TARTU ÜLIKOOL

Arvutiteaduse instituut

Informaatika õppekava

Kristina Katarina Kaljumäe

Suurte keelemudelite kasutamine sõnade semantiliseks klassifitseerimiseks

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur DSc. (Tech)

Tartu 2025

Suurte keelemudelite kasutamine sõnade semantiliseks klassifitseerimiseks

Lühikokkuvõte:

Keele uurimiseks ja keeletehnoloogiliste rakenduste loomiseks on oluline omada võimalikult palju teavet sõnade tähenduse kohta. Üheks oluliseks aspektiks on sõna kuuluvus teatud semantilisse klassi. Siiski puudub hetkel tõhus süsteem selliste klasside märgendamiseks. Manuaalne märgendamine on ajamahukas ning automaatseid lahendusi veel ei eksisteeri. Käesolevas uurimistöös kasutati erinevaid GPT-mudeleid, et hinnata nende sobivust sõnade semantiliseks märgendamiseks. Esiteks määrati mudelite abil sõnade semantilised alamklassid ning nende põhjal püüti järeldada, kas sõna kuulub üldisemasse kategooriasse „füüsiline koht“. Samuti uuriti, kas sõna tähendus sõltub selle esinemise kontekstis. Töö tulemused näitasid, et alamklasside kasutamine füüsilise koha määramiseks on paljulubav, kuid nõuab täpsemat päringute koostamist. Lisaks ilmnes, et sõna kuulumine semantilisse klassi sõltub pigem sõna enda tähendusest kui selle esinemise kontekstis.

Võtmesõnad: loomuliku keele töötlus, suured keelemudelid, semantilised klassid, korpused

CERCS: P176 Tehisintellekt

Using large language models for semantic word classification

Abstract:

To study language and develop language technology applications, it is essential to have as much information as possible about word meanings. One important aspect of meaning is if a word belongs to a specific semantic class. However, there is currently no effective system for annotating such classes. Manual annotation is time-consuming, and automated solutions do not exist yet. This study employed various GPT models to evaluate their suitability for semantic annotation of words. First, the models were used to determine the semantic subclasses of words, and based on these, the study attempted to determine whether a word belongs to the broader category of “physical location”. Additionally, the research examined whether a word’s meaning depends on its context of use. The results showed that using subclasses to identify physical locations is promising but requires more precise prompt

engineering. Furthermore, it was found that a word's semantic class is influenced more by its inherent meaning than by the context in which it appears.

Keywords: natural language processing, large language models, semantic classes, corpora

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus

Märgendamine on oluline protsess, mis toetab keeleteadlaste tööd ning võimaldab arendada keeletehnoloogilisi rakendusi. Eesti keeles on senini märgendatud andmeid kuni süntaktilise tasandini. Selles etapis on sõnu liigitatud muu hulgas asjaolumäärusteks, mis võivad väljendada kohta, aega, hulka, viisi või seisundit [1].

Keele paremaks mõistmiseks on vajalik edasi märgendada neid üksusi ka semantiliselt. Asjaolumääruste semantiline märgendamine on aga keerukas, sest näiteks lausetes „Ma pihtisin isale“ ja „Ma jooksin väljakule“ on vormiliselt sarnased väljendid „isale“ ja „väljakule“ tähenduslikult erinevad. Selliste nüansside eristamine reeglite põhiselt on keeruline, mistõttu automaatsete märgendajate loomine, olgu reeglitele tuginedes või manuaalsel märgendusel treenitud masinõppe mudeleid kasutades, on seni olnud ressursimahukas ning piiratud eduga. Alternatiiviks on täielik manuaalne märgendamine, kuid see on äärmiselt ajakulukas.

Käesolevas töös uuritakse, kas suurte keelemudelite rakendamine võib aidata lihtsustada asjaolumääruste semantilist märgendamist. Töö keskendub kohamääruste alla kuuluvatele väljenditele, mis tähistavad füüsilisi kohti, ning toetab Kertu Sauli doktoritöö uurimisvaldonda. Eesmärk on analüüsida, kas keelemudelid suudavad eristada, kas antud sõna viitab füüsilisele kohale või mitte. Selle uurimiseks koostati mudelitele kahte tüüpi päringud (ingl *prompt*). Esimesel juhul hindas mudel sõna tähendust ilma kontekstita, teisel juhul lisati päringule ka kontekst, milles sõna esines, et analüüsida konteksti mõju märgendamise täpsusele.

Töö esimeses peatükis antakse ülevaade teoreetilisest taustast. Lähemalt käsitletakse märgendamise protsessi, tekstikorpuseid, millele tugineti, hindamismõõdikuid ning suuri keelemudeleid, keskendudes eeskätt GPT-põhistele mudelitele, mida kasutati eksperimentides. Teises peatükis kirjeldatakse tööprotsessi, esitatakse uurimishüpoteesid ning selgitatakse metoodikat ja andmestike koostamist. Kolmandas peatükis analüüsitakse ja tõlgendatakse saadud tulemusi, võrreldes erinevate GPT-mudelite täpsust olukordades, kus sõnu märgendati ilma kontekstita ja kontekstiga.

1. Teoreetiline ülevaade

Käesolevas peatükis antakse ülevaade kolmest teemavaldkonnast, mis on olulised töö taustaks. Esiteks selgitatakse märgendamise mõistet ning kirjeldatakse, kuidas semantilisi klasse on antud töös määratletud ja märgendatud. Teiseks tutvustatakse tekstikorpuste olemust ning käsitletakse täpsemalt neid konkreetseid korpuseid, mis on töö aluseks. Kolmandaks antakse ülevaade suurtest keelemudelitest, keskendudes eelkõige GPT-mudelitele, mida on selles töös rakendatud.

1.1 Märgendamine

Teksti märgendamine on keeletehnoloogias vajalik protsess, et antud teksti objektiivselt analüüsida ning erinevaid keelenähtusi uurida [2, 3]. Märgendamine on alustekstile lisainfo lisamine [4]. Info lisamine võib toimuda käsitsi, automaatselt arvutiprogrammi poolt, või olla mõlemast kombineeritud protsess, kus inimene kontrollib käsitsi üle arvutiprogrammi märgendamise [2]. Info, mida märgendamisel tekstile lisatakse oleneb sellest, mis tüüpi märgendamisega on tegu [4]. Enim levinud märgendustasand eesti keeles on morfoloogiline märgendamine [2, 4]. Morfoloogilise märgendamise puhul lisatakse igale tekstis leiduvale sõnale info tema algvormi, sõnaliigi ja sõna grammatilise vormi kohta [2, 4, 5]. Populaarsust kogub ka lisaks morfoloogilistele märgendustele tekstile süntaktiliste märgendite lisamine ehk tekstile süntaktilise info lisamine [2, 4]. Antud töös keskendatakse semantilisele märgendamisele ehk sõnadele info lisamisele nende tähenduse põhjal [4].

1.1.2 Semantiline märgendamine

Eelnevalt mainitud morfoloogiline ja süntaktiline märgendamine on vajalikud, et mõista teksti struktuuri. Semantiline märgendamine keskendub sõnade tähendusele, seega võimaldab see analüüsida teksti sisu. Seda meetodit on varasemates uurimustes kasutatud näiteks riikidevaheliste kultuuriliste, sotsiaalsete ja ajalooliste erinevuste ja/või sarnasuste analüüsimiseks ning inimeste suhtumise uurimiseks tragöödiatesse [6, 7].

Harilikult kasutatakse sellistes uurimustes automaatset märgendamist, sest see aitab hoida märgendamise võimalikult objektiivsena, kuna märgendamine ei põhine uurijate enda subjektiivsel tõlgendusel sõnadest [6, 7]. Selle lähenemise negatiivne külg on, et tihti tuleb automaatse märgendaja leidmiseks uurijatel ise mudeleid treenida, et tagada võimalikult hea

täpsus märgendamisel ja see protsess võib osutuda väga kulukaks [8]. Suurte keelemudelite kasutamine semantiliseks märgendamiseks võiks pakkuda lahenduse sellele probleemile, vähendades treenimise vajadust ja sellega seotud kulusid.

1.1.3 Nimisõnade semantilised klassid

Selles töös märgendatakse, kas nimisõna tähistab füüsilist kohta. Selle otsuse tegemiseks on vaja määrata ka kitsamad semantilised klassid, mis aitavad eristada füüsilisi kohti teistest tähenduslikest rühmadest. Käesolevas peatükis selgitatakse, mida mõistetakse *koha* all ning milliseid semantilisi klasse kasutatakse märgendamisel.

Keeleteaduses on koht tavaliselt lauses nimisõnafraas, mis kuulub kohamääruse koosseisu ja väljendab tegevuse kohta lauses [1]. Koht võib funktsioneerida sihtkohana, asukohana, lähtekohana, liikumisteena või piirina [1]. Näiteks lauses *"Ma jõudsin koju"* on nimisõna *"kodu"* koha tähendusega.

Kohtade semantika ei piirdu vaid füüsiliste ruumidega. Nimisõnad, mis täidavad lauses koha rolli, võivad viidata ka [1]:

- kollektiividele (*"Ta töötab riigikogus"*),
- sündmustele (*"Käisin kontserdil"*),
- abstraktsetele mõistetele (*"Ta on tugev matemaatikas"*).

