University of Ljubljana
Faculty of Computer and Information Science

Romana Grilj

Performance comparison of open source and commercial information extraction tools

master's thesis

the 2nd cycle master's study programme
computer and information science

Supervisor: doc. dr. Slavko Žitnik

Ljubljana, 2023

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Romana Grilj

Primerjava uspešnosti odprtokodnih in komercialnih orodij za luščenje podatkov

magistrsko delo

magistrski študijski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Slavko Žitnik

Ljubljana, 2023

Acknowledgments

Worth mentioning in the acknowledgment is everyone who contributed to your thesis.

Romana Grilj, 2023

To all the flowers of this world.

"The only reason for time is so that everything doesn't happen at once."

— Albert Einstein

Contents

List of used acronmys

acronym	meaning
CA	classification accuracy
DBMS	database management system
SVM	support vector machine
...	...

Abstract

Title: Performance comparison of open source and commercial information extraction tools

This sample document presents an approach to typesetting your BSc thesis using LaTeX. A proper abstract should contain around 100 words which makes this one way too short. A good abstract contains: (1) a short description of the tackled problem, (2) a short description of your approach to solving the problem, and (3) (the most successful) result or contribution in your thesis.

Keywords

Data analysis, Information Retrieval, structural data, Web Mining

Povzetek

Naslov: Primerjava uspešnosti odprtokodnih in komercialnih orodij za luščenje podatkov

Magistrsko delo obravnava področja obdelave naravnega jezika ter zaznavo objektov. Primerjali smo različne oblačne storiteve kot so: Vertex AI, AWS SageMaker, Azure Cognitive Services ter odprtokodno rešitev Hugging Face Transformers. Cilj naloge je raziskati in analizirati ter primerjati njihove zmogljivosti, značilnosti ter ustreznost na različnih področjih uporabe obdelave naravnega jezika, kot so prepoznava imenskih entit, analiza sentimenta, prepoznava objektov, povzemanje besedila, klasifikacija ter izvleček besedne zveze.

V delu bodo podrobno predstavljene storitve treh največjih oblačnih ponudnikov: Vertex AI je Googlova platforma, Amazonova storitev SageMaker ter Microsoftova storitev Azure Cognitive Service so trenutno največje platforme za strojno učenje, obdelavo podatkov ter razvoj modelov, ki omogočajo integracijo funkcionalnosti obdelave naravnega jezika ter zaznavo objektov ter primerjava z odprtokodno platformo Hugging Face Transformers.

V raziskavi so bila preučena naslednje naloge obdelave naravnega jezika, kot je prepoznavanje imenskih entitet, kot so imena oseb, krajev, datumov in organizacij v besedilu. Analiza sentimenta je naloga za določanje čustvenega naboja besed ali besednih zvez, ki je lahko pozitiven, negativen ali nevtralen. Povzemanje zajema ustvarjanje krajšega povzetka daljšega besedila. Izvleček besedne zveze obravnava metodologije za ekstrakcijo ključnih besed ali besednih zvez v besedilu. Klasifikacija besedila za avtomatsko razvrščanje be-

sedila v različne kategorije. Zaznava objektov preučuje algoritme in tehnike za prepoznavanje objektov ali entitet na slikah.

Za primerjavo uspešnosti modelov so bile uporabljene različne metrike kot so odzivnost, natančnost, F1 ocena, ROUGE ter Accuracy.-SLO Uporabljeni so bili naslednji korpusi za evaluiranje modelov:CoNLL 2003 za prepoznavo imenskih entit, IMdb Reviews za analizo sentimenta, COCO za prepoznavo objektov v slikah, CNN/Daily Mail za povzemanje besedila ter semeval-2017 za klasifikacijo ter izvleček besedne zveze.

Skupaj s predstavitvijo ponudnikov oblačnih storitev, njihovih zmogljivosti in rezultatov evalvacije na različnih področjih uporabe, je magistrsko delo prispevalo k boljšemu razumevanju primernosti ter učinkovitosti omenjenih orodij za različne naloge obdelave naravnega jezika.

Ključne besede

analiza podatkov, ekstrakcija podatkov, strukturni podatki, spletno rudarjenje

Chapter 1

Uvod

Obdelava naravnega jezika je eno izmed najhitreje razvijajočih se področij v umetni inteligenci, ki se ukvarja z obdelavo, razumevanjem in generiranjem naravnega jezika, kot ga uporabljamo ljudje za komunikacijo. Napredne tehnike obdelave naravnega jezika in modeli so v zadnjem desetletju privedli do prebojev v številnih storitvah, ki segajo od avtomatskega prevajanja in analize čustev do odgovarjanja na vprašanja in samodejnega povzemanja besedil.

Cilj te magistrske naloge je raziskati in analizirati različne pristope k modelom obdelave naravnega jezika, razpoložljive korpuse, uporabljene metrike za evaluacijo modelov ter številna področja uporabe.

V nadaljevanju se bomo osredotočili na pomembnost kakovostnih korpusov za uspešno učenje modelov. Pregledali bomo obstoječe in priljubljene korpuse, ki se uporabljajo za različne naloge. Kot ključen element bomo preučili metrike za ocenjevanje učinkovitosti modelov, kot so natančnost, F1-ocena, ROUGE in druge. Razložili bomo, kako se uporabljajo za različne naloge in kako lahko z njimi ocenimo zmogljivost modelov.

Nato se bomo posvetili raznolikim področjem uporabe tehnologij obdelave naravnega jezika. Raziskali bomo, kako se tehnologije uporabljajo v analizi čustev, povzemanje, iskanju informacij in še več. Preučili bomo tudi izzive in omejitve, s katerimi se srečujejo modeli pri uporabi na različnih področjih.

V poglavju **??** bomo podrobneje spoznali izbrane funkcionalnosti, ki jih omogočajo modeli, ki temeljijo na umetni inteligenci in strojnem učenju ter so zasnovane za obdelavo, razumevanje in generiranje naravnega jezika. Različni modeli se uporabljajo za reševanje različnih nalog, povezanih z obdelavo jezika, kot so avtomatsko prevajanje, analiza čustev, razumevanje besedil, odgovarjanje na vprašanja, izluščevanje informacij iz besedil, klasifikacija besedil in še več.

V poglavju **??** bomo podrobneje spoznali uporabljene korpuse, ki so ključnega pomena za uspešen razvoj, učenje in evalvacijo modelov. Korpusi so zbirke podatkov, ki so ročno označeni ali označeni s pomočjo algoritmov za različne naloge.

V poglavju **??** bomo temeljito raziskali različne metrike, ki omogočajo oceno učinkovitosti modelov glede na njihove specifične naloge. Raznolikost nalog v področju obdelave naravnega jezika zahteva prilagodljivost pri izbiri metrik. V nadaljevanju so navedene ključne metrike, ki jih bomo uporabljali pri evalvaciji modelov.

V predzanjem poglavju **??** so predstavljeni rezultati različnih ponudnikov ter njihovih storitev.

Sledijo še **??**.

Pridobili como pregled nad uporabo modelov ter njihovo uporabo na različnih področij. Cilj je prispevati k razumevanju tehnologij obdelave naravnega jezika ter predstaviti rezultate različnih storitev.

Chapter 2

Opis ponudnikov in storitev

2.1 Hugging Face

Hugging Face je platforma na področju obdelave naravnega jezika in strojnega učenja. Njihova rešitev je postala zelo uporabna na področju raziskav, razvoja in uporabe strojnega učenja.

Njihova glavna odprtokodna knjižnica "Transformers" je postala temelj raziskav obdelave naravnega jezika, saj ponuja obsežen nabor naprednih modelov, kot so GPT, BERT, RoBERTa in drugih, ki so ključni za različne naloge, vključno z razumevanjem jezika, strojnim prevajanjem, analizo čustev in generiranjem besedila.

Hugging Face Hub je platforma, ki spodbuja sodelovanje in izmenjavo med raziskovalci, razvijalci in navdušenci nad strojnim učenjem. Ta platforma omogoča enostavno deljenje in odkrivanje modelov, kar olajša razvoj novih aplikacij in omogoča dostop do že pripravljenih modelov. Področje obdelave naravnega jezika je prvič bilo na voljo leta 2017. [**?**]

2.1.1 Hugging Face Transformers

Hugging Face Transformers je odprtokodna knjižnica, ki je postala ena najpomembnejših orodij za obdelavo naravnega jezika in strojnega učenja. Njen cilj je ponuditi razvijalcem enostaven dostop do najnovejših arhitektur in

modelov. Zgrajena je na osnovi Pythona in je postala ključno orodje za reševanje izzivov na področju obdelave naravnega jezika in razvoja aplikacij ter storitev.

Ena od ključnih prednosti Hugging Face Transformers je enostavnost uporabe. Razvijalci lahko z nekaj vrsticami kode dostopajo do pred-treniranih modelov in jih takoj uporabljajo. Poleg tega knjižnica omogoča tudi prilagajanje modelov in ponuja odprtokodno skupnost, ki nenehno prispeva z novimi modeli, izboljšavami in rešitvami.

