Annotation guideline

**Adapted Guidelines for Croatian Sentiment Analysis**

Ovaj dokument sadrži prilagođene smjernice za procjenu označavanja sentimenta u očišćenim tekstualnim podacima – specifično recenzijama odabranih filmova. Primarni cilj je procijeniti točnost oznaka sentimenata dodijeljenih pojedinačnim rečenicama, uzimajući u obzir inherentne izazove određivanja sentimenta u hrvatskom jeziku ovisnom o složenosti konteksta i korelacije rečenica.

Dok anotatori mogu obraditi kontekst i zaključiti suptilnosti u značenju pojedinih rečenica (s razlikom u viđenju, ovisno o prethodnoj – narednoj rečenici), modeli strojnog učenja često se suočavaju s poteškoćama s ovim zadacima što uvelike utječe na sam proces određivanja sentimenta i pri manualnoj obradi.

**Kriteriji označivanja sentimenata – upute (Sentiment rating instructions)**

Sentiment se određuje prema 5-stupanjskoj ljestvici (0-4) pri čemu je:

* **Pozitivnom sentimentu** dodijeljena oznaka **0**
* **Neutralnom sentimentu** dodijeljena oznaka **1**
* **Negativnom sentimentu** dodijeljena oznaka **2**
* **Mješovitom (mixed) sentimentu** dodijeljena oznaka **3**
* **Sarkazmu/ironiji** dodijeljena oznaka 4

**Pozitivan sentiment (oznaka: 0)**

Rečenice se kategoriziraju kao pozitivne kada izražavaju osjećaje podrške, divljenja, pozitivnih stavova, oprosta, ohrabrenja, uspjeha ili pozitivnih osjećaja. Ove rečenice koriste pozitivne izraze koji prenose povoljnu perspektivu, odražavajući konstruktivno ili optimistično stajalište.

Primjer: *osobno sam veliki ljubitelj kreativnosti, posebice po tom pitanju, tako da mi je upravo takav tajanstveni teaser dovoljno zaokupio pažnju, uz neizostavnu želju za daljnjim zavođenjem.* (iskazan optimizam)

**Neutralan sentiment (oznaka: 1)**

Neutralni sentiment pripisuje se rečenicama koje prenose činjenične, objektivne informacije bez ikakvih emocionalnih ili procjeničkih izraza. Ove rečenice ne pokazuju jasnu pozitivnu ili negativnu pristranost, već umjesto toga iznose neutralne činjenice, brojke ili općenite izjave.

Primjer: *žanrovski gledano film spada u znanstveno-fantastični triler s određenom dozom humora i romantike.* (neutralna činjenica, općenita izjava)

**Negativan sentiment (oznaka: 2)**

Negativan sentiment pripisuje se rečenicama koje izražavaju nepovoljnu ocjenu, kritiku, razočaranje ili druge negativne emocije. Ove rečenice obično izražavaju osjećaj neuspjeha, općenite tuge - nesreće ili neodobravanja. Kontekstualno razumijevanje ključno je za prepoznavanje negativnog osjećaja, osobito kada se negativan aspekt podrazumijeva, a ne izričito naveden.

Primjer: *u 51. se našla na prekretnici koja ju je uvijek toliko mučila – odjednom je opet postala solerica, i to ne zbog debljine, neumjerenog uživanja u alkoholu, brzopletosti ili bilo koje osobine za koju je uvjerena da je koči.*

Primjer 2: *no, može li to uspjeti na duge staze?* (relevantan kontekst za prepoznavanje negativnog sentimenta)

**Mješoviti sentiment (oznaka: 3)**

Rečenice koje se označavaju s mješovitim (mixed) sentimentom većinom uključuju i pozitivne i negativne elemente, često unutar jedne izjave. Takve rečenice mogu predstavljati suprotne perspektive ili dvosmislene situacije, što dovodi do kombinacije emocija.

Primjer: *međutim, iako je stvarna bol prilično uspješno obradila temu koju je trebala obraditi i u cjelini nam je donijela zabavno i simpatično iskustvo, treba istaknuti kako se ne radi o savršenom filmu, što ćemo objasniti u pasusima koji slijede.* (kombinacija pozitivnog i negativnog)

**Sarkazam/ironija (oznaka: 4)**

Rečenice koje sadrže sarkazam, ironiju, ismijavanje ili slične oblike izražavanja kategorizirane su ovim sentimentom – oznakom 4. Sarkazam i ironija često se izražavaju jezikom koji se površno čini pozitivnim, ali unutar konteksta prenosi suprotan ili kritičan osjećaj. Prepoznavanje sarkazma ili ironije zahtijeva nijansirano razumijevanje šireg konteksta u kojem je izjava dana (\* potencijalan izazov za strojno određivanje).

Primjer: *no, snaga šarma i seksipila (da, da i seksipila) te i takve šašave bridget prenijela se i na narednu generaciju (uz napomenu da nju voli i muška i ženska populacija), a uz ovaj četvrti, posljednji nastavak ju je definitivno zavoljela i generacija z, koja se ipak znači može i zna odlijepiti od tik toka i ekrana svojih smartphonea – ima nade i za “mladež”, dakle*.

**Zaključno: ključna razmatranja**

**Kontekstualna ovisnost:** Sentiment rečenice ne može se uvijek odrediti izolirano. Kod anotiranja je potrebno uzeti u obzir okolni kontekst - i prethodne i sljedeće rečenice (posebno kritično u slučajevima kada se rečenica može činiti neutralnom, ali poprima pozitivnu ili negativnu konotaciju kada se stavi u kontekst).

**Ograničenje u vidu određivanja sarkazma i ironije:** Ljudski anotatori mogu prepoznati „nijanse“ i suptilne promjene na temelju konteksta. Inherentna dvosmislenost i složenost prirodnog jezika zahtijevaju ljudsku prosudbu za točnu dodjelu sentimenta (problem tj. ograničenje u vidu automatizirane analize sentimenta)