Selles töös keskendutakse aga eelkõige nimisõnadele, mis tüüpiliselt viitavad füüsilisele asukohale (nt *"maja"*, *"mets"*, *"linn"*). Nimisõnu, mis suurema osa juhtudest ei tähista füüsilist asukohta, ei käsitleta siin "kohtadena", isegi kui nad mõnes kontekstis seda rolli täidavad.

Selleks et täpsemalt määratleda, milliseid sõnu käesolevas töös käsitletakse kohtadena, koostati 14 semantilist alamklassi. Neist seitse viitavad sellele, et sõna tähistab enamikes kontekstides füüsilist kohta; viis klassi viitavad vastupidiselt sellele, et sõna ei tähista füüsilist kohta. Üks klass kajastab juhtumeid, kus sõna tähendust ei saa ilma kontekstita

üheselt määrata, ning üks klass hõlmab sõnu, mille tähendus jääb ebaselgeks – enamasti seetõttu, et tegemist ei ole eestikeelsete sõnadega.

Kõik alamklassid on esitatud tööprotsessi peatükis, kuid siin on näide ühest. Kui objekti, mida sõna tähistab saab inimene sisse minna (nt *„koolimaja“*, *„buss“*), loetakse seda selles töös füüsiliseks kohaks. Alamklasside määratlemiseks ja märgendamiseks kasutati käesolevas töös manuaalset märgendamist, kuna vastavat ülesannet täitvaid eestikeelseid automaatseid tööriistu ei ole praegu olemas.

1.2 Tekstikorpused

Märgendatavad sõnad on tavaliselt osa ulatuslikumatest tekstidest. Selliseid tekstikogumeid nimetatakse korpusteks. Järgnev lõik tugineb Zaki Abdulfattah artiklile [9], mis väidab, et tekstikorpused on mahukad tekstikogud, mis sisaldavad suures koguses nii kirjutatud kui ka räägitud keelt. Korpused koostatakse sageli eesmärgiga kasutada neid keele uurimiseks või keeletehnoloogilistes rakendustes, mistõttu on oluline, et need oleksid masinloetavad. Korpusi liigitatakse üldkorpusteks ja spetsiifilisteks korpusteks. Üldkorpuste eesmärk on hõlmata keele võimalikult paljusid aspekte ning pakkuda mitmekülgset ülevaadet keelest. Spetsiifilised korpused keskenduvad seevastu mõnele kindlale keeleosale, pakkudes sellest detailset ülevaadet. Mõlemat tüüpi korpustes kasutatakse enamasti elektrooniliselt ligipääsetavaid kirjalikke tekste, kuna neile on lihtsam ligipääs.

1.2.1 Eesti keele koondkorpus

Antud töös tehtavate märgenduste aluseks on võetud Eesti keele koondkorpus. Järgnev materjal on refereeritud Eesti keele koondkorpuse veebilehelt [10]. Korpuse loomise eesmärgiks oli koostada võimalikult mahukas ja mitmekesine eesti keele korpus, mida saaks kasutada arvutilingvistika ja keeleteaduslike statistiliste analüüside tarbeks. Korpus koosneb terviklikest tekstidest, mis on võetud kirjalikust keelekasutusest. Iga korpuses sisalduva sõna kohta on esitatud teave selle algvormi ning morfoloogiliste kategooriate kohta. Suurem osa tekstidest pärineb erinevatest ajakirjadest. Lisaks sisaldab koondkorpus ka teadus- ja seadustekste. Ilukirjandust esineb korpuses vähesel määral, peamiselt autoriõigustega seotud piirangute tõttu.

1.2.2 Verbifraaside andmebaas

Eesti keele koondkorpuse põhjal on Katrin Tsepelina ja Sven Laur koostanud andmebaasi [11], mida kasutatakse käesolevas töös edasiste andmestikkude loomiseks. Andmebaas keskendub Eesti keele koondkorpusest kogutud verbidele ja nende alluvatele. Täpsemalt vaadatakse lihtverbe (nt kirjutama) ning ühendverbe (nt kirjutama maha). Välja on jäetud verbid, mis olid umbisikulised ning millel puudusid alluvad. Andmebaasi jäeti alles ainult verbid, mis on lihtminevikus (nt jooksis), olevikus (nt jookseb) või täisminevikus (nt on jooksnud). Andmebaas koosneb kahest tabelist.

- **transaction_head** — sisaldab 30078992 rida ja annab ülevaate andmebaasis esinevatest verbidest.

- **transaction_row** — sisaldab 54050499 rida ja keskendub verbi alluvatele ehk verbidega seotud fraasidele või sõnadele.

Lisaks põhikorpusele on loodud täiendav andmebaas, mis talletab kõik laused, milles eelmainitud andmebaasis sisalduvad sõnad esinevad. Kokku on andmebaasis 16120745 lauset. See võimaldab analüüsida sõnade tähendusi ka kontekstis, mitte ainult isoleeritult.

1.3 Suured keelemudelid

Suured keelemudelid on mudelid, mis hetkel kasutavad närvivõrkudel põhinevat transformer-arhitektuuri (ingl *transformer architecture*), mis koosneb suurtel tekstikorpustel eeltreenitud kümnetest kuni sadadest miljarditest parameetritest [12]. Parameetrid on treenimise käigus saadud kaalud, mille põhjal mudel ennustusi teeb [13]. Suurem parameetrite hulk vähendab suure keelemudeli poolt tehtud semantilisi vigu ning parandab mudeli oskust lahendada ülesandeid [14]. Selline ülesehitus on vajalik, et võimalikult täpselt ennustada sõnade, lausete või isegi paragraafide tõenäosust antud kontekstis, kus keelemudelid kasutavad seda tõenäosuslikku hinnangut, et genereerida kõige asjakohasem väljund vastavalt päringule [13]. Käesolevas töös on oluline, et keelemudel on juhendatav ehk, et keelemudelile saab esitada päringu, mille tulemusel sisendit semantiliselt analüüsitakse.

1.3.2 GPT perekond

Suured keelemudelid jaotuvad perekondadeks, millest tuntumad on GPT, LLaMA, Claude, Qwen ja DeepSeek. Käesolevas töös keskendatakse GPT perekonnale selle tuntuse ja kerge kättesaadavuse tõttu. GPT ehk generatiivse eeltreenitud transformeri (ingl *generative pre-trained transformer*) perekond on OpenAI poolt arendatud kogum keelemudelitest [12].

Selles töös kasutati OpenAI mudeleid GPT-3.5 Turbo, GPT-4 ja GPT-4o, mis olid kättesaadavad Azure platvormi kaudu. Lisaks rakendati ka OpenAI ametlikku veebiliidest, mille kaudu kasutati GPT-4o põhist ChatGPT-d. Azure platvorm võimaldas päringuid teostada API kaudu tänu, millele sai mugavalt esitada mudelitele suurtes kogustes päringuid. Samas pakkus ChatGPT liides võimalust üksikute päringute kiireks testimiseks ja tulemuste vahetuks hindamiseks.

GPT-mudelid genereerivad väljundi nendele kasutaja poolt antud päringu põhjal [12]. Seega GPT-mudelitele erinevate päringute andmine, isegi kui päringu põhimõte on sama, tagab keelemudelitelt teisiti sõnastatud vastuse. Selle tõttu on parima vastuse saamiseks oluline koostada läbimõeldud päring ning katsetada näiteks sünonüümide kasutamist. Päringut on võimalik ka arendada (ingl *prompt engineering*) ehk lisada päringusse täpsemaid juhendeid, mis muudavad vastuse struktuuri [12].

Järgnev lõik on kirjutatud tuginedes Daniel Jurafsky ja James H. Martini raamatule [15]. Viise kuidas päringuid arendada on mitmeid, kuid üheks näiteks oleks väheste näidetega päringute koostamise (ingl *few-shot prompting*) eelistamine üle ilma näideteta päringute koostamise (ingl *zero-shot prompting*). Väheste näidetega päringute koostamine tähendab, et mudelile antavale päringule lisatakse näited oodatud tulemusest lisaks juhendile endale. Lisatud näidete arv ei pea olema üldse suur, pigem on oluline, et vähemalt üks näide eksisteeriks. Oluline on arvestada ka kontekstipõhise õppimisega (ingl *in-context learning*), mille puhul mudel kasutab vastuse genereerimisel ära eelnevaid samas vestluses esitatud päringuid. Selline konteksti kasutamine võimaldab mudelil paremini mõista kasutaja kavatsusi ning pakkuda sisukamaid ja täpsemaid vastuseid.

Käesolevas töös loodi iga päringu jaoks keelemudeliga uus vestlus, et vähendada varasemate päringute mõju tulemustele. Samuti ei antud mudelile kaasa ühtegi näidet. Tulevikus saaks proovida ka näidete lisamist päringule, et hinnata, kas see parandab vastuste kvaliteeti.

Varasemalt kirjeldatud päringud on olnud kasutaja päringud (ingl *user prompt*), mida esitab kasutaja keelemudelile [16]. Need päringud on tavaliselt selged küsimused või juhised, et anda mudelile teada, millist vastust või tegevust oodatakse [16]. Kasutaja päringud võivad ulatuda lihtsatest küsimustest, nagu „Mis on Eesti pealinn?“, kuni keerukamate ülesanneteni, näiteks „Kuidas kirjutada teadusartikli kokkuvõtet?“.

