

PËRPUNIMI I GJUHËS NJERËZORE PËRMES ALGORITMAVE NË PYTHON - RASTI I GJUHËS SHQIPE

Marijon PANO
Blerina ÇELIKU
Realb KUSHE

Departamenti i Informatikës,
Fakulteti i Shkencave Natyrore dhe Shkencave Humane,
Universiteti “Fan S. Noli”, Korçë

Abstrakt

Përpunimi i Gjuhës Natyrore (NLP) është një fushë e inteligjencës artificiale që shqyrton mënyrën se si kompjuterë tmund të kuptojnë dhe të përpunojnë gjuhën njerëzore. NLP përpiket të imitojë aftësitë e komunikimit të njeriut për të ndihmuar makinën të ndërveprojë me njerëzit në mënyrë natyrale dhe të kuptueshme. Ky integrim i teknologjive përfshin analizën gramatikore, interpretimin dhe gjenerimin e teksteve, sintetizimin apo klasifikimin e tyre. Kypunim synon të përdorë modelet gjuhësore të zhvilluara në Python, të trajnuara në gjuhën shqipe për të analizuar sentimentin e fjalive në gjuhën shqipe. Gjithashtu në të ilustrohen pjesë të programeve që ndihmojnë në zhvillimin e asistentëve virtuale në gjuhën shqipe (rasti ichatboteve) si dhe klasifikimine informacionit sipas tematikave (NER Recognition).

Fjalë kyçe: NER, NLP, Replit, Analizë Sentimenti. Chatbot, tokenizimi gjuhësor

NATURAL LANGUAGE PROCESSING THROUGH ALGORITHMS IN PYTHON – THE CASE OF THE ALBANIAN LANGUAGE

Abstract

Natural Language Processing (NLP) is a field of artificial intelligence that examines how computers can understand and process human language. NLP strives to mimic human communication abilities to enable machines to interact with humans in a natural and comprehensible way. This integration of technologies includes grammatical analysis, text interpretation and generation, as well as text synthesis or classification. This paper aims to use language models developed in Python, trained in the Albanian language, to analyze the sentiment of sentences in Albanian. Additionally, it illustrates parts of programs that assist in the development of virtual assistants in Albanian (e.g., chatbots), as well as the classification of information by topics (NER Recognition).

Keywords: NER, NLP, Replit, Sentiment Analysis, Chatbot, language tokenization

1. Një hyrje mbi fushën e procesimit të gjuhës njerëzore

Aplikimet e NLP synojnë zhvillimin e sistemeve që janë në gjendje të interpretojnë, analizojnë dhe përkthejnë gjuhën e shkruar dhe të folur. Qëllimi kryesor është që aplikimi të mundësojë ndërveprimin natyror ndërmjet njerëzve dhe makinave, duke përdorur gjuhën si ndërfaqe kryesore. Kjo përfshin krijimin e sistemeve që mund të kuptojnë kontekstin e bisedës, të përgjigjen në mënyrë të kuptueshme dhe të ofrojnë informacione



Figura 1: NLP si ndërfaqe në komunikimin mes individit dhe makinës

të sakta bazuar në inputin e përdoruesit. NLP synon të bëjë komunikimin me teknologjinë më të lehtë dhe më të intuitiv, duke e bërë atë të aksesueshëm për një gamë të gjerë përdoruesish dhe situatash.

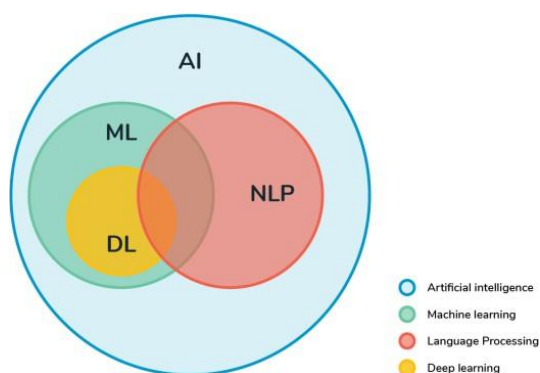
Përdorimi i NLP-së është shumë i gjerë dhe përfshin aplikime të shumta si përkthimi automatik, klasifikimi i tekstit, përvetësimi i njohurive dhe më shumë. Aplikimet e NLP-së ndihmojnë në automatizimin e shumë proceseve dhe rritin efikasitetin në menaxhimin e informacionit. Përsëmbull, përkthimi automatik mund të ndihmojë në tejkalimin e barrierave gjuhësore në komunikimin ndërkombëtar, ndërsa klasifikimi i tekstit mund të ndihmojë në organizimin dhe analizën e informacionit të madh në dokumente dhe mesazhe.

