LATVIJAS UNIVERSITĀTE
DATORIKAS FAKULTĀTE

TEKSTA SENTIMENTA KLASIFICĒŠANA LATVIEŠU VALODĀ, IZMANTOJOT LIELO VALODAS MODEĻU VAICĀJUMUS UN REDDIT DATU KOPU

BAKALAURA DARBS

Autors: **Pauls Purviņš**
Studentu apliecības Nr.: pp19026
Darba vadītājs: Phd. Comp. Sc. Ēvalds Urtāns

RĪGA, 2023

ANOTĀCIJA

Šajā bakalaura darbā tiek apskatīta lielo valodu modeļu (LVM) izmantošana sentimenta analīzē, un tiek piedāvāta jauna pieeja latviešu valodas datu kopu veidošanai, izmantojot Reddit foruma datus. Validācijas datu kopā tika sasniegta vairāk nekā 82% pareizība ar nulles šāviena metodi, izstrādājot vaicājumus GPT-3.5-turbo modelim, kas vairāk nekā divas reizes uzlabojot iepriekšējos rezultātus trīs klašu sentimenta analīzē. Pētījums parāda, ka LVM var daļēji aizstāt cilvēku marķētājus, padarot lielu datu kopu veidošanu ekonomiski izdevīgāku. Rezultāti liecina, ka LVM var apgūt dziļāku valodas sapratni, tomēr tie dažreiz novirzās no atbilžu veidnēm, kas sarežģī atbilžu analīzi un rada kļūmes datu apstrādes procesā. Turpmākie pētījumi varētu aplūkot arī citus modeļus sentimenta analīzei latviešu valodā, analizēt dažādu valodas iezīmju ietekmi uz vaicājumiem un izpētīt LVM pielietojumu datu kopas ģenerēšanā, lai pielāgotu esošos modeļus. Šis darbs veicina sentimenta analīzes attīstību, izmantojot LVM dotās iespējas. Izveidotā datu kopa satur vairāk nekā 90000 paraugus, kļūstot par lielāko pieejamo marķēto sentimenta datu kopu latviešu valodā.

Darba pamattekstā ir 42. lappuses.

Atslēgvārdi: lielie valodas modeļi, dabiskās valodas apstrāde, mašīnmācīšanās, neironu tīkli, sentimenta analīze

ABSTRACT

This bachelor thesis explores the possibility of using large language models (LLMs) in sentiment analysis and presents a new approach for creating a Latvian language dataset using Reddit data. By engineering prompts for the GPT-3.5-turbo model, we achieved over 82% accuracy using the zero-shot method, surpassing previous research on used validation set more than two times. We demonstrate that LLMs can partially replace human labelers, making dataset creation more cost-effective, especially for larger datasets. Our findings confirm the LLM's more profound understanding of language. However, LLMs occasionally deviate from response templates, making parsing challenging. Future research should investigate alternative models for sentiment analysis in Latvian, analyze language patterns in prompts, and explore LLM-generated datasets to fine-tune existing models. This work contributes to advancing sentiment analysis in non-English languages, leveraging the power of LLMs. The created dataset contains over 90000 samples making it the largest available labeled sentiment data set for the Latvian language. The main text of the bachelor thesis consists of 42. pages

KEYWORDS: large language models, natural language processing, machine learning, neural networks, sentiment analysis

Saturs

Apzīmējumu saraksts

API (Application Programming Interface) – Lietojumprogrammas saskarne

JSON (JavaScript object notation) – JavaScipt objektu notācija, datu faila formāts

Korpuss (corpus) – datu kopa, tekstu kopums

LSTM (Long Short-term memoy) - Rekurentā neironu tīkla paveids

LVM (LLM - Large Language Model) - Lielais Valodas Modelis

MI (AI - Artificial intelligence) - Mākslīgais intelekts

RRN (Recurent Neural Network) - Neironu tīkla arhitektūra

SOTA (State Of The Art) - Modernākais

SVM (Support Vector Machine) - Pārraudzītās mācīšanās modelis lietots klasifikācijas un regresijas uzdevumos

Ievads

Cilvēki jau sen ir pierakstījuši savas domas un pieredzes, sākot ar zīmējumiem uz alu sienām līdz mūsdienām - sociālajiem tīkliem. Jau sen ir aktuāls jautājums, kā pareizi un automātiski interpretēt dažādu tekstu pausto emocionālo nokrāsu, jeb sentimentu, un līdz ar mašīnmācīšanās un sociālo tīklu attīstību šī joma ir guvusi arvien lielāku uzmanību. Uzņēmumiem interesē, ko par tiem saka klienti, politiķiem interesē, kāda ir publikas attieksme pret tiem, un ziņu portāliem interesē, vai tajos publiskoto rakstu autori neizplata savus uzskatus, bet gan pauž objektīvu un neitrālu skatījumu par notikumiem.

Pēdējo gadu laikā ir strauji attīstījušies transformeru arhitektūrā balstītie lielie valodas modeļi, un šajā darbā tiek apskatīta iespēja tos lietot gan kā aizstājējus sentimenta analīzes modeļiem, gan kā rīku lielāku un kvalitatīvāku datu kopu iegūšanai.

Darbā tiek apskatīta lielo valodas modeļu vēsture, arhitektūra un ierobežojumi, sentimenta analīzes vēsture, pielietojumi, kā arī iepriekš veiktie pētījumi latviešu valodā. Tiek apskatītas latviešu valodā pieejamās sentimenta analīzes datu kopas, dokumentēts jaunas datu kopas izstrādes process un prezentēti rezultāti un secinājumi - vai ar LVM vaicājumu un nulles šāviena metodes palīdzību ir iespējams iegūt augstākus rezultātus sentimenta analīzes jomā, kā ar specializēto klasifikācijas modeļu palīdzību.

Zinātniskais raksts "Using Large Language Models to Improve Sentiment Analysis in Latvian language" pieņemts publicēšanai "7th international conference on innovations and creativity"[1] konferencē.

[1] https://icic.liepu.lv/

1 Lielie valodas modeļi (LVM)

Lielie valodas modeļi ir uz transformeru arhitektūras balstīti modeļi ar vairākiem miljoniem un miljardiem parametru, kas apmācīti nevis uz konkrētu dabiskās valodas apstrādes uzdevumu, bet gan uz plašu teksta korpusu bez konkrēta uzdevuma, un uzdevumu risināšanas spēja tajos izriet no padziļinātas valodas sapratnes. Mūsdienās lielākā daļā valodas uzdevumu tiek risināti pamatā izmantojot kādu no lielajiem valodas modeļiem, un, ja nepieciešams, tam pievienojot papildus neironu slāņus un datu apstrādes loģiku rezultātu uzlabošanai. Šajā nodaļā apskatīta lielo valodas modeļu vēsture, arhitektūra un apmācības procesi.

1.1 Vēsture

Vēl pirms pirmo neironu tīklu izstrādes Alans Tjūrings (Alan Turing) 1950. gadā izstrādāja Tjūringa testu, kas ilgi tika izmantots valodas modeļu spējas atdarināt cilvēku komunikāciju.[55] Mūsdienās gan tiek debatēts par šī testa precizitāti, jo tajā ir daudz mainīgo, kas nav fiksēti, piemēram, vērtētāja zināšanas jomā un vai vērtētājs zina, ka sarunājas ar robotu. Tādēļ mūsdienās modeļu novērtēšanai vairāk tiek lietotas lielas datu kopas, kā, piemēram, GLUE (General Language Understanding Evaluation), kas sastāv no 9 dažādiem teksta izpratnes uzdevumiem, tādiem kā sentimenta analīzes, jautājumu un atbilžu noteikšanas un teksta semantiskās līdzības analīzes[58] un SQuAD (The Stanford Question Answering Dataset), kurā modelim jāatbild uz kādu jautājumu, kura atbilde atrodama dotajā Wikipedia rakstā[40], kas pārbauda modeļu spēju analizēt tekstu un atbildēt uz jautājumiem. 1954. gadā Franks Rosenblats (Frank Rosenblatt) izstrādāja pirmo neironu tīklu - vienslāņa perceptronu, kas spēja risināt binārās klasifikācijas problēmas, izmantojot lineāro regresiju. Pēc perceptrona radīšanas neironu tīklu un mākslīgā intelekta nozarē izstrāde ievērojami palēninājās, līdz 20.gs. beigām un 21.gs sākumam, kad ne tikai datorsistēmām pieauga skaitļošanas jauda, bet arī palielinājās pieejamais datu daudzums. Datu digitalizācija un vieglāka pieejamība ļāva izstrādāt labāk ģeneralizējamas sistēmas, pievēršot mazāku uzmanību labāku algoritmu atrašanai.[43]

2017. gadā *Google Brain* izpētes komanda publicēja rakstu "Attention is All you Need", iepazīstinot pasauli ar transformera tipa modeļiem, kas ļāva ievērojami paātrināt dabiskās valodas apstrādes risinājumus un uz kuriem balstījās lielo valodas modeļu izstrāde.[56] Transformeru arhitektūra padziļinātāk aplūkota 1.2. nodaļā.

Gadu vēlāk, 2018. gadā, Google mākslīgā intelekta komanda publicēja BERT (Bidirectional Encoder Representations from Transformers) modeli, kura bāzes versija saturēja 110 miljonus parametru, bet lielā versija saturēja 340 miljonus parametru. BERT tika izstrādāts kā bāzes modelis ar spēju veidot teksta reprezentācijas vektorus. Tam pievie-

nojot vēl vienu papildus slāni un to apmācot, var iegūt modernu dabiskās valodas modeli lielai daļai dabisko valodu apstrādes uzdevumu.[8] 2020. gadā veiktais pētījums atklāj, ka BERT modelis un tā atvasinājumi ir kļuvuši par neatņemamu daļu no dabisko valodu apstrādes pateicoties tā dziļajai valodas izpratnei.[41] 2023. gadā BERT publikācija[8] ir citēta vairāk kā 49000 reižu. 2018. gadā *OpenAI* publicēja arī savu valodas modeli GPT (Generative Pre-trained Transformer) ar 117 miljoniem parametru. GPT modelis līdzīgi kā BERT spēja mācīties no neattīrītiem datiem un tad ar nelielu papildus apmācīšanu specifiskam uzdevumam iegūt nozīmīgus rezultātus tādos uzdevumos kā jautājumu atbildēšanā, loģiskās spriešanas, eksāmenu jautājumu atbildēšanā un tekstu secības noteikšanā.[37]

2019. gadā *OpenAI* publicēja GPT-2, otro versiju GPT modelim, kas sasniedza vēl augstākus rezultātus, pateicoties savam izmēram - 1,5 miljardiem parametru. Tā pat kā GPT, GPT-2 tika apmācīts uz nemarķētu datu kopu un netika apmācīts konkrētu problēmu risināšanai, parādot modeļa spēju apgūt dažādu uzdevumu risinājumus tikai no gana liela daudzuma ar nemarķēta teksta datiem. Modelis tika apmācīts uz *OpenAI* iekšējās WebText datu kopas, un autori secināja, ka modelis nav spējis apgūt visu informāciju, ko varēja no šīs datu kopas, un parametru skaita palielināšana varētu uzlabot modeļa veiktspēju. WebText ir interneta tekstu datu kopa, kas sastāv no 40GB teksta datu, kas iegūti no 8 miljoniem interneta lapu, OpenAI šo datu kopu nav publicējuši plašākai lietošanai.[38]

Sekojot GPT-2 secinājumiem, ka, palielinot modeļa izmērus un apmācības laiku, pieaug to veiktspēja, 2020. gadā *OpenAI* publicēja GPT-3 - 175 miljardu parametru modeli. Tā autori secina, ka gana liels modelis ir spējīgs risināt lielu daļu dabiskās valodas uzdevumu (teksta tulkošana, klasificēšana, jautājumu atbildēšana) bez papildus apmācības, uzdodot uzdevumus brīva teksta formā. Kopā ar GPT-3 publicēšanu *OpenAI* aktualizēja jautājumu par šāda veida spējīgu modeļu sekām, brīdinot, ka lielie valodas modeļi varētu tikt izmantoti dezinformācijas, misinformācijas, akadēmiskā negodīguma un citu sabiedrību negatīvi ietekmējošu procesu veicināšanā.[6] 2021. gadā uz GPT-3 bāzes tika izstrādāts InstructGPT modelis, kas ir tas pats GPT-3 modelis, bet ar pārraudzītās mācīšanas palīdzību apmācīts precīzāk sekot lietotāja instrukcijām un atbildēt uz jautājumiem.[31] 2022. gadā *OpenAI* publicēja jaunu versiju GPT-3 modelim, vēlāk sauktu par GPT-3.5, uz kura bāzes tika veidota populārā ChatGPT platforma, kas pievērsa pasaules uzmanību lielo valodas modeļu spējām.[29] GPT-3.5 (arī saukts par GPT-3.5-Turbo) tika lietots šī darba rezultātu ieguvei.

Dažādas citas organizācijas publicēja līdzīgus modeļus GPT3, piemēram, LaMDA[25], BLOOM [45] un *Meta* izstrādātais LLaMA[54]. Daļa no tiem tika publicēti nevis tikai aiz slēgta API, bet tika publicēts arī to pirmkods ar atvērtā pirmkoda licenci un apmācītie svari, padarot šos modeļus pieejamākus zinātniskajai kopienai.

