



Finalna naloga

Verzija 1

Dokument:64764

Kazalo:

1	UVOD	1
1.1	Namen in ozadje projekta	1
1.2	Vloga umetne inteligence	1
2	OPIS IZZIVA	2
2.1	Kaj se od tekmovalcev pričakuje	2
2.2	Cilji izziva.....	2
2.2.1	Zajem podatkov s spletnih strani VIZ	2
2.2.2	Zajem podatkov iz spletnih medijev	2
2.2.3	Analiza sentimenta	2
2.2.4	Možne smeri reševanja in omejitve.....	3
3	OPIS IZZIVA	4
3.1	Skupne metodološke osnove.....	4
3.1.1	Metodologija spletnega zajema (scraping, API, RSS)	4
3.1.2	Normalizacija in čiščenje podatkov	5
3.1.3	Priprava strukturirane zbirke objav.....	5
3.2	Modul 1: Zajem podatkov s spletnih strani VIZ	6
3.2.1	Opis in namen modula	6
3.2.2	Posebnosti spletnih strani VIZ.....	6
3.2.3	Zajem statičnih in dinamičnih vsebin	6
3.2.4	Prepoznavanje ključnih informacij	7
3.2.5	Izvoz strukturiranih podatkov v skupno zbirko	7
3.3	Modul 2: Zajem podatkov iz spletnih medijev	7
3.3.1	Opis in namen modula	7
3.3.2	Identifikacija medijskih virov	7
3.3.3	Iskanje po ključnih besedah	8
3.3.4	Prepoznavanje entitet v člankih	8
3.3.5	Vzpostavitev zbirke člankov za nadaljnjo sentimentno analizo.....	8
3.4	Modul 3: Analiza sentimenta.....	9
3.4.1	Opis in namen modula	9
3.4.2	Uporabljeni model	9
3.4.3	Postopek analize	9
3.4.4	Merjenje uspešnosti.....	9
3.4.5	Vizualizacija rezultatov	10
3.5	Integracija prototipov	10
4	MERILA ZA OCENJEVANJE	11
4.1	Modul 1 – Zajem podatkov s spletnih strani VIZ	11
4.2	Modul 2 – Zajem podatkov iz spletnih medijev	11
4.3	Modul 3 – Analiza sentimenta	11
4.4	Dodatne točke – Celovitost in predstavitev	12

1 UVOD

1.1 Namen in ozadje projekta

Projekt je vzpostavljen z namenom digitalizacije, poenotenja in izboljšanja kakovosti postopkov nadzora ter upravljanja v vzgojno-izobraževalnem sistemu Republike Slovenije.

1.2 Vloga umetne inteligence

Umetna inteligenca (AI) ima v projektu ključno vlogo pri **avtomatizaciji analitičnih procesov** in **izboljšanju kakovosti odločanja** na podlagi podatkov. Njena uporaba omogoča prehod od tradicionalnih metod spremljanja in poročanja k **podatkovno podprtemu, proaktivnemu in napovednemu modelu ocenjevanja kakovosti vzgojno-izobraževalnih zavodov (VIZ)**.

Ena od pomembnih aplikacij AI v sistemu je **analiza sentimenta** spletnih in medijskih objav, povezanih z delovanjem posameznih VIZ. Z uporabo **naravnega jezikovnega procesiranja (NLP)** in modelov, kot je **croSloBERTa**, sistem omogoča prepoznavanje čustvenega tona (pozitiven, nevtralen, negativen) ter kvantitativno oceno sentimenta v razponu od 0 (močno negativen) do 100 (močno pozitiven). Takšna analiza dopolnjuje klasične kazalnike kakovosti z vpogledom v **javnomenjsko percepcijo zavoda** in s tem prispeva k bolj celovitemu spremljanju kakovosti izobraževalnega okolja.

AI bo v prihodnje v sistemu služila tudi širšim analitičnim namenom – od **prepoznavanja vzorcev tveganj, avtomatiziranega vrednotenja poročil**, do **semantičnega iskanja in povezovanja vsebin** znotraj dokumentacijskih baz. S tem projekt vzpostavlja temelje za **inteligentno podporo odločanju**, ki povezuje podatke iz uradnih virov z informacijami iz odprtih spletnih virov, kar omogoča celovit in sodoben pristop k spremljanju kakovosti VIZ.