Knjižnica ponuja tudi funkcionalnosti za preprosto prenosljivost modelov med različnimi platformami in orodji za povečanje učinkovitosti uporabe modelov na stvarnih sistemih. Poleg tega Hugging Face Transformers omogoča tudi preprosto združevanje z drugimi knjižnicami za strojno učenje in obdelavo podatkov. [**?**]

Podpora različnih jezikov je pogojena v prvi vrsti z izbiro modela. Transforjerji ponujajo širok nabor storitev za obdelavo naravnega jezika in sicer: klasifikacija besedila, prevajanje, povzemanje, znakovna klasifikacija, tabela vprašanj odgovorov, odgovori na vprašanja, razvrščanje brez vzorcev, klepet, generiranje besedila, dodajanje mankajočih besed, podobnost besed

2.2 Google Cloud

Google Cloud je celovita platforma za računalništvo v oblaku, ki jo zagotavlja Goofgle. Ponuja različne storitve ko so: shranjevanje, bazami podatkov, strojnim učenjem, kar omogoča podjetjem, da gradijo in razvijajo aplikacije ter storitve v globalnem obsegu. Google Cloud zagotavlja prilagodljivo in razširljivo infrastrukturo, ki organizacijam omogoča inovacije ter optimizacijo operacij prek rešitev v oblaku. S svojimi podatkovnimi centri po vsem svetu Google Cloud zagotavlja zanesljivo zmogljivost, varnost in dostopnost za podjetja vseh velikosti. [**?**]

Področje obdelave naravnega jezika je bilo dodano v Google cloud leta 2015.

2.2.1 Vertex AI

Vertex AI je napredna platforma za umetno inteligenco v oblaku, ki jo ponuja Google Cloud. S podporo za številne priljubljenena ogrodja (framework) za strojno učenje, kot so TensorFlow ter PyTorch je Vertex AI odlična izbira za razvijalce z različnimi potrebami in izkušnjami. Platforma Vertex AI ponuja tudi številne napredne storitve in orodja za razvoj in optimizacijo modelov. Vključuje integrirano orodje, ki omogoča hitro in enostavno oceno uspešnosti modelov na različnih primerih rabe. Prav tako ponuja samodejno prilagajanje hiperparametrov, kar omogoča avtomatsko iskanje najboljših hiperparametrov za izboljšanje zmogljivosti modelov. Prvotno je bila izdana leta 2020 in ima že več kot 100.000 uporabnikov po vsem svetu. Uporabljajo jo raznolika podjetja, od majhnih startupov do velikih korporacij, kot so Walmart, Pfizer in Coca-Cola. Podpirajo kar 11 različnih jezikov in sicer: Angleščina, Francoščina, Nemščina, Španščina, Kitajščina (poenostavljena), Kitajščina (tradicionalna), Japonščina, Korejščina,Portugalščina ter Ruščina.

Vertex AI ponujajo širok nabor storitev za obdelavo naravnega jezika in sicer: sentimentalna analiza, analiza entitete, analiza sentimenta entitete, analiza sintakse, vsebinska klasifikacija. Vertex Ai lahko uporabljamo z več programskimi jeziki kot so: Go, Java, Node.js, Python.

Ena od ključnih funkcij Vertex AI je tudi funkcija Vertex Data Labeling, ki omogoča enostavno označevanje podatkov za učenje modelov. [**?**]

2.3 Amazon Web Services

f Amazon Web Services (AWS) je ponudnik storitev v oblaku, ki jih ponuja Amazon. Uporabnikom omogoča najem računalniških virov, kot so strežniki in shramba. To omogoča organizacijam, da prilagodljivo in učinkovito gradijo, upravljajo ter skalirajo svoje aplikacije in storitve brez potrebe po fizični strojni opremi. AWS je mednarodno priznan za svojo zanesljivost in širok nabor storitev za obdelavo podatkov, analitiko, umetno inteligenco ter druge poslovne potrebe. [**?**]

Področje obdelave naravnega jezika je bilo dodano leta 2017. V Evropi pa je bil dostopen šele leto kasneje.

Podpora različnih jezikov je pogojena v prvi vrsti z izbiro modela.

2.3.1 Amazon SageMaker

Amazon SageMaker je storitev za strojno učenje, ki jo ponuja Amazon Web Services (AWS). Omogoča hitro in enostavno izgradnjo, usposabljanje in razporejanje zmogljivih modelov, kar omogoča razvoj naprednih rešitev in izboljšanje procesov. Ponuja intuitiven uporabniški vmesnik in API-je, ki omogočajo hitro postavitev in upravljanje. SageMaker ponuja tudi integrirano okolje Jupyter, ki omogoča uporabo interaktivnih beležnic za raziskovanje in analizo podatkov. [**?**]

Amazon SageMaker ponujaja širok nabor storitev za obdelavo naravnega jezika in sicer: klasifikacija besedila, analiza sentimenta, prepoznavanje imenskih entitet, prevanjanje, povzetek ter druge.

Uporabljamo ga lahko z pomočjo dveh različnih programskih jezikov R ter Python.

2.4 Microsoft Azure

Azure je oblačna platforma, ki jo ponuja Microsoft, namenjena podjetjem za razvoj, upravljanje in gostovanje njihovih aplikacij in storitev prek interneta v oblaku. Ponuja širok nabor storitev, vključno s spletnim gostovanjem, shranjevanjem podatkov, analitiko, umetno inteligenco in spletnimi storitvami. Uporabniki lahko ustvarjajo virtualne strežnike in omrežja ter jih prilagajajo glede na svoje potrebe. Azure zagotavlja visoko stopnjo varnosti in skladnosti, kar je ključno za zaščito podatkov in zagotavljanje zasebnosti strank. [**?**]

Področje obdelave naravnega jezika je bilo dodano leta 2018.

Azure ponuja široko podporo različnim jezikom, skupno več kot 96.Nekateri izmed njih so: Angleščina, Finščina, Francoščina, Danščina , ter druge.

Kar je zelo pomembno omeniti podpira tudi Slovenski jezik.

2.4.1 Azure Cognitive Services

Azure Cognitive Services je celovita in napredna platforma za umetno inteligenco (AI), ki jo ponuja Microsoftova platforma Azure. Te storitve omogočajo analizo in razumevanje naravnega jezika v besedilu, kar omogoča razvoj aplikacij za avtomatsko razvrščanje besedil, odgovarjanje na vprašanja, prevajanje besedil in analizo sentimenta. Ponuja širok nabor naprednih storitev in API-jev, ki omogočajo prepoznavanje, razumevanje in generiranje naravnega jezika, prepoznavanje obrazov, prepoznavanje govora, analizo besedil, prevajanje med jeziki in še veliko več. Te storitve omogočajo razvoj pametnih aplikacij, ki temeljijo na umetni inteligenci, in reševanje različnih izzivov na področju razumevanja in analize podatkov.

Azure Cognitive Services ponujaja širok nabor storitev za obdelavo naravnega jezika in sicer: prepoznava entitet, sentimentalna analiza, odgovarjanje na vprašanja, prevajanje. [**?**]

Uporabljamo ga lahko z pomočjo več različnih programskih jezikov C#, Java, JavaScript, Python.

Chapter 3

Izbrana področja uporabe

3.1 Prepoznavanje imenskih entitet

Prepoznavanje imenskih entitet je tehnika na področju obdelave naravnega jezika, ki se uporablja za prepoznavanje in klasifikacijo besed v besedilu. Te posebne vrste so imenovane entitete, kot so imena oseb, organizacij, lokacij, datumov, številk, denarnih zneskov in drugih specifičnih poimenovanj.

Cilj je prepoznati in določiti začetek in konec posameznih entitet v besedilu ter jim pripisati ustrezno kategorijo.[**?**]

Številne praktične uporabe:

1. Avtomatsko označevanje imenskih entitet v novicah, člankih in drugih besedilnih vsebinah.

2. Razumevanje strukture in vsebine dokumentov za informacijsko iskanje in kategorizacijo.

3. Pomoč pri analizi sentimenta, kjer želimo razumeti, kako se osebe, organizacije ali druge entitete omenjene v besedilu nanašajo na določeno temo ali izdelek.

Primer:

V stavku "Janez Novak je rojen 10. avgusta 1985 v Ljubljani" bi sistem prepoznal "Janez Novak" kot ime osebe, "10. avgust 1985" kot datum in "Ljubljana" kot lokacijo.

3.2 Analiza sentimenta

Analiza sentimenta je proces določanja čustvenega odziva, nagnjenosti ali stališča zapisanega besedila. Cilj analize sentimenta je ugotoviti, ali je določeno besedilo pozitivno, negativno ali nevtralno. To je lahko koristno pri analizi mnenj strank, razumevanju čustvenega odziva na izdelke, blagovne znamke, dogodke in druge. [**?**]

Obstaja več pristopov k analizi sentimenta:

1. Pravilni pristopi: Uporabljajo se predvsem pravila in vzorci za identifikacijo pozitivnih in negativnih izrazov v besedilu. Na primer, besede, kot so "dobro", "fantastično", "radostno" itd., bi bile označene kot pozitivne, medtem ko bi bile besede, kot so "slabo", "žalostno", "neznosno" in tako dalje označene kot negativne.

2. Strojno učenje na podlagi besedila: Ta pristop vključuje uporabo algoritmov strojnega učenja, ki so naučeni prepoznati čustveni naboj besed v besedilu na podlagi velikega števila označenih podatkov (besedil s čustvenimi oznakami).