2023. gada 14. martā *OpenAI* publicēja GPT-4. Modeļa izstrādātāji atsakās sniegt detalizētu informāciju par modeļa izmēru un implementāciju, taču publicētajā rakstā minēts, ka līdzīgi InstructGPT, tas ir apmācīts uz nemarķēta teksta kopas un tad ar pārraudzītās mācīšanās palīdzību pietrenēts (fine-tuned). Pietrenēšanai tika izmantota ar cilvēku un automātisku sistēmu palīdzību radītas datu kopas, kas iemāca modelim precīzāk atbildēt uz vaicājumiem. Šis modelis spēj ne tikai strādāt ar teksta datiem, bet arī kā ievaddatus spēj saņemt attēlus un apstrādāt tajos esošo vizuālo informāciju. Šī spēja ļauj vel labāk veidot vaicājumus modelim, piemēram, kopā ar vaicājumu par putnu sugu, pievienot attēlu lai precīzāk to identificētu.[30]

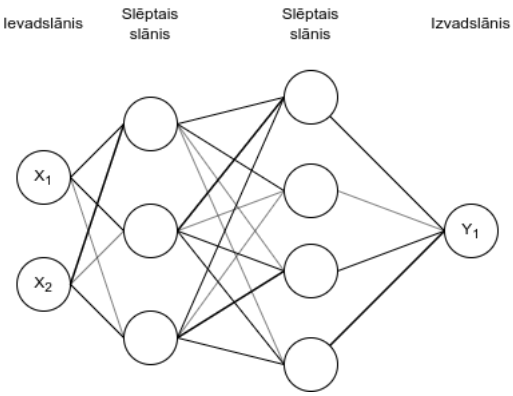
2023.gada 22. martā vietnē futureoflife.org tika publicēta atvērtā vēstule ar nosaukumu "Pause Giant AI Experiments: An Open Letter", kurā izteikts aicinājums mākslīgā intelekta kopienai uz laiku pārtraukt lielāku modeļu par GPT-4 izstrādi un ļaut kopienai un valstu pārvaldei izstrādāt drošības protokolus mākslīgā intelekta sistēmu uzraudzībai un kontrolei. [9]

1.2 Uzbūve

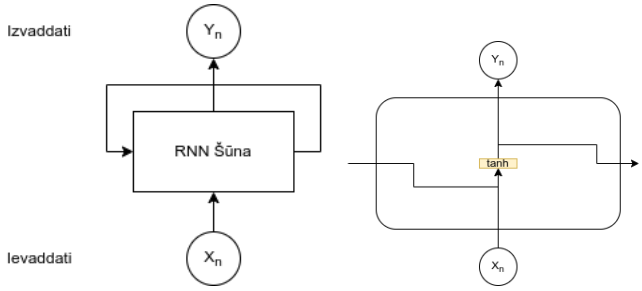
Lai arī sākotnējie neironu tīkli sastāvēja no viena slāņa vai dažiem pilnsaistes slāņiem (neironu slāņi, kurā katrs iepriekšējā slāņa neirons ir savienots ar katru nākošā slāņa neironu) (skat 1. attēlu), jau 1972. gadā Shun-ichi Amari publicēja pirmo rekurento neironu tīklu (RNN).[2] Lai arī pilnsaistes neironu tīkli spēj efektīvi apstrādāt datus, tie nesniedz vēlamos rezultātus datu sekvenču apstrādē, piemēram, dabiskās valodas vārdu virknēm. Rekurentie neironu tīkli ir paredzēti šādām datu sekvencēm, tādēļ tie kopā ar to paveidu LSTM (Long Short-term memory)[16] (skat. 3. attēlu) ilgu laiku sasniedza augstākos rezultātus dabiskās valodas apstrādē un bija pamatā lielai daļai modernāko risinājumu. Rekurentie neironu tīkli apstrādā datus secīgi (skat 2. un 3. attēlu), katrā no iterācijām kā ievaddatus saņemot gan jaunos ievaddatus, gan informāciju no iepriekšējajām iterācijām. Šī arhitektūra ļauj modelim apstrādāt nezināma garuma datu virknes, veicot audio, video, teksta un citu datu ģenerēšanu, tulkošanu un analīzi.[13][44][17]

RNN un LSTM neironu tīkli bija modernākais pieejamais risinājums līdz 2017. gada novembrim, taču tiem bija arī vairākas problēmas, kas liedza sasniegt rezultātus, kādus sasniedza modernākie transformera tipa modeļi. No tām ievērojamākā ir resursi, kas nepieciešami šo modeļu apmācībai, jo, ņemot vērā to arhitektūru, to apmācība ir slikti paralelizējama un salīdzinoši lēna.[56]

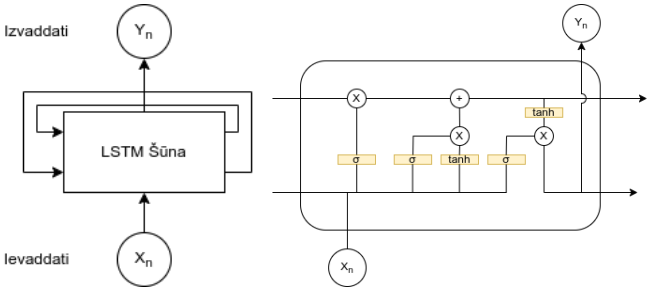
Transformeru modeļi datu sekvenču apstrādi risina citādāk - tie neapstrādā datus rekursīvi un neglabā atmiņā informāciju par jau apskatītajiem datiem. Tā vietā šie modeļi katrā solī saņem visus datus un, izmantojot uzmanības galvas (attention heads), izvēlas, kura informācija ir svarīga konkrētā laika soļa kontekstā. Sīkāk konstrukciju skatīt 4. at-



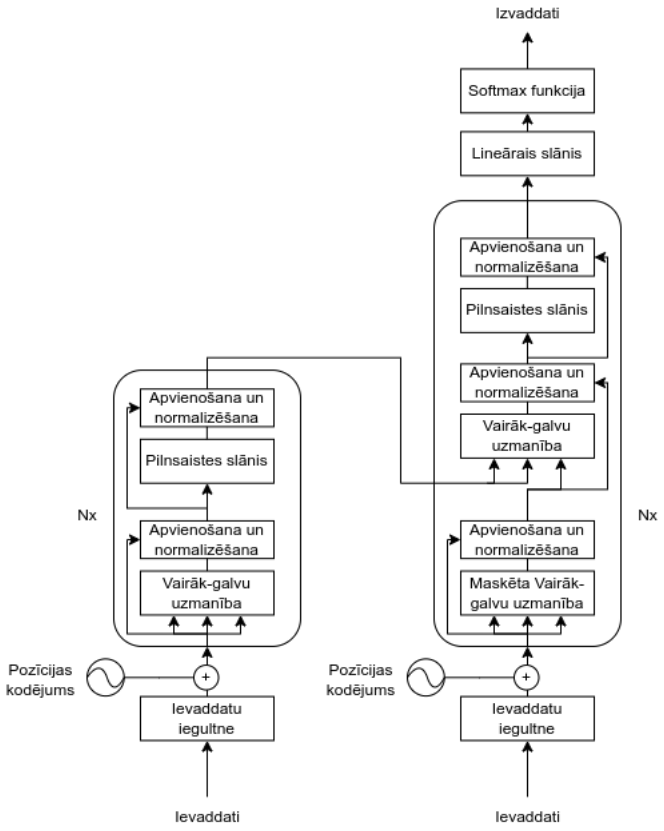
1. att. Pilnsaistes neironu tīkla modelis[12]



2. att. Rekurentā neironu tīkla un šūnas modelis[12]



3. att. LSTM tīkla un šūnas modelis[12]



4. att. Transformera modeļa arhitektūra[56]

tēlā. Pateicoties tam, ka transformeri nav iteratīvi modeļi, to apmācība un lietošana ir ievērojami vieglāka un ātrāka, bet tiem ir ierobežots "logs" ar informāciju, ko iespējams apstrādāt (GPT-3.5 atmiņa ir aptuveni 4,096 tekstvienības).[56]

1.3 Apmācība

Lielo valodas modeļu apmācība ir ļoti dārgs un resursietilpīgs process, kurā ietilpst datu kopu izveide, modeļu izstrāde un apmācība. 2020. gadā veikts pētījums liecina, ka 1,5 miljardu parametru modeļa apmācība izmaksā aptuveni 1,6 miljonus ASV dolāru.[47]

Tā kā lielie valodas modeļi informāciju apgūst no neanotētiem un nemarķētiem datiem, tad to apmācībai tiek izmantots gandrīz viss, kas ir pieejams internetā, tostarp grāmatas, ziņu raksti, populārzinātniskā literatūra un pat koda repozitoriji un dažādu valodu avoti. LVM tiek apmācīti ar nepārraudzītās mācīšanas palīdzību, visbiežāk, uzdodot modelim paredzēt nākamo vārdu teikumā, taču dažkārt, piemēram, BERT gadījumā, modelis mācās noteikt trūkstošu vārdu teikuma vidū vai arī noteikt, vai 2 teikumi loģiski seko viens otram. Visām šīm metodēm kopīgais ir tas, ka modelis apgūst valodas, to īpatnības un sakarības bez konkrēta mērķa.[18]

Dažkārt, piemēram, InstructGPT, GPT-3.5 un GPT-4 gadījumos, kā arī daudzos

BERT atvasinājumos, pēc sākotnējās apmācības seko pārraudzītās mācīšanās posms, kura laikā modelis vai nu tiek sagatavots konkrētu uzdevumu veikšanai, vai apmācīts ģenerēt loģiskākus un lietotājam vēlamākus rezultātus nenovirzoties no uzdotā jautājuma un samazinot liekās informācijas apjomu. Šāda optimizācija ne tikai padara modeļa lietošanu ērtāku un patīkamāku, bet arī samazina izmaksas kas rodas, kad modelis nesniedz pareizu atbildi vai uzģenerē lieku tekstu. [18][30][31]

1.4 Emerģentās prasmes

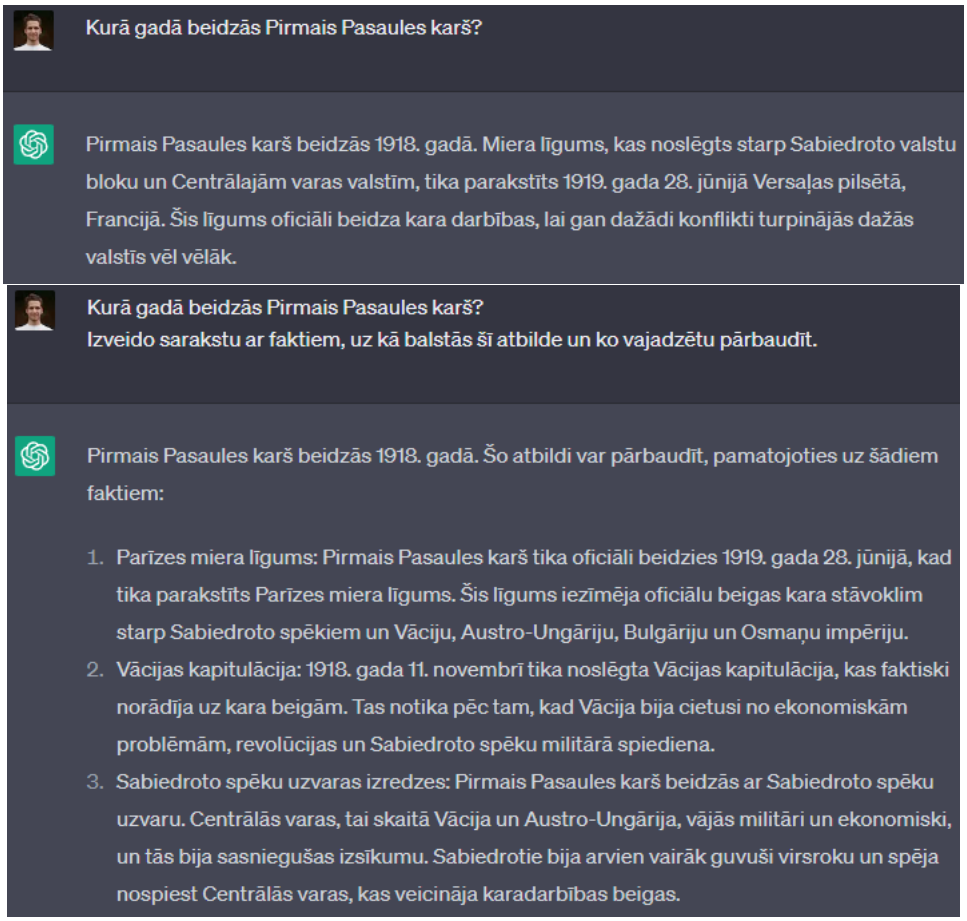
Līdz ar LVM attīstību tiek novērotas arvien jaunas šo modeļu īpatnības un neparedzētas prasmes, kā, piemēram, spēja veidot loģiskus domu pavedienus, veikt aritmētiskas darbības soli pa solim un identificēt vārdu pārnestās vai netiešās nozīmes kontekstā. Šīs prasmes tiek dēvētas par emerģentajām prasmēm (Emergant abilities), jo to rašanās nav paredzamas, aplūkojot mazāku modeļu iespējas. Šīs spējas modeļiem rodas, pieaugot to parametru skaitam un apgūtās informācijas apjomam.[62]

1.5 Ierobežojumi

Kā jau 1.3. minēts, LVM apmācība un uzturēšana ir dārga un bieži vien neizdevīga, salīdzinot ar mazāku transformeru vai citu metožu lietošanu.