2 OPIS IZZIVA

2.1 Kaj se od tekmovalcev pričakuje

Od sodelujočih ekip se pričakuje, da v okviru študentskega izziva pripravijo **funkcionalen prototip**, ki demonstrira zmožnost **avtomatskega zajema, obdelave in analize podatkov** o vzgojno-izobraževalnih zavodih (VIZ) iz javno dostopnih spletnih virov. Tekmovalci morajo samostojno zasnovati arhitekturo prototipa, razviti delujoče module ter zagotoviti **osnovno dokumentacijo in predstavitev rezultatov**.

Pri delu se spodbuja uporaba **odprtokodnih tehnologij, metod umetne inteligence in sodobnih programskih orodij**, ki omogočajo hitro eksperimentiranje, deljenje rezultatov in ponovljivost postopkov.

Ker gre za **kompleksen izziv**, ki se po zahtevnosti približuje **raziskovalnemu delu in razvoju strokovnjakov s področja umetne inteligence**, se od študentov in dijakov **ne pričakuje popolna implementacija vseh zahtev**. Tekmovalci naj se **okvirno držijo navodil in izvedejo tiste dele projekta, ki jih obvladajo**.

Tudi **delne rešitve** (npr. samo zajem podatkov, brez sentimentne analize, ali osnovna analiza sentimenta brez popolne integracije) so **popolnoma sprejemljive** in bodo ocenjene glede na kakovost izvedenega dela, razumevanje problema in inovativnost pristopa. Ključno je, da ekipe pokažejo **razmišljanje, tehnično radovednost in sposobnost ustvarjanja uporabnih prototipov**, tudi če niso celoviti ali popolnoma optimizirani.

2.2 Cilji izziva

2.2.1 Zajem podatkov s spletnih strani VIZ

- Samodejno pridobivanje osnovnih informacij (npr. novice, obvestila, dogodki) s spletnih strani izbranih VIZ.
- Priprava mehanizmov za prepoznavanje strukture strani in filtriranje nerelevantnih vsebin.

2.2.2 Zajem podatkov iz spletnih medijev

- Iskanje in zbiranje člankov, ki omenjajo posamezne VIZ, iz slovenskih spletnih medijev.
- Prepoznavanje entitet (npr. ime šole, kraj, dogodek) ter povezava člankov z ustreznim VIZ.

2.2.3 Analiza sentimenta

- Uporaba ali prilagoditev obstoječega jezikovnega modela (npr. **croSloBERTa**) za oceno sentimenta besedil v slovenskem jeziku.
- Pretvorba rezultatov v **numerično lestvico od 0 do 100**, kjer 0 pomeni močno negativen in 100 močno pozitiven sentiment.
- Priprava povzetkov in vizualizacij, ki prikazujejo sentiment po posameznih VIZ, časovnih obdobjih ali virih.

2.2.4 Možne smeri reševanja in omejitve

Tekmovalci imajo pri izbiri **tehnologij, orodij in metodologij popolno svobodo**, saj je izziv raziskovalno-prototipne narave. Dovoljena je uporaba kateregakoli programskega jezika (npr. Python, JavaScript, R) ter knjižnic za obdelavo besedil (npr. HuggingFace, spaCy, NLTK) ali zajem podatkov (npr. BeautifulSoup, Scrapy, Selenium, Playwright).

Pri zasnovi rešitve pa morajo upoštevati naslednje omejitve in priporočila:

- **Viri podatkov morajo biti javno dostopni** in ne smejo zahtevati prijave ali kršiti pogojev uporabe spletnih strani.
- **Zajem podatkov naj bo etičen in spoštljiv do zasebnosti**, brez shranjevanja osebnih podatkov.
- Modeli za analizo sentimenta morajo biti **učinkoviti (čas analize < 1 s na objavo)** in prilagojeni slovenskemu jeziku.
- Vse rešitve morajo biti predstavljene v **reproducibilni obliki** (npr. Jupyter Notebook, spletna aplikacija ali predstavitevni video).
- Rezultat izziva je **prototip**, ki ni namenjen neposredni vključitvi v produkcijski sistem, temveč kot **razvojna in raziskovalna referenca**.

3 OPIS IZZIVA

3.1 Skupne metodološke osnove

V sledečih poglavjih so opisane **podrobne metodološke osnove** za zajem, obdelavo in strukturiranje podatkov iz spletnih virov. Ta poglavja predstavljajo **celovit tehnični okvir** možne implementacije, ki pa je zaradi svoje **kompleksnosti in zahtevnosti** namenjen predvsem kot **usmeritev** za tekmovalce, ne kot zavezujoča specifikacija.