Evolucioni i NLP-së ka kaluar përmes disa fazave, duke filluar nga algoritmet e thjeshta të bazuara në rregulla deri tek modelet më të avancuara të mësimin të thellë. Ndryshimi i këtyre teknologjive është reflektuar në përmirësimin e saktësisë dhe kapacitetit të sistemeve NLP për të kuptuar kontekstin dhe semantikën e gjuhës në mënyra më të avancuara. Modelet e

mësimi të thellë, si rrjetet neuronale të shumëfishta, kanë revolucionarizuar fushën duke ofruar performancë më të mirë dhe rezultate më të sakta në analizën dhe përpunimin e teksteve.

Rëndësia e NLP-së në Biznes është shumë e madhe. Shumë kompani përdorin NLP për të analizuar feedback-un e klientëve, për të përmirësuar përvojën e përdoruesve dhe për të automatizuar procese të ndryshme të biznesit. Kjo ndihmon në krijimin e marrëdhënieve më të mira me klientët dhe në përmirësimin e shërbimeve dhe produkteve që ofrojnë. Ndikimi në Shoqëri është po ashtu i rëndësishëm. NLP ndihmon në përkthimin e dokumenteve dhe analizimin e lajmeve, duke bërë informacionin më të aksesueshëm për një audiencë më të gjerë. Po ashtu, ajo ofron shërbime të mbështetjes për klientët në shumë gjuhë dhe dialekte, duke kontribuar në përmirësimin e komunikimit dhe përkthimin në një nivel global.

Figura 2: Ndërveprimi i fushave të ndryshme për të mundësuar procesimin e gjuhës njerëzore



2. Metodatat e Përpunimit të Gjuhës Njerëzore

2.1. Tokenizim

Tokenizimi është një nga hapat e parë në përpunimin e gjuhës natyrore dhe përfshin ndarjen e tekstit në njësi më të vogla që quhen "token". Këto njësi mund të jenë fjalë, fraza, ose fjalitë, dhe ndihmojnë në ndarjen e tekstit në përbërës që mund të analizohen më lehtë. Procesi i tokenizimit është esencial për çdo analizë të mëtejshme të tekstit, pasi krijon bazën për analizimin gramatikor dhe klasifikimin e tekstit. (Bird, Klein, & Loper, 2019)

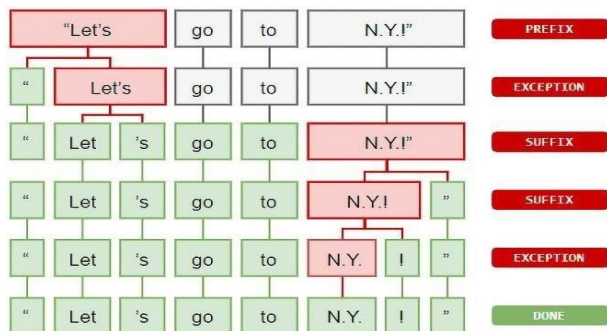


Figura3:

Tokenizimi i fjalëve dhe fjalive

Tokenizimi i Fjalëve është një teknikë që ndan tekstin në fjalë individuale, duke e bërë të mundur analizimin dhe përpunimin e secilës fjalë veçmas. Ky proces ndihmon në identifikimin e frekuencës së fjalëve dhe në krijimin e modeleve të përfaqësimit të fjalëve.

Tokenizimi i Fjalive, nga ana tjetër, ndan një tekst më të madh në fjalitë e tij përbërëse, duke lejuar një analizë më të thellë të strukturës së fjalive dhe marrëdhënieve ndërmjet tyre. Tokenizimi është një hap i rëndësishëm në përpunimin e gjuhës natyrore, pasi ndihmon në ndarjen e informacionit në mënyrë që të mund të analizohen dhe interpretohen më mire (BERT Team,2018).

2.2. Pastrimi i tekstit

Pastrimi i tekstit është një proces kyç në përpunimin e gjuhës natyrore, që përfshin disa hapa të ndryshëm për të siguruar që teksti të jetë në një format të pastër dhe të standardizuar. Heqja e Karaktereve të Panevojshme është një hap që përfshin eliminimin e simboleve dhe karaktereve që nuk kontribuojnë në analizën e tekstit, siç janë ndonjëherë simbolet speciale ose shenjat e pashqyrtuara (Brown et al., 2020).