Tā kā transformera tipa arhitektūrā ir ierobežots ievaddatu apjoms, tad tas nosaka modeļa spēju un ierobežojumus saistībā ar apstrādājamo teksta garumu.[56]

Lielajiem valodas modeļiem nav morālo vērtību un spējas pateikt "Es nezinu atbildi uz šo jautājumu", kaut arī modelis var "izvēlēties" uzģenerēt šo tekstu, tas bieži mēģinās izveidot loģisku atbildi. Šo ierobežojumu dēļ LVM var viegli radīt viltus informāciju un lietotājs, kas centīsies balstīt savas zināšanas vai rakstu darbus uz LVM ģenerēto tekstu, var viegli paust nepatiesu informāciju. Viens no veidiem, kā risināt šo problēmu, ir ar vaicājumu inženierijas (prompt engineering) palīdzību. Piemēram, vaicājumā iekļaujot prasību modelim izvadīt ne tikai ģenerējamo tekstu, bet arī faktus, kuru patiesumu lietotājam patstāvīgi būtu jāpārbauda (piemērs attēlots 5. attēlā).[63]

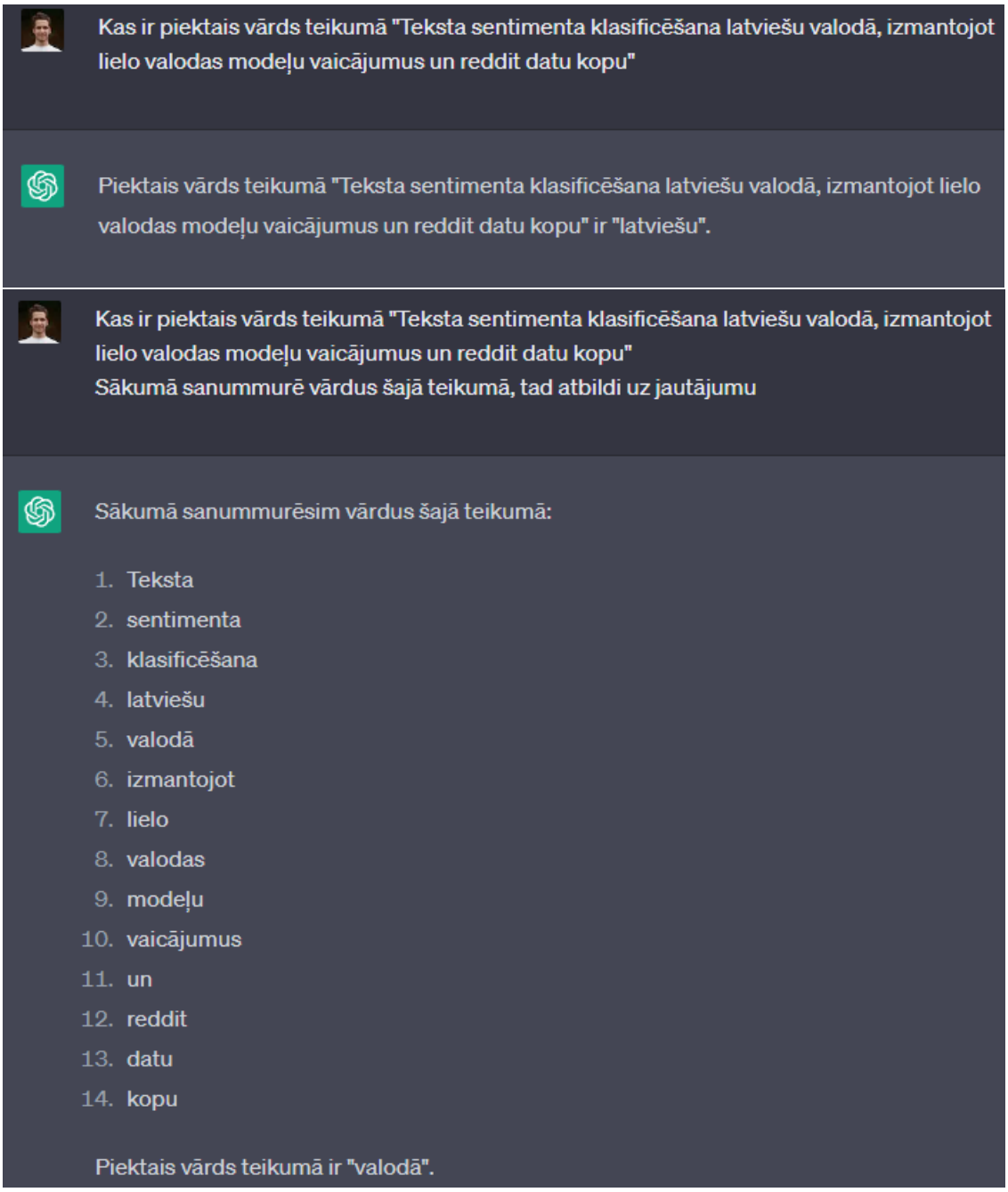


5. att. Piemērs kā iegūt pārbaudāmus faktus uz kā balstīta GPT modeļa sniegtā atbilde.

Analizējot cilvēku problēmu risināšanas un domāšanas metodes, jau sen ir izdalītas 2 smadzeņu ”sistēmas”:

- ”Sistēma 1” (citreiz saukta par ”ātro domāšanu”), kas pieņem ātrus un intuitīvus lēmumus, taču, kuru viegli var apmānīt, ja pareizās atbildes iegūšanai ir nepieciešams iedziļināties jautājumā;
- ”Sistēma 2” (saukta arī par ”lēno domāšanu”), kas patērē vairāk enerģijas un nestrādā tik ātri, taču, pieņemot lēmumus, balstās pamatotos lēmumos un problēmas risina soli pa solim nevis intuitīvi.[52][19]

LVM darbojas kā ”Sistēma 1”, tie nespēj dabiski domāt soli pa solim un veidot sakarīgu domu pavedienu (chain of thought), bet atbild ar ”pirmo atbildi, kas ienāk prātā” kā redzams 6. attēlā.



6. att. Piemērs kā ar pareizu vaicājuma izveidi atrisināt modeļa ”ātro domāšanu” un veicināt padziļinātu domu ķēdi.

Šīs LVM vājības risināšanai ir piedāvāts likt modelim veikt secīgus spriedumus par to, kā tas nonāca pie savas atbildes (skatīt 6. attēlu). Šī tehnika vaicājumu veidošanai ne tikai uzlabo modeļa spriešanas spējas, bet tā arī ļauj lietotājam izprast modeļa ”domu gājienu” un fiksēt loģikas kļūdas tā spriedumos pat gadījumos, kad gala atbilde šķiet loģiska. [59][61][63]

Vēl viens veids kā uzlabot LVM sniegtos rezultātus, ir prasot tam pašam analizēt problēmas ar iepriekšējo atbildi un izlabot kļūdas tajā.[20]

2 Sentimenta analīze

Nodaļā aprakstīts, kas ir sentimenta analīze, biežāk lietotās pieejas, risinājumu novērtēšana kā arī populārākās angļu valodas un latviešu valodas risinājumi.

Sentimentu definē kā subjektivitātes izpausmi caur pozitīvu, neitrālu vai negatīvu viedokli.[21][53]

Sentimenta analīze tiek lietota, lai analizētu sentimenta izpausmi dažādos tekstos to starpā atsauksmēs, sociālajos medijos, ziņās un citās entītijās, par ko autors var izteikt viedokli. Sentimenta analīze ir vispārinātāka forma teksta emocionālās nokrāsas analīzei un klasificē tekstu kā pozitīvi, neitrāli vai negatīvi noskaņotu, neanalizējot konkrētas emocijas kā dusmas, prieku, bēdas, sajūsmu un citas.[53]

Sentimenta analīze ir izaicinošs uzdevums ne tikai automatizētām sistēmām, bet arī daļai cilvēku dažādu valodas izteiksmju. Piemēram, sarkasma, ironijas un negācijas dēļ, var tekstam vai tā fragmentam, atkarībā no konteksta, var tikt piešķirta pretēja nozīme.[34]

2.1 Pielietojumi

Līdz ar Web2.0 (interneta "laikmets", kurā lielu daļu satura veido interneta vietņu lietotāji nevis īpašnieki)[42], uzņēmumiem radās interese par to, kas par tiem tiek teikts sociālajos tīklos. Viens veids, kā to risināt un novērtēt lietotāju un klientu apmierinātību, ir skatīties uz ziņu sentimentu, ko viņi publicē.[26]

Ar sentimenta analīzes palīdzību var analizēt arī sabiedrības viedokli par dažādiem politiskiem lēmumiem, notikumiem, vēlēšanu rezultātiem un indivīdu attieksmi vienam pret otru[36]. Sentimenta analīze arī ir nozīmīgs posms neitrālu un objektīvu ziņu nodrošināšanā.[26]

Sentimenta analīze var palīdzēt arī citos dabisko valodu apstrādes uzdevumos, dodot papildus kontekstu par tekstā pausto informāciju.[26]

2.2 Pieejas

Sentimenta analīzei tiek lietotas divas pieejas: klasificēšana ar mašīnmācīšanās palīdzību un leksikonā balstīta klasificēšana. Ar leksikonu klasificējot tekstu, tiek skatīta atsevišķu vārdu nozīme, piemēram, "jauks" lielākoties norāda uz pozitīvu sentimentu, kamēr "drausmīgs" norāda uz negatīvu. Leksikons parasti tiek iegūts, izmantojot marķētus datus un analizējot vārdu biežumu pozitīvos un negatīvos tekstos, un bieži var tikt tulkots, tādējādi ļaujot veikt sentimenta analīzi citās valodās.[53] Latviešu valodas uzbūves

un brīvā rakstura dēļ leksikonu izveide ir sarežģīta, un dažkārt vārdu nozīmi ir iespējams pateikt tikai kontekstā.[5][32] Mašīnmācīšanā balstīti risinājumi strādā, apmācot neironu tīklu simbolu virkņu klasificēšanai. Tam izmanto korpusu ar marķētiem teksta fragmentiem (bieži mikrobloginga vietņu ierakstiem[10][11][28][35][36][50]), taču bieži to precizitāte slikti ģeneralizējas ārpus sākotnējā apmācību domēna.[53] Sentimenta klasifikatori var būt bināri (pozitīvs vai negatīvs) vai terciāri (pozitīvs, negatīvs vai neitrāls).[53] Klasificēšanas uzdevumos ir tendence izlaist neitrālo klasi pieņemot, ka tā atrodas starp pozitīvo un negatīvo, taču pētījumi liecina, ka neitrālās klases ieviešana var uzlabot rezultātu precizitāti par vairākiem procentiem, ja pieejamā datu kopa ir gana liela.[23]

Sentimentu var analizēt veselu dokumentu vai bloga rakstu līmenī, kā arī atsevišķu teikumu vai aspektu/atribūtu līmenī. Dokumenta un teikuma līmeņa analīzes pamats ir vienāds, jo teikums ir arī uzskatāms par īsu dokumentu. Visa dokumenta analīze var dot precīzāku rezultātu, apstrādājot plašāku kontekstu, kamēr teikuma līmeņa analīze var sniegt detalizētāku informāciju par to, kurās dokumenta vietās pausts kāds sentiments. Aspektu līmeņa analīze fokusējas uz vēl smalkāku analīzi, kurā tiek izdalīts katrs paustais sentiments, piemēram, teikumu "Šodien ir jauks laiks, taču vējš ir nepatīkami dzestrs." var sadalīt divos sentimenta apgabalos - pozitīvs par laikapstākļiem kopumā un negatīvs par vēju. Šis sentimenta analīzes veids ir sarežģītāks, jo pieprasa vel padziļinātāku teksta analīzi un izpratni, tādēļ tā veikšanai gandrīz nekad netiek lietotas metodes, kas nav balstītas dziļajos neironu tīklos. [46]

2.3 Risinājumu novērtēšana

Lai novērtētu dažādu problēmu risinājumus, to starpā mākslīgo neironu tīklu rezultātus, ir nepieciešamas dažādas metrikas. Klasifikācijas uzdevumos metrikas bieži balstās uz pārpratumu matricas (skat 1. tabulu). Pārpratumu matrica attēlo, kā atšķiras divu klasificētāju rezultāti, visbiežāk par patieso vērtību tiek uzskatīti cilvēku marķējumi un paredzētais ir modeļa novērtējums, taču ar šo metodi var salīdzināt arī divu marķētāju rezultātu sakritību.