3. Analiza sentimenta s čustvenimi slovarji: Ta pristop vključuje uporabo slovarjev z besedami in izrazoslovjem, ki so povezani z določenimi čustvi. Besedilo se nato preveri in oceni glede na prisotnost pozitivnih ali negativnih besed iz čustvenih slovarjev.

4. Algoritmi globokega učenja.

Primer:

Če imamo naslednji stavek: "Ta film je fantastičen, vreden ogleda!", bi analiza sentimenta prepoznala, da je izraz pozitiven. Ta analiza temelji na uporabi naravnojezikovnega procesiranja in strojnega učenja.

3.3 Povzemanje besedila

Pri povzermanju besedila gre za postopek ustvarjanja krajšega in jedrnatega povzetka iz daljšega besedila, kot je članek ali dokument. Namen povzemanja je izluščiti ključne informacije in ideje iz izvornega besedila ter jih predstaviti na bolj pregleden in krajši način. To je zelo koristno pri velikih količinah podatkov, ko želimo hitro pridobiti bistvo informacij, ne da bi brali celotno besedilo.

Tehnike za povzemanje uporabljajo različne algoritme in metode, ki vključujejo strojno učenje in obdelavo naravnega jezika, da bi učinkovito izluščile ključne besede, stavke ali odstavke, ki predstavljajo osrednje ideje v izvornem besedilu. Rezultat je običajno kratek povzetek, ki ohranja pomembne informacije iz izvirnega besedila. Ta tehnologija ima širok spekter uporab, kot so samodejno povzemanje novic, generiranje opisov izdelkov, izdelava povzetkov raziskovalnih člankov in še veliko več. [**?**]

3.4 Luščenje ključnih besed

Nanaša se na besede ali izraze, ki so najpomembnejši ali najbolj značilni za določeno besedilo ali dokument. Te besede so običajno tiste, ki nosijo ključne informacije ali so bistvene za razumevanje vsebine.

Je pomembna naloga, saj nam omogoča, da hitro ugotovimo, o čem govori določeno besedilo. Te besede so lahko uporabne tudi za avtomatsko indeksiranje dokumentov, iskanje relevantnih informacij in razumevanje teme besedila brez potrebe po branju celotnega besedila. [**?**]

3.5 Klasifikacija besedila

Klasifikacija besedil je tehnika, pri kateri avtomatizirano določimo kategorijo ali razred določenega besedila na podlagi vsebine besedila. To je lahko zelo uporabno, saj nam omogoča razvrščanje besedil v različne skupine glede na njihovo vsebino.

Postopek klasifikacije besedil se običajno začne s pripravo in čiščenjem besedil. To vključuje odstranjevanje nepotrebnih znakov, šumnikov, posebnih znakov, pretvorbo vseh črk v male črke, lahko pa tudi odstranjevanje pogostih besed, ki nimajo velikega pomena za klasifikacijo (npr. "in", "ali", "je", "na", "s", itd.).

Nato se besedila predstavijo v obliki, ki jo lahko uporabimo za učenje modela. Pogosto se uporablja metoda imenovana vreča besed ("Bag-of-Words"), kjer se besedilo pretvori v nabor besed, ki se pojavljajo v njem, in število pojavitev teh besed. Ta postopek lahko ponazorimo s pomočjo vektorja. [**?**]

Primer:

Lahko klasificiramo e-poštna sporočila kot "spam" ali "ne-spam", novice glede na tematiko, uporabniške komentarje glede na ton (pozitiven, negativen, nevtralen), itd.

3.6 Zaznava objektov

Je tehnika, ki se uporablja za avtomatsko zaznavanje in identifikacijo objektov na digitalnih slikah ali video posnetkih. Namen te tehnike je, da prepozna in označi različne objekte v podobi ter jih loči od ozadja ali drugih objektov.

Postopek objektnega zaznavanja običajno vključuje naslednje korake:

1. Zaznavanje: Model preučuje sliko ali video posnetek in identificira regije, kjer bi se lahko nahajali objekti.

2. Zaznava lokacije: Po tem, ko so bile regije prepoznane, algoritem določi omejitveno okvirje (bounding boxes), ki natančno označujejo položaje in mejne točke objektov na sliki.

3. Klasifikacija: Ko so objekti omejeni z omejitvenimi okviri, analizira vsebino znotraj teh okvirov in jih razvrsti v različne kategorije (npr. avto, pes, zgradba, itd.).

4. Sledenje: V video posnetkih je lahko zaželeno, da algoritem sledi objektom skozi različne kadre in tako beleži njihovo gibanje.

Zaznava objektov se uporablja v številnih aplikacijah, kot tudi v samovozečih vozilih za zaznavanje drugih vozil in pešcev, identifikacija prometnih znakov, nadzorne kamere, prepoznavanje obrazov, analiza medicinskih slik in še veliko drugega. Gre za enega ključnih elementov umetne inteligence. [**?**]

Chapter 4

Korpusi

4.1 Kaj je korpus?

Je zbirka podatkov, ki so organizirani in shranjeni v strukturirani ali nestrukturirani obliki ter označeni za namen analize, raziskav, učenja modelov ali drugih postopkov obdelave podatkov. Korpusi vsebujejo različne vrste podatkov, od številk, besedil, slik, zvokov, videoposnetkov do drugih tipov informacij. V kontekstu računalniškega znanstvenega modeliranja in strojnega učenja so korpusi ključnega pomena, saj služijo kot osnova za razvoj, treniranje in evalvacijo modelov. Modeli se učijo na teh podatkih, tako da prepoznajo vzorce in povezave med vhodnimi podatki in ciljnimi izhodi. Na primer, v naravnojezikovni obdelavi korpusov vsebuje besedilne podatke, ki so lahko članki, knjige, novičarski članki ali socialni mediji.

Nekatere ključne točke o uporabi korpusov:

Učenje modelov: Korpusi se uporabljajo za učenje modelov, pri čemer modeli na osnovi teh podatkov pridobivajo razumevanje jezika in njegove strukture. Čeprav obstajajo tudi nespremljani pristopi, večina uspešnih modelov zahteva velike, kakovostne in označene korpuse za doseganje najboljših rezultatov.

Razvoj in optimizacija: Razvijalci modelov uporabljajo različne korpuse

za optimizacijo modelov in prilagajanje hiperparametrov. Z vzorci podatkov iz korpusov se preizkušajo različne arhitekture modelov in strategije učenja.

Evaluacija: Korpusi se uporabljajo za evalvacijo modelov. Preizkušajo se na ločenem testnem korpusu, ki modelom omogoča, da se oceni, kako dobro delujejo na novih, nevidenih podatkih.

Nadzor kakovosti: Kvaliteta korpusov je ključnega pomena za uspešno delovanje modelov. Zato je pomembno, da so korpusi natančno označeni in urejeni. Nadzor kakovosti pomaga prepoznati morebitne napake ali pristranskosti v korpusih.

Prilagajanje specifičnim aplikacijam: Včasih so potrebni specializirani modeli za določene aplikacije ali domene. V takih primerih je morda potrebno ustvariti ali prilagoditi korpuse, ki se bolje prilegajo ciljni uporabi.

Razvoj modelov za redke jezike: Razvoj modelov obdelave naravnega jezika za redke jezike zahteva ustrezne korpuse v ciljnem jeziku, kar je lahko izziv, saj so ti korpusi pogosto omejeni ali pa jih sploh ni na voljo.

Pomembno je, da so korpusi pravilno pripravljeni, imajo ustrezne metapodatke in so primerni za ciljno nalogo, da bi omogočili kakovostno analizo in doseganje uporabnih rezultatov.

4.2 Uporabljeni korpusi

4.2.1 CoNLL 2003

Je zbirka podatkov, ki se uporablja za razvoj in evalvacijo sistemov za obdelavo naravnega jezika, prevsem za nalogo imenskih entitet. Imenuje se po konferenci CoNLL (Conference on Computational Natural Language Learning) leta 2003, kjer je bil ta nabor podatkov predstavljen v okviru tekmovanja za prepoznavanje imenovanih entitet. Korpus CoNLL 2003 je priljubljen referenčni korpus za prepoznavanje poimenovanih entitet naravnega jezika v obdelavi naravnega jezika. Uporabljen je bil v skupni nalogi na konferenci o računalniškem učenju naravnega jezika (CoNLL) leta 2003.

Poimenovane entitete so razdeljene v štiri glavne kategorije:

1. Oseba (PER): Posamezna imena ljudi.
2. Organizacija (ORG): Imena podjetij, ustanov ali organizacij.
3. Lokacija (LOC): Imena geografskih lokacij, kot so mesta, države ali regije.
4. Razno (MISC): Druge poimenovane entitete, ki ne spadajo v zgoraj navedene kategorije, na primer datumi, odstotki ali denar.