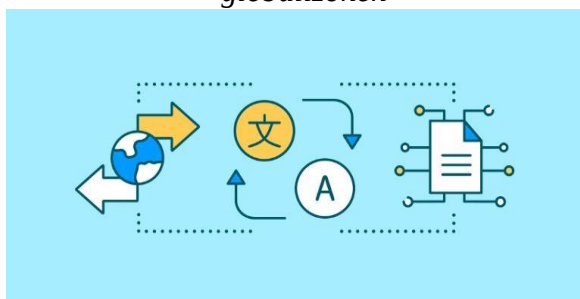
Normalizimi i Tekstit është procesi i transformimit të tekstit në një format të standardizuar, si konvertimi i të gjitha fjalëve në shkronja të vogla, që ndihmon në ruajtjen e konsistencës dhenë përmirësimin e rezultateve të analizës. Eliminimi i Ndërprerësve përfshin heqjen e fjalëvetë zakonshme që nuk ofrojnë shumë informacion për analizën, si artikujt dhe ndihmësit.

Stemming dhe Lemmatization janë procese të ndryshme për shkurtesën e fjalëve në formën e tyre themelore; stemming thjesht shkurton fjalët deri në një formë të thjeshtë, ndërsa lemmatization përfshin konvertimin e fjalëve në formën e tyre të saktë morfologjike. Këto hapa ndihmojnë në pastrimin dhe përpunimin e tekstit për analizë të mëtejshme (Goldberg, 2016).

2.3. Përkthimi automatik

Përkthimi automatik është procesi i përkthimit të tekstit nga një gjuhë në një tjetër përmes përdorimit të algoritmeve dhe modeleve të NLP-së. Teknikat e Përkthimit përfshijnë përdorimin e modeleve të mësimit të thellë dhe të rjetit neuronale për të siguruar që përkthimi të jetë i saktë dhe i natyrshëm

Figura 4: Përkthimi automatik i ndihmon platformat E-Commerce të globalizohen



3. Analiza e sentimentit

3.1. Çfarë përfshin kjo disiplinë?

Analiza e sentimentit është procesi i vlerësimit të ndjenjave të shprehura në një tekst dhe ndihmon në përcaktimin e tonit të përgjithshëm si pozitiv, negativ, ose neutrale. Ky process përfshin përdorimin e metodave të mësimit të makinerisë dhe modeleve të avancuara për të analizuar tekstin dhe për të identifikuar ndjenjat që shprehen në të. Aplikimet Praktike të analizës së sentimentit janë të shumta dhe përfshijnë analizën e feedback-ut të klientëve, monitorimin e reputacionit të markës, dhe identifikimin e tendencave në rrjetet sociale. Megjithatë, sfida në analizën e sentimentit përfshin identifikimin e ironisë dhe konteksteve që mund të ndërlikojnë përcaktimin e saktë të sentimentit, pasi ndjenjat mund të shprehen në mënyra të ndryshme dhe ndonjëherë të jenë të fshehura pas një gjuhe të

ndërlikuar (Devlin, Chang, Lee, & Toutanova, 2018).

3.2. Repl-it, mjedisi i punës për ekzekutimin e skriptave në Python për NLP.

Replit është një platformë zhvillimi në internet që ofron një mjedis të integruar për kodimin dhe testimin e kodit në një shumëllojshmëri gjuhësh programimi.

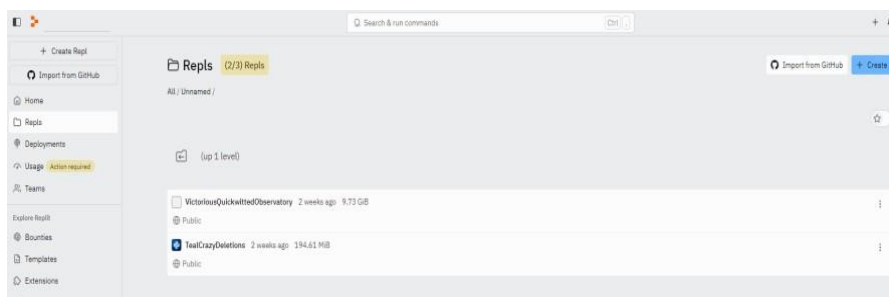
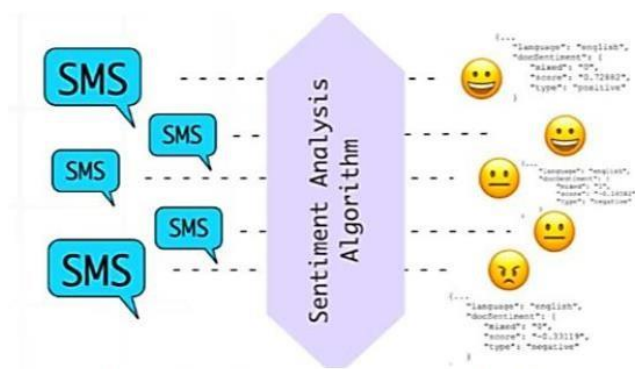


Figura 5: Repositori i Replit.