1. tabula. Pārpratumu tabula trīs klašu sentimenta klasifikācijas uzdevumam

	Prognozētais Poz	Prognozētais Nei	Prognozētais Neg
Patiesais Poz	Patiesais pozitīvais (T_P)	Nepatiesais neitrālais (F_{IP})	Nepatiesais negatīais (F_{GP})
Patiesais Nei	Nepatiesais pozitīvais (F_{PI})	Patiesais neitrālais (T_I)	Nepatiesais negatīais (F_{GI})
Patiesais Neg	Nepatiesais pozitīvais (F_{PG})	Nepatiesais neitrālais (F_{IG})	Patiesais negatīvais (T_G)

Balstoties uz pārpratumu matricā redzamajiem datiem, varam iegūt dažādas metrikas:

- Pareizība (accuracy): Kāda daļa no visiem datiem paredzēta pareizi. Šo metriku var uztvert arī kā varbūtību, ka nejauši izvēlēts piemērs no datu kopas tiks prognozēts pareizi.

$$A = \frac{T_P + T_I + T_G}{T_P + T_I + T_G + F_{PI} + F_{PG} + F_{IP} + F_{IG} + F_{GP} + F_{GI}} \tag{1}$$

- Precizitāte (precision): Kāda daļa no klases prognozēm bija pareizas. Tiek aprēķināta katrai klasei atsevišķi:

$$P_P = \frac{T_P}{T_P + F_{PI} + F_{PG}} \tag{2}$$

$$P_I = \frac{T_P}{T_I + F_{IP} + F_{IG}} \tag{3}$$

$$P_G = \frac{T_P}{T_G + F_{GP} + F_{GI}} \tag{4}$$

$$P = \frac{P_P + P_I + P_G}{3} \tag{5}$$

- Pārklājums (recall): Kāda daļa no klases tika prognozēta pareizi. Tāpat kā pareizību, šo metriku jārēķina katrai klasei atsevišķi.

$$R_P = \frac{T_P}{T_P + F_{IP} + F_{GP}} \tag{6}$$

$$R_I = \frac{T_P}{T_I + F_{PI} + F_{GI}} \tag{7}$$

$$R_G = \frac{T_P}{T_G + F_{PG} + F_{IG}} \tag{8}$$

$$R = \frac{R_P + R_I + R_G}{3} \tag{9}$$

- F1: Apvieno precizitāti un pārklājumu vienā formulā, kas arī apraksta modeļa paredzējumu pareizību, taču vairāk ņem vērā kļūdainos paredzējumus.

$$F1 = 2 \times \frac{P \times F}{P + R} \tag{10}$$

2.4 Līdzšinējie pētījumi

Sentimenta analīzes pētniecība mašīnmācīšanās jomās aizsākās ap 2002. gadu[33], taču tā ir jau sen pētīta joma no sociālo zinātņu skata punkta[1], tādēļ tālāk apska-

tīta daļa pētījumu latviešu valodai 2. tabulā, kā arī SOTA angļu valodā, balstoties uz paperswithcode.com[1] datiem 3. tabulā.

Kā redzams 2. un 3. tabulās, latviešu valodā sentimenta analīze ir ievērojami atpalikusi no modernākajiem modeļiem angļu valodā un daļa problēmas ir labu datu kopu trūkums, ko varētu risināt modeļi, kam nav nepieciešams tik daudz konkrētā uzdevuma treniņdatu. Angļu valodas risinājumi vairs neizmanto teksta virkņu apstrādi bez lielo valodas modeļu palīdzības, un, lai arī latviešu valodā ir pieejams AiLab izstrādātais LVBERT[67], lielo valodas modeļu pielietojums sentimenta analīzē latviešu valodā vēl nav pētīts.

[1]https://paperswithcode.com/task/sentiment-analysis

2. tabula. Sentimenta analīzes pētījumi latviešu valodai

Autors	Darba nosaukums	Gads	Saturs	Analīzes domēns	Rezultāti
Igors Gulbinskis	Digitālo tekstu sentimenta analīze[14]	2010	Izmantoja statistikā balstītu PMI-IR algoritmu 2 klašu klasifikācijai	Blogi, ziņas un Twitter	Ieguva 75% pareizību, kā arī izstrādāja sistēmu latviešu sentimenta monitoringam no sociālajiem tīkliem
Kārlis Gediņš	Automātiskā teksta emocionālās noskaņas noteikšana latviešu valodās[11]	2013	Izmantoja naivā Baiesa metodi 2 klašu klasifikācijai	Twitter	Ieguva 85.7% pareizību
Ginta Garkāje, Evelīna Zilgalve un Roberts Darģis	Normalization and Automatized Sentiment Analysis of Contemporary Online Latvian Language[10]	2014	Izmantoja naivā Baiesa metodi 2 klašu klasifikācijai	Ziņu portālu komentāri	Ieguva 72.2% pareizību
Jānis Peisenieks	Mašīntulkošanas iespējas Twitter sīkziņu sentimenta analīzē[35]	2014	Apvienoja mašīntulkošanu ar sentimenta analīzi angļu valodai	Twitter	Labākos rezultātus sniedz Bing mašīntulkošanas un AlchemyAPI sentimenta analīzes rīku kombinācija sasniedzot 76% pareizību uz 2 klasēm un 35.5% pareizību uz 3 klasēm.

2. tabula. Sentimenta analīzes pētījumi latviešu valodai

Autors	Darba nosaukums	Gads	Saturs	Analīzes domēns	Rezultāti
Gatis Špats	Application of Opinion Mining for written content classification in Latvian text[49]	2015	Salīdzināja leksikonā balstītu, SVM un naivā Baiesa metodes 3 klašu analīzei	Twitter	Leksikonā balstīta analīze: 51% (ar uzlabotu datu kopu: 83%), Naivā Baiesa metode: 61%, SVM: 66%
Gatis Špats un Ilze Birzniece	Opinion Mining in Latvian Text Using Semantic Polarity Analysis and Machine Learning Approach[50]	2016	Salīdzināja leksikonā balstītu un naivā Baiesa metodes 3 klašu analīzei	Twitter	Leksikonā balstīta analīze: 73% Naivā Baiesa metode: 62% (55% netīrai datu kopai)
Dāvis Nicmanis	Sabiedrības attieksmes modelēšana, izmantojot sentimenta analīzi[28]	2017	Izstrādāja LSTM tīklu sentimenta analīzei	Twitter	Izstrādāts korpuss latviešu sentimenta analīzei, Sasniegta 82.2% precizitāte ar LSTM uz angļu valodas tweetiem
Rinalds Vīksna	Emocionālās ekspresijas noteikšana sīkziņās latviešu valodā[57]	2018	Izstrādāja korpusu un salīdzināja dažādas sentimenta analīzes pieejas un augmentācijas	Twitter	Sasniegta 74% precizitāte izmantojot loģistisko regresiju.
Mārcis Pinnis	Latvian Tweet Corpus and Investigation of Sentiment Analysis for Latvian[36]	2018	Izstrādāja korpusu un salīdzināja dažādas sentimenta analīzes pieejas	Twitter	Izstrādāts korpuss latviešu sentimenta analīzei, Sasniegta 76.6% precizitāte izmantojot perceptronu klasifikatoru.

2. tabula. Sentimenta analīzes pētījumi latviešu valodai

Autors	Darba nosaukums	Gads	Saturs	Analīzes domēns	Rezultāti
Uga Sprocgis, Matīss Rikters	What Can We Learn From Almost a Decade of Food Tweets[51]	2020	Izstrādāja korpusu un salīdzināja Perceptronu klasifikatoru[36] pret naivā Baiesa metodi	Twitter	Izstrādāts korpuss latviešu sentimenta analīzei, Sasniegta 61.2% precizitāte izmantojot naivā Baiesa metodi.

3. tabula. SOTA sentimenta analīzes pētījumi angļu valodā

Darba nosaukums	Gads	Saturs	Rezultāti
XLNet: Generalized Autoregressive Pretraining for Language Understanding[65]	2019	Prezentē jaunu prims-apmācītu transformeru modeli, kas izlabo BERT vājības	Izmantojot dažādas klasifikatoru galvas un prezentēto XLNet, modelis sasniedz SOTA dažādos uzdevumos, to starpā sentimenta analīzē gūstot 96.2% pareizību (SOTA) uz IDBM[24] datu kopas un 97% pareizību us SST-2[48] datu kopas
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[39]	2019	Prezentē satvaru ar kā palīdzību dabisko valodu apstrādes uzdevumus var pārveidot par tādu, ko var risināt transformera tipa modelis, saņemot ievaddatos tekstu un izdodot citu tekstu.	sasniedz vairākus SOTA rezultātus, to starpā 97.5% pareizību uz SST-2 datu kopas.

3. tabula. SOTA sentimenta analīzes pētījumi angļu valodā

Darba nosaukums	Gads	Saturs	Rezultāti
Unsupervised Data Augmentation for Consistency Training[64]	2019	Apskata kā ievaddatu augmentāciju uzlabošana, pievienojot vairāk datu trokšņu, palīdz lielajiem modeļiem apgūt informāciju un uzlabo rezultātus.	Sasniedz 97.4% pareizību Amazon Review Polarity[27] datu kopā, 65.8% pareizību Amazon Review Full[27] datu kopā un 95.8% pareizību IMDb datu kopā
An Algorithm for Routing Vectors in Sequences[15]	2022	Prezentē veidu, kā apstrādāt ievaddatus, samazinot ne-nepieciešamās informācijas apjomu, tādējādi ļaujot iegūt mazākus un ātrāk apmācāmus modeļus.	Sasniegta 96.2% pareizība IDBM datu kopā, 96% pareizība SST-2 datu kopā un jauns SOTA rezultāts - 59.8% pareizība SST-5[48] datu kopā
XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond[3]	2021	Izstrādā vairākvalodu modeli, kas trenēts uz vairāk kā 30 valodām ieskaitot latviešu	Sasniedz 79.5 F1 TweetEval[4] datu kopā angļu valodai. Neizvērtē latviešu valodas pareizumu.
Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study[60]	2023. 10. Aprīlis	Apskata iespēju izmantot GPT-3.5-Turbo (ChatGPT) modeli sentimenta klasificēšanas un citiem līdzīgiem uzdevumiem	Sasniedz 93% pareizību SST-2 datu kopai neapsteidzot pašreizējo SOTA bet sasniedot tai tuvu rezultātu. Analīze strādā GPT modeļa pēdējās reprezentācijas vērtību, klasificējot ar lineāru klasifikatoru.

3 Korpuss

Līdzšinējie pētījumi latviešu valodā lieto Twitter un ziņu portālus kā datu kopas avotus. Šajā darbā tiek apskatīta iespēja iegūt datus no foruma Reddit, kā arī izveidot vēl nebijušu datu kopu, balstoties uz LVM rezultātiem.

Līdzšinējās datu kopas latviešu valodā (sīkāk apskatītas 4. tabulā) ir salīdzinoši nelielas un nepietiekamas, lai uz tām apmācītu modeļus, kas līdzvērtīgi angļu valodas modernākajiem risinājumiem (skatīt 3. tabulu) kā to pierāda Dāvis Nicmanis[28].

4. tabula. Latviešu sentimenta datu kopas

Nosaukums	Izmērs	Pozitīvi piemēri	Neitrāli piemēri	Negatīvi piemēri	Datu avots
latvian-tweet-sentiment-corpus[35]	1177	383	627	167	Twitter
LV-twitter-sentiment-corpus[28]	2272	797	1223	252	Twitter
Latvian Twitter Eater Corpus cilvēka marķētais[51]	5420	1631	2507	1282	Twitter
Latvian Twitter Eater Corpus automātiski marķētais[51]	18130	2976	14926	228	Twitter
sikzinu_analize[57]	3682	935	2208	539	Twitter
OM[50]		3104	2617	506	Twitter

3.1 Datu iegūšana

Par datu kopas avotu tika izvēlēts forums Reddit[1] ar mērķi apskatīt līdz šim neizmantotus datu avotus. Tika izvēlēts /r/latvia[2] apakšforums, kurā tiek apspriesti ar Latviju saistītie temati gan latviešu, gan angļu valodās. Šis ir arī lielākais latviešu apakšforums Reddit vietnē.

Reddit dod piekļuvi tajā publicētajiem datiem caur API, taču tam ir ierobežojumi, piemēram, caur šo API iespējams piekļūt tikai pēdējiem 1000 publicētajiem rakstiem. Lai apietu šo ierobežojumu tika izvēlēts alternatīvs API, ko piedāvā pushsift.io[3] vietne, kas ļauj piekļūt arī senākiem datiem. Tika izstrādāta programma, kas iteratīvi pieprasīja Reddit vietnē publicētos rakstus no API[4]. Katram rakstam, no oficiālā Reddit API[5] tika pieprasīti tam piesaistītie komentāri.