Tekmovalci naj pri razvoju svojih prototipov **upoštevajo tista priporočila, ki jih lahko realno implementirajo** glede na razpoložljiv čas, znanje in orodja. Dovoljena je uporaba poenostavljenih pristopov, alternativnih algoritmov ali delnih rešitev posameznih korakov (npr. samo osnovni scraping ali analiza sentimenta na omejenem naboru podatkov).

Cilj ni popolna implementacija celotne metodologije, temveč **funkcionalen in razumljiv prototip**, ki demonstrira razumevanje koncepta, uporabo ustreznih tehnologij in sposobnost smiselne integracije med moduli. Celotna specifikacija zato **ni zavezujoča**, ampak služi kot **priporočena referenčna smer** za razvoj.

3.1.1 Metodologija spletnega zajema (scraping, API, RSS)

Spletne strani vzgojno-izobraževalnih zavodov (VIZ) niso standardizirane, zato je treba uporabiti prilagodljivo metodologijo za zajem nestrukturiranih podatkov. Cilji:

- samodejno odkrivanje relevantnih vsebin (npr. novice, obvestila, kontaktne strani),
- ekstrakcija informacij (naslov, datum objave, vsebina),
- po možnosti prepoznavanje ključnih entitet (npr. ime ravnatelja, ime šole, kraj, vrsta objave).

Primarni zajem (HTML scraping)

- Uporaba orodij **BeautifulSoup**, **Scrapy** ali **Playwright** za pridobivanje HTML vsebine.
- Iskanje semantičnih oznak (npr. <article>, <h1>, <meta name="author">) in ekstrakcija besedila.
- Po potrebi uporaba **XPath** ali **CSS selektorjev** za ciljno zajemanje elementov.

Alternativni zajem (API, RSS)

- Če šola uporablja **RSS vire** ali CMS platforme (WordPress, Joomla ipd.), je priporočljiv neposreden zajem teh strukturiranih virov, saj omogočajo čistejše in hitrejše pridobivanje podatkov.

Avtomatska identifikacija informacij z LLM (Large Language Model)

- Zaradi različne strukture strani se lahko uporabi **jezikovne modele (LLM)** (npr. **GPT**, **SloBERTa**, **OpenLLaMA-SI**) za **semantično razumevanje vsebine**.
- Model prejme besedilo spletne strani in navodilo v obliki t. i. »prompt predloge«, npr.:

Navodilo modelu: Iz spodnjega besedila spletne strani šole izlušči:

- ime šole
- ime in priimek ravnatelja
- e-naslov šole
- kraj ali regijo

Vrni rezultat v JSON obliki.

- LLM semantično analizira vsebino in vrne **strukturiran izhod**, tudi če podatki niso označeni z HTML atributi.
- Implementacija je mogoča z orodji **LangChain**, **HuggingFace Transformers** ali **OpenAI API**, pri čemer model deluje kot **semantični parser** nestrukturiranega besedila.

3.1.2 Normalizacija in čiščenje podatkov

Zajeti podatki s spletnih strani in medijev so praviloma nestrukturirani, zato je pred nadaljnjo obdelavo potrebno zagotoviti **enotno, čisto in analizabilno obliko**. Postopek obsega:

- **Odstranjevanje HTML oznak, oglasnih in navigacijskih elementov** ter drugih nerelevantnih vsebin.
- **Standardizacijo besedila** (pretvorba v UTF-8, usklajevanje šumnikov, odstranjevanje odvečnih presledkov in znakov).
- **Deduplikacijo** objav glede na naslov, URL ali vsebino.
- **Usklajevanje časovnih zapisov** (pretvorba v ISO 8601 format) in kodnih oznak virov.
- **Jezikovni filter**, ki ohrani le slovensko besedilo; po potrebi uporaba knjižnic za zaznavanje jezika (*langdetect*, *fastText*).
- **Prepoznavanje in normalizacija entitet** (npr. imen šol, oseb, lokacij) s pomočjo LLM ali orodij za **Named Entity Recognition (NER)**.

Rezultat postopka je **enotno strukturiran, očiščen in semantično označen nabor podatkov**, pripravljen za nadaljnjo analizo in izračun sentimenta.