Lisaks kasutajapäringutele on võimalik keelemudeliga suhtlemisel kasutada ka süsteemi päringuid (ingl *system prompts*), mida esitatakse tavaliselt API kaudu [16]. Süsteemi Päringud on spetsiaalsed juhised, mis võivad mõjutada või määrata, kuidas mudel kasutaja esitatud küsimusele vastab [16]. Need päringud on kõrgema prioriteediga kui kasutaja päringud ning võimaldavad suunata mudeli käitumist või vastuste vormi [16]. Näiteks võib süsteemi päringuks olla juhis „Hoia vastused alla 50 sõna“ või „Vasta kui Itaalias töötav kokk“. Sellisel juhul, kui kasutaja esitab küsimuse „Seleta, kuidas valmistada maitsev pastaroog“, suunab süsteemi päring mudelit vastama lühidalt ning itaalia köögi baasil.

1.3.3 ChatGPT

ChatGPT on juturobot, mis töötab GPT perekonda kuuluva GPT-4o põhjal [12, 17]. ChatGPT suudab aidata kasutajal teha mitmeid ülesandeid nagu näiteks leida vastuseid küsimustele või teha pikkadest tekstidest kokkuvõtteid [12].

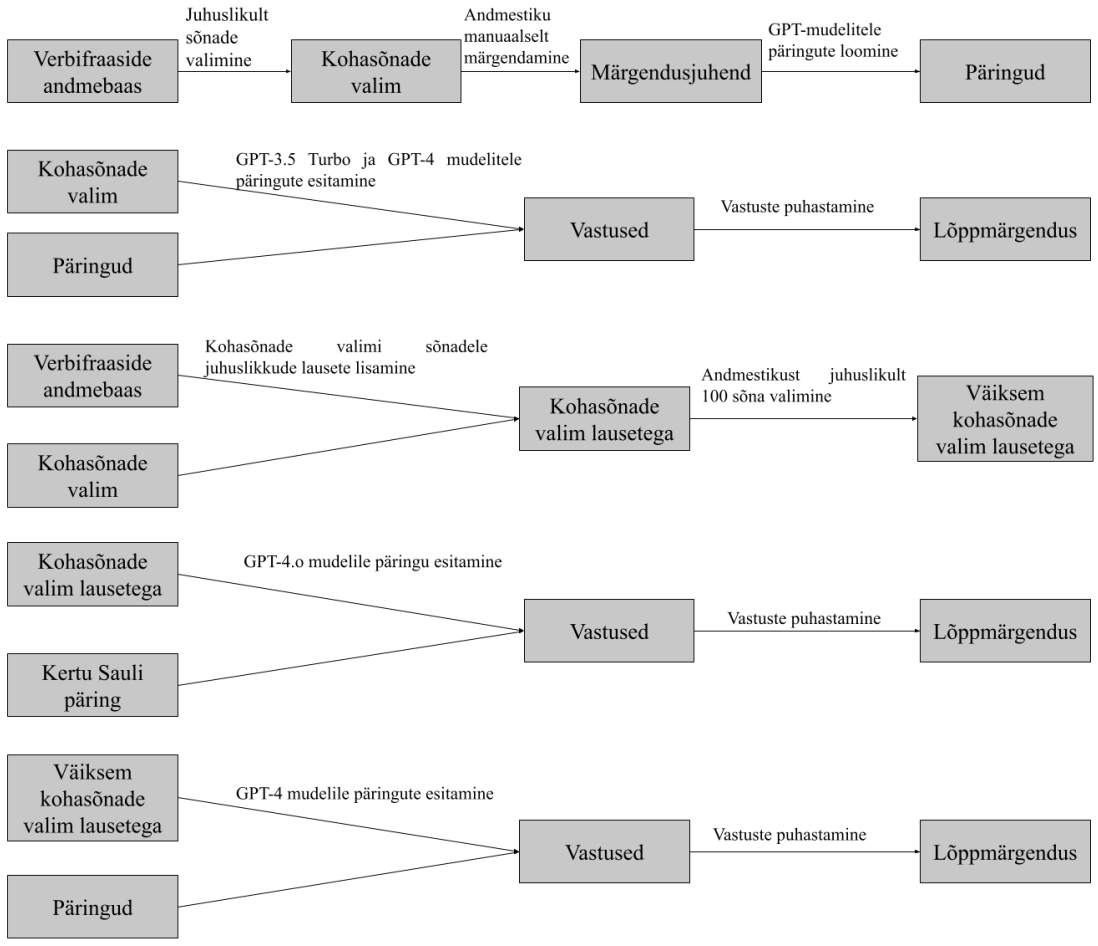
Kuigi ChatGPT on kujunenud väga populaarseks juturobotiks, on oluline meeles pidada, et genereeritud tekst ei pruugi olla alati tõene. Seda, kui keelemudelid nagu ChatGPT tagastavad valeinformatsiooni justkui see oleks tõsi, nimetatakse hallutsineerimiseks (ingl *hallucination*) [18]. Hallutsineerimine toimub, sest keelemudelid kasutavad tokenite tõenäosuseid, et luua kõige realistlikum lause [18]. Kuna keelemudelid ei ole võimelised mõistma, mis loodud lause tähendus on, siis ei saa need ka kontrollida, kas genereeritud tekst on tõene või väär [18]. Seega keelemudeli poolt saadud kõige tõenäolisemad laused võivad küll kõlada usutavalt, aga ei pruugi alati olla tõesed [18]. Seetõttu võib ka käesolevas töös esineda olukordi, kus keelemudeli vastused on ekslikud või ei vasta soovitud formaadile. Sellised juhtumid võivad olla põhjustatud keelemudeli hallutsineerimisest või ebaselgelt sõnastatud päringutest, mida mudel võib mitmeti tõlgendada.

2. Tööprotsess

Käesolevas peatükis antakse ülevaade kasutatud andmestikkudest ning GPT-mudelite rakendamisest sõnade märgendamisel. Töö eesmärgiks oli testida järgmisi hüpoteese.

1. Nimisõnafraasi semantiline füüsilise koha määratlus ei sõltu lause kontekstist.
2. Füüsilise koha määratlust saab leida, määrates nimisõnafraasile täpsemad semantilised klassid.
3. GPT-mudeleid saab kasutada täpsemate semantiliste klasside määramiseks.
4. GPT-mudeleid saab kasutada füüsilise koha määramiseks.
5. Täpsemad semantilised klassid ei sõltu lause kontekstist.

Tööprotsess jagunes kaheks: sõnade märgendamine ilma kontekstita ja sõnade märgendamine koos kontekstiga. Üldise ülevaate tööprotsessi ülesehitusest annab joonis 1.



Joonis 1. Tööprotsessi kirjeldus.

Märgendamisel ilma kontekstita keskenduti üksnes uuritavale sõnale, määrates sellele kõige tõenäolisem tähendus. Ka GPT-mudelite hinnang sõnale tulenes ainult sõnast endast. Sellise lähenemisviisi oluliseks puuduseks on, et mitmetel sõnadel võib erinevates kontekstides olla erinev tähendus. Näiteks sõna "*tee*" võib lauses "*Kass seisis keset teed.*" tähistada kohta, kuid lauses "*Ma jõin magusat teed.*" viidata joogile. Seega võib ilma kontekstita olla keeruline määratleda sõna täpset tähendust.

Sellest tulenevalt uuriti ka sõnade märgendamist olukorras, kus GPT-mudelile anti lisaks sõnale ette ka lause, milles sõna esineb. Kontekstiga märgendamist viidi läbi kahel viisil. Esiteks hinnati sarnaselt kontekstita juhtumitele, kuidas mudel klassifitseerib sõnu erinevatesse semantilistesse alamklassidesse lause põhjal. Teiseks küsiti mudelilt otseselt, kas antud sõna konkreetses lauses tähistab füüsilist kohta või mitte.

Töös kasutatud kood on kättesaadaval GitHubis (Lisa 1).

2.1 Andmestikud

Selles alapeatükis tutvustatakse andmestikke, mille põhjal koostati kohasõnade valim GPT-mudelite hindamiseks. Kõik andmestikud põhinevad Eesti keele koondkorpuse andmetel koostatud verbifraaside andmebaasil, mida käsitleti lähemalt peatükis 1.2.2.

2.1.1 Andmestik 1: kohasõnade valim

Kohasõnade valimi loomisel võeti aluseks Kertu Sauli doktoritöö raames loodud kohasõnade andmestik [19]. Kohasõnade andmestiku aluseks on Eesti keele koondkorpuse põhjal loodud andmebaasi fail, millest on välja võetud verbide kohakäänetes nimisõnafraasi alluvad [19]. Kohakäändes alluvad on esinemissageduse põhjal grupeeritud ehk iga sõna esineb tabelis täpselt üks kord [19]. Sõnadest on välja filtreeritud ka sagedased ajasõnad ning väljendverbide osised [19]. Andmestikust eemaldati ka 1000 kõige kõrgema sagedusega sõna, ehk lõpliku valimi moodustasid pigem harva esinevad sõnad.