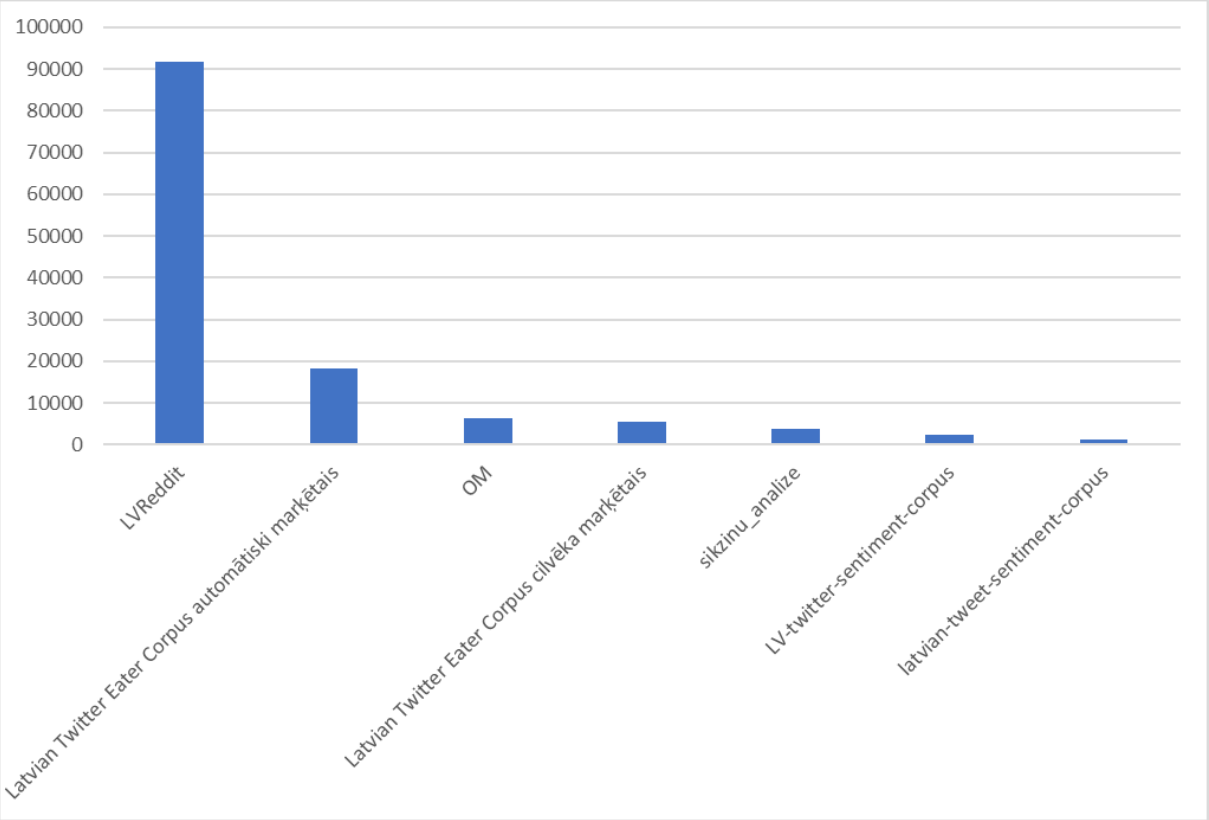
[1] https://reddit.com
[2] https://reddit.com/r/latvia
[3] https://github.com/pushshift/api
[4] https://api.pushshift.io/reddit/search/submission
[5] https://www.reddit.com/r/latvia/comments/post_id.json

Lai izveidotu datu kopu latviešu valodā tika izmantota publiski pieejama Python programmēšanas valodas bibliotēka langdetect[1], kas katram teksta piemēram noteica valodu. datu ieguves un analīzes pirmkods pieejams GitHub repozitorijā[2].

3.2 Datu analīze

Izstrādātā datu kopa satur 3028 rakstus latviešu valodā kopā ar 88793 komentāriem latviešu valodā un tālāk darbā tiek saukta par LVReddit datu kopu. Tās izmēra salīdzinājums ar līdzšinējajām datu kopām attēlots 7. attēlā.

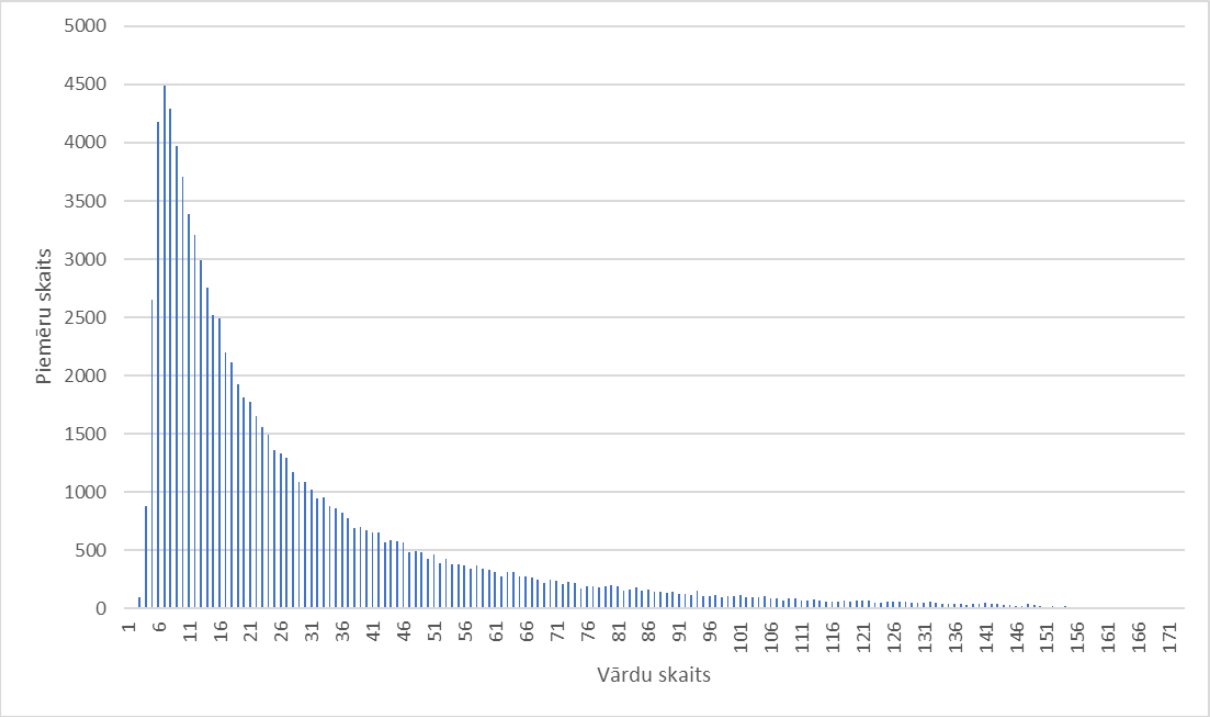
7. att. Datu kopas izmēra salīdzinājums ar līdzšinējajām datu kopām



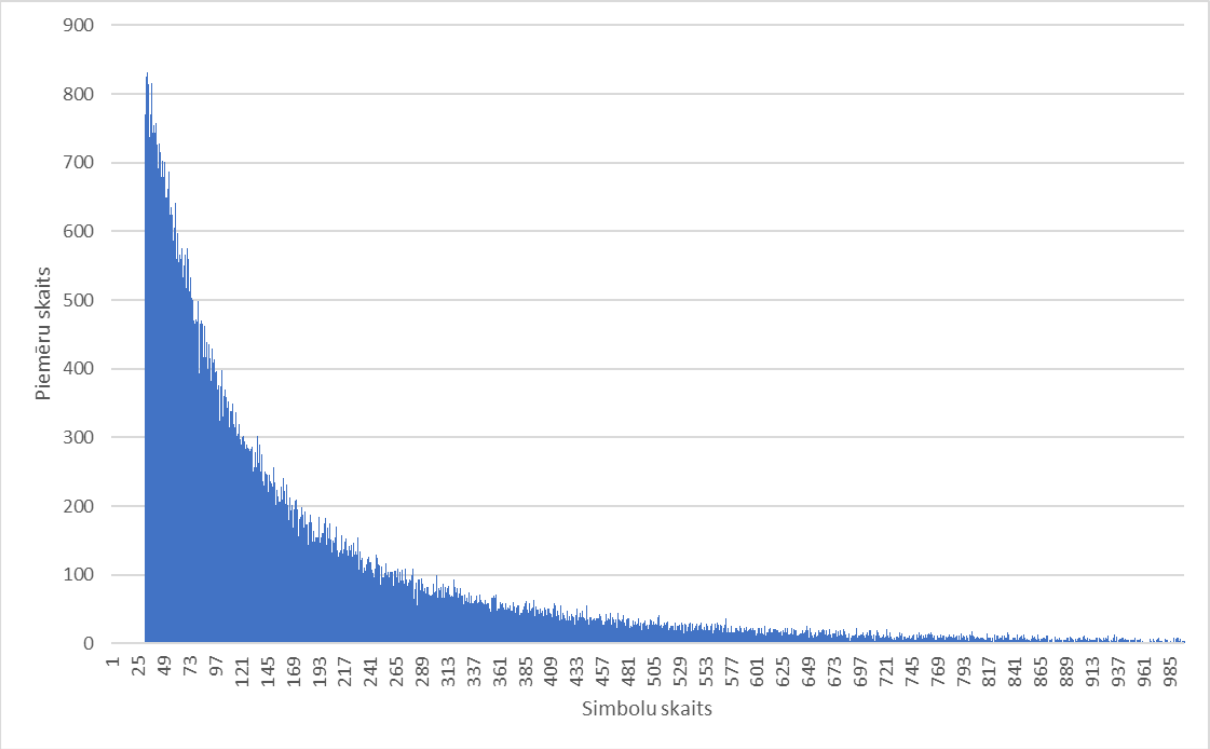
Tā satur 2621970 vārdus (183625 unikālus vārdus, skaitā ietverot arī ciparus un emocijzīmes, no kuriem 170890 sastāv tikai no burtiem) (skat. 8. att) un 16712879 simbolus (739 unikālus simbolus) (skat. 9. att).

[1]https://pypi.org/project/langdetect/
[2]https://github.com/Puupuls/LVRedditCorpus

8. att. Vārdu skaits datu kopas piemēros



9. att. Simbolu skaits datu kopas piemēros



641 no 739 (97.9%) simboliem, kas iekļauti datu kopā, nav latviešu alfabēta simboli vai latviešu valodā lietotas rakstzīmes, tie veido 0.32% no datu kopā iekļautā teksta garuma.

Pirmēri no datu kopas:

1. Izskatās labi! Dievinu pixel art veidīgas spēles
2. Man šķiet gadalaiki mums ir lieliski, un tagad arī ziemas nav tik aukstas. Man patīk!
3. Man patika aptauja! Lika uz brīdi aizdomāties, bezsmadzeņu skrollēšanas laikā.
4. VISIEM LIELS PALDIES PAR ATSAUCĪBU! Tiešām šeit ir labākie!
5. Visvairāk priecē tas, ka tev ir laba un laicīga paškritika.
6. Elektroenerģijas patēriņš dzīvoklī 133,590 KWh - 16,81 eur Jaudas obligātā iepirkuma komponente par ampēriem - 12,00 eur
7. Par ko tur vispār ir? Centos iesākt skatīties, bet kkā palika borong. Vot One Punch Man ir kaut kas pavisam cits...
8. Vectēva pasakas :) un kādā Bolderājas klubā var būt.
9. DPD kurjers ar tevi sazināsies ;)
10. Parunā arī ar saviem skolotājiem, viņi iespējams varēs palīdzēt.
11. Cool story, nacist. 21. gadsimts saka čau, tev rīt uz skolu.
12. Labi, ka neaicina ārstēt kovidu ar Rigviru. Mildronātu gan pieteica konkursam uz potenciālajām kovidzālēm. Besī man šitais personāžs.
13. Cereju atbraukt uz Laviju izbaudīt ziemu. Forša ziema kā Anglijā. Pizdec
14. Nu ja, Jelgavas Tehnikums jau tur pat ir.. neko labāku negaidīti no tā rajona :D
15. Liels paldies okupantiem, ka atņēma!

4 Metodoloģija

Vaicājumu izstrāde tika veikta iteratīvā procesā, tos novērtējot izmantojot latvian-tweet-sentiment-corpus[35] datu kopu.

Novērtēšanai netika izmantota pilna datu kopa, bet gan tika izveidota validācijas apakškopa, kas saturēja vienādu skaitu katras klases piemēru, samazinot datu kopu no 1177 uz 501 piemēru.

Balstoties uz nesenajiem pētījumiem jomā ([7][63][22][66][60]) tika izstrādāti 7 vaicājumi angļu valodā, kuru veiktspēja tika novērtēta uz validācijas kopas. Atbildes, kuru sentimentu nevarēja noteikt izmantojot regulārās izteiksmes, tika uzskatītas par neitrālām (5. tabulā uzrādīts cik procentus no atbildēm varēja apstrādāt).

Labākie vaicājumi tikai ietverti vaicājumā ChatGPT par to kā tos uzlabot, kas deva vel 6 vaicājumus angļu valodā, kā arī 3 vaicājumus latviešu valodā.

Vaicājumi tika atkārtoti novērtēti un un tika veikta vel viena iterācija vaicājumu uzlabošanai ar ChatGPT rīka palīdzību, sniedzot vel 4 vaicājumus angļu un 4 vaicājumus latviešu valodā.

Visi rezultāti tika apkopoti, attēloti 5. nodaļā. Tika izvēlēts vaicājums, kas ieguva labāko rezultātu uz validācijas datu kopas un veikts eksperiments, lai noskaidrotu, kāds ieguvums ir vairaku iterāciju veikšanai un vidējās atbildes iegūšanai.

Izvēlētais vaicājums tika pielietots LVReddit datu kopas apstrādei kā arī pārējo datu kopu analīzei apskatot metodes ieguvumus pār autoru lietotajām metodēm. No LVReddit datu kopas tika atdalīta apakškopa saturot 100 piemērus katrā klasē. Apakškopu neatkarīgi novērtēja divi cilvēki, nosakot katra piemēra sentimentu, lai novērtētu izstrādātās datu kopas pareizumu.

5 Rezultāti

5. tabulā attēloti izstrādātie vaicājumi un to gūtie rezultāti. Jāņem vērā, ka šie vaicājumi tika izmantoti ievades datiem latviešu valodā, bet modelis, dažkārt, iegūst labākus rezultātus ar vaicājumiem angļu valodā, kurus tas saprot ievērojami labāk, jo tā apmācību kopā ir bijis ievērojami vairāk datu angļu valodā. Šie paši dati attēloti arī 10. attēlā. Vislabākos rezultātus sasniedza vaicājums "Based on the tone of the text, what is your overall impression? Choose one of the following: Positive, Negative, or Neutral.", sasniedzot 82% pareizību. Vissliktāko rezultātu sasniedza vaicājums "Skalā no 1 līdz 10, kur 1 ir ļoti negatīvs un 10 ir ļoti pozitīvs, kā vērtētu noskaņu, kas izpaužas šajā teikumā?", sasniedzot 34.33% un tikai par 1% pārsniedzot rezultātus kas būtu sagaidāms, un pielīdzināms nejauši izvēlētām atbildēm.