3.1.3 Priprava strukturirane zbirke objav

Po izvedenem zajemu in čiščenju podatkov je treba rezultate pretvoriti v **enotno, strojno berljivo in analizabilno obliko**. Namen je oblikovati **strukturirano zbirko objav**, ki omogoča nadaljnjo analizo sentimenta, povezovanje z obstoječimi evidencami VIZ ter enostavno vizualizacijo podatkov.

Postopek vključuje:

- **Standardizacijo polj** (npr. id, viz_ime, datum_objave, vir, vsebina, tip_vsebine, sentiment).
- **Uporabo enotnih podatkovnih formatov** – priporočljivo **JSON** ali **CSV**, s kodiranjem **UTF-8**.
- **Povezovanje entitet** (npr. šole, kraja, ravnatelja) z **šifranti VIZ** iz sistema za zagotavljanje enolične identifikacije.
- **Dodajanje metapodatkov**, kot so čas zajema, vir (spletna stran ali medij), vrsta vsebine, zaznana entiteta.

- **Strukturiranje vsebine** tako, da omogoča učinkovito filtriranje, iskanje in združevanje po različnih dimenzijah (čas, vir, VIZ, sentiment).

Primer zapisa posamezne objave v zbirki:

```
{
  "id": "VIZ-01234-2025-10-17-001",
  "viz_ime": "OŠ Poljane",
  "ravnatelj": "Tanja Novak",
  "datum_objave": "2025-10-16",
  "vir": "https://os-poljane.si/novice/nova-oprema-za-laboratorij",
  "vsebina": "Naša šola je prejela novo opremo za naravoslovni laboratorij.",
  "tip_vsebine": "novica",
  "vir_tip": "spletna stran VIZ",
  "sentiment": null,
  "čas_zajema": "2025-10-17T09:30:00Z"
}
```

Tako pripravljena zbirka predstavlja **enoten vhod** za **modul analize sentimenta** in omogoča **povezavo z drugimi podatkovnimi viri** v okviru sistema.

Za prototip se priporočata **preprosta lokalna zbirka** (npr. **SQLite**, **JSON datoteka**) ali **lahki strežniški repozitorij** (npr. **MongoDB**, **PostgreSQL**), ki omogoča hitro testiranje in deljenje rezultatov med moduli.

3.2 Modul 1: Zajem podatkov s spletnih strani VIZ

3.2.1 Opis in namen modula

Namen modula je razviti **preprost prototip za samodejni zajem podatkov** s spletnih strani vzgojno-izobraževalnih zavodov (VIZ). Modul naj omogoča pridobivanje osnovnih informacij, kot so **naslovi novic, obvestila, datumi objav in vsebina člankov**, ter pripravo podatkov za nadaljnjo obdelavo. Glavni cilj je pokazati **razumevanje spletnega zajema** in sposobnost **pretvorbe neurejenih vsebin v strukturirano obliko**.

3.2.2 Posebnosti spletnih strani VIZ

Spletna strani šol in vrtcev so zelo raznolike – uporabljajo različne **CMS sisteme** (WordPress, Joomla, Drupal ipd.), imajo različno strukturo menijev in pogosto objavljajo novice na različnih podstraneh.

Tekmovalci naj zato:

- izberejo **manjši nabor šol (npr. 3–5)** za testiranje,
- analizirajo osnovno strukturo njihovih strani,
- in se osredotočijo na **enoten tip vsebine** (npr. novice ali obvestila).

3.2.3 Zajem statičnih in dinamičnih vsebin

Za pridobivanje vsebine lahko uporabijo:

- **preprost HTML scraping** z orodji, kot so *BeautifulSoup* (Python) ali *Cheerio* (JavaScript),
- **zajem dinamičnih vsebin**, če se stran nalaga prek JavaScript-a, z orodji, kot so *Selenium* ali *Playwright*.

Ni potrebno, da zajem deluje za vse šole – pomembno je, da prototip pokaže **koncept delovanja** in **zmožnost pridobivanja osnovnih podatkov**.

3.2.4 Prepoznavanje ključnih informacij

Po zajemu vsebine naj program poskuša iz besedila izluščiti **ključne informacije**, kot so:

- naslov objave,
- datum objave,
- besedilo članka,
- povezava (URL),
- ime šole ali ravnatelja (če se pojavlja v besedilu).