Antud kohasõnade andmestikust oli tarvis juhuslikult valida hulk sõnu, mida täpsemalt edasi uurida. Selleks jäeti andmestikust alles üksnes sõna sisaldav veerg ning sellest valiti juhuslikult 1000 rida, mis salvestati eraldi andmestikuna. Kuna aluseks kasutati kohasõnade andmestikku, sisaldas valitud 1000 sõna keskmisest enam füüsilisi kohti, mille uurimine on

käesoleva töö keskne eesmärk. Seega võimaldas see valim saada sisukama ülevaate erinevatest eestikeelsetest kohtadest võrreldes mõne muu andmestiku kasutamisega.

2.1.2 Andmestik 2: kohasõnade valim lausetega

Selleks et testida suurte keelemudelite võimekust füüsiliste kohtade tuvastamisel konteksti põhjal oli tarvis luua teine andmestik, mis sisaldaks ka lauset, milles sõna esineb.

Aluseks võeti taas koondkorpuse põhjal koostatud andmebaasifail, millest eraldati verbide kohakäänetes nimisõnafraasi alluvad. Nendest koostatud andmestikku lisati kolm täiendavat veergu.

‘case’ – sõna käänet tähistav märge (abl/ad/all/adt/el/ill/in),

‘proper’ – näitaja, kas sõna on formaalses keeles kasutatav (0 = ei, 1 = jah),

‘number’ – kas sõna esineb ainsuses või mitmuses (sg/pl).

Iga sõna kohta valiti juhuslikult viis erinevat lauset, milles see sõna esines, ning nende põhjal moodustati kontekstipõhine andmestik. Alles jäeti ainult need sõnad, mis esinesid ka esimeses andmestikus (vt 2.1.1). Kui mõni sõna ei esinenud vähemalt viies lauses, lisati kõik laused, milles see sõna siiski esines, tagamaks, et igal sõnal oleks vähemalt üks esinemine andmestikus. Lõpuks saadud andmestik koosneb 1876 reast. Loodud kood töötab mitte ainult valitud sõnadega, vaid suudab töödelda ka kogu verbiandmebaasi.

2.1.3 Andmestik 3: väiksem kohasõnade valim lausetega

Teise andmestiku põhjal koostati kolmas, kitsendatud andmestik, kuhu jäeti alles ainult need sõnad, mis esinesid vähemalt viies erinevas lauses. Sellist esinemissagedus peeti piisavaks, et võimaldada sõna tähenduse ja kasutusviiside usaldusväärset hindamist mitmes erinevas kontekstis. Nendest sõnadest valiti juhuslikult 100, mida analüüsiti edasi põhjalikumalt. Tulemuseks saadud andmestik koosneb 500 kirjest ning sisaldab iga valitud kohakäändes sõna kohta viis näitelauset.

2.2 Päringute koostamine

Selleks et koostada suurtele keelemudelitele päringud, mille põhjal tuvastada, kas sõna tähistab kohta, oli esmalt vaja koostada märgendusjuhend. Märgendusjuhend on reeglite kogum, mis annab ülevaate sellest, millist tüüpi sõnasid käsitletakse antud töö raames kohtadena ning milliseid mitte.

Juhendi koostamiseks kasutati eelnevalt mainitud kohasõnade valimit, milles märgendati käsitsi esimesed 500 sõna. Iga sõna kohta määrati üks neljast märgendist.

- **‘1’** – sõna tähistab enamus ajast füüsilist kohta.
- **‘0’** – sõna ei tähista enamus ajast füüsilist kohta.
- **‘?’** – ilma kontekstita ei ole võimalik otsustada.
- **‘–’** – sõnal puudub tähendus eesti keeles või ei õnnestu tähendust kindlaks teha.

Protsessi käigus kujunes välja arusaam, millist tüüpi sõnasid kuidas märgendada. Selle põhjal jaotati sõnad tähenduse järgi gruppidesse ning töötati välja märgendusreeglid. Järgnev loetelu annab ülevaate märgendusjuhendis kasutatud reeglitest.

1. Nimi, mis viitab geograafilisele asukohale, on koht (nt Tartu).
2. Nimisõna, mis viitab füüsilisele esemele, mis ei ole suurem kui inimene ja mida inimene saab omada, on koht (nt seljakott).
3. Füüsiline asukoht, kuhu inimene saab siseneda, on koht (nt maja).
4. Kui inimene saab selle peal seista, on see koht (nt lava).
5. Taimed on kohad (nt puu).
6. Suunad on kohad (nt all).
7. Piirkonnad on kohad (nt elamurajoon).
8. Isikunimed ei ole kohad (nt Anne).
9. Organisatsioonid, üritused ja ettevõtted ei ole kohad (nt näitus).
10. Ajaperioodid ei ole kohad (nt keskaeg).
11. Joogid ja toidud ei ole kohad (nt pannkook).
12. Riided ei ole kohad (nt särk).
13. Asutuste puhul ei saa ilma kontekstita otsustada (nt spordiklubi).
14. Osa sõnasid ei oma eesti keeles tähendust (nt glur).

Iga reegli jaoks koostati ingliskeelne päring. Selleks valiti igale reeglile kümme treeningsõna, millest viis vastasid reeglile ja viis mitte. Neid sõnasid kasutades koostati päring, mille puhul ChatGPT-4o märgendas vähemalt kaheksa sõna kümnest õigesti. Seejärel koostati igale reeglile testhulk uuest kümnest sõnast samas jaotuses (5 vastavad, 5 mitte), millele päringut rakendades suutis mudel taas märgendada vähemalt kaheksa sõna õigesti. Nii treeningsõnade kui ka testsõnade valimisel kasutati 1000 realist kohasõnade valimit.

Näiteks esimese märgendusjuhendi reegli “Nimi, mis viitab geograafilisele asukohale, on koht” põhjal tehtud päringu puhul märgendas ChatGPT-4o kõik treening- ning testsõnad õigesti. Selle näite puhul saadud tulemusi on võimalik näha tabelites 1 ja 2. Kõiki test- ja treeningandmete tulemusi on võimalik näha GitHubis (Lisa 1).

Tabel 1. Näide treeningandmete tulemustest.

Sõna	ChatGPT-4o märgendus	Manuaalne märgendus	Sõna	ChatGPT-4o märgendus	Manuaalne märgendus
Nõugaste	1	1	Lembit	0	0
turbaauk	0	0	rumm	0	0
lillepaviljon	0	0	kambja	1	1
Saksa	0	0	Toronto	1	1
Svalbard	1	1	Lublino	1	1

Tabel 2. Näide testandmete tulemustest.

Sõna	ChatGPT-4o märgendus	Manuaalne märgendus	Sõna	ChatGPT-4o märgendus	Manuaalne märgendus
liuväli	0	0	maahaigla	0	0
Belesta	1	1	Rootsi	1	1

Sõna	ChatGPT-4o märgendus	Manuaalne märgendus
J	0	0
hankija	0	0
Priština	1	1

Sõna	ChatGPT-4o märgendus	Manuaalne märgendus
Abhaasia	1	1
kettaheitesektor	0	0
Philadelphia	1	1

Kõik loodud päringud olid sarnasel kujul ning neid on võimalik näha GitHubis (Lisa 2). Esimese märgendusjuhendi reegli põhjal loodud päring oli järgmine.

Provide the output in JSON format, where:
'1' indicates that the given Estonian word is a place name referring to a geographical location.
'0' indicates that the word does not meet these criteria.
Evaluate the following words accordingly:

Päring esitati mudelitele GPT-3.5 Turbo ja GPT-4 süsteemi päringuna ning märgendatavad sõnad anti kasutaja päringuna. Lisaks koostati iga reegli kohta sarnane päring, kus GPT-mudelile anti lisainfona ka sõna kasutuskontekst.

Provide the output in JSON format, where:

'1' indicates that the given Estonian word is a place name referring to a geographical location.

'0' indicates that the word does not meet these criteria.

Evaluate the following words according to whether they serve that purpose in the given context.

Ka kontekstiga päringu puhul anti GPT-4 mudelile päring süsteemi päringuna. Sõna ja lause, mida hinnati anti kasutaja päringuna kujul *[{"word": "", "context": ""}]*.

2.3 Päringute kasutamine

Ilma kontekstita päringuid esitati nii Azure GPT-3.5 Turbo kui ka GPT-4 mudelitele. Mõlemale mudelile edastati kõik 14 märgendusjuhendi reegli alusel koostatud päringut. Iga päring sisaldas kümmet sõna, mille märgendamist GPT-mudelilt oodati. Päringuid esitati järjestikku kuni kõik 500 eelnevalt käsitsi märgendatud sõna olid mõlema mudeli poolt töödeldud. Päringud GPT-mudelitele esitas Sven Laur, kellel oli ligipääs vastavatele tasulistele mudelitele. Päringute esitamise kood on kättesaadav GitHubis (Lisa 3).

Märgendamine toimus nii, et iga sõna vaadeldi eraldi 14 erineva kategooria lõikes — see tähendab, et igat sõna hinnati igas reeglirühmas vastavalt konkreetsele märgendusreeglile. Selline lähenemine võimaldas hinnata mudelite suutlikkust tuvastada sõna tähendust ilma kontekstita erinevates semantilistes tingimustes ning võrrelda mudeleid omavahel.