3.3. Shkalla e vlerësimit të sentimentit

Shkalla e vlerësimit të sentimentit, është një mënyrë për të matur intensitetin ose forcën e sentimentit të shprehur në një tekst. Kjo shkallë ofron një vlerësim më të hollësishëm të ndjenjave që përmban një tekst (Klein et al, 2017).

Figura 6: Algoritma të analizës së sentimentit synojnë identifikimin e ndjesive brenda fjaltive.



Sentimenti mund të jetë: Pozitiv: Kur një tekst shpreh ndjenja të mira, kënaqësi, mirënjohje ose optimizëm. Për shembuj, komente si "Shërbimi

ishte i shkëlqyer" ose "Unë jam shumë i kënaqur me këtë produkt" janë të klasifikuara si pozitive. Negativ: Kur një tekst shpreh ndjenja të këqija, pakënaqësi, kritikë ose dëshpërim. Për shembuj, komente si "Ky produkt është shumë i keq" ose "Kam pasur një përvojë të tmerrshme" janë të klasifikuara si negativeNeutral: Kur një tekst është i paanshëm dhe nuk shpreh ndjenja të forta pozitive ose negative. Për shembuj, komente si "Ky produkt është në dispozicion" ose "Sot është një ditë e zakonshme" janë të klasifikuara si fjali neutral.

3.4. Skripti për analizimin e sentimentit të fjalive në shqip.

Në Replit, ne mund të përdorim Python për të ekzekutuar kodin dhe për të instaluar paketat e nevojshme. Për të kryer analizën e sentimentit, ne do të përdorim bibliotekat transformers për modelet e trajnuara dhe matplotlib për vizualizimin. Këto hapa janë thelbësorë për të siguruar që ambienti ynë të jetë i gatshëm për të realizuar analizën dhe për të krijuar grafikë të qarta. Libraria “transformers” instalohet përmes komandës: `pip install`



```
~/VictoriousQuickwittedObservatory$ pip install transformers torch
Requirement already satisfied: transformers in ./pythonlibs/lib/python3.11/site-packages (4.44.2)
Requirement already satisfied: torch in ./pythonlibs/lib/python3.11/site-packages (2.4.0)
Requirement already satisfied: filelock in ./pythonlibs/lib/python3.11/site-packages (from transformers) (3.15.4)
Requirement already satisfied: huggingface-hub<1.0,>=0.23.2 in ./pythonlibs/lib/python3.11/site-packages (from transformers) (0.24.6)
Requirement already satisfied: numpy>=1.17 in ./pythonlibs/lib/python3.11/site-packages (from transformers) (2.1.0)
Requirement already satisfied: packaging>=20.0 in ./pythonlibs/lib/python3.11/site-packages (from transformers) (24.1)
Requirement already satisfied: pyyaml>=5.1 in ./pythonlibs/lib/python3.11/site-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in ./pythonlibs/lib/python3.11/site-packages (from transformers) (2024.7.24)
```

transformers torch

Figura7: Shkarkimi dhe instalimi i "transformers" përmes Shell

Në fillim të kodit, importohet funksionaliteti i nevojshëm nga bibliotekat transformers dhe matplotlib:

```
from transformers import pipeline
import matplotlib.pyplot as plt

sentiment_analyzer=pipeline('sentiment-analysis',model='nlpTown/bert-base-multilingual-uncased-sentiment')
```

Ky model është një model shumëgjuhësh BERT, i cili është trajnuar për të analizuar sentimentin në disa gjuhë, përfshirë shqipen. Një listë e teksteve në gjuhën shqipe është përcaktuar për të analizuar:

```
texts = [ "Ky është një film shumë i bukur!",
          "Nuk më pëlqeu kjo ngjarje.",
          "Shumë e këndshme dhe e ndihmueshme.",
          "I keq, nuk do të kthehem më këtu."]
```

Këto tekste përfaqësojnë mendimet e ndryshme për një film dhe ngjarjedhe do të analizohen përtëpërcaktuar sentimentin e tyre. Fjalitë e dhëna tek analizuesi gjuhësor janë të ngjyrave të ndryshme emocionale, nga ato që shprehin emocione të forta pozitive, si për shembull *“Ky film është një film i shkëlqyer!”*, tek ato të cilat shprehin emocione të forta negative *“Nuk më pëlqeu fare kjo ngjarje. E urrej kujtimin rreth saj!”*.