Za to se lahko uporabijo **preproste metode iskanja po ključnih besedah**, ali pa **LLM modeli** (npr. GPT ali SloBERTa), ki besedilo razumejo semantično in vrnejo strukturiran odgovor.

3.2.5 Izvoz strukturiranih podatkov v skupno zbirko

Zbrane informacije naj se pretvorijo v **preprost strukturiran format**, na primer **JSON** ali **CSV**, s polji kot so:

```
{
  "viz_ime": "OŠ Trnovo",
  "datum_objave": "2025-10-15",
  "naslov": "Učenci osvojili prvo mesto",
  "vsebina": "Učenci so osvojili prvo mesto na tekmovanju...",
  "vir": "https://os-trnovo.si/novice/tekmovanje"
}
```

Tako pripravljen nabor bo služil kot **vhod za naslednje module** (zajem medijskih člankov in analiza sentimenta).

Pomembno je, da rezultat pokaže **delujoč prototip**, ki zna prepoznati osnovno strukturo vsebine in jo predstaviti v enotni obliki, tudi če je nabor podatkov omejen.

3.3 Modul 2: Zajem podatkov iz spletnih medijev

3.3.1 Opis in namen modula

Namen modula je razviti **prototip za samodejno zbiranje člankov iz slovenskih spletnih medijev**, ki omenjajo vzgojno-izobraževalne zavode (VIZ). Modul naj omogoča iskanje člankov po ključnih besedah, zajem osnovnih podatkov o članku (naslov, vsebina, vir, datum) ter pripravo enotne zbirke besedil za nadaljnjo **analizo sentimenta**. Cilj je prikazati, kako se lahko iz odprtih spletnih virov pridobi relevantne informacije o VIZ.

3.3.2 Identifikacija medijskih virov

Tekmovalci naj sami izberejo **nabor slovenskih spletnih medijev**, ki redno objavljajo novice o šolah in izobraževanju (npr. 24ur.com, RTV Slovenija, Siol.net, Delo, Lokalno.si). Za zajem člankov lahko uporabijo:

- **RSS vire**, če so na voljo (npr. <https://www.rtv slo.si/rss>),
- **javne API-je**, kjer obstajajo,
- ali **spletni scraping**, če vsebine niso dostopne drugače.

Priporoča se omejitev na **manjši nabor virov (2–3)** za demonstracijo koncepta.

3.3.3 Iskanje po ključnih besedah

Za filtriranje člankov naj se uporabljajo **ključne besede**, povezane z vzgojno-izobraževalnimi zavodi, kot so:

- ime šole ali vrtca (npr. *OŠ Trnovo*, *Gimnazija Bežigrad*),
- ime občine ali kraja,
- besede, kot so *ravnatelj*, *učitelj*, *dijaki*, *vrtec*, *izobraževanje*.

Rezultate iskanja naj modul shrani skupaj z osnovnimi metapodatki članka (naslov, datum, vir, URL).

Za izboljšano prepoznavanje ustreznih člankov se lahko uporabi **semantično iskanje** z LLM ali **tekstovno podobnost (cosine similarity)**, če tekmovalci to želijo raziskati.

3.3.4 Prepoznavanje entitet v člankih

Iz vsebine člankov naj modul poskuša izluščiti **ključne entitete**, kot so:

- ime VIZ (npr. *OŠ Škofljica*),
- ime ravnatelja ali učitelja,
- kraj ali občina,
- tema dogodka (npr. *nagrada*, *incident*, *projekt*).

Za ta korak zadostuje uporaba **ključnih besed** ali **preprostih regularnih izrazov**. Naprednejši pristop vključuje uporabo **modelov za prepoznavanje imenovanih entitet (NER)** ali **LLM modela**, ki lahko besedilo analizira semantično in prepozna, katera imena pripadajo šolam in osebam.

3.3.5 Vzpostavitev zbirke člankov za nadaljnjo sentimentno analizo

Zbrani članki naj se shranijo v **strukturirani zbirki**, podobno kot v modulu 1, v formatu **JSON** ali **CSV**, s polji, kot so:

```
{
  "naslov": "Starši zadovoljni z novim ravnateljem OŠ Trnovo",
  "viz_ime": "OŠ Trnovo",
  "datum_objave": "2025-10-12",
  "vir": "https://www.rtv slo.si/slovenija/starsi-zadovoljni",
  "vsebina": "Starši učencev so izrazili podporo novemu ravnatelju ...",
  "krajevna_enota": "Ljubljana",
  "sentiment": null
}
```

Ta zbirka predstavlja **vhodne podatke za Modul 3 (analiza sentimenta)**.