Kontekstiga päringuid esitati 100 sõna kohta, kus iga sõna esines 5 lauses (Andmestik 3). Ühes päringus esitati kõik viis lauset, milles konkreetne sõna esines. Päringuid koostati kõigi 14 märgenduskategooria lõikes ja neid esitati üksnes GPT-4 mudelile. Ka need päringud esitas Sven Laur ning vastav kood on samuti leitav GitHubis (Lisa 4).

Lisaks esitas Kertu Saul ka päringu andmestiku 2 kohta mudelile GPT-4o. Esitatud päring on loodud Kertu sauli poolt ning see koosneb süsteemi ja kasutaja päringust. Süsteemi päring on *Sa oled lingvist, kes aitab tuvastada füüsilisi kohti.* ning kasutaja päring on järgmine.

Ma olen lingvist. Määra, kas järgmises lauses 'lause' olev sõna "sõna" on füüsiline koht vastavalt järgmistele kategooriatele:

Füüsilised kohad:

1. Kohanimed (nt Bristol, Sepphoris)

2. Ehitised/äride füüsilised asukohad (pangamaja, multimeediastuudio, Kuku klubi, arvutifirma)

3. Füüsilised objektid, kaasa arvatud elusolendid (esikohapoodium, Kuu, varundusseade, sadul, pilv)

4. Alad, mille geograafiline asukoht on defineeritav (põlengupaik, põhjapoolus, kaldapealne, tolmupilv)

Mittefüüsilised kohad:

1. Abstraktsed kohad, mille geograafilist asukohta pole võimalik määrata (nt Wifi, arvutiturg, õhuruum, digitaalplatvorm).

2. Tegevused ja sündmused (kleidiproov, värbamine).

3. Elusolend, kes on tegevuse tegija (rüselejal käisid sussid).

4. Seisund (jooksevad jalad rakku, istun sitas).

5. Viisimäärus (teravaimalt, käsikäes).

6. Põhjuslikud määrused (hävitamisel, protsessori olemasolul).

7. Ajamäärused (aasta, hommik).

8. Konstruktsioonid ja stampväljendid (vaatamata hoiakule, käib jutt kehtivusest).

Lauses: lause

Sõna: sõna

Vasta kujul:

- Kui sõna on füüsiline koht: "lause;sõna;LOC"

- Kui sõna ei ole füüsiline koht: "lause;sõna;NONE"

NB! Vasta ALATI AINULT kujul LOC või NONE

Päringu tulemusel saadi andmestik, mis näitab, kas keelemudel klassifitseeris antud sõna igas konkreetses lauses füüsilise kohana või mitte. Päringu esitamise kood on kättesaadav GitHubis (Lisa 5).

2.4 Tulemuste hindamine

Tulemuste hindamiseks võrreldi esmalt GPT-mudelite poolt teostatud märgendust käsitsi märgendatud andmestikuga. Selleks koondati mudelite vastused kaheks eraldiseisvaks andmestikuks: üks GPT-3.5 Turbo ja teine GPT-4 kohta. Mõlemas andmestikus oli iga rida seotud konkreetse märgendatud sõnaga ning veerud vastasid märgendusjuhendis määratletud reeglitele. Andmestiku loomiseks oli tarvis GPT-mudelitelt saadud andmed puhastada. Kuigi päringus oli öeldud, et vastus tuleb esitada JSON formaadis oli saadud väljundites palju varieeruvust. Regexi abil leiti väljundist vaja minevad tulemused, mis koondati andmestikuks.

Iga lahter sisaldas väärtust ‘1’ või ‘0’, sõltuvalt sellest, kas sõna vastas vastavale atribuudile. GPT-mudelid pidasid osasid sõnu ebasobivaks ning keeldusid neid märgendamast. Sellisel juhul tähistati vastav lahter väärtusega NaN ning jäeti edasisest tulemuste analüüsist välja. Ühte näidisrida saadud andmestikust, kus veeru number vastab märgendusjuhendi reeglile saab näha tabelis 3.

Tabel 3. Näidis rida GPT-4 märgendamisest.

sõna	1	2	3	4	5	6	7	8	9	10	11	12	13	14
peasaal	0	0	1	0	0	0	0	0	0	0	0	0	0	0

Seejärel võrreldi mudelite loodud andmestikke käsitsi märgendatud andmestikuga, hinnates igas veerus lahtrite vastavust sõnade kaupa. Selle põhjal arvutati iga atribuudi kohta järgmised mõõdikud: õigsus (ingl *accuracy*), saagis (ingl *recall*), täpsus (ingl *precision*) ja F1-skoor.

Mõõdikute parema mõistmise huvides on esmalt vajalik selgitada nende tausta. Lähtuvalt sellest, kuidas lahtrid olid käsitsi märgendatud ning kuidas GPT-mudel nende väärtusi ennustas, saab tulemused jagada nelja kategooriasse: õige positiivne (ingl *true positive*), valepositiivne (ingl *false positive*), õige negatiivne (ingl *true negative*) ja valenegatiivne (ingl *false negative*).

- **Õige positiivne**: olukord, kus käsitsi märgendus ja mudeli ennustus ühtivad ning väärtus on ‘1’ ehk lahter vastab märgendusjuhendi reeglile.

- **Õige negatiivne**: olukord, kus käsitsi märgendus ja mudeli ennustus ühtivad, kuid väärtus on ‘0’ ehk lahter ei vasta märgendusjuhendi reeglile.

- **Valepositiivne**: olukord, kus käsitsi märgendati lahtri väärtuseks ‘0’, kuid mudel ennustas väärtuseks ‘1’.

- **Valenegatiivne**: olukord, kus käsitsi märgendati lahtri väärtuseks ‘1’, kuid mudel ennustas väärtuseks ‘0’.

Nimetatud kategooriate alusel on võimalik arvutada eelmainitud mõõdikud.

- **Õigsus** näitab õigete ennustuste osakaalu kõikidest ennustustest.

$$Õigsus = \frac{õige\ positiivne + õige\ negatiivne}{õige\ positiivne + õige\ negatiivne + valepositiivne + valenegatiivne}$$

- **Saagis** kirjeldab, kui suur osa kõigist tegelikest positiivsetest juhtumitest suudeti õigesti tuvastada.

$$Saagis = \frac{õige\ positiivne}{õige\ positiivne + valenegatiivne}$$

- **Täpsus** näitab, kui suur osa positiivseks ennustatud juhtumitest oli tegelikult positiivne.

$$Täpsus = \frac{õige\ positiivne}{õige\ positiivne + valepositiivne}$$

Kuna saagis ja täpsus võivad üksteisega vastuollu minna, siis on nende tasakaalustatud hindamiseks kasutusel F1-skoor, mis on nende kahe mõõdiku harmooniline keskmine.

$$F1\text{–}skoor = 2 \times \frac{täpsus \times saagis}{täpsus + saagis}$$

Kuna F1-skoor on tasakaalustatud mõõdik, mis võtab arvesse nii täpsust kui ka saagist, on see käesolevas töös valitud peamiseks hindamiskriteeriumiks tulemuste võrdlemisel.

3. Tulemused

Käesolevas peatükis esitatakse Azure GPT-3.5 Turbo, GPT-4 ja GPT-4o mudelite märgendamise tulemused. Analüüs keskendub esmalt sõnade märgendamisele ilma kontekstita, seejärel hinnatakse mudelite täpsust kontekstiga märgendamisel, kus lisaks sõnale anti ette ka sõna sisaldav lause.

Ilma kontekstita sõnade märgendamisel kasutati mudeleid GPT-3.5 Turbo ning GPT-4. Kontekstiga sõnade märgendamisel kasutati sematiliste alamklasside märgendamiseks mudelit GPT-4 ning füüsilise koha märgendamiseks mudelit GPT-4o.

3.1 Ilma kontekstita sõnade märgendamise tulemused

Tulemusi võrreldakse käsitsi märgendatud andmestikuga. Võrdlusi tehti iga märgendusjuhendi reegli ehk andmestiku veeru lõikes, võimaldades hinnata mudelite õigsust, täpsust, saagist ja F1-skoori.

Lisaks treeniti käsitsi märgendatud andmete põhjal randomForestClassifier-mudel, mille eesmärk oli ennustada sõnale määratav märgend (’0’, ’1’, ’-’ või ’?’) semantiliste alamklasside põhjal. Seda mudelit kasutati ka selleks, et hinnata, kuivõrd hästi GPT-mudelite tulemuste põhjal saab sõnu märgendada. Nii oli võimalik tuvastada GPT-mudelite rakenduslik potentsiaal sõnade automaatses märgendamises, hinnates mudeli üldist täpsust, õigsust, saagist ja F1-skoori.

GPT-mudelid keeldusid osasid sõnu märgendamast seega on GPT-3.5 Turbo valimi suurus 448 sõna ning GPT-4 valimi suurus 478 sõna. Kuna märgendatud sõnade jaotus ei olnud semantiliste alamklasside lõikes tasakaalus, on kõige sobivamaks hindamismõõdikuks F1-skoor, mis arvestab nii täpsust kui ka saagist. Tabelitest 4 ja 5 on näha, et üldjoontes saavutas GPT-4 kõrgemad F1-skoorid kui GPT-3.5 Turbo. Mõlema mudeli puhul jäid mõnede semantiliste alamklasside, näiteks *suund* ja *taim*, F1-skoorid väga madalaks. Selle üheks võimalikuks põhjuseks on asjaolu, et vastavaid sõnu esines valimis väga vähe ja seega on skoorid statistiliselt ebatäpsed.