5. tabula. Izstrādātie vaicājumi un to rezultāti

Nr.	Vaicājums	Pareizība	F1	Apstrādāto atbilžu %
1.	On the scale from negative to neutral to positive, the sentiment of this sentence is as follows:	70,4%	0,713	98,0%
2.	List the most notable things that show the sentiment (no more than 2). Then state the sentiment in one word (negative, neutral, positive) defaulting to neutral.	70,0%	0,742	95,2%
3.	List the most notable things that show the sentiment of the text (no more than 2). Describe your confidence (in percent each on a new line) that this being positive, negative, and neutral text.	71,1%	0,713	94,2%
4.	Translate this text and then list the most notable things that show the sentiment that show the sentiment of the text (no more than 2). Describe your confidence (in percent each on a new line) that this being positive, negative, and neutral text.	72,3%	0,726	97,6%

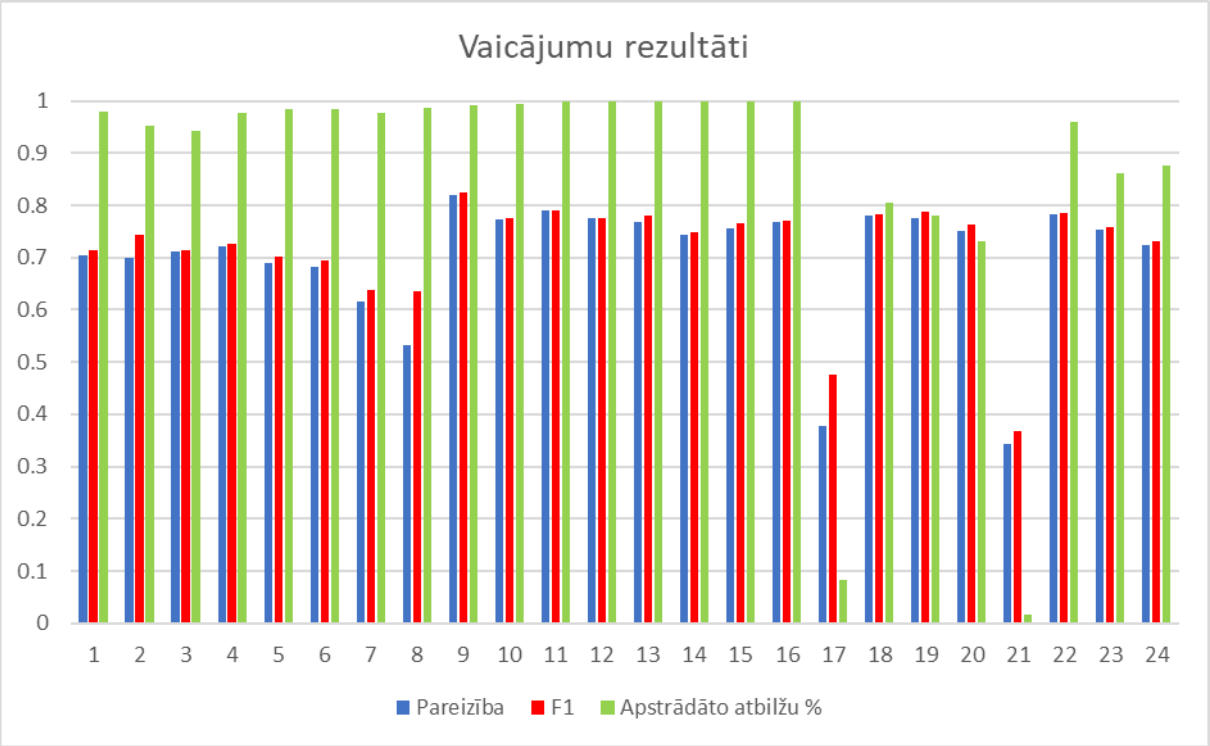
5. tabula. Izstrādātie vaicājumi un to rezultāti

Nr.	Vaicājums	Pareizība	F1	Apstrādāto atbilžu %
5.	First, list the most notable things that show the sentiment of the text (no more than 2). Then replace NUMBER with your sentiment prediction (between 0=negative and 10=positive) in the following sentence: "The sentiment is: NUMBER."	69,0%	0,701	98,4%
6.	Translate this text and then list the most notable things that show the sentiment of the text (no more than 2). Then replace NUMBER with your sentiment prediction (between 0=negative and 10=positive) in the following sentence: "The sentiment is: NUMBER."	68,2%	0,695	98,4%
7.	Finish the sentence "The sentiment is..." with a sentiment value between 0=negative and 10=positive. Do not add anything else except one number.	61,7%	0,638	97,6%
8.	How does this sentence make you feel? Choose one of the following: Positive, Negative, or Neutral.	53,3%	0,636	98,8%
9.	Based on the tone of the text, what is your overall impression? Choose one of the following: Positive, Negative, or Neutral.	**82,0**%	**0,825**	99,2%
10.	How likely is it that the author of this text has a positive or negative attitude towards the subject matter? Choose one of the following: Positive, Negative, or Neutral.	77.2%	0,776	99.4%
11.	What is the general sentiment of this sentence? Choose one of the following: Positive, Negative, or Neutral.	79.0%	0,790	**100%**
12.	How would you describe the author's overall mood or attitude in this text? Choose one of the following: Positive, Negative, or Neutral.	77.4%	0,775	99.8%

5. tabula. Izstrādātie vaicājumi un to rezultāti

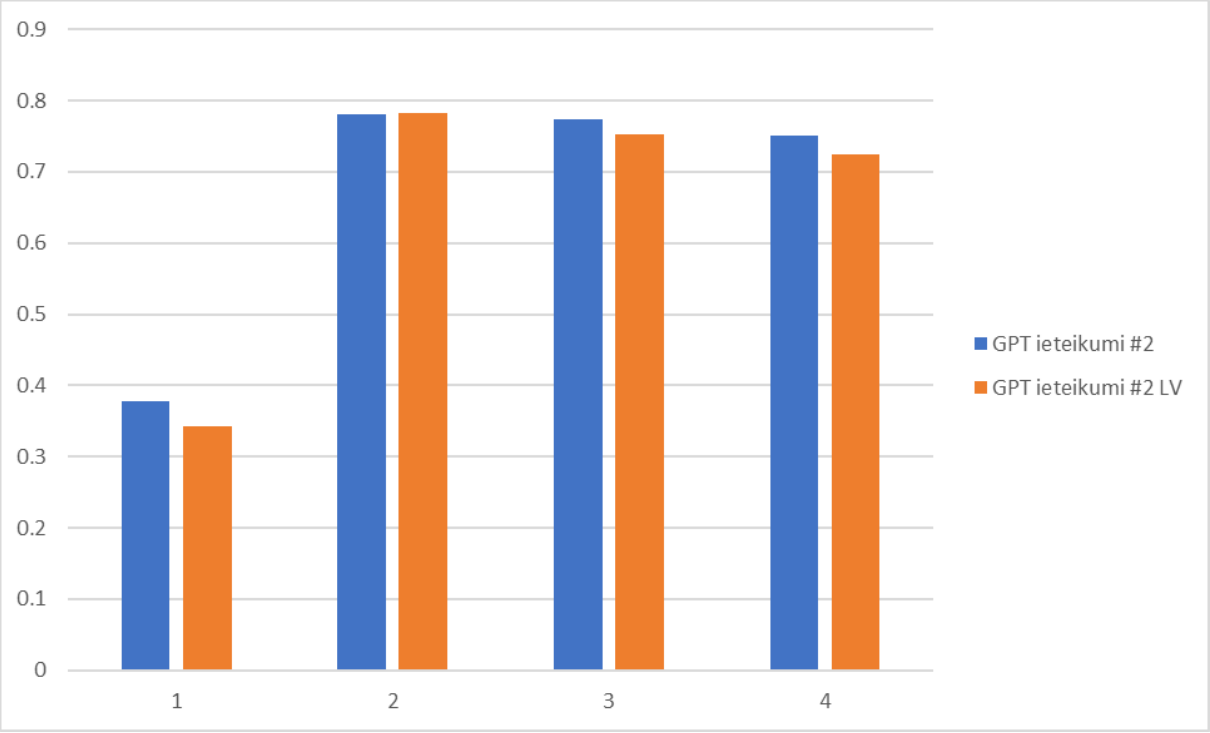
Nr.	Vaicājums	Pareizība	F1	Apstrādāto atbilžu %
13.	What is your overall impression of the sentiment in this sentence? Choose one of the following: Positive, Negative, or Neutral.	76.8%	0,781	**100%**
14.	Kāda ir šī teksta emocionālā nokrāsa? Izvēlies vienu no: Pozitīva, Negatīva vai Neitrāla.	74,3%	0,749	**100%**
15.	Pamatojoties uz teksta toni, kāds iespaids par to rodas? Izvēlies vienu no: Pozitīva, Negatīva vai Neitrāla.	75,6%	0,766	**100%**
16.	Kādu attieksmi pauž teksta autors? Izvēlies vienu no: Pozitīva, Negatīva vai Neitrāla.	76,8%	0,770	**100%**
17.	On a scale of 1 to 10, where 1 is highly negative and 10 is highly positive, how would you rate the sentiment displayed in this sentence?	37.7%	0,475	8.4%
18.	Based on the tone of the text, would you categorize the overall sentiment as positive, neutral, or negative?	78,0%	0,783	80.4%
19.	Does the author's language in this sentence indicate a positive, neutral, or negative sentiment?	77.4%	0,788	78.0%
20.	What emotional tone is exhibited in this sentence? Would you categorize it as positive, neutral, or negative?	75,0%	0,764	73.1%
21.	Skalā no 1 līdz 10, kur 1 ir ļoti negatīvs un 10 ir ļoti pozitīvs, kā vērtētu noskaņu, kas izpaužas šajā teikumā?	34.33%	0,368	1.6%
22.	Balstoties uz teksta toni, kāda ir kopējā noskaņa - pozitīva, neitrāla vai negatīva?	78.2%	0,786	96.0%
23.	Vai autora valoda šajā teikumā norāda uz pozitīvu, neitrālu vai negatīvu noskaņu?	75,2%	0,759	86.2%
24.	Kāda emocionālā nokrāsa izpaužas šajā teikumā? Vai to varētu kategorizēt kā pozitīvu, neitrālu vai negatīvu?	72,5%	0,731	87.6%

10. att. Izstrādāto vaicājumu rezultāti

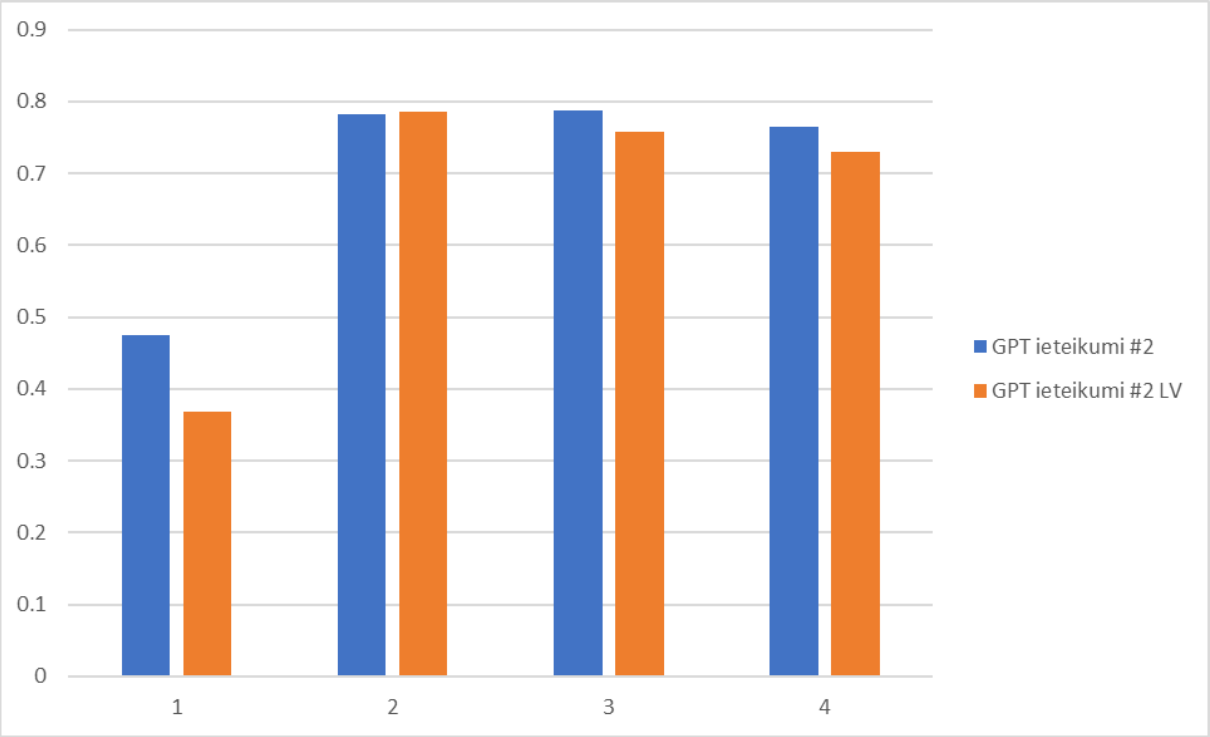


Pēdējajā iterācijā veidotie 4 angļu un 4 latviešu vaicājumi veidoti vienādā formātā, tādējādi ir iespējams salīdzināt GPT-3.5-Turbo veiktspēju vaicājumiem, kas uzdoti angļu un latviešu valodās, šie salīdzinājumi attēloti 11., 12. un 13. attēlos.