Podatki v korpusu so predstavljeni v obliki ene besede na vrstico, kjer vsaka vrstica predstavlja besedo in pripadajočo oznako v stavku. Besede in oznake so ločene z presledkom. Korpus CoNLL 2003 se pogosto uporablja za evalvacijo zmogljivosti modelov za prepoznavanje poimenovanih entitet in že več let je standardno merilo (benchmark) za raziskovalce in strokovnjake v skupnosti obdelave naravnega jezika. Ostaja dragocen vir za razvoj in preizkušanje novih algoritmov in sistemov za prepoznavo imenskih intitet. [**?**]

Table 4.1: Primer CoNLL 2003 korpusa

He PRP B-NP O
will MD B-VP O
probably RB I-VP O
be VB I-VP O
replaced VBN I-VP O
by IN B-PP O
Shearer NNP B-NP B-PER
's POS B-NP O
Newcastle NNP I-NP B-ORG

CoNLL2003 podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 14.000 vrsticami primerov

2. validacijski (validation) z 3.250 vrsticami primerov

3. preizkusni (test) z 3.450 vrsticami primerov

4.2.2 IMDb Reviews

IMDB podatkovna zbirka, znana tudi kot IMDB Movie Reviews Dataset. Sestavljen iz pregledov filmov, ki so jih prispevali uporabniki na spletni strani IMDb (Internet Movie Database).

Podatki vsebujejo ocene in besedilne komentarje, ki jih je ustvarila skupnost uporabnikov IMDb. Vsak pregled vsebuje besedilni komentar in oceno filma, ki se giblje med 1 (najslabša) in 10 (najboljša). Cilj te podatkovne zbirke v naravnem jeziku je razviti modele, ki lahko avtomatsko analizirajo besedilne komentarje in napovedo, ali je pregled pozitiven ali negativen glede na oceno in besedilo. [**?**]

Table 4.2: Primer IMDB korpusa

review	sentiment
If you like original gut wrenching laughter you will like movie.	positive
A rating of "1", depressing and relentlessly bad this movie is.	negative

IMDB podatkovna zbirka je običajno razdeljena na dva sklopa:

1. učni (train) z 25.000 vrsticami primerov

2. preizkusni (test) z 25.000 vrsticami primerov

Vsak sklop vsebuje tisoče pregledov filmov. To je idealna podatkovna zbirka za naloge analize čustvenega tona besedil (sentiment analysis), kjer modeli ocenjujejo, ali je mnenje v besedilu pozitivno, negativno ali nevtralno.

4.2.3 COCO

COCO (Common Objects in Context) je nabor podatkov v področju računalniškega vida in zaznave objektov. Namenjen je zagotavljanju celovite in raznolike

zbirke slik za različne naloge, vključno z zaznavo objektov, segmentacijo in podnaslavljanjem. Nabor podatkov naj bi odražal scenarije iz resničnega sveta in vsebuje slike, ki so kompleksne ter vključujejo več objektov v različnih kontekstih.

Nabor podatkov COCO je obsežen in vsebuje deset tisoče slik z milijoni označenih posameznih objektov. Slike prihajajo iz različnih virov, zajemajo raznolike prizore, ozadja, svetlobne pogoje in velikosti objektov.

Tukaj je nekaj ključnih značilnosti nabora podatkov COCO:

1. Kategorije slik: Nabor podatkov COCO vsebuje slike, ki zajemajo 80 različnih kategorij objektov, od splošnih objektov, kot so "oseba," "avto" in "pes," do bolj specifičnih objektov, kot so "mobilni telefon," "zobna ščetka" in "zmaj."

2. Anotacije: Vsaka slika v naboru podatkov COCO je opremljena z oznakami na ravni objekta in koordinatami okvirja. To pomeni, da je vsak posamezen objekt določene kategorije znotraj slike označen, okoli njega pa je narisano območje z okvirjem, ki označuje njegovo lokacijo. Informacije o anotacijah so ključnega pomena za usposabljanje modelov za detekcijo objektov in segmentacijo.

3. Segmentacija objektov: Poleg anotacij območja z okviri nabor podatkov COCO prav tako zagotavlja maske segmentacije na ravni slikovnih pik za vsak posamezen objekt. To pomeni, da so objekti ne le lokalizirani z okviri, ampak so natančno določene tudi meje objektov na ravni slikovnih pik. [**?**]

COCO podatkovna zbirka je običajno razdeljena na dva sklopa:

1. učni (train) z 117.000 primeri

2. preizkusni (test) z 4.950 primeri

Figure 4.1: 000000502136.jpg


```
[{
    "license": 3,
    "file_name": "000000502136.jpg",
    "coco_url": "http://images.cocodataset.org/val2017/000000502136.jpg",
    "height": 423,
    "width": 500,
    "date_captured": "2013-11-15 17:08:30",
    "flickr_url": "http://farm3.staticflickr.com/2253/1755223462_fabbeb8dc3_z.jpg",
    "id": 502136
},
{
    "segmentation": [
        [
            54.74,
            350.34,
            53.75,
            353.33,
                ...
            349.35
        ]
    ],
    "area": 4651.359250000001,
    "iscrowd": 0,
    "image_id": 502136,
    "bbox": [
        3.98,
        289.63,
        120.43,
        103.51
    ],
    "category_id": 64,
    "id": 21011
}]
```


Figure 4.2: COCO .json primer

4.2.4 CNN/Daily Mail

CNN/Daily Mail je zbirka novičarskih člankov skupaj s povzetki, ki se uporablja za usposabljanje in preizkušanje modelov za povzemanje besedil. Ta nabor podatkov vsebuje različne novičarske članke in njihove povzetke, zaradi česar je primeren za naloge abstraktivnega povzemanja, kjer se ustvarijo povzetki v lastnih besedah, ne le izbirajo stavke iz izvornega besedila. Nabor podatkov vsebuje na tisoče člankov s pripadajočimi povzetki, kar omogoča raziskovalcem obsežno treniranje in evaluiranje modelov.[**?**]

Tukaj je nekaj ključnih značilnosti nabora podatkov CNN/Daily Mail:

1. Novičarski Članki in Povzetki: Nabor podatkov vsebuje novičarske članke iz medijskih virov, kot sta CNN (Cable News Network) in Daily Mail, skupaj s pripadajočimi povzetki. Ti članki pokrivajo različne teme in dogodke ter so različnih dolžin.

2. Abstraktno Povzemanje: Vključuje ustvarjanje povzetka v povsem novih besedah.

Table 4.3: Primer cnn_dailymail korpusa

label	text	highlights
002509a...	Fears are growing that Britain's jails are becoming...	Athens pushes through...
7526a1...	It was a farce that would lead to...	AZ Alkmaar were playing....

CNN/Daily Mail podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 287.000 vrsticami primerov

2. validacijski (validation) z 13.400 vrsticami primerov

3. preizkusni (test) z 11.500 vrsticami primerov

4.2.5 SemEval -2017

SemEval podatkovna zbirka je zbirka besedilnih podatkov, ki je anotirana za različne naloge na področju obdelave naravnega jezika.

Tukaj je nekaj ključnih značilnosti SemEval podatkovnih zbirk:

1. Anotacije: Podatki v SemEval podatkovnih zbirkah so običajno anotirani, kar pomeni, da so označeni z dodatnimi informacijami. Na primer, v podatkovni zbirki za naloge razreševanja sentimenta bi bili vzorci besedil označeni s pozitivnimi, negativnimi ali nevtralnimi sentimenti.

2. Raznolikost: SemEval podatkovne zbirke zajemajo širok spekter nalog, jezikov in domen. To omogoča raziskovalcem primerjavo modelov in pristopov na različnih področjih.

Raziskovalna skupnost: SemEval podatkovne zbirke so postale pomemben del naravnega jezika raziskovalne skupnosti, saj omogočajo primerjavo najnovejših pristopov in tehnologij na enotnem naboru podatkov. [**?**]

Table 4.4: Primer Semeval-2017 korpusa

label	`text`
`-1`	I missed the Barcelona game yesterday. http://t.co/AqpknXC
`0`	I'm bout to just listen to nicki minaj all night
`1`	One Night like In Vegas I make dat Nigga Famous

f

SemEval podatkovna zbirka je običajno razdeljena na tri sklope:

1. učni (train) z 49.547 vrsticami primerov
2. validacijski (dev) z 12.285 vrsticami primerov
3. preizkusni (test) z 12.285 vrsticami primerov

Chapter 5

Metrike

5.1 Opis spemenljivk za izračun metrik

Pravilno pozitivni (True Positive)

Je izraz, ki se uporablja v statistiki in strojnem učenju za opis primerov, kjer je model pravilno napovedal pozitiven rezultat za določeno skupino. To pomeni, da je model prepoznal pozitiven pojav, ko je bil dejansko prisoten. [**?**]

Primer:

Predpostavimo, da razvijamo model za prepoznavanje spam sporočil v elektronski pošti. Model pravilno prepozna 25 sporočil kot nezaželena (spam), ki dejansko vsebujejo nezaželeno vsebino. To pomeni, da imamo 25 primerov "pravih pozitivnih". Te primere model pravilno prepozna kot spam, ker resnično vsebujejo neželeno vsebino.

Napačno pozitivni (False Positive)

Označuje situacijo, ko model napačno napove, da je nekaj pozitivno, medtem ko je v resnici negativno. Gre za vrsto napake, kjer model napačno identificira primer kot pripadajoč pozitivnemu razredu, čeprav dejansko pripada negativnemu razredu.[**?**]

Primer:

Predpostavimo, da imamo model za prepoznavanje spam sporočil v elek-

tronski pošti. Če model označi sporočilo kot "spam", čeprav ni dejansko spam, imamo situacijo lažno pozitivnega primera. Drugače povedano, model je napačno napovedal pozitiven primer (spam), ko je dejansko negativen primer (ni spam).

