NAPREDNO PREPOZNAVANJE ENTITET Z VELIKIMI JEZIKOVNIMI MODELI. ŠTUDIJA PRIMERA POVEZOVANJE PROCESNIH ODPADKOV S TAKSONOMIJO EU

Igor Dimovski[1], Rok Vinder[1], Rok Štembergar[1], Stevanče Nikoloski[1,2]
[1] Result, d.o.o., Celovška 182, 1000 Ljubljana
[2] Univerza v Novem mestu Fakulteta za ekonomijo in informatiko, Na Loko 2, 8000 Novo mesto
igor.dimovski@result.si,rok.vinder@result.si,rok.stembergar@result.si,stevance.nikoloski@result.si

V dobi digitalizacije in hitrega napredka znanstvenih raziskav postaja uporaba velikih jezikovnih modelov (LLM) in generativne umetne neizogibna in ključna za obvladovanje ter analizo obsežnih zbirk znanstveno-strokovnih dokumentov. Ta študija uvaja napredno aplikacijo, ki izkorišča te tehnologije za semantično analizo in natančno prepoznavanje ključnih entitet, kot so procesni odpadki omenjeni v znanstveno-strokovnoj dokumentih. Naša aplikacija za semantično analizo, ki se opira na obdelavo naravnega jezika (NLP), omogoča globlje razumevanje in povezovanje relevantnih informacij, kar prinaša novo dimenzijo avtomatizirane inteligence v analizo znanstvenih vsebin.

Ključna inovacija aplikacije je algoritem, zasnovan za povezovanje oz. preslikava procesnih odpadkov iz znanstvenih člankov na EU taksonomijo odpadkov, ki se naslanja na napredne zmogljivosti, kot sta RAG arhitektura in Mistral 7B LLM, ter vektorsko bazo podatkov ChromaDB. Ta sistem ne le da dokazuje visoko stopnjo uspešnosti z 85-odstotno natančnostjo kartiranja, temveč se tudi spopada z izzivom 15% neklasificiranih odpadkov, ki zahtevajo dodatno preučitev zaradi nizkih ocen podobnosti.

Ta pristop k preslikavi odpadkov ne samo izboljšuje strategije za upravljanje z odpadki s poudarkom na prepoznavanju procesov, ki odpadke ustvarjajo, ampak tudi razkriva možnosti za ponovno uporabo teh odpadkov kot surovin v drugih procesih. Zagotavljanje teh ključnih informacij gospodarskim akterjem v pravem času krepi trajnostne prakse, kar je izjemnega pomena v luči doseganja globalnih okoljskih ciljev. Preko analize 27 znanstvenih člankov aplikacija izkazuje svoj potencial za pomembne prispevke k bolj učinkovitemu in okolju prijaznemu upravljanju z odpadki. Poleg tega pristop ne le povečuje natančnost in zanesljivost preslikave odpadkov, ampak tudi vključuje sistematično verifikacijo teh preslikav s strani strokovnjakov na področju upravljanja z odpadki. Ti strokovnjaki ne samo, da potrjujejo pravilnost preslikav, ampak tudi zagotavljajo dragoceno povratno informacijo, ki se uporablja za natančno prilagoditev (fine-tuning) algoritmov velikega jezikovnega modela (LLM). Ta iterativni proces ne samo izboljša točnost in relevantnost rezultatov, ampak tudi zagotavlja, da se sistem neprestano uči in prilagaja novim ugotovitvam in trendom v industriji, kar še dodatno krepi njegovo zmožnost podpiranja trajnostnih praks in inovativnih rešitev v upravljanju z odpadki.

Ta študija poudarja ne le tehnološke dosežke aplikacije, ampak tudi njen širši pomen za napredek v okoljski trajnosti. Z izboljšanjem klasifikacije in upravljanja odpadkov prispeva k zmanjšanju okoljskega odtisa in podpira globalne prizadevanja za ohranjanje našega planeta. V zaključku, raziskava odpira nove perspektive v izkoriščanju umetne inteligence za analizo znanstveno-strokovnih podatkov, hkrati pa naslavlja ključne izzive v okoljskem sektorju, s čimer postavlja temelje za prihodnji razvoj na tem področju.

Ključne besede: veliki jezikovni modeli; prepoznavanje entitet; semantična analiza; umetna inteligenca; EU taksonomija odpadkov; zmanjševanje odpadkov

AN ADVANCED ENTITY RECOGNITION WITH LARGE LANGUAGE MODELS. A CASE STUDY ON PROCESS WASTE MAPPING TO EU TAXONOMY

This research presents an advanced AI application designed to improve semantic analysis and entity recognition within scientific documents, employing Large Language Models (LLMs) like Mistral 7B, RAG architecture, and ChromaDB vector database. The application's core innovation is an algorithm capable of mapping process wastes identified in scientific texts to the EU's waste taxonomy with an 85% success rate, demonstrating a significant advance in waste management and sustainability. This process involves sophisticated entity identification, tagging, and linking, facilitating an automated, intelligent analysis of scientific documents. By studying 27 scientific papers, the application's effectiveness in waste reduction and resource optimization is validated. This approach not only showcases the potential of AI in enhancing environmental sustainability but also marks a step forward in the intelligent processing and understanding of scientific data.

Keywords: large language models; entity recognitions; semantic analysis; artificial intelligence; EU waste taxonomy; waste reduction