obavezna literatura (u trenutku prijave prijedloga studijskog programa)

1. Dan Jurafsky, James H. Martin, Speech and Language Processing, Prentice Hall (3rd edition), 2023. https://web.stanford.edu/~jurafsky/slp3/
2. Jacob Eisenstein, Introduction to Natural Language Processing, MIT Press, 2019. https://mitpress.mit.edu/books/introduction-natural-language-processing
3. Yoav Goldberg, Neural Network Methods in Natural Language Processing (Synthesis Lectures on Human Language Technologies), Morgan & Claypool Publishers, 2017. https://www.morganclaypool.com/doi/10.2200/S00762ED1V01Y201703HLT037
4. C., Manning, H. Shütze: Foundations of Statistical Natural Language Processing, MIT Press, 1999. http://nlp.stanford.edu/fsnlp/

Dopunska literatura (u trenutku prijave prijedloga studijskog programa)

1. François Chollet, Deep Learning with Python, Manning Pub. 2017. https://www.manning.com/books/deep-learning-with-python
2. S. Bird, E. Klein, E. Loper: Natural Language Processing with Python, O'Riley, 2009. http://nltk.org/book/
3. Bing Liu, Web Data Mining, Springer, 2011. http://www.cs.uic.edu/~liub/WebMiningBook.html
4. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. http://nlp.stanford.edu/IR-book/information-retrieval-book.html
5. Lewis Tunstall, Leandro von Werra, Thomas Wolf, Natural Language Processing with Transformers, O'Reilly Media, Inc. 2022

Načini praćenja kvalitete koji osiguravaju stjecanje izlaznih znanja, vještina i kompetencija

Predviđa se periodičko provođenje evaluacije s ciljem osiguranja i kontinuiranog unapređenja kvalitete nastave i studijskog programa (u okviru aktivnosti Odbora za upravljanje i unapređenje kvalitete Fakulteta informatike i digitalnih tehnologija). U zadnjem tjednu nastave provodit će se anonimna evaluacija kvalitete održane nastave od strane studenata. Provest će se i analiza uspješnosti studenata na predmetu (postotak studenata koji su položili predmet i prosjek njihovih ocjena).

Jezik izvođenja nastave	Hrvatski jezik
Mogućnost izvođenja na stranom jeziku	NE

OBVEZE, PRAĆENJE RADA I VREDNOVANJE STUDENATA

Konstruktivno povezivanje

VRSTA AKTIVNOSTI	ECTS	ECTS - PRAKTIČNI RAD	ISHODI UČENJA	SPECIFIČNA AKTIVNOST	METODA PROCJENJIVANJA	BODOVI MAX.
Pohađanje nastave i aktivnosti u nastavi	2	1	I1-I7	Prisutnost studenata i odgovaranje na pitanja nastavnika	Evidencija prisustva	0
Aktivnost na vježbama	1	0.75	I2-I7	Zadaci na vježbama, periodički testovi, domaće zadaće	Periodički zadaci i domaće zadaće	30
Seminarski rad	2	2	I2-I7	Praktični projektni rad	Priprema podataka (10) Rješenje NLP problema (15) Evaluacija rezultata (5) Prezentacija (5) Tehnička dokumentacija (5)	40
Završni ispit	1		I1-I7	Teorija		30
UKUPNO	**6**	**3.75**				**100**

Obveze i vrednovanje studenata – puna nastavna satnica

1. Pohađanje nastave i aktivnosti u nastavi

Nastava se odvija prema mješovitom modelu u kombinaciji klasične nastave u učionici i *online* nastave uz pomoć sustava za e-učenje prema rasporedu koji je prikazan je tablicom u nastavku. Studenti su dužni koristiti sustav za e-učenje Merlin (https://moodle.srce.hr/) gdje će se objavljivati informacije o predmetu, materijali za učenje, zadaci za vježbu, zadaci za domaće zadaće te obavijesti vezane za izvođenje nastave (putem foruma Obavjesti).

Studenti koji studiraju u punoj nastavnoj satnici dužni su redovito pohađati nastavu, aktivno sudjelovati tijekom nastave te izvršavati aktivnosti predmeta u okviru sustava Merlin koje će nastavnici najavljivati putem foruma.