Napačno negativni (False Negatives)

Označuje napako, ki se pojavi v kontekstu klasifikacije ali analize besedila, ko model napačno napove, da je nekaj negativno, čeprav je v resnici pozitivno. To je vrsta napake, kjer model spregleda ali ne prepozna pozitivnih primerov. V primeru analize besedila v naravni jezikovni obdelavi, false negative se zgodi, ko model ne uspe zaznati pozitivnega elementa v besedilu, ki bi ga moral prepoznati. Na primer, če imamo model za prepoznavanje pozitivnih izjav v komentarjih in model spregleda pozitivno izjavo, to bi bil primer false negative. [**?**]

Primer:

Predpostavimo, da imamo napreden sistem za filtriranje neželenih sporočil (spam), ki ga uporabljamo za preverjanje prihajajočih e-poštnih sporočil. Sistem je zasnovan tako, da prepoznava in premika neželena sporočila v mapo za neželeno pošto.

Vendar pa se pojavi napačno negativen rezultat, ko sistem napačno presodi e-poštno sporočilo kot varno (ne-spam), čeprav vsebuje vse znake neželene vsebine. Na primer, če e-poštno sporočilo vsebuje povezave do nerealnih ponudb ali oglasev za sumljive izdelke, bi bila takšna sporočila številčno gledano ena od "False Negatives".

V tem primeru je sistem spregledal prepoznavo neželene vsebine, kar je povzročilo, da je sporočilo pristalo v glavnem predalu prejete pošte namesto v mapi za neželeno pošto. To lahko predstavlja težavo, saj se takšni neželeni vsebini lahko izognemo le, če sistem zanesljivo prepozna vse takšne primere. [**?**]

5.2 Natančnost (Precision)

Natančnost je pomembna metrika za ocenjevanje uspešnosti modelov v različnih nalogah. Povdarja natančnost pozitivnih napovedi, torej tistih primerov, ki jih model prepozna kot pozitivne. Visoka preciznost pomeni, da so pozitivne napovedi modela zanesljive in imajo malo lažno pozitivnih napak.

V kontekstu naravne jezikovne obdelave, natančnost igra ključno vlogo pri razumevanju besedila. Na primer, pri analizi sentimenta želimo natančno ugotoviti, ali je izraz pozitiven ali negativen. Visoka preciznost v tem primeru pomeni, da so napovedi modela o sentimentu točne in se malo zmotijo.[**?**]

Formula za izračun:

$$\text{Natančnost} = \frac{PravilnoPozitivni}{PravilnoPozitivni + PravilnoPozitivni}$$

5.3 Odzivnost (Recall)

Nanaša se na eno od metrik uspešnosti pri vrednotenju modelov za obdelavo naravnega jezika. Meri kot razmerje med številom pravilno prepoznanih relevantnih primerov in celotnim številom dejansko obstoječih relevantnih primerov. Višja odzivnost pomeni, da je model bolje usposobljen za iskanje in pridobivanje vseh relevantnih informacij, vendar to lahko vodi tudi v več lažno pozitivnih rezultatov. Zato je pomembno doseči uravnoteženost med odzivnostjo in natančnostjo (precision) pri oceni uspešnosti modelov. Primer uporabe odzivnosti je v iskalnih sistemih, kjer želimo zagotoviti, da se relevantni dokumenti ali informacije ne izpustijo pri iskanju. S pravilno optimizacijo modelov lahko dosežemo visoko kakovostno izluščevanje informacij iz besedil, kar je ključno za številne aplikacije, kot so avtomatizirano odzivanje na povratne informacije strank, analiza sentimenta in razumevanje besedil v različnih jezikih. [**?**]

Formula za izračun:

$$\text{Odzivnost} = \frac{PravilnoPozitivni}{PravilnoPozitivni + NapacnoPozitivni}$$

5.4 F1 ocena (F1-score)

F1 ocena je pomembna metrika za ocenjevanje uspešnosti modelov v obdelavi besedil. Združuje natančnost in odzivnost v eno metriko, ki odraža ravnotežje med tema dvema metrikama. Pri nalogah, kot so klasifikacija besedil, luščenje informacij ali identifikacija entitet, sta tako natančnost kot odzivnost ključni. Visoka natančnost pomeni pravilno identifikacijo relevantnih elementov, medtem ko visoka odzivnost zagotavlja prepoznavanje vseh resnično pozitivnih primerov. Izračuna se kot povprečje med natančnost in odzivnost, dajeta pa ji enako težo. To omogoča, da ocenimo, kako dobro model obvladuje oba cilja hkrati. Visoka vrednost F1 ocene kaže, da je model uspešno uskladil identifikacijo pravih pozitivnih primerov z izogibanjem napačno pozitivnim rezultatom. Uporaba metrike je zlasti smiselna, ko sta metriki natančnost in odzivnost pomembni za končni rezultat in ko želimo doseči optimalno uravnoteženost med tema dvema vidikoma.

Formula za izračun:

$$\text{F1 ocena} = 2 \text{ x } \frac{Natancnost \times Odzivnost}{Natancnost + Odzivnost}$$

5.5 Accuracy

Accuracy je metrika, ki se pogosto uporablja za ocenjevanje uspešnosti modelov v strojnem učenju, vključno z modeli uporabljenimi v obdelavi naravnega jezika. Ta metrika meri, kako pravilno model napove razrede ali kategorije za vhodne podatke v primerjavi z dejanskimi vrednostmi.

V kontekstu obdelave naravnega jezika se natančnost uporablja, na primer, pri nalogah klasifikacije besedil. Predpostavimo, da imamo model, ki se uči razvrščati besedila v določene kategorije, kot so "pozitivno", "negativno" ali "neutralno." Za vsako besedilo ima model svojo napoved, kateri kategoriji pripada.[**?**]

Formula za izračun:

$$\text{Accuracy} = \frac{TruePositive+TrueNegative}{TruePositive+TrueNegative+FalsePositive+FalseNegative}$$

5.6 ROUGE

ROUGE je metrika, ki se uporablja za ocenjevanje kakovosti generiranih besedil v primerjavi z referenčnimi besedili. Gre za kratico, ki označuje "Recall-Oriented Understudy for Gisting Evaluation". Metrika ROUGE je pogosto uporabljena v področju obdelave naravnega jezika, še posebej v nalogah avtomatskega povzemanja besedil.

Primerja generirano besedilo z referenčnim besedilom (običajno človeško ustvarjenim besedilom) in oceni, kako dobro so se ujemale. Metrika upošteva različne vidike, kot so prekrivanje besed, besedni nizi in skupna dolžina besedil. Glavni cilj metrike je merjenje stopnje, do katere je generirano besedilo sposobno pravilno povzeti pomembne informacije iz referenčnega besedila.

Obstajajo različne različice metrike ROUGE, kot so ROUGE-1, ROUGE-2, ROUGE-L itd. Vsaka različica meri različne vidike podobnosti med generiranim besedilom in referenčnim besedilom. Na primer, ROUGE-1 meri prekrivanje eno-besednih nizov med generiranim in referenčnim besedilom, medtem ko ROUGE-2 meri prekrivanje dvo-besednih nizov.

Metrika ima širok nabor uporabe v raziskavah in nalogah, ki vključujejo avtomatsko povzemanje besedil, strojno prevajanje in druge naloge, kjer je pomembno oceniti kakovost generiranih besedil v primerjavi z referenčnimi besedili. Metrika lahko pomaga raziskovalcem in razvijalcem oceniti učinkovitost svojih modelov in tehnik ter izboljšati rezultate pri generiranju besedil.

Chapter 6

Analiza raziskave

Pri analizi raziskave so bili izbrani trije največji ponudniki storitev v oblaku: Google Cloud, Amazon Web Services ter Microsoft Azure. Pri izbiri odprtokodne rešitve je bilo težko izbrati najboljšega, saj so v tem času tri najboljše odprtokodne rešitve zelo tesno skupaj kakor tudi povezane. Na podlagi pregledanih funkcionalnosti ter uporabe je bil izbran Hugging Face - Transformers. Pomembno je omeniti, da so vse oblačne storitve po rezultatih tesno skupaj v nekaterih primerih, kot je razvidno iz tabele analize je za klasifikacijo besedila odprtokodna rešitev Transformers je imela najboljši rezultat.

Analiza napak

Pri raziskovanju modelov in ter korpusov, hitro spoznamo da so lahko modeli celo preveč prilagojeni(overfit) določenemu področju, kar pomeni da preveč podrobno pozna eno področje, na katerem je bil model treniran in ne more dobro generalizirati novih podatkov/primerov, kakor tudi da so premalo podrobni ali premalo raznoliki.

Zato so najbolj poznani in razširjeni modeli, učeni na širokem naboru različnih podatkov, da je možnost napake manjša.

Vrste napak, ki jih lahko prepoznamo pri NLP modelih:

Nezadostni podatki: Za gradnjo natančnega modela je potrebna velika količina podatkov. Če ni dovolj podatkov, bo model morda težko naučil vzorce v podatkih.