Pomembno je, da prototip pokaže, kako se lahko članki najdejo, filtrirajo in shranijo v enotni obliki, tudi če obseg podatkov ni velik. Namen modula ni popolnost, temveč **razumevanje procesa in sposobnost povezovanja med podatki in VIZ.**

3.4 Modul 3: Analiza sentimenta

3.4.1 Opis in namen modula

Namen modula je razviti **prototip za oceno sentimenta** (pozitivnost ali negativnost) člankov, ki so bili zbrani v prejšnjih moduli. Modul naj omogoča, da se vsakemu članku pripiše **številčna vrednost sentimenta** na lestvici od 0 (močno negativen) do 100 (močno pozitiven). Cilj ni popolna natančnost, temveč **demonstracija uporabe modela umetne inteligence** za analizo besedil v slovenskem jeziku.

3.4.2 Uporabljeni model

Za izvedbo analize sentimenta se priporoča uporaba **predtreniranega modela croSloBERTa**, ki je posebej prilagojen za slovenski jezik in omogoča dobre rezultate pri klasifikaciji sentimenta.

Tekmovalci lahko:

- uporabijo **croSloBERTa** preko knjižnice *HuggingFace Transformers*,
- ali izberejo **alternativni model (LLM)**, če ocenijo, da je primernejši, vendar morajo svojo izbiro **kratko utemeljiti** (npr. boljša hitrost, lažja uporaba, dostopnost).

Za poenostavitev je dovolj, da model ločuje tri osnovne razrede sentimenta (*negativen, nevtralen, pozitiven*) in jih pretvori v **številčno oceno**.

3.4.3 Postopek analize

Postopek analize sentimenta obsega tri osnovne korake:

1. **Priprava besedila** – odstranitev nepotrebnih znakov, URL-jev in HTML oznak.
2. **Vhod v model** – pretvorba besedila v obliko, primerno za obdelavo z modelom (tokenizacija).
3. **Izračun sentimenta** – model vrne verjetnosti za posamezne razrede (npr. pozitiven 0.75, negativen 0.10, nevtralen 0.15).
4. **Pretvorba v lestvico 0–100**, npr.:
 - a. močno negativen: 0–30
 - b. nevtralen: 40–60
 - c. močno pozitiven: 70–100

Pomembno je, da je rešitev **enostavna in razumljiva**, ne pa popolna ali produkcijsko optimizirana.

3.4.4 Merjenje uspešnosti

Ker gre za raziskovalni prototip, se uspešnost lahko meri le okvirno. Tekmovalci lahko navedejo:

- **ocenjeno natančnost ali F1 rezultat** na majhnem testnem naboru (če ga ustvarijo sami),
- **čas obdelave posamezne objave** (v sekundah),
- po želji pa tudi **primerjavo dveh različnih pristopov** (npr. croSloBERTa vs. enostavni leksikonski pristop).

Hitrejše rešitve bodo lahko dodatno nagrajene, vendar hitrost **ni obvezen kriterij**.

3.4.5 Vizualizacija rezultatov

Rezultate sentimentne analize naj modul prikaže na **pregleden in razumljiv način**, na primer:

- grafični prikaz sentimenta po posameznih VIZ (npr. stolpčni graf),
- prikaz povprečnega sentimenta po obdobjih (npr. časovna vrstica),
- prikaz deležev pozitivnih, nevtralnih in negativnih objav.

Vizualizacija je lahko narejena v **Jupyter Notebooku**, **Excelu** ali s pomočjo **spletne knjižnice** (npr. *Matplotlib*, *Plotly*, *Streamlit*).

3.5 Integracija prototipov

V okviru izziva **ni zahtevano**, da so vsi trije moduli popolnoma povezani v enotno aplikacijo. Pomembno je le, da je med njimi **osnovna logična povezava** – torej, da se lahko **izhod enega modula uporabi kot vhod v naslednjega**.

Tekmovalci lahko vsak modul razvijajo **ločeno** (npr. kot samostojen skript, notebook ali aplikacijo), pri čemer morajo zagotoviti, da se podatki iz prvega modula (zajem s spletnih strani VIZ) in drugega modula (zajem iz spletnih medijev) lahko uporabijo za tretji modul (analizo sentimenta).