Od studenta se očekuje redovito pohađanje nastave, sudjelovanje u svim aktivnostima predmeta te praćenje obavjesti vezanih uz nastavu u sustavu za e-učenje.

2. Aktivnost na vježbama

Student je obvezan izraditi zadatke tijekom semestra na vježbama te domaće zadaće za kontinuirano praćenje studentskog rada.

3. Seminarski rad

Praktična primjena usvojenih znanja obuhvaća razradu i izradu odabranog samostalnog projektnog rada koji uključuje rješavanje nekog od standardnih zadataka računalne analize prirodnog jezika poput: klasifikacije tekstova, pretraživanje informacija u nestrukturiranim podacima, automatskog sažimanja dokumenta, ekstrakcije informacija (npr. entiteta i ključnih riječi), izlučivanje tema iz tekstova, razvoj sustava za praćenje mišljenja u komentarima, otkrivanje toksičnog diskursa ili emocija iz korisničkih komentara, otkrivanje lažnih vijesti, razvoj dijaloških sustava, generiranja tekstova, analiza semantike, parafraziranja i razumijevanja prirodnog jezika te drugih zadataka.

Student je dužan izraditi i predstaviti samostalni praktični projektni rad koji obuhvaća prezentaciju i tehničku dokumentaciju.

4. Završni ispit

Teorijski dio predmeta se polaže na završnom ispitu s najmanje postignutih 50% bodova.

Obveze i vrednovanje studenata – prilagođena nastavna satnica

1. Pohađanje nastave i aktivnosti u nastavi

Nastava se odvija prema mješovitom modelu u kombinaciji klasične nastave u učionici i *online* nastave uz pomoć sustava za e-učenje prema rasporedu koji je prikazan je tablicom u nastavku. Studenti su dužni koristiti sustav za e-učenje Merlin (https://moodle.srce.hr/) gdje će se objavljivati informacije o predmetu, materijali za učenje, zadaci za vježbu, zadaci za domaće zadaće te obavijesti vezane za izvođenje nastave (putem foruma Obavijesti).

Studenti koji studiraju u sklopu prilagođene nastavne satnice mogu izostati s najviše 50% sati nastave (predavanja i vježbi), a dužni su aktivno sudjelovati tijekom nastave (u učionici ili *online*) te izvršavati aktivnosti predmeta u okviru sustava Merlin koje će nastavnici najavljivati putem foruma.

2. Obveze i aktivnosti vrednovanja

Obveze i vrednovanje studenata koji studiraju u sklopu prilagođene nastavne satnice, jednake su onima studenata koji studiraju u sklopu pune nastavne satnice.

Ocjenjivanje

Kontinuiranim radom tijekom semestra na prethodno opisani način studenti mogu ostvariti najviše 70 ocjenskih bodova, a da bi mogli pristupiti ispitu predmeta moraju ostvarili 50% i više bodova (minimalno 35).

Ispit nosi udio od maksimalno 30 ocjenskih bodova, a smatra se položenim samo ako na njemu student postigne minimalno 50%-ni uspjeh (ispitni prag je 50% uspješno riješenih zadataka).

Ako je ispit prolazan, skupljeni bodovi će se pribrojati prethodnima i prema ukupnom rezultatu formirat će se pripadajuća ocjena. U suprotnom, student ima pravo pristupa ispitu još 2 puta (ukupno do 3 puta tijekom akademske godine).

Konačna ocjena ostvarenosti ishoda učenja na predmetu

Konačna ocjena ostvarenosti ishoda učenja na predmetu je zbroj ocjenskih bodova postignutih u kontinuiranom praćenju i vrednovanju i ocjenskih bodova postignutih na ispitu, a donosi se na sljedeći način:

A – 90% - 100%	(ekvivalent: izvrstan 5, slovna ocjena A)
B – 75% - 89,9%	(ekvivalent: vrlo dobar 4, slovna ocjena B)
C – 60% - 74,9%	(ekvivalent: dobar 3, slovna ocjena C)
D – 50% - 59,9%	(ekvivalent: dovoljan 2, slovna ocjena D)
F – 0% - 49,9%	(ekvivalent: nedovoljan 1, slovna ocjena F)

Ispitni termini

06.02.2024.
20.02.2024.
18.03.2024.
03.09.2024.

SATNICA IZVOĐENJA NASTAVE – zimski (III.) semestar akademske godine 2023./2024.