Table 6.1: Tabela analize

		Transformers	Vertex AI	SageMaker	Cognitive Services
Prepoznavanje imenskih entitet	Odzivnost	0.919	0.919	0.961	0.824
	Natančnost	0.923	0.920	0.954	0.858
	F1 ocena	0.921	0.919	**0.958**	0.841
Analiza sentimenta	Odzivnost	0.926	0.936	0.891	0.981
	Natančnost	0.928	0.924	0.862	0.862
	F1 ocena	0.929	**0.930**	0.876	0.876
Povzetek	ROUGE-L	0.192	0.306	0.201	**0.330**
Izvleček besedne zveze	Odzivnost	0.573	0.543	0.622	0.530
	Natančnost	0.475	0.637	0.513	0.670
	F1 ocena	0.519	0.586	0.562	**0.592**
Klasifikacija besedila	Odzivnost	0.926	0.962	0.763	0.920
	Natančnost	0.930	0.957	0.858	0.880
	F1 ocena	**0.928**	0.907	0.808	0.900
Zaznava objektov	Accuracy	0.940	0.977	**0.980**	0.965

Nezadostna raznolikost podatkov: Pomembno je, da imajo podatki za usposabljanje dobro razpršenost. Če so podatki preveč homogeni, bo model morda težko naučil vzorce, ki veljajo za splošne primere.

Nekvalitetni podatki: Pomembno je, da so podatki za usposabljanje kakovostni. Če so podatki napačni ali pristranski, bo model morda težko naučil natančen model.

Napačen algoritem: Obstaja veliko različnih algoritmov za stojno učenje, zato je pomembno, da izberete algoritem, ki je primeren za specifično nalogo. Če izberete napačen algoritem, bo morda težko zgraditi natančen model.

Napačna nastavitev parametrov: Večina algoritmov stojnega učenja ima parametre, ki jih je mogoče prilagoditi za izboljšanje natančnosti modela. Če niso pravilno nastavljeni parametri, bo morda težko zgraditi natančen model.

Pri testitranju imenovanje imenskih entitet je najvišjo uspešnost dosegla storitev AWS Sage Maker. Na drugem mestu se je uvrstila odprtokodna storitev Hugging Face Transformers, medtem ko je tretje mesto zasedla storitev Vertex AI ponudnika Google Cloud.

Pri testiranju sentimentalne analize je najboljše rezultate dosegla storitev AWS SageMaker. Na drugem mestu se je uvrstila odprtokodna rešitev Hugging Face Transformers, medtem ko je tretje mesto zasedla storitev Vertex

AI podjetja Google Cloud.

Najboljši rezultati so bili doseženi pri povzemanju z uporabo storitve Azure Cognitive Services. Na drugem mestu se je uvrstil Vertex AI, medtem ko je tretje mesto pripadlo AWS SageMaker.

V ocenjevanju izvlečka besednih zvez je najvišje mesto osvojila storitev Azure Cognitive Services. Na drugem mestu se je uvrstil Vertex AI, medtem ko je tretje mesto zasedel AWS SageMaker.

Pri klasifikaciji besedila se je najboljša učinkovitost pokazala pri odprtokodni platformi Transformers. Na drugem mestu je bila storitev Vertex AI, medtem ko je tretje mesto pripadlo storitvi Azure Cognitive Services.

V zaznavanju objektov je najvišje mesto zasedla storitev AWS Cognitive Services, takoj za njo je sledil Vertex AI, medtem ko je tretje mesto pripadlo odprtokodni rešitvi Transformers.

V nadaljevanju so predstavljeni podrobnejši rezultati treh iteracij z povprečnimi vrednostmi pripadajočih metrik ter standardni odklon kateri je merilo razpršenosti podatkov. Pomaga nam razumeti, koliko se podatki razlikujejo od povprečne vrednosti. Rezultati v vseh tabelah so zaokroženi na tri decimalna mesta.

6.1 Prepoznavanje imenskih entitet (Named Entity Recognition)

Pri analizi imenskih entitet je bil uporabljen CONLL-2003 korpus.

Za prepoznavanje oseb (PER) in organizacij (ORG) se je najbolje izkazala Vertex AI storitev. Na splošno pa je bil v vseh področjih najboljši AWS SageMaker.

Table 6.2: Prepoznavanje imenskih entitet

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	Odzivnost	0.929	0.912	0.917	0.919	0.009
	Natančnost	0.924	0.945	0.901	0.923	0.022
	F1 ocena	0.927	0.928	0.909	0.921	0.011
Google Vertex AI	Odzivnost	0.946	0.914	0.896	0.919	0.025
	Natančnost	0.895	0.923	0.942	0.920	0.024
	F1 ocena	0.920	0.918	0.918	0.919	0.001
Amazon SageMaker.	Odzivnost	0.981	0.962	0.941	**0.961**	0.020
	Natančnost	0.962	0.980	0.921	**0.954**	0.030
	F1 ocena	0.971	0.971	0.931	**0.958**	0.023
Azure Cognitive Services	Odzivnost	0.821	0.831	0.821	0.824	0.006
	Natančnost	0.831	0.841	0.903	0.858	0.039
	F1 ocena	0.826	0.836	0.860	0.841	0.017

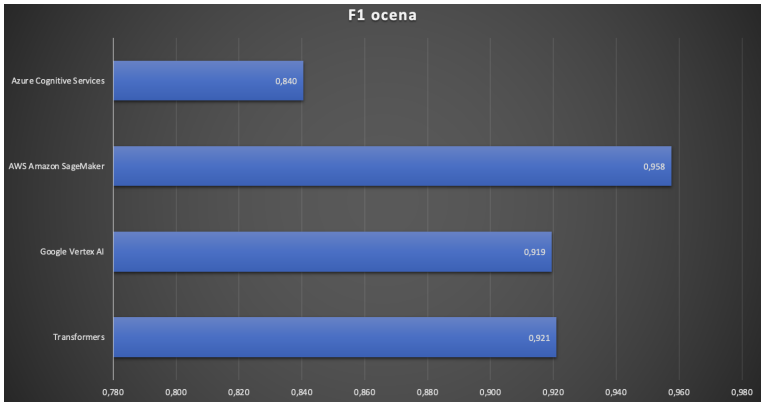

Figure 6.1: Prepoznavanje imenskih entitet F1 rezultat

Analiza napak

Najpogosteje opažene napake pri prepoznavi imenskih entitet:

1. Napake imen: To so napake v imenu imenske entitete, na primer napačna črka ali napačen zlog, zapletene besede ali tuje besede. To lahko oteži identifikacijo imena in njegovo kategorizacijo.

2. Napake v tipu: napaka v tipu imenske entitete, na primer napačno označitev osebe kot kraja ali obratno. Ime je zapleteno ali dvoumno zato, ker imajo zapletena imena lahko več kot eno pomensko področje.

3. Izpuščanje/nekategorizacija: napaka, pri katerih se imenska entiteta izpusti iz besedila.

4. Ponavljanje:napaka, pri katerih se imenska entiteta ponovi v besedilu.

6.2 Analiza sentimenta (Sentiment Analaysis)

Table 6.3: Analiza sentimenta

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	Odzivnost	0.937	0.912	0.930	0.926	0.013
	Natančnost	0.938	0.987	0.860	0.928	0.064
	F1 ocena	0.937	0.952	0.893	0.929	0.031
Vertex AI	Odzivnost	0.921	0.948	0.938	0.936	0.014
	Natančnost	0.942	0.888	0.943	0.924	0.031
	F1 ocena	0.931	0.917	0.940	0.930	0.012
Amazon SageMaker.	Odzivnost	0.901	0.882	0.891	0.891	0.010
	Natančnost	0.821	0.853	0.912	0.862	0.046
	F1 ocena	0.859	0.867	0.901	0.876	0.022
Azure Cognitive Services	Odzivnost	0.881	0.905	0.887	0.981	0.012
	Natančnost	0.852	0.884	0.851	0.862	0.019
	F1 ocena	0.866	0.894	0.869	0.876	0.015

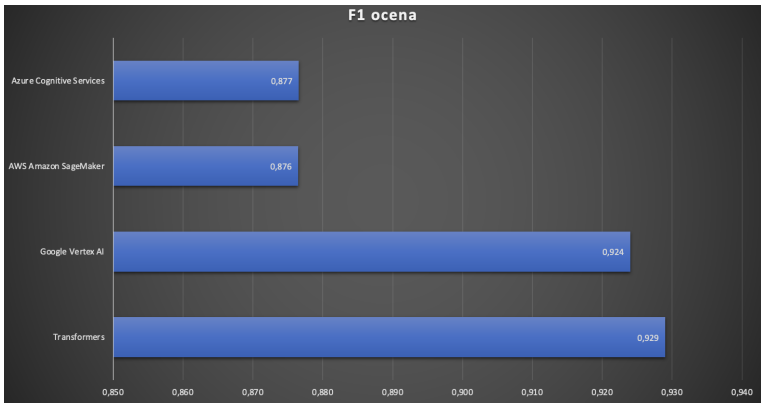

Figure 6.2: Analiza sentimenta F1 rezultat

Pri analizi imenskih entitet je bil uporabljen IMDb Reviews korpus. Za analizo sentimenta je bila najboljša Vertex AI storitev.