Tabel 4. GPT-3.5 Turbo tulemused semantiliste alamklasside lõikes.

semantiline alamklass	õigsus	täpsus	saagis	F1-skoor
tähenduseta sõna	0,82	**0,40**	0,88	**0,55**
ajaperiood	**0,90**	0,29	0,73	0,41
isikunimi	0,88	0,25	0,89	0,39
geograafiline lokatsioon	0,73	0,25	0,88	0,38
saab peal seista	0,77	0,25	0,77	0,37
saab sisse minna	0,67	0,22	0,87	0,35
organisatsioon, üritus või ettevõte	0,63	0,16	0,89	0,28
ese	0,65	0,13	0,92	0,23
piirkond	0,78	0,11	0,92	0,20
asutus	0,7	0,08	0,92	0,14
toit või jook	0,83	0,04	**1,00**	0,07
riideese	0,85	0,03	0,50	0,05
taim	0,79	0,02	0,67	0,04
suund	0,85	0,00	0,00	0,00

Tabel 5. GPT-4 tulemused semantiliste alamklasside lõikes.

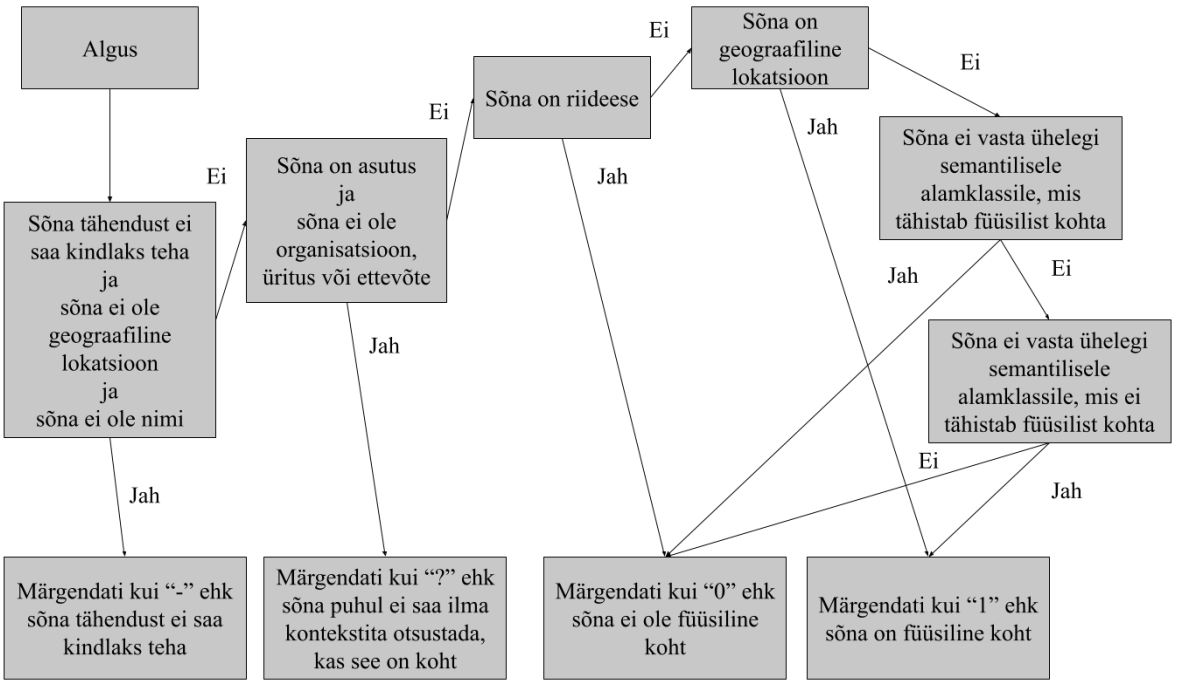
semantiline alamklass	õigsus	täpsus	saagis	F1-skoor
geograafiline lokatsioon	0,92	0,55	0,98	**0,71**

semantiline alamklass	õigsus	täpsus	saagis	F1-skoor
ajaperiood	0,96	**0,56**	0,86	0,68
saab peal seista	0,88	0,40	0,85	0,54
tähenduseta sõna	0,78	0,36	0,89	0,52
isikunimi	0,92	0,32	0,95	0,48
organisatsioon, üritus või ettevõte	0,81	0,28	0,87	0,42
ese	0,85	0,26	0,92	0,40
saab sisse minna	0,71	0,24	0,88	0,38
riideese	**0,98**	0,25	0,75	0,38
piirkond	0,89	0,21	0,81	0,34
toit või jook	0,96	0,14	**1,00**	0,25
asutus	0,80	0,11	**1,00**	0,20
taim	0,96	0,06	0,33	0,11
suund	0,91	0,02	**1,00**	0,05

Parimad tulemused saavutas GPT-4 mudel *geograafiliste lokatsioonide* ja *ajaperioodide* ära tundmisel, samas kui GPT-3.5 Turbo puhul oli parim tulemus *tähenduseta sõnade* klassifitseerimisel. *Tähenduseta sõnade* alamklass oli ainus, mille puhul GPT-3.5 Turbo F1-skoor ületas GPT-4 oma. Selle põhjuseks võib olla asjaolu, et GPT-4 kaldus klassifitseerima ka *isikunimesid* ja *geograafiliste lokatsioonide nimesid* tähenduseta sõnadeks, mida ei tehtud ei manuaalses märgenduses ega GPT-3.5 Turbo puhul. Sellise segaduse vältimiseks saaks tulevikus päringutes täpsemalt määratleda, kas nimesid tuleb käsitleda tähenduslike või tähenduseta sõnadena.

Semantiliste alamklasside märgendamise tulemused näitavad üldiselt, et mudelite täpsus ja F1-skoor on suhteliselt madalad, samas kui õigsus ja saagis on kõrged. Madal F1-skoor tuleneb eelkõige madalast täpsusest, mis mõjutab otseselt F1-skoori väärtust selle arvutamise valemi kaudu. Need tulemused viitavad sellele, et GPT-mudelid suudavad sageli õigesti tuvastada, kui sõna kuulub mingisse semantilisse alamklassi (kõrge saagis), kuid samas kipuvad nad ekslikult märgendama ka sõnu, mis sinna tegelikult ei kuulu (madal täpsus). Seega võib järeldada, et GPT-mudelid kalduvad liigselt klassifitseerima sõnu semantiliste alamklasside esindajatena ka juhul, kui see ei ole põhjendatud. Tulevikus saaks täpsuse parandamise eesmärgil katsetada päringute täiustamist, lisades neisse selgemaid näiteid ja kirjeldusi sõnadest, mis kindlasti ei kuulu vastavasse semantilisse klassi, et suunata mudelit vähendama valepositiivseid märgendusi.

Manuaalselt märgendatud andmestiku põhjal treeniti vaikeväärtustega RandomForestClassifier-mudel, mille eesmärk oli ennustada sõna märgendust selle semantiliste alamklasside alusel. Seejärel kasutati treenitud mudelit sõnade märgendamiseks GPT-mudelite genereeritud andmete põhjal ning saadud tulemusi võrreldi manuaalse märgendusega. Selle eesmärgiks oli tuvastada, kas manuaal märgendusel töötav mudel on üle viidav GPT-mudelite märgendusele. Lisaks loodi ka reeglistik, mis ennustas iga andmestiku rea puhul, kuidas vastavat sõna märgendada. Reeglistik loodi seoste põhjal, mis manuaalselt märgendatud andmestikust välja tulid. Täpsemalt kirjeldab reeglistiku joonis 2.



Joonis 2. Reeglistik, mille põhjal sõnu märgendati.

Mõlema ennustus viisi tulemuste põhjal arvutati füüsilise koha määramise täpsus, õigsus, saagis ja F1-skoor. Tabelis 6 on esitatud RandomForestClassifier-mudeli, reeglipõhise märgenduse ning juhusliku märgendamise tulemused. Kuna märgendatav klass ei olnud enam binaarne, vaid koosnes neljast võimalikust väärtusest ('0', '1', '-' ja '?'), siis ei olnud otstarbekas kasutada varasemalt toodud kaheklassilise klassifikatsiooni valemeid. Selle asemel kasutati mõõdikute arvutamiseks teeki sklearn.metrics, täpsemalt funktsioone precision_score, recall_score, f1_score ja accuracy_score, kus keskmiseks oli märgitud "*macro*".

