

VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ BAKALAURO STUDIJŲ PROGRAMA

Teksto suvokimas pasitelkiant neuroninius tinklus

Language Understanding Using Neural Networks

Baklauro baigiamasis darbas

Atliko:	4 kurso, 6 grupės studentas	
	Mindaugas Barčauskas	(parašas)
Darbo vadovas:	asist. Oleg Mirzianov	(parašas)
Recenzentas:	asist. Linas Petkevičius	(parašas)

Vilnius
2020

TURINYS

SANTRAUKA	4
SUMMARY	5
IVADAS	6
1. UŽDAVINIŲ PASIRINKIMAS	8
1.1. Sentimento analizė	9
1.2. Skyrybos klaidų aptikimas.....	9
2. DUOMENŲ RINKINIAI	10
2.1. Kriterijai duomenų rinkinių sudarymui	10
2.2. Duomenų rinkinių surinkimui naudojami įrankiai	10
2.3. Duomenų rinkinių sąrašas	10
2.3.1. Prekių aprašymų rinkinys	10
2.3.2. Lietuvių kalbos skyrybos klaidų rinkinys	11
3. SPRENDIMO BŪDŲ PASIRINKIMAS	12
3.1. Santrauka	12
3.2. Natūralios kalbos apdorojimo algoritmų raidos apžvalga	12
3.3. Dirbtinių neuroninių tinklų architektūrų įvertinimas	13
3.3.1. Rekurentiniai dirbtiniai neuroniniai tinklai	13
3.3.2. Rekurentinių neuroninių tinklų pranašumai	14
3.3.3. Rekurentinių dirbtinių neuroninių tinklų trūkumai	14
3.3.4. transformeriais grįstų dirbtinių neuroninių tinklų architektūros	15
3.3.5. Kiti transformeriais grįsti dirbtiniai neuroniniai tinklai	15
3.3.6. Kaip veikia BERT	15
3.3.7. BERT pritaikymas specifinėms užduotims	16
3.3.8. BERT pranašumai	16
3.3.9. BERT trūkumai	17
3.4. Sprendimo būdų palyginimas	17
4. SENTIMENTO ANALIZĖS LIETUVIŲ KALBA TYRIMAS	19
4.1. Sentimento analizė pasitelkiant rekurentinius dirbtinius neuroninius tinklus	19
4.2. Sentimento analizė pasitelkiant transformerių tipo dirbtinius neuroninius tinklus	21
4.3. Dirbtinių neuroninių tinklų sprendimo būdų palyginimas su esamais sprendimo būdais sentimento analizei nustatyti	21
4.4. Rezultatų apibendrinimas	21
5. SKYRYBOS KLAIDŲ APTIKIMO TYRIMAS	22
5.1. Tyrimo apibendrinimas	22
5.2. Duomenys	22
5.3. Duomenų paruošimas neuroniniams tinklams	22
5.4. Tyrimo būdai	23
5.5. Skyrybos klaidų aptikimo tyrimas pasitelkiant suprogramuotas taisykles	23
5.6. Tyrimas pasitelkiant rekurentinius dirbtinius neuroninius tinklus	23
5.7. Tyrimas pasitelkiant transformeriu grįstus dirbtinius neuroninius tinklus	24
5.8. Skyrybos tyrimas pasitelkiant mišrų metodą	25
5.9. Tyrimo metu patirti sunkumai	25
5.10. Tyrimo rezultatai	25
5.11. Skyrybos klaidų aptikimo palyginimas su esamais įrankiais.....	25
5.12. Plėtros galimybės	26

REZULTATAI	27
IŠVADOS	28
LITERATŪRA	29
SAVOKŲ ŽODYNAS	31
PRIEDAI	31
1 priedas. Kalbos dalių žodyno pavyzdys	32
2 priedas. Prekių aprašymų duomenų rinkinio pavyzdys	33
3 priedas. Skrybos klaidų aptikimo duomenų rinkinio pavyzdys.....	34
4 priedas. Duomenų rinkiniai ir kodo pavyzdžiai	35

Santrauka

Per kelis pastaruosius metus, padaryta daug atradimų kalbos analizės srityje. Šios naujos remiasi įvairiomis dirbtinių neuroninių tinklų architektūromis. Tačiau šie atradimai dar nėra pritaikyti lietuvių kalboje. Taip yra dėl mažo kiekio standartizuotų duomenų rinkinių lietuvių kalboje. Šiame darbe bus tiriami patys efektyviausi būdai spręsti sentimentų analizės bei skyrybos taisymo užduotis lietuvių kalboje. Sukurti dirbtiniai neuroniniai tinklai gali padėti verslui lengviau nustatyti žmonių nuomones apie prekes, o skyrybos taisymas padėtų jaunimui geriau išmokti lietuvių kalbos skyrybos taisyklių. Taip pat padėtų puoselėti lietuvių kalbą, kuri vis juda į antrą planą pasauliui globalizuojantis. Taip pat tai padėtų lietuvių kalbos mokytojams lengviau ištaisyti mokinių rašinius. Šiame darbe taip pat buvo surinkti struktūrizuoti duomenų rinkiniai dirbtinių neuroninių tinklų apmokymui. Sukurtas prekių aprašymų duomenų rinkinys lietuvių kalba. Taip pat sukurtas skyrybos klaidų duomenų rinkinys lietuvių kalba. Išanalizuotos rekurentinių dirbtinių neuroninių tinklų bei transformeriais grįstų neuroninių tinklų savybės, pranašumai bei silpnosios pusės. Atlikti tyrimai bei aprašyti rezultatai sprendžiant sentimentų analizės bei skyrybos klaidų taisymo užduotis.

Summary

This work analyzes most popular and effective neural