Analiza napak

1. Nepravilna identifikacija sentimenta: To je najpogostejša vrsta napake, ki se je pojavila pri analizi sentimenta, kar je lahko ko je ali besedna zveza je dvoumna in lahko pomeni tako pozitiven kot negativen sentiment.LAhko je tudi napaka v zapisu same besede ali besedne zveze, kot tudi da je model premalo naučen za določeno področje

2. Nepravilna kategorizacija sentimenta: Ta vrsta napake se pojavi, ko analiza sentimenta pravilno identificira sentiment, vendar ga napačno kategorizira. To lahko povzroči, da analiza sentimenta ne bo uporabna za namen, za katerega je bila namenjena.

3. Napaka v kontekstu: Ta vrsta napake se pojavi, ko analiza sentimenta pravilno identificira sentiment in ga pravilno kategorizira, vendar ga napačno razvrsti v kontekstu. To lahko povzroči, da analiza sentimenta ne bo uporabna za namen, za katerega je bila namenjena.

4. Napaka v viru: Ta vrsta napake se pojavi, ko analiza sentimenta pravilno identificira sentiment, ga pravilno kategorizira in ga pravilno razvrsti v kontekstu, vendar ga napačno dobi iz vira. To lahko povzroči, da analiza sentimenta ne bo uporabna za namen, za katerega je bila namenjena.

6.3 Povzetek (Summarisation)

Table 6.4: Povzetek

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	ROUGE-L	0.178	0.187	0.212	0.192	0.018
Vertex AI	ROUGE-L	0.312	0.291	0.315	0.306	0.013
Amazon SageMaker.	ROUGE-L	0.203	0.184	0.216	0.201	0.016
Azure Cognitive Services	ROUGE-L	0.387	0.318	0.284	0.330	0.052

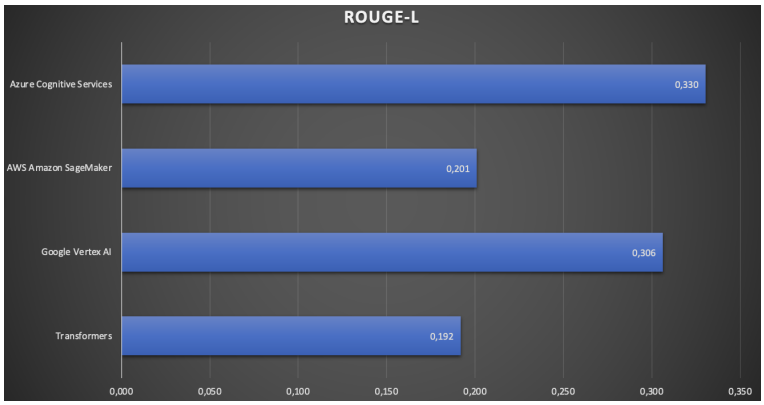

Figure 6.3: Povzetek F1 rezultat

Pri izdelavi povzetka je bil uporabljen korpus CNN/Daily Mail.

Kot vrhunska izbira za ustvarjanje povzetkov pa se je izkazala storitev Vertex AI.

Analiza napak

Najpogosteje opažene napake pri povzemanju besedila:

1. Izpuščanje pomembnih informacij: kar pomeni, da je povzetek netočen ali nepopoln. To se lahko zgodi iz več razlogov, na primer zaradi tega, da model ne prepozna pomembnih informacij ali pa ne more pravilno razumeti pomena besedila.

2. Napaka v kontekstu: Ta vrsta napake se pojavi, ko analiza sentimenta pravilno identificira sentiment in ga pravilno kategorizira, vendar ga

napačno razvrsti v kontekstu. To lahko povzroči, da analiza sentimenta ne bo uporabna za namen, za katerega je bila namenjena.

3. Napaka v viru: Ta vrsta napake se pojavi, ko analiza sentimenta pravilno identificira sentiment, ga pravilno kategorizira in ga pravilno razvrsti v kontekstu, vendar ga napačno dobi iz vira. To lahko povzroči, da analiza sentimenta ne bo uporabna za namen, za katerega je bila namenjena.

6.4 Izvleček besedne zveze (Key Phrases)

Table 6.5: Izvleček besedne zveze

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	Odzivnost	0.523	0.640	0.556	0.573	0.060
	Natančnost	0.398	0.499	0.528	0.475	0.068
	F1 ocena	0.452	0.561	0.542	0.519	0.058
Vertex AI	Odzivnost	0.499	0.541	0.589	0.543	0.045
	Natančnost	0.688	0.635	0.589	0.637	0.050
	F1 ocena	0.578	0.584	0.589	0.586	0.005
Amazon SageMaker.	Odzivnost	0.675	0.605	0.587	0.622	0.046
	Natančnost	0.520	0.492	0.526	0.513	0.018
	F1 ocena	0.587	0.543	0.555	0.562	0.023
Azure Cognitive Services	Odzivnost	0.532	0.559	0.500	0.530	0.030
	Natančnost	0.674	0.648	0.689	0.670	0.021
	F1 ocena	0.595	0.600	0.579	0.592	0.011

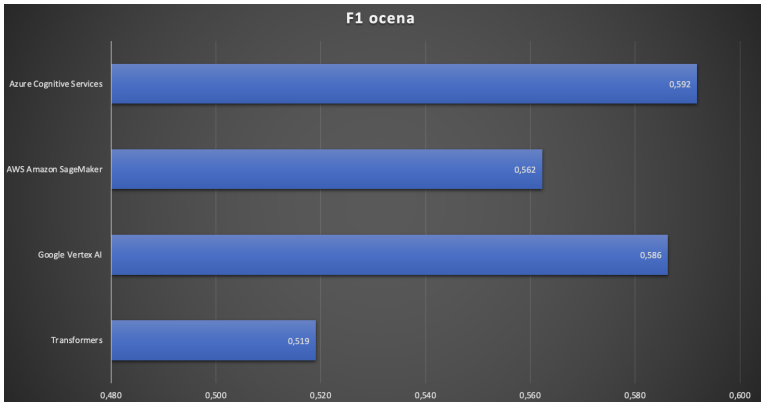

Figure 6.4: Izvleček besedne zveze F1 rezultat

Pri izvajanju naloge zvlečka besedne zveze je bil uporabljen podatkovni niz semeval-2017.

Kot najboljša rešitev za izvleček besedne zveze pa se je izkazala storitev Azure Cognitive Services.

Analiza napak

Najpogosteje opažene napake pri izvlečeku besedne zveze:

1. Napačne oznake: napake v oznaki besednih zvez, na primer napačno označitev besedne zveze kot pomembne, čeprav ni pomembna ali obratno.

2. Izpuščanje: napake, pri katerih se besedna zveza izpusti iz izvlečka, na primer zaradi napake pri prepoznavanju besednih zvez ali zaradi napake pri razdelitvi besedila na stavke.

3. Ponavljanje: To so napake, pri katerih se besedna zveza ponovi v izvlečku, na primer zaradi napake pri razdelitvi besedila na stavke ali zaradi napake pri ohranitvi besedne zveze.

6.5 Klasifikacija besedila (Text Classification)

Table 6.6: Klasifikacija besedila

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	Odzivnost	0.933	0.948	0.896	0.926	0.027
	Natančnost	0.898	0.935	0.958	0.930	0.030
	F1 ocena	0.915	0.941	0.926	0.928	0.013
Vertex AI	Odzivnost	0.842	0.901	0.844	0.962	0.034
	Natančnost	0.989	0.924	0.959	0.957	0.033
	F1 ocena	0.910	0.912	0.989	0.907	0.008
Amazon SageMaker.	Odzivnost	0.789	0.802	0.697	0.763	0.057
	Natančnost	0.879	0.799	0.895	0.858	0.051
	F1 ocena	0.832	0.800	0.784	0.808	0.024
Azure Cognitive Services	Odzivnost	0.935	0.925	0.900	0.920	0.018
	Natančnost	0.827	0.888	0.925	0.880	0.049
	F1 ocena	0.878	0.906	0.912	0.900	0.018

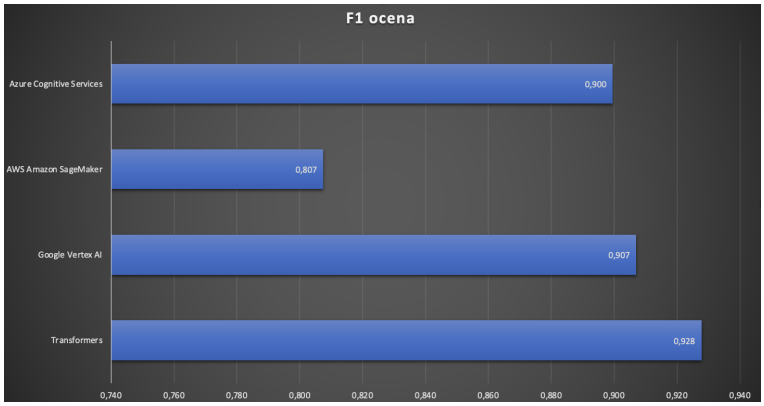

Figure 6.5: Klasifikacija besedila F1 rezultat

Pri izvajanju naloge klasifikacije besedila je bil uporabljen podatkovni niz IMDb Reviews.

Kot najboljša rešitev za naloge klasifikacije besedila pa se je izkazala storitev Transformers.