Tabel 6. Füüsilise koha märgendamise tulemused.

märgendamise viis	mudel	õigsus	täpsus	saagis	F1-skoor
RandomForestClassifier	GPT-4	**0,66**	**0,49**	**0,54**	**0,49**
Reeglipõhine	GPT-4	0,64	0,48	0,49	0,48
Reeglipõhine	GPT-3.5 Turbo	0,59	0,44	0,47	0,45

märgendamise viis	mudel	õigsus	täpsus	saagis	F1-skoor
RandomForestClassifier	GPT-3.5 Turbo	0,58	0,47	0,52	0,44
Juhuslikult pakkumine	-	0,25	0,25	0,25	0,25

Tulemustest ilmnes selgelt, et kõigi mõõdikute lõikes ületasid GPT-4 tulemused GPT-3.5 Turbo omi. Siiski ei ole kummagi mudeli F1-skoor eriti kõrge. Kuna GPT-mudelid tegid mitmete semantiliste klasside puhul ebatäpseid ennustusi, ei saavutanud ka nende põhjal loodud ennustusmudelid kõrget märgendustäpsust. Reeglipõhise ja RandomForestClassifier-mudeli tulemuste vahel märkimisväärset erinevust ei esinenud. Üldiselt on saadud tulemused liiga nõrgad, et neid praktikas märgendamiseks kasutada. Samas on tulemused siiski paremad kui juhuslik pakkumine, mis viitab potentsiaalile saavutada tulevikus paremaid tulemusi. Üheks võimalikuks lähenemiseks võiks olla näidete lisamine päringusse, mis aitaks GPT-mudelil paremini mõista, millist vastust oodatakse.

3.2 Kontekstiga sõnade märgendamise tulemused

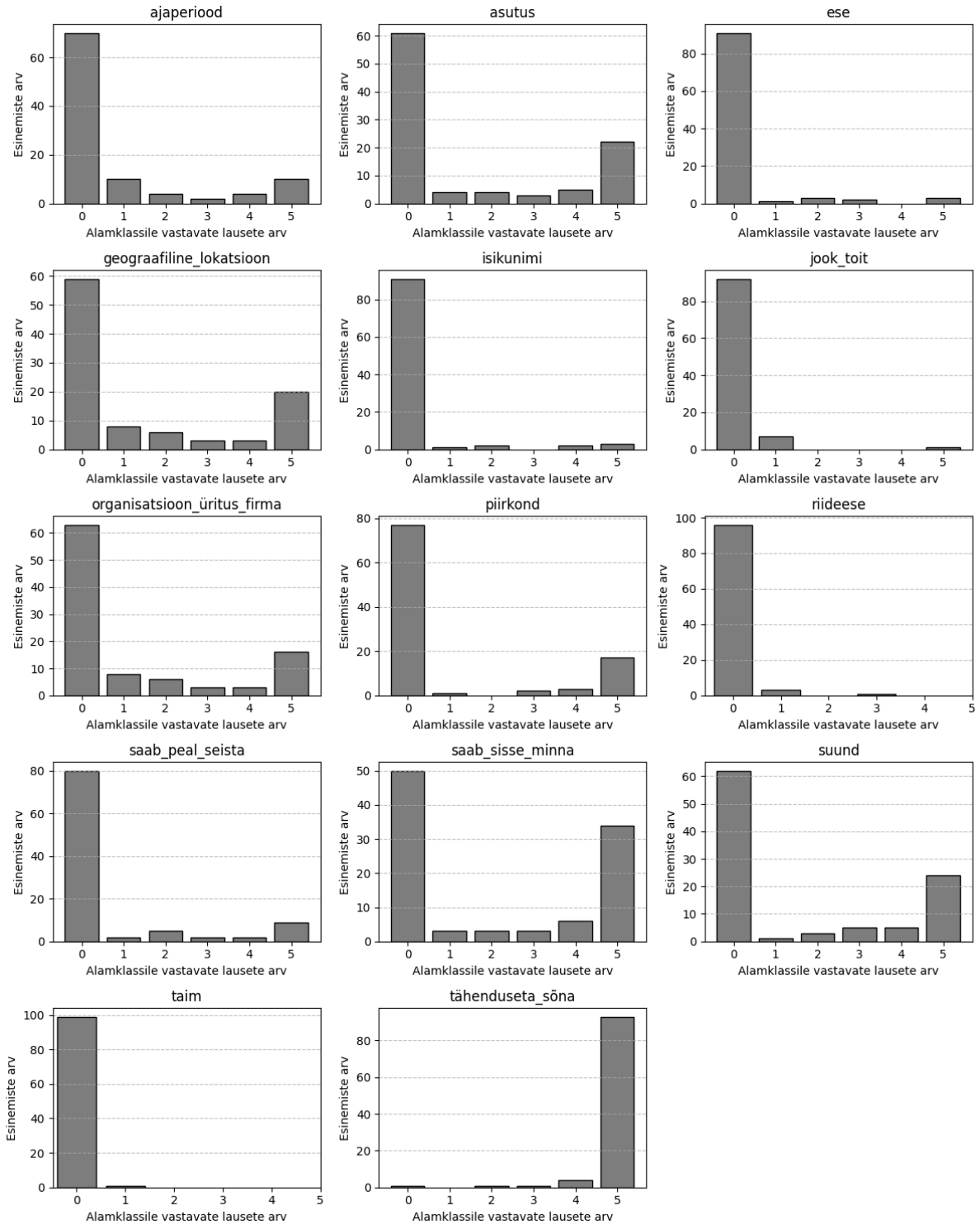
Mudelile GPT-4 esitatud kontekstiga päringute vastuste põhjal koostati andmestik, mis näitas, mitu korda iga sõna esines lausetes, kus see kuulus vastavasse semantilisse alamklassi. Ühte näidisrida saadud andmestikust, kus veeru number vastab märgendusjuhendi reeglile saab näha tabelis 7.

Tabel 7. Näidis rida GPT-4 märgendamisest.

sõna	1	2	3	4	5	6	7	8	9	10	11	12	13	14
veefestival	1	0	5	0	0	0	0	0	5	2	0	0	5	0

Saadud andmestiku põhjal koostati histogramm, mis võimaldas hinnata, kuivõrd semantilise alamklassi märgendamine sõltub kontekstist (joonis 3). Histogrammil kujutati, kui sageli sama sõna viies erinevas lauses märgendati vastavaks konkreetsele semantilisele alamklassile. Alamklassid, mille puhul enamik sõnu olid märgendatud kas kõigis (5) või mitte üheski (0) lauses, viitavad sellele, et sõna tähendus ei sõltu oluliselt kontekstist, sest sõna kas alati vastab kategooriale või mitte. Seevastu alamklassid, kus paljusid sõnu

märgendati vastavaks ainult 2–3 korral, viitavad sellele, et semantilise alamklassi määramine sõltub oluliselt kontekstist. Kui sõna sobib kategooriasse vaid pooltel juhtudel, näitab see, et kontekst mängib selle tähenduse tõlgendamisel märkimisväärset rolli.

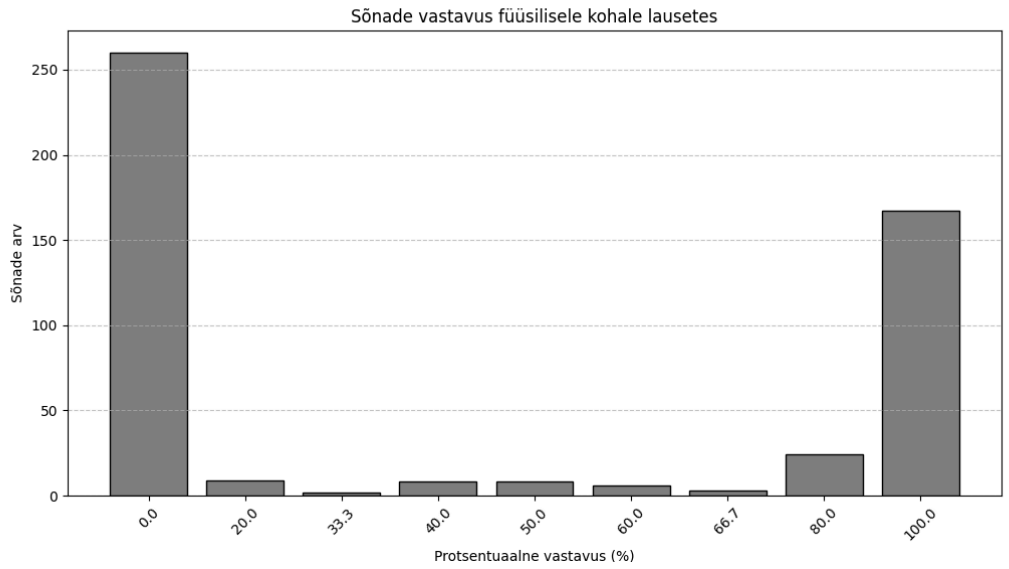


Joonis 3. Mudeli GPT-4 semantiliste alamklasside märgendamine.

Kuna valim koosnes vaid sajast juhuslikult valitud sõnast, on mitmete semantiliste alamklasside puhul keeruline teha kindlaid järeldusi kontekstitundlikkuse kohta. Näiteks märgendati sõnu kategooriatesse nagu taim, riideese, isikunimi, ese, tähenduseta sõna ning jook või toit väga harva, mistõttu puudub piisav andmestik nende alamklasside kohta usaldusväärsete järelduste tegemiseks.

See-eest alamklassid nagu geograafiline lokatsioon, asutus, piirkond, suund, kohad, kuhu saab siseneda (nt hooned, sõidukid) ning organisatsioon, üritus või firma näivad olevat suhteliselt sõltumatud kontekstist. Samas viitavad tulemused sellele, et sõnad, mis kuuluvad alamklassidesse ajaperiood või pinnad, millel saab seista, on kontekstitundlikumad. Nende puhul esines sagedamini olukordi, kus sõna vastavus semantilisele alamklassile sõltus selgelt lause tähendusest.