Nastava će se na predmetu odvijati u zimskom semestru prema sljedećem rasporedu:

predavanja: utorkom u 358 od 14-16 sati

vježbe: petkom u 365 od 14-16 sati

Tj.	Datum	Vrijeme	Prostor*	Tema	Nastava	Izvođač
1	03.10.23.	14-16	358	Uvod u predmet. Uvod u NLP.	P1	SMI
1	06.10.23.	14-16	365	Priprema podataka. Korpusi. Pravni i etički aspekti	P2	SMI
2	10.10.23.	14-16	358	Uvod u strojno i duboko učenje za NLP.	P3	SMI
2	13.10.23.	14-16	365	Web scraping (kreiranje dataseta)	V1	KB
3	17.10.23.	14-16	358	Modeli reprezentacije teksta. Vreća riječi. Embedings.	P4	SMI
3	20.10.23.	14-16	365	Čišćenje teksta, tokeniziranje, statistika	V2	KB
4	24.10.23.	14-16	358	Statistični jezični modeli. Veliki jezični modeli.	P5	SMI
4	27.10.23.	14-16	365	Reprezentiranje teksta (vreća riječi, embeddings)	V3	KB
5	31.10.23.	14-16	online	Pretraživanja informacija, Sličnost. Evaluacija.	P6	SMI
5	03.11.23.	14-16	365	TFIDF, information retrieval	V4	KB
6	07.11.23.	14-16	358	Metode klasifikacije teksta.. Principi evaluacije.	P7	SMI
6	10.11.23.	14-16	365	Klasifikacija teksta 1 (klasični machine learning)	V5	KB
7	14.11.23.	14-16	358	Zadaci klasifikacije teksta: otkrivanje mišljenja, stavova, emocija, toksičnih komentara, lažnih vijesti.	P8	SMI
7	17.11.23.	14-16	365	Klasifikacija teksta 2 (klasični machine learning)	V6	KB
8	21.11.23.	14-16	358	Problemi klasifikacije s većim brojem klasa (multiclass) i labela (multilabel). Interpretacija dobivenih modela. Rad s neuravnoteženim klasama.	P9	SMI
8	24.11.23.	14-16	365	Evaluacije Rok: prijava teme seminar	V7	KB
9	28.11.23.	14-16	358	Modeli za duboko učenje: Deep feed-forward network. RNN, LSTM, Bi LSTM, GRU.	P10	SMI
9	01.12.23.	14-16	365	Klasifikacija teksta 1 (deep learning: feed-forward neural network)	V8	KB
10	05.12.23.	14-16	358	Modeliranje dugih sljedova. Označavanje vrste riječi i imenovanje entiteta.	P11	SMI
10	08.12.23.	14-16	365	Klasifikacija teksta 2 (deep learning: recurrent neural network)	V9	KB
11	12.12.23.	14-16	358	Mehanizmi pažnje (attention). Transformeri. Učenje principima transfera zadataka (transfer learning), principi učenja s jednim (one-shot learning) ili nekoliko primjera (few-shoots learning).	P12	SMI
11	15.12.23.	14-16	365	Transformeri	V10	KB
12	19.12.23.	14-16	358	Ekstrakcija informacija. Ekstrakcija ključnih riječi. Ekstrakcija relacija. Principi evaluacije ekstrakcije.	P13	SMI
12	22.12.23.	14-16	365	Ekstrakcija ključnih riječi	V11	KB
13	09.01.24	14-16	358	Ekstraktivno i apstraktivno sažimanje teksta, generiranje teksta. Principi evaluacije generiranog teksta.	P14	SMI

13	12.01.24	14-16	365	Generativni modeli 1	V12	KB
14	16.01.24	14-16	358	Automatsko otkrivanje tema u tekstu. Latentne reprezentacije teksta. Principi evaluacije latentnih modela.	P15	SMI
14	19.01.24	14-16	365	Generativni modeli 2 Rok: za predaju seminar	V13	KB
15	23.01.24	14-16	358	Obrane seminara	P16	
15	26.01.24	16-18	358	Obrane seminara	V14	

*Napomena: upisati broj prostorije ili *online*

P – predavanja
V – vježbe
06.02.2024.
20.02.2024.
18.03.2024.
03.09.2024.