Za izmenjavo podatkov med moduli se priporoča **enostaven skupni format** (npr. **JSON** ali **CSV**), v katerem so prisotna osnovna polja, kot so naslov, vsebina, datum_objave, viz_ime, vir in sentiment.

Če se ekipe odločijo, lahko pripravijo tudi **preprosto prototipno predstavitev** rezultatov – npr. *dashboard*, *Jupyter Notebook poročilo* ali *demo aplikacijo*, ki prikaže, kako moduli med seboj sodelujejo.

Ključno je, da prototip kot celota **jasno pokaže potek obdelave podatkov**:

zajem → čiščenje → analiza sentimenta → prikaz rezultatov.

Način povezovanja med moduli ni zavezujoč – šteje predvsem **razumljivost in delovanje osnovnega koncepta**.

4 MERILA ZA OCENJEVANJE

Rešitve študentskega izziva bodo ocenjene glede na **funkcionalnost posameznih modulov** in **kakovost izvedbe**. Vsak tekmovalec lahko pridobi **največ 100 točk + 10 dodatnih točk**.

Vrednotenje je razdeljeno po treh glavnih moduli, znotraj katerih se ocenjujejo ključni kvalitativni kriteriji.

4.1 Modul 1 – Zajem podatkov s spletnih strani VIZ

Modul 1 – Zajem podatkov s spletnih strani VIZ lahko tekmovalcu prinese do 30 točk.

Kriterij	Opis	Točke
Funkcionalnost zajema	Ali prototip uspešno pridobi vsebino z dejanskih spletnih strani VIZ	0–10
Struktura in čistost podatkov	Ali so podatki logično strukturirani (npr. JSON/CSV), brez nepotrebnih oznak in podvojenih zapisov	0–8
Prilagodljivost in stabilnost	Ali rešitev deluje za več različnih spletnih strani / CMS sistemov	0–6
Inovativnost in pristop	Uporaba zanimivih tehnik (npr. LLM za razumevanje strani) ali kreativnih rešitev	0–6

4.2 Modul 2 – Zajem podatkov iz spletnih medijev

Modul 2 – Zajem podatkov iz spletnih medijev lahko tekmovalcu prinese do 30 točk.

Kriterij	Opis	Točke
Iskanje in filtriranje člankov	Ali sistem uspešno prepozna članke, ki se nanašajo na VIZ	0–10
Prepoznavanje entitet	Natančnost in ustreznost povezovanja člankov z VIZ (npr. ime šole, kraj, oseba)	0–8
Uporabnost in kakovost zbirke	Jasnost, preglednost in enotnost strukturiranih podatkov	0–6
Ustvarjalnost pristopa	Inovativne ideje (semantično iskanje, uporaba LLM, združevanje virov ipd.)	0–6

4.3 Modul 3 – Analiza sentimenta

Modul 3 – Analiza sentimenta lahko tekmovalcu prinese do 40 točk.

Kriterij	Opis	Točke
Delujoč model sentimenta	Ali prototip pravilno izračuna in prikaže osnovni sentiment (negativen, nevtralen, pozitiven)	0–20

Kakovost rezultatov	Smiselnost ocen (npr. pravilna porazdelitev vrednosti 0–100, razumljiv izhod)	0–10
Učinkovitost in hitrost	Povprečen čas analize, odzivnost sistema (hitrejše rešitve se nagradijo)	0–5
Vizualizacija rezultatov	Jasnost grafov, poročil ali prikaza sentimenta po VIZ	0–5

4.4 Dodatne točke – Celovitost in predstavitev

Modul 4 – Celovitost in predstavitev lahko tekmovalcu prinese do 10 točk.

Kriterij	Opis	Točke
Integracija in logična povezanost modulov	Ali se moduli smiselno dopolnjujejo in delujejo v zaporedju	0–5
Dokumentacija in predstavitev	Jasna razlaga rešitve, opis pristopa, predstavitev rezultatov	0–5

Opomba:

Popolna tehnična implementacija vseh modulov ni pogoj. Tudi **delno delujoči prototipi** bodo ocenjeni pozitivno, če pokažejo razumevanje problema, logično zasnovo in ustvarjalnost.

Ekipe, ki bodo pokazale **inovativne pristope ali izvirne ideje**, lahko prejmejo dodatne pohvale in priporočila za sodelovanje pri nadaljnjem razvoju sistema.