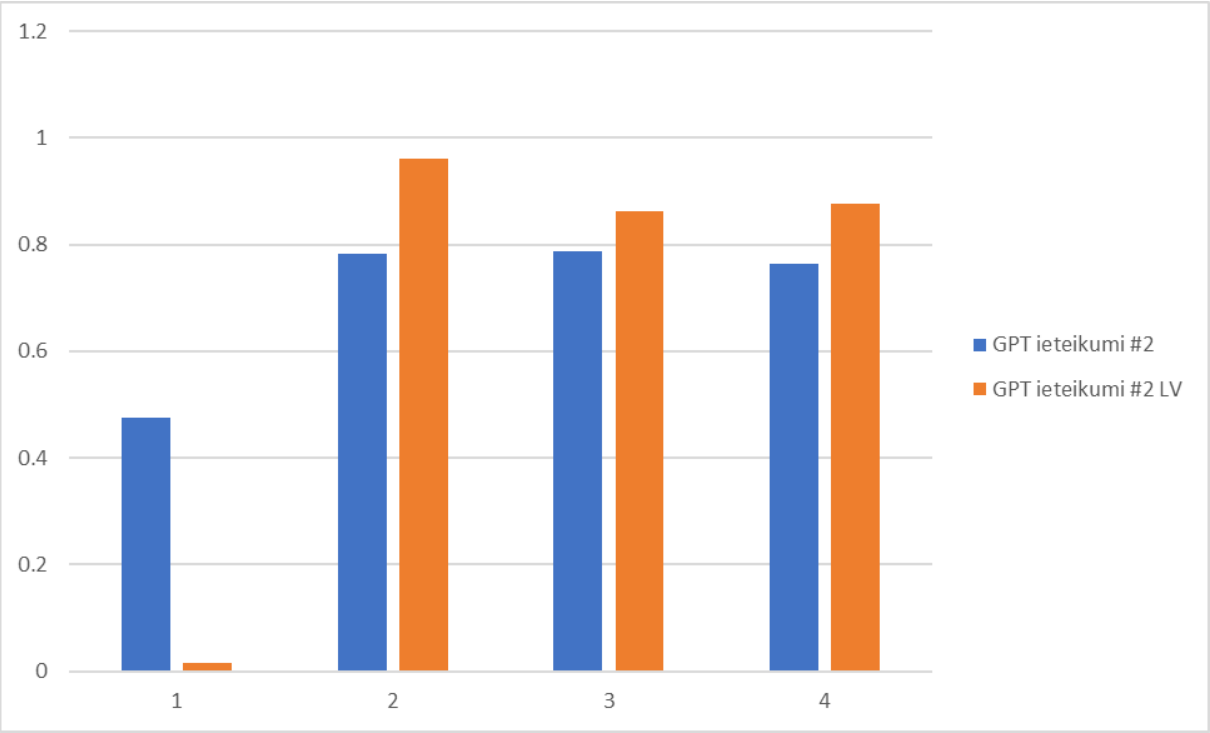
11. att. Latviešu un angļu vaicājumu pareizības salīdzinājums



12. att. Latviešu un angļu vaicājumu f1 salīdzinājums

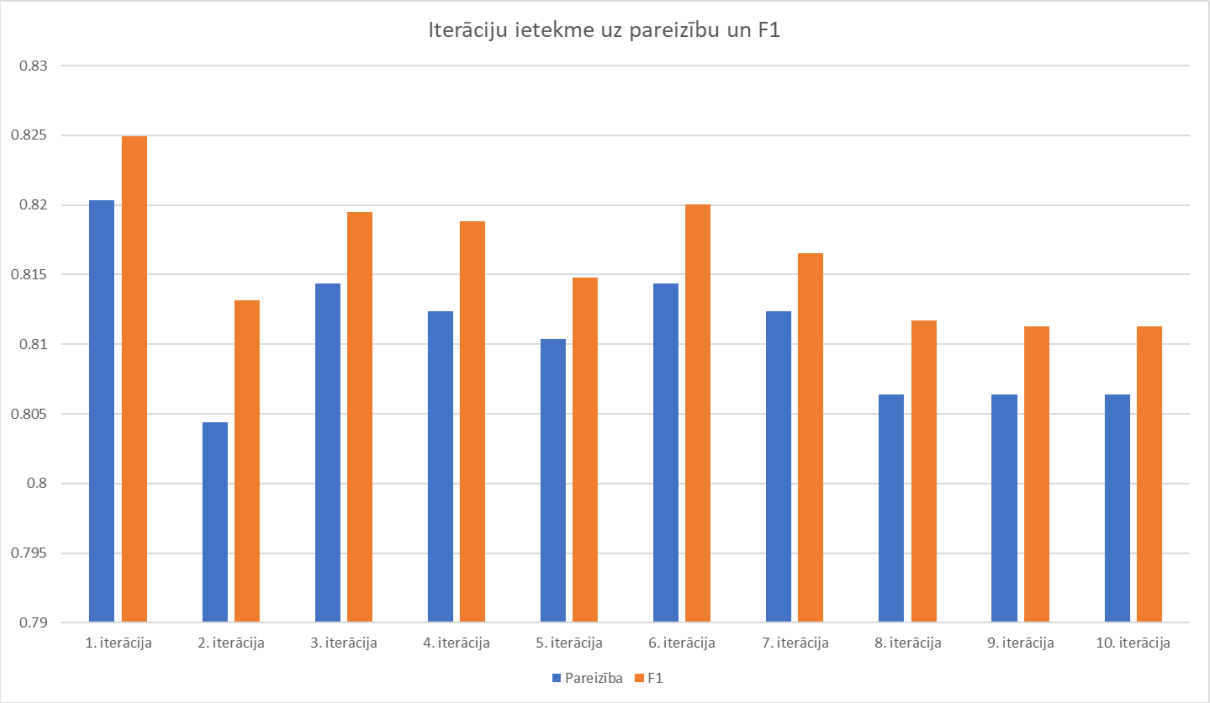


13. att. Latviešu un angļu vaicājumu apstrādāto atbilžu % salīdzinājums



14. attēlā attēlota atkārtotu vaicājumu veikšanas ietekme uz rezultātu pareizību. Katrai iterācijai, par marķēto vērtību tika pieņemta visu līdzšinējo iterāciju rezultātu vidējā vērtība.

14. att. Atkārtojumu ietekme uz rezultātiem



6. tabulā attēlota labākā vaicājuma pareizība, salīdzinot ar dažādu datu kopu autoru sākotnēji publicētajiem rezultātiem. LVM vaicājums pārspēja datu kopu autoru publicētos rezultātus 3 no 5 gadījumiem, OM datu kopā sasniedzot līdzvērīgu rezultātu un par 10% atpaliekot sikzinu_analize datu kopā.

6. tabula. Rezultātu salīdzinājums ar līdzšinējajām datu kopām

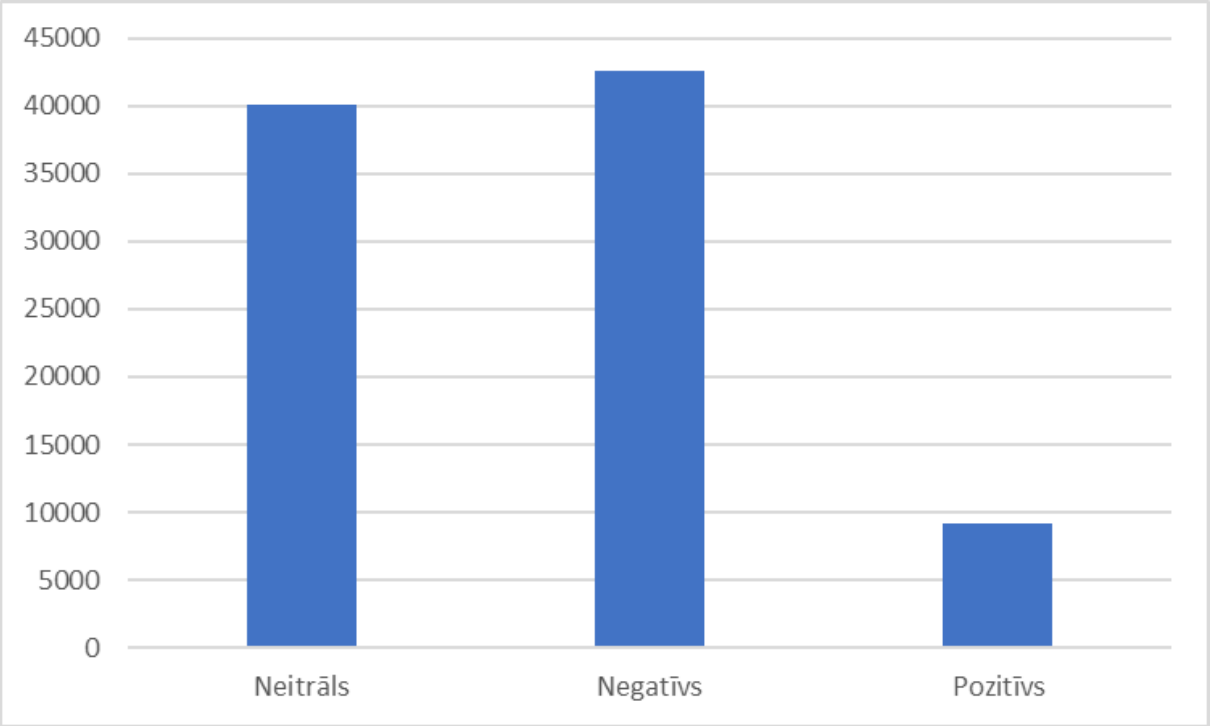
Datu kopa	LVM Vaicājums	OM leksikona rezultāti	Oriģinālais rezultāts
LVReddit	**70,4**%*	43,0%*	-
latvian-tweet-sentiment-corpus	**82,0**%	60,4%	35,5%[35]
LV-twitter-sentiment-corpus	**62,2**%	53,4%	-**[28]
Latvian Twitter Eater Corpus validācijas kopa	**65,1**%	49,7%	53,9%[51]
sikzinu_analize	62.0%	54,0%	**72,6**%[57]
OM	72,6%	**73**%[50]	62%[50]

* Rezultāts iegūts izvērtējot apakškopu

** Darbā netika iegūti rezultāti

Iegūtā LVReddit datu kopa sastāv no 91821 piemēriem, kuru sadalījums klasēs redzams 15. attēlā.

15. att. Klašu sadalījums LVReddit datu kopā

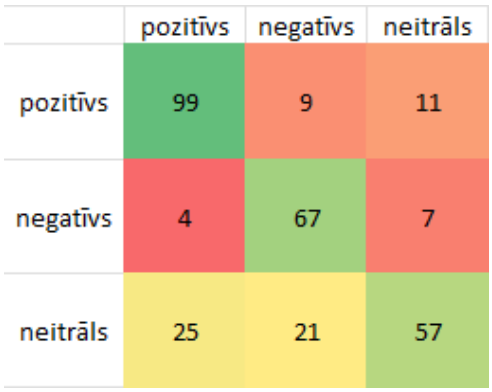


Datu kopā tika salīdzināta gan modeļa pareizība salīdzinot ar cilvēku marķējumu(skat. 16. attēlu), gan cilvēku marķējumi viens pret otru (skat. 17. attēlu). Cilvēku un modeļa rezultātu salīdzināšanai tika ņemti piemēri, kuros cilvēku balsojumi sakrita.

16. att. Cilvēku marķējumu un GPT-3.5-turbo marķējumu pārpratumu matrica

	pozitīvs	negatīvs	neitrāls
pozitīvs	74	7	18
negatīvs	2	52	13
neitrāls	12	14	31

17. att. Cilvēku marķējumu pārpratumu matrica



	pozitīvs	negatīvs	neitrāls
pozitīvs	99	9	11
negatīvs	4	67	7
neitrāls	25	21	57

Cilvēku savstarpējā pareizība bija 74,3% bet cilvēku un GPT-3.5-Turbo savstarpējā pareizība bija 70,4%. Individuālie rezultāti katram cilvēkam pret modeli bija 64,3% un 61,7%.

6 Secinājumi

Lielie valodas modeļi ir pielietojami dažādu uzdevumu risināšanā un piedāvā ievērojamu uzlabojumu arī dabisko valodu apstrādē valodās, kuras nav angļu valoda. Salīdzinot ar iepriekš gūtajiem rezultātiem, lielais valodas modelis GPT-3.5-turbo ar nulles šāviena metodi sniedz ievērojamu uzlabojumu 3 klašu sentimenta analīzē, sasniedzot 82% pareizību latvian-tweet-sentiment-corpus datu kopā un vairāk kā dubultojot Jāņa Peisenieka iegūto 76% pareizību divu klašu analīzē un 35,5% pareizību 3 klašu sentimenta analīzē.

Rezultātu pareizību uzlabot varētu speciāli izstrādāts modelis ar atbilžu klasifikāciju, kurš netika apskatīts šī pētījuma ietvaros [60].

Salīdzinot latviešu un angļu valodās vaicājumus, latviešu valodā veidotie vaicājumi, sasniedza līdzvērtīgus un dažkārt labākus rezultātus kā angļu valodā veidotie vaicājumi. Šie rezultāti varētu būt izskaidrojami ar pašu piemēru valodu, kas ļauj modelim darboties vienā valodā. Šī rezultātu līdzība parāda modeļa dziļo valodas izpratni, neatkarīgi no valodas.