Analiza në Python, përmes librarive përkatëse të NLP, i vendos sentimentet e analizuara të fjalive në një shkallë dhjetore, nga 0.0 deri tek 1.0, ku 0.0 përfaqëson një emocion të fortë negativ, ndërsa 1.0 përfaqëson një emocion të fortë pozitiv.

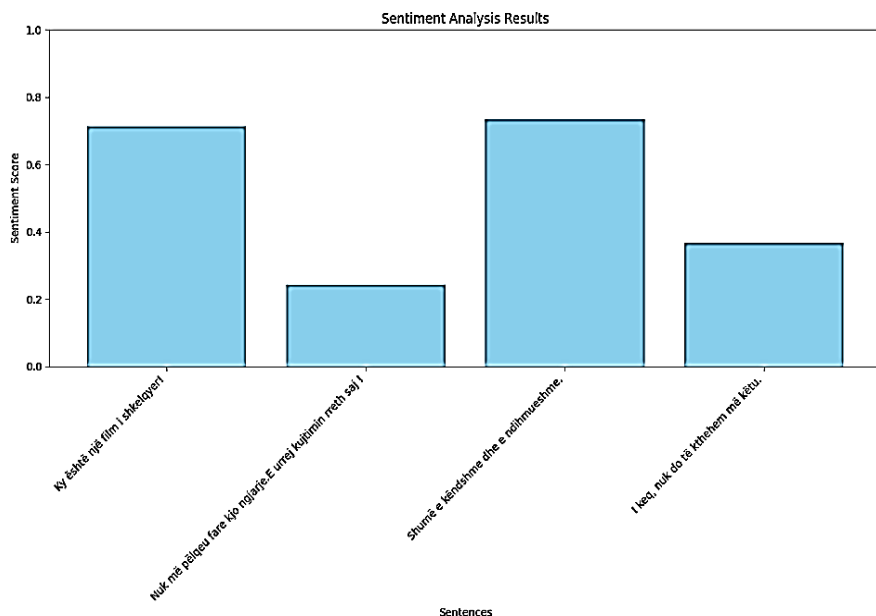


Figura 8: Rezultatet grafike të analizës së sentimenteve të fjalive në s hqip.

4. Chatbot-et në gjuhën shqipe

Chatbotet kanë një gamë të gjerë përdorimesh. Ato janë të njohura për shërbimin ndaj klientëve, ku ndihmojnë përdoruesit në zgjidhjen e problemeve, ofrojnë informacion mbi produkte dhe shërbime, dhe ndihmojnë në procesin e blerjes. Gjithashtu, ato përdoren në aplikacione të ndryshme për të ofruar informacion të shpejtë dhe të saktë. Chatbotet ndahen në dy kategori kryesore.

Të parat janë chatbotet me rregulla (rule-based), të cilat ndjekin një set të caktuar rregullash dhe ofrojnë përgjigje të paracaktuara në bazë të inputeve të përdoruesve. Të dytat janë chatbotet që përdorin inteligjencën artificiale (AI), të cilat janë në gjendje të mësojnë nga



ndërveprimet e tyre me përdoruesit dhe të përmirësojnë përgjigjet e tyre me kalimin e kohës.

Figura 9: TOBI dhe RONA- dy asistentë virtuale në gjuhën shqipe.

Në vijim jepet një pjesë

kodpërkrijimit të një chatboti të thjeshtë në gjuhën shqipe.

```
class AlbanianChatbot: def init (self):  
  
    self.model_name = 'microsoft/DialogPT-medium'  
    self.tokenizer=AutoTokenizer.from_pretrained(self.model_name)  
    self.model=AutoModelForCausalLM.from_pretrained(self.model_name)  
    self.chat_history_ids = torch.tensor([]).long()  
    self.chat_history = []  
  
    def respond (self, user_input):  
  
        new_user_input_ids = self.tokenizer.encode(user_input +  
        self.tokenizer.eos_token, return_tensors='pt') chat_output =  
        self.model.generate( self.chat_history_ids, max_length=1000,  
        pad_token_id=self.tokenizer.eos_token_id,top_k=50, top_p=0.95,  
        do_sample=True)
```

```
chat_response=self.tokenizer.decode(chat_output[:self.chat_history_ids.shape[-1]:][0], skip_special_tokens=True) self.chat_history_ids = chat_output

return chat_response
```

Figura 10: Pjesë nga kodi i ChatBotit.