Analiza napak

Najpogosteje opažene napake pri klasifikaciji besedila:

1. Napačna klasifikacija razredov: zaradi pomanjkanja jasnih ločnic med razredi ali zaradi podobnosti med besedili različnih razredov.

2. Nezaznavanje: kadar se izpustijo pomembne informacije iz besedila, kar la povzroči, da je klasifikacija netočna ali nepopolna.

   Paragraph locked by Gregor Kocmut

6.6 Zaznava objektov (Object Detection)

Table 6.7: Zaznava objektov

		Iteracija 1	Iteracija 2	Iteracija 3	Povprečje	St. odklon
Transformers	Accuracy	0.900	0.970	0.940	0.940	0.018
Vertex AI	Accuracy	0.963	0.991	0.977	0.977	0.014
Amazon SageMaker.	Accuracy	0.995	0.963	0.982	0.980	0.016
Azure Cognitive Services	Odzivnost	0.960	0.950	0.985	0.965	0.017

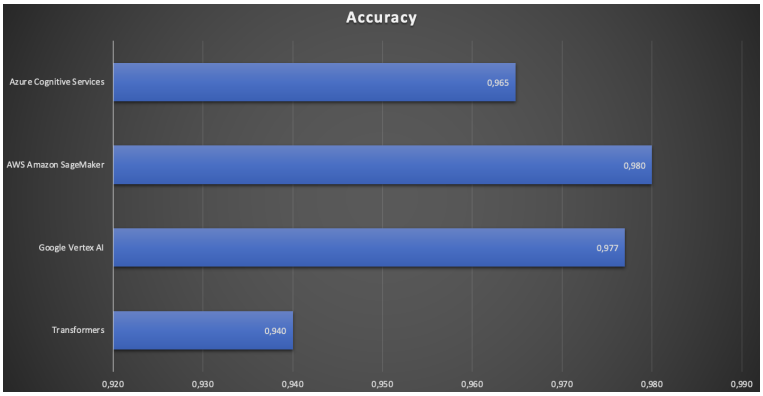

Figure 6.6: Zaznava objektov Accuracy rezultat

Pri zaznavi objektov je bil uporabljen COCO korpus.

Kot najboljša izbira za zaznavanje objektov pa se je izkazala storitev AWS SageMaker.

Analiza napak

Do napak pri zazvanju objektov prihaja zaradi različnih razlogov, na primer zaradi napak v algoritmu za zaznavanje objektov, zaradi slabe kakovosti slik ali zaradi prisotnosti motenj v okolju.

Najpogosteje opažene napake pri zaznavi objektov:

1. Napačne oznake: napake v oznaki besednih zvez, na primer napačno označitev besedne zveze kot pomembne, čeprav ni pomembna ali obratno.

2. Izpuščanje: napake, pri katerih se besedna zveza izpusti iz izvlečka, na primer zaradi napake pri prepoznavanju besednih zvez ali zaradi napake pri razdelitvi besedila na stavke.

3. Ponavljanje: To so napake, pri katerih se besedna zveza ponovi v izvlečku, na primer zaradi napake pri razdelitvi besedila na stavke ali zaradi napake pri ohranitvi besedne zveze.

Chapter 7

Zaključek

Magistrska naloga je obravnavala širok spekter področij obdelave naravnega jezika z uporabo oblačnih ter odprtokodne rešitev Hugging Face Transformers. Cilj raziskave je bil razumeti, kako se različni ponudniki odzivajo na različne izzive in naloge ter določiti njihovo uspešnost na posameznih področjih. Analiza je zajemala prepoznavanje imenskih entitet, analizo sentimenta, povzemanje besedil, luščenje ključnih besed, klasifikacijo besedila ter zaznavo objektov.

Na podlagi raziskave smo ugotovili, da se je vsak ponudnik specializiral in izkazal za najboljšega na določenem področju. Pri prepoznavanju imenskih entitet je izstopal AWS SageMaker s svojo natančnostjo. Google Cloud Vertex AI je blestel pri analizi sentimenta, kar je ključno za razumevanje čustev in mnenj v besedilu. Pri povzemanju besedila je Azure Cognitive Services pokazal najboljšo oceno, kar poudarja njegovo sposobnost za ustvarjanje povzetkov besedilnih vsebin. Pri luščenju ključnih besed je bila rešitev Azure Cognitive Services v ospredju, saj je omogočala najboljši izvleček bistvenih informacij iz besedil. Hugging Face je prevzel vodilno vlogo pri analizi sentimenta, saj ponuja široko paleto prednatreniranih modelov za natančno določanje čustvenega tona besedil. AWS pa se je izkazal kot močan v zaznavanju objektov.

Vendar pa je pomembno poudariti, da ima vsak ponudnik v oblaku svoje

prednosti in omejitve ter da izbira med njimi temelji na specifičnih potrebah in zahtevah uporabnika. Pri izbiri pravega ponudnika je treba upoštevati faktorje, kot so natančnost, hitrost, stroške, prilagodljivost in integracija z obstoječimi sistemi.

Dodati je potrebno, da so rezultati analize odvisni od specifičnih nalog in podatkovnih zbirk, ki so bili uporabljeni v tej študiji. Prav tako se tehnologije in zmogljivosti ponudnikov v oblaku nenehno razvijajo, zato je pomembno, da podjetja in raziskovalci ostanejo pozorni na nove in izboljšane metode za obdelavo naravnega jezika.

V sklepni fazi je jasno, da noben ponudnik v oblaku ne izstopa kot absolutno najboljši na vseh področjih. Različni ponudniki imajo svoje prednosti in posebne zmogljivosti glede na določene naloge obdelave naravnega jezika. Izbiro ustrezne platforme je torej smiselno prilagoditi specifičnim potrebam in zahtevam projekta. Hkrati pa je obetavno opaziti, kako odprtokodne rešitve, kot je Hugging Face Transformers, pridobivajo na veljavi in omogočajo raziskovalcem in razvijalcem, da izkoristijo najboljše iz več različnih tehnologij.

Zaključno lahko rečemo, da je obdelava naravnega jezika v oblaku zelo dinamično in obetavno področje, ki bo še naprej oblikovalo način, kako interaktiramo s tehnologijo in kako razumemo ter uporabljamo jezikovne vsebine v digitalnem svetu.

Bibliography

[1] Hugging Face. Dostopno na: https://huggingface.co/learn/nlp-course/chapter1/4 [Dostopano 10. 06. 2023].

[2] Transformers. Dostopno na: https://huggingface.co/learn/nlp-course/chapter2/1?fw=pt [Dostopano 10. 06. 2023].

[3] Google Cloud. Dostopno na: https://cloud.google.com/natural-language#section-1 [Dostopano 10. 06. 2023].

[4] Vertex AI. Dostopno na: https://cloud.google.com/vertex-ai [Dostopano 10. 06. 2023].

[5] Amazon Web Services (AWS). Dostopno na: https://aws.amazon.com/ [Dostopano 10. 06. 2023].

[6] Amazon SageMaker. Dostopno na: https://aws.amazon.com/sagemaker/ [Dostopano 10. 06. 2023].

[7] Azure. Dostopno na: https://azure.microsoft.com/en-us [Dostopano 10. 06. 2023].

[8] Azure Cognitive Services. Dostopno na: https://azure.microsoft.com/en-gb/products/cognitive-services [Dostopano 10. 06. 2023].

[9] Named Entity Recognition. Dostopno na: https://www.shaip.com/blog/named-entity-recognition-and-its-types/ [Dostopano 10. 06. 2023].

[10] Sentiment Analysis. Dostopno na: https://aws.amazon.com/what-is/sentiment-analysis/ [Dostopano 10. 06. 2023].

[11] Summarization. Dostopno na: https://huggingface.co/tasks/summarization [Dostopano 10. 06. 2023].

[12] Keyphrase Extraction. Dostopno na: https://www.geeksforgeeks.org/keyphrase-extraction-in-nlp/ [Dostopano 10. 06. 2023].

[13] Text Classification. Dostopno na: https://huggingface.co/tasks/text-classification [Dostopano 10. 06. 2023].

[14] Object Detection. Dostopno na: https://huggingface.co/tasks/object-detection [Dostopano 10. 06. 2023].

[15] True vs. False and Positive vs. Negative. Dostopno na: https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative [Dostopano10. 06. 2023].

[16] Erik F. Tjong Kim Sang and Fien De Meulder "Introduction to the CoNLL-2003". Dostopno na: https://aclanthology.org/W03-0419.pdf [Dostopano 10. 06. 2023].

[17] IMDB Dataset Reviews. Dostopno na: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews [Dostopano 10. 06. 2023].

[18] COCO 2017 Dataset. Dostopno na: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset [Dostopano 10. 06. 2023].

[19] CNN dailymail Dataset. Dostopno na: https://huggingface.co/datasets/cnn_dailymail [Dostopano 10. 06. 2023].

[20] SemEval-datasetst. Dostopno na: https://www.kaggle.com/datasets/azzouza2018/semevaldatadets?resource=download [Dostopano 10. 06. 2023].

[21] Accuracy. Dostopno na: `https://developers.google.com/machine-learning/crash-course/classification/accuracy` [Dostopano 10. 06. 2023].

[22] Precision and Recall. Dostopno na: `https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall` [Dostopano 10. 06. 2023].