LVM attīstība atvieglo jaunu datu kopu izstrādi, ļaujot izstrādāt marķētas datu kopas ar cilvēku marķētājiem pielīdzināmu pareizību par ievērojami zemāku cenu. Vietnē upwork[1], kas ļauj par samaksu nolīgt dažādu darbu veicējus, par marķēšanas stundu jāmaksā 10-25 dolārus stundā. Vaicājumu salīdzināšana un visas LVReddit datu kopas marķēšana izmaksāja 63 dolārus.

Salīdzinot vaicājuma gūtos rezultātus ar līdzšinējo autoru datu kopu rezultātiem, LVM vaicājums pārspēja līdzšinējos rezultātus trīs no piecām datu kopām. Divās datu kopās rezultāts bija zemāks par 0.4% un 10,6%.

Ievērojamā atšķirība dažādās datu kopās var būt ziņu garumu un satura atšķirībās, dažādās datu atlases pieejās, kā arī patieso vērtību noteikšanas metodikā.

[1] https://www.upwork.com/

7 Turpmākie pētījumi

Darbā aplūkotā tēma raisa padziļinātus jautājumus dažādās jomās, kurus aplūkot turpmākos pētījumos. Pirmkārt, varētu apskatīt angļu valodas SOTA (State-of-the-Art) pielietojumu latviešu valodā, izmantojot internacionālos modeļus, piemēram, RoBERTa. Tas ļautu labāk izprast, kā šie modeļi varētu tikt pielāgoti latviešu valodai un kādi uzlabojumi būtu nepieciešami.

Vēl, varētu pētīt esošo fundamentālo lielo valodas modeļu pielāgošanu latviešu valodai, lai uzlabotu to veiktspēju un precizitāti. Tas varētu ietvert gan modeļu arhitektūras izmaiņas, gan papildu apmācību datu kopu izmantošanu.

Iespējams, pētījums varētu būt arī būtu lielo un mazo burtu lietojuma ietekmes uz vaicājumiem aplūkošana, lai noteiktu, vai tas ietekmē modeļa veiktspēju un kādas korekcijas būtu nepieciešamas.

Arī, varētu padziļināti apskatīt valodas ietekmi uz vaicājumiem, lai saprastu, kāda ir modeļa jutību pret valodas specifiskām īpatnībām un kā tas varētu ietekmēt rezultātus.

Varētu izstrādāt attīrītu datu kopu, kas ļautu uzlabot modeļa veiktspēju un precizitāti, novēršot kļūdas un neprecizitātes treniņa datu kopā.

Kā arī, apskatīt lielo valodas modeļu lietošanu aspekta līmeņa sentimenta analīzē, lai labāk saprastu, kā šie modeļi varētu tikt pielāgoti šāda veida uzdevumiem.

Vēl, varētu izpētīt lielo valodas modeļu izmantošanu datu kopu ģenerēšanai, ne tikai marķēšanai. Tas ļautu izveidot plašākas un daudzveidīgākas datu kopas, kas varētu uzlabot modeļa veiktspēju.

Vēl, varētu izpētīt klasifikatora modeļa izveidi, balstoties uz fundamentālā valodas modeļa pēdējās izvades iegultņu vērtībām. Tas varētu uzlabot modeļa veiktspēju un precizitāti klasifikācijas uzdevumos, kā arī, izstrādāt LVBERT balstītu sentimenta analīzes modeli, kas būtu pielāgots latviešu valodai un ļautu veikt efektīvu sentimenta analīzi. Ar šādu modeli būtu iespējams atlasīt paraugus, kurus nepieciešams pārbaudīt cilvēku marķētājiem.

Visbeidzot, varētu atkārtot eksperimentu ar vēl jaunākiem un spējīgākiem modeļiem, piemēram, GPT4, lai redzētu kādus uzlabojumus tie piedāvā latviešu valodas apstrādē un analīzē.

Bibliogrāfija

[1] Oskar Ahlgren. "Research on Sentiment Analysis: The First Decade". *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (2016), 890.—899. lpp.

[2] Shun-ichi Amari. "Learning Patterns and Pattern Sequences by Self-Organizing Nets of Threshold Elements". *IEEE Transactions on Computers* C-21 (1972), 1197.—1206. lpp.

[3] Francesco Barbieri, Luis Espinosa Anke un José Camacho-Collados. "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond". *International Conference on Language Resources and Evaluation*. 2021.

[4] Francesco Barbieri u. c. "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification". *ArXiv* abs/2010.12421 (2020).

[5] Guntis Barzdins u. c. "Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order". *Nordic Conference of Computational Linguistics*. 2007.

[6] Tom B. Brown u. c. "Language Models are Few-Shot Learners". *ArXiv* abs/2005.14165 (2020).

[7] DAIR.AI. 2023. g. maijs. URL: `https://www.promptingguide.ai/`.

[8] Jacob Devlin u. c. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". *ArXiv* abs/1810.04805 (2019).

[9] futureoflife. 2023. g. marts. URL: `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`.

[10] Ginta Garkaje, Evelina Zilgalve un Roberts Dargis. "Normalization and Automatized Sentiment Analysis of Contemporary Online Latvian Language". *Baltic HLT*. 2014.

[11] Kārlis Gediņš un Pēteris Paikens. "Automātiskā teksta emocionālās noskaņas noteikšana latviešu valodā". *LU Archive* (2013). URL: `https://dspace.lu.lv/dspace/handle/7/21072`.

[12] Ian Goodfellow, Yoshua Bengio un Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[13] Alex Graves u. c. "A Novel Connectionist System for Unconstrained Handwriting Recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (2009), 855.—868. lpp.

[14] Igors Gulbinskis un Darja Šmite. "Digitālo tekstu sentimenta analīze". *LU Archive* (2010). URL: `https://dspace.lu.lv/dspace/handle/7/18978`.

[15] Franz A. Heinsen. “An Algorithm for Routing Vectors in Sequences”. *ArXiv* abs/2211.11754 (2022).

[16] Sepp Hochreiter un Jürgen Schmidhuber. “Long Short-Term Memory”. *Neural Computation* 9 (1997), 1735.—1780. lpp.

[17] Harijs Ijabs un Ē. Urtāns. “Bidirectional Long Short-Term Memory Networks for Automatic Crop Classification at Regional Scale using Tabular Remote Sensing Time Series”. *Balt. J. Mod. Comput.* 10 (2022).

[18] Dan Jurafsky un James H. Martin. “Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition”. *Prentice Hall series in artificial intelligence*. 2000.

[19] Daniel Kahneman. “Thinking, Fast and Slow”. 2011.

[20] Geunwoo Kim, Pierre Baldi un Stephen McAleer. “Language Models can Solve Computer Tasks”. *ArXiv* abs/2303.17491 (2023).

[21] Soo-Min Kim un Eduard H. Hovy. “Determining the Sentiment of Opinions”. *International Conference on Computational Linguistics*. 2004.

[22] Takeshi Kojima u. c. “Large Language Models are Zero-Shot Reasoners”. *ArXiv* abs/2205.11916 (2022).

[23] Moshe Koppel un Jonathan Schler. “THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT”. *Computational Intelligence* 22 (2006).

[24] Andrew L. Maas u. c. “Learning Word Vectors for Sentiment Analysis”. *Annual Meeting of the Association for Computational Linguistics*. 2011.

[25] Warren S. McCulloch un Walter H. Pitts. “A Logical Calculus of the Ideas Immanent in Nervous Activity (1943)”. *Ideas That Created the Future* (2021).

[26] Walaa Medhat, Ahmed Hussein Hassan un Hoda Korashy. “Sentiment analysis algorithms and applications: A survey”. *Ain Shams Engineering Journal* 5 (2014), 1093.—1113. lpp.

[27] Jianmo Ni, Jiacheng Li un Julian McAuley. “Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects”. *Conference on Empirical Methods in Natural Language Processing*. 2019.

[28] Dāvis Nicmanis un Pēteris Paikens. “Sabiedrības attieksmes modelēšana, izmantojot sentimenta analīzi”. *LU Archive* (2017). URL: `https://dspace.lu.lv/dspace/handle/7/35260`.

[29] OpenAI. 2022. g. nov. URL: `https://openai.com/blog/chatgpt`.

[30] OpenAI. “GPT-4 Technical Report”. *ArXiv* abs/2303.08774 (2023).

[31] Long Ouyang u. c. "Training language models to follow instructions with human feedback". *ArXiv* abs/2203.02155 (2022).

[32] Peteris Paikens. "LEXICON-BASED MORPHOLOGICAL ANALYSIS OF LATVIAN LANGUAGE". 2007.

[33] Bo Pang, Lillian Lee un Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques". *Conference on Empirical Methods in Natural Language Processing*. 2002.

[34] Sohana Parvin, Murugan Sumathi un Caroline Mohan. "Challenges of Sentiment Analysis - A Survey". *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)* (2021), 781.—786. lpp.

[35] Janis Peisenieks un Raivis Skadins. "Uses of Machine Translation in the Sentiment Analysis of Tweets". *Baltic HLT*. 2014.

[36] Marcis Pinnis. "Latvian Tweet Corpus and Investigation of Sentiment Analysis for Latvian". *Baltic HLT*. 2018.

[37] Alec Radford un Karthik Narasimhan. "Improving Language Understanding by Generative Pre-Training". 2018.

[38] Alec Radford u. c. "Language Models are Unsupervised Multitask Learners". 2019.

[39] Colin Raffel u. c. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *ArXiv* abs/1910.10683 (2019).

[40] Pranav Rajpurkar, Robin Jia un Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD". *Annual Meeting of the Association for Computational Linguistics*. 2018.

[41] Anna Rogers, Olga Kovaleva un Anna Rumshisky. "A Primer in BERTology: What We Know About How BERT Works". *Transactions of the Association for Computational Linguistics* 8 (2020), 842.—866. lpp.

[42] Margaret Rouse. 2020. g. apr. URL: `https://www.techopedia.com/definition/4922/web-20`.

[43] Stuart Russell un Peter Norvig. *Artificial Intelligence: A Modern Approach, 4th Global ed.* `http://aima.cs.berkeley.edu/global-index.html`. 2021.

[44] Hasim Sak, Andrew W. Senior un Françoise Beaufays. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". *Interspeech*. 2014.

[45] Teven Le Scao u. c. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model". *ArXiv* abs/2211.05100 (2022).

[46] Kim Schouten un Flavius Frasincar. “Survey on Aspect-Level Sentiment Analysis”. *IEEE Transactions on Knowledge and Data Engineering* 28 (2016), 813.—830. lpp.

[47] Or Sharir, Barak Peleg un Yoav Shoham. “The Cost of Training NLP Models: A Concise Overview”. *ArXiv* abs/2004.08900 (2020).

[48] Richard Socher u. c. “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”. *Conference on Empirical Methods in Natural Language Processing*. 2013.

[49] Gatis Spats. “Application of Opinion Mining for written content classification in Latvian text”. *RTU Noslēgumu darbu reģistrs* (2015).

[50] Gatis Spats un Ilze Birzniece. “Opinion Mining in Latvian Text Using Semantic Polarity Analysis and Machine Learning Approach”. *Complex Syst. Informatics Model. Q.* 7 (2016), 51.—59. lpp.

[51] Uga Sprocgis un Matīss Rikters. “What Can We Learn From Almost a Decade of Food Tweets”. *Baltic HLT*. 2020.

[52] Keith E. Stanovich un Richard F. West. “Individual differences in reasoning: Implications for the rationality debate?”: *Behavioral and Brain Sciences* 23 (2000), 645.—665. lpp.

[53] Maite Taboada. “Sentiment Analysis: An Overview from Linguistics”. 2016.

[54] Hugo Touvron u. c. “LLaMA: Open and Efficient Foundation Language Models”. *ArXiv* abs/2302.13971 (2023).

[55] Alan M. Turing. “Computing Machinery and Intelligence”. *Mind* LIX (1950), 433.—460. lpp.

[56] Ashish Vaswani u. c. “Attention is All you Need”. *ArXiv* abs/1706.03762 (2017).

[57] Rinalds Vīksna. “Emocionālās ekspresijas noteikšana sīkziņās latviešu valodā”. *RTU Noslēgumu darbu reģistrs* (2018). URL: `https://github.com/RinaldsViksna/sikzinu_analize`.

[58] Alex Wang u. c. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. *BlackboxNLP@EMNLP*. 2018.

[59] Xuezhi Wang u. c. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. *ArXiv* abs/2203.11171 (2022).

[60] Zengzhi Wang u. c. “Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study”. *ArXiv* abs/2304.04339 (2023).

[61] Jason Wei u. c. “Chain of Thought Prompting Elicits Reasoning in Large Language Models”. *ArXiv* abs/2201.11903 (2022).

[62] Jason Wei u. c. “Emergent Abilities of Large Language Models”. *ArXiv* abs/2206.07682 (2022).

[63] Jules White u. c. “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT”. *ArXiv* abs/2302.11382 (2023).

[64] Qizhe Xie u. c. “Unsupervised Data Augmentation for Consistency Training”. *arXiv: Learning* (2019).

[65] Zhilin Yang u. c. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. *Neural Information Processing Systems*. 2019.

[66] Yongchao Zhou u. c. “Large Language Models Are Human-Level Prompt Engineers”. *ArXiv* abs/2211.01910 (2022).

[67] Arturs Znotins un Guntis Barzdins. “LVBERT: Transformer-Based Model for Latvian Language Understanding”. *Baltic HLT*. 2020.