Konkluzione

Përpunimi i Gjuhës Njerëzore (NLP) për gjuhën shqipe, duke përdorur teknologjitë e Inteligjencës Artificiale (AI) dhe mjetet e zhvillimit në Python krijon mundësi të shumta për avancimin e bizneseve. Ky proces synon të lehtësojë analizën dhe përpunimin e të dhënave tekstuale në shqip, me aplikacione që përfshijnë analizën e sentimentit, identifikimin e entiteteve të veçanta dhe përkthimin automatik. Modelet e mësimi të rrjetave neurale si BERT dhe GPT, ofrojnë një bazë të fuqishme për trajtimin e gjuhës shqipe. Këto modele janë trajnuar me datase të mëdha teksti dhe janë të afta të përballen me detyra të ndryshme.

Python është një gjuhë programimi e njohur dhe e fuqishme për zhvillimin e aplikacioneve NLP, duke ofruar biblioteka si spaCy, NLTK, dhe Hugging Face Transformers. Këto biblioteka ofrojnë mjete dhe modele të gatshme që ndihmojnë në analizën sintaksore dhe semantike të tekstit, si dhe në detyra të tjera të përpunimit të gjuhës. SpaCy, për shembull, ofron mbështetje për analizën gramatikore dhe morfologjike, ndërsa Hugging Face Transformers ofron një gamë të gjerë modelesh për identifikimin e entiteteve të veçanta (NER) dhe aplikacione të tjera. Replit, si një platformë e fuqishme për zhvillimin në re, lehtëson krijimin dhe testimin e aplikacioneve në Python. Ajo ofron një ambient të thjeshtë dhe të aksesueshëm për zhvilluesit, duke eliminuar nevojën për konfigurime komplekse të mjedisit të zhvillimit. Kjo e bën Replit një mjet të vlefshëm për hulumtime dhe eksperimente në përpunimin e gjuhës shqipe, duke inkurajuar përdoruesit të eksplorojnë dhe të zhvillojnë projekte inovative.

Megjithatë, përkundër përparimeve të dukshme, përpunimi i gjuhës shqipe përmes AI ka ende disa sfida të rëndësishme. Gjuha shqipe ka nevojë për modele të personalizuar që mund të kuptojnë nuancat dhe kontekstin kulturor të gjuhës shqipe. Pjesëmarrja e komunitetit dhe përpjekjet për të krijuar më shumë burime të hapura janë thelbësore për të përmirësuar këtë fushë.

Në përfundim, metodat dhe teknologjitë për përpunimin e gjuhës njerëzore në shqip, përmes AI, Python dhe Replit, ofrojnë mundësi të

jashtëzakonshme për të avancuar kuptimin dhe analizën e teksteve në këtë gjuhë. Ndërsa përparimi në këtë fushë është premtues, është e nevojshme të vazhdohet me përpjekjet për të kapërcyer sfidat dhe për të zhvilluar zgjidhje më të avancuara që do të mundësojnë një përpunim më të saktë dhe të hollësishëm të gjuhës shqipe. Angazhimi në këto teknologji dhe përdorimi i mjeteve të disponueshme janë hapa të rëndësishëm për të kontribuar në zhvillimin e mëtejshëm të NLP-së për gjuhën shqipe.

Referenca

1. Bird, S., Klein, E., & Loper, E. (2019). Natural language processing with Python: Analyzing text with the natural language toolkit. O'Reilly Media.
2. BERT Team. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.
3. Brown, T. B., Mann, B., Ryder, N., & Subbiah, M. (2020). Language models are fewshot learners. arXiv. <https://arxiv.org/abs/2005.14165>
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). Association for Computational Linguistics.
5. Goldberg, Y. (2016). A primer on neural network models for natural language processing. Journal of Artificial Intelligence Research, 57, 345-420. <https://doi.org/10.1613/jair.5104>
6. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In Proceedings of the 3rd International Conference on Learning Representations (pp. 1- 15). <https://arxiv.org/abs/1412.6980>
7. Klein, G., Kim, Y., Deng, Y., & Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 67-72). Association for Computational Linguistics.