GPT-4o mudelit küsiti andmestiku 2 põhjal, kas igas lauses esinev uuritav sõna tähistab füüsilist lokatsiooni. Tulemuste alusel koostati andmestik, mis näitas iga sõna kohta, mitmes lauses seda esitati ning mitmel juhul mudel sõna füüsilise kohana märgendas. Selle põhjal arvutati iga sõna puhul protsentuaalne osakaal, mis näitab, kui sageli sõna esines füüsilise kohana. Protsentide alusel loodi histogramm, mis visualiseerib, millisel määral sõnad kogu andmestiku lõikes tähistasid füüsilisi kohti (joonis 4).



Joonis 4. Mudeli GPT-4o füüsiliste kohtade märgendamine kontekstiga.

Tulemused näitavad, et enamikul juhtudel määratles GPT-4o mudel antud sõna kas kõigis lausetes füüsilise kohana (100%) või mitte üheski lauses füüsilise kohana (0%). See viitab sellele, et sõna määratlemine füüsilise kohana ei sõltu suuresti lause kontekstist.

Siiski on oluline rõhutada, et need tulemused põhinevad GPT mudeli hinnangutel, mitte inimese poolt tehtud märgendusel, mistõttu võib nende täpsus olla piiratud. Lisaks analüüsiti iga sõna maksimaalselt viies erinevas lauses, mis on suhteliselt väike proovi hulk ning ei pruugi hõlmata sõna tähenduslikku varieeruvust. Kuna laused valiti juhuslikult, võivad need esindada pigem haruldasi või mittetüüpilisi kasutusjuhte. Seetõttu kehtivad saadud tulemused eelkõige analüüsitud andmestiku piires ning vajavad täiendavat kontrolli, et teha suuremaid üldistusi.

Kokkuvõte

Käesolevas bakalaureusetöös uuriti, kas suurte keelemudelite abil on võimalik määrata, kas antud sõna viitab füüsilisele kohale või mitte. Eesmärgi täitmiseks analüüsiti esmalt GPT-mudelite võimekust semantiliste alamklasside märgendamisel. Tulemused näitasid, et GPT-4 oli selles ülesandes edukam kui GPT-3.5 Turbo, kuid mitmete klasside puhul esines siiski märkimisväärseid raskusi. Edasiste uuringute käigus saaks proovida näidete lisamist päringusse, et parandada mudeli arusaamist klasside sisust. Samuti tuleks kasutada ulatuslikumat andmestikku, kus semantilised klassid oleksid paremini esindatud, et tagada usaldusväärsem analüüs.

Kuigi semantiliste alamklasside ennustamise täpsus jäi tagasihoidlikuks, kasutati saadud märgendusi hüpoteesi kontrollimiseks, kas füüsilise koha määratlust saab leida, määrates nimisõnafraasile täpsemad semantilised klassid. Analüüs näitas, et sellisel lähenemisel on potentsiaali, kuid selle rakendamiseks praktikas peab alamklasside märgendustäpsus paranema.

Lisaks uuriti GPT-4 mudeli abil, kas semantiliste klasside määratlemine sõltub lause kontekstist. Tulemused jäid andmestiku piiratud mahu tõttu osaliselt ebamääraseks. Siiski ilmnes, et enamik semantilisi alamklasse ei olnud kontekstitundlikud. See võis olla kas selle tõttu, et sõnal oli alati kindel tähendus, või seetõttu, et üks tähendus esines märgatavalt sagedamini kui teised. Samas objekte, mille peal saab seista, ning ajaperioode kirjeldavad klassid näitasid suuremat sõltuvust kontekstist.

Töö viimases osas rakendati GPT-4o mudelit, et hinnata, kas nimisõnafraasi füüsilise koha määratlus sõltub lause kontekstist. Selleks esitati mudelile uuritav sõna koos lausega, kus sõna esines. Analüüsi tulemused näitasid, et enamiku sõnade puhul jäi märgendus eri kontekstides ühtseks. Sõna kas märgendati igas lauses füüsilise kohana või ei märgendatud sõna üheski lauses füüsilise kohana. Seega võib järeldada, et vähemalt käesolevas andmestikus ei olnud nimisõnafraasi semantiline tähendus märkimisväärselt kontekstitundlik.

Viidatud Kirjandus

[1] Erelt M., Metslang H. Eesti keele süntaks. Eesti keele varamu III. Tartu: Tartu Ülikooli Kirjastus, 2017.

[2] Ariva L., Eskor L. Mis on arvutilingvistika? *Oma Keel*, 2004. https://www.emakeeleselts.ee/omakeel/2004_1/Ariva.pdf (30.11.2024)

[3] EKI teatmik: Eesti õigekeelsuskäsiraamat. https://eki.ee/teatmik/sonaliigid/ (30.11.2024)

[4] Muischnek K. Keelekorpused – sama mitmekesised kui keel ise. *Oma Keel*, 2015. https://www.emakeeleselts.ee/omakeel/2015_1/OK_2015-1_05.pdf (30.11.2024)

[5] Eesti Keele Instituut: EESTI KEELE KÄSIRAAMAT 2007. https://arhiiv.eki.ee/books/ekk09/index.php?p=6&p1=2.

[6] Hou Z. Using semantic tagging to examine the American Dream and the Chinese Dream. *Semiotica*, 2019. https://research-ebsco-com.ezproxy.utlib.ut.ee/linkprocessor/plink?id=85671096-3d90-34fe-845a-d3b159e217d0 (05.12.2024)

[7] Ong T. T., McKenzie R. M., Amand M. The narrative of human suffering: using automated semantic tagging to analyse news articles and public attitudes towards the MH370 air tragedy. *Asian Englishes*, 2023. https://doi.org/10.1080/13488678.2021.1927564 (05.12.2024)

[8] Li J., Li Y., Wang X., Tan W.-C. Deep or Simple Models for Semantic Tagging? It Depends on your Data. *PVLDB*, 2020. https://www.vldb.org/pvldb/vol13/p2549-li.pdf (05.12.2024)

[9] Abdulfattah Z. Corpus Linguistics: What Is a Corpus. https://www.academia.edu/36355275/Corpus_Linguistics_What_Is_a_Corpus (23.04.2025)

[10] Eesti keele koondkorpus. https://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et (02.05.2025)

[11] Tsepelina, K; Laur, S. Verbide ja verbi alluvate kogumine koondkorpusest. GitHub. https://github.com/estnltk/syntax_experiments/blob/verb_templates/workflows/001_verb_transactions/v33/v33.ipynb (02.05.2025).

[12] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M. A., Socher, R., Amatriain, X., Gao, J. Large Language Models: A Survey. *ArXiv*, 2024. https://arxiv.org/html/2402.06196v2 (05.12.2024)

[13] Google for Developers. https://developers.google.com/machine-learning/resources/intro-llms (06.12.2024)

[14] Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E., Le Q. V., Zhou D. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf (06.12.2024)

[15] Jurafsky D., Martin J.H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models. Third Edition draft. January 12, 2025. https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf (02.05.2025).

[16] System message design. 2025 https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/advanced-prompt-engineering (08.05.2025)

[17] Ortiz S. What is ChatGPT? How the world's most popular AI chatbot can benefit you. *Zdnet*, 2024. https://www.zdnet.com/article/what-is-chatgpt-how-the-worlds-most-popular-ai-chatbot-can-benefit-you/ (07.12.2024)

[18] The University of Arizona Library https://ask.library.arizona.edu/faq/407990 (07.12.2024)

[19] Saul K. Kohasõnade verbi kaupa otsimine. GitHub. https://github.com/estnltk/syntax_experiments/blob/semantic_labelling/physical_location_lab

elling/physical_location_by_context/code/v01_experiment_1/v01_locations_by_verb.ipynb (07.05.2025).

Lisad

1 Koodi repositoorium

https://github.com/KristinaKatarina/kohafraaside-tuvastamine/tree/main

2 Ilma kontekstita sõnade märgendamiseks loodud päringud

https://github.com/estnltk/syntax_experiments/tree/semantic_labelling/physical_location_labelling/physical_locations_by_word/code/01_gpt_semantic_subcategorisation/prompts

3 Sven Lauri GPT-3.5 Turbo ja GPT-4 mudelitele päringute esitamise kood

https://github.com/estnltk/syntax_experiments/tree/semantic_labelling/physical_location_labelling/physical_locations_by_word/code/01_gpt_semantic_subcategorisation

4 Sven Lauri GPT-4 mudelile kontekstiga päringute esitamise kood

https://github.com/estnltk/syntax_experiments/tree/semantic_labelling/physical_location_labelling/physical_locations_by_word/code

5 Kertu Sauli GPT-4o mudelile kontekstiga päringute esitamise kood

https://github.com/estnltk/syntax_experiments/blob/semantic_labelling/physical_location_labelling/physical_location_by_context/code/v03_gpt_annotation/v03_gpt_annotation.ipynb

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristina Katarina Kaljumäe,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose “Suurte keelemudelite kasutamine sõnade semantiliseks klassifitseerimiseks”, mille juhendaja on Sven Laur, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kristina Katarina Kaljumäe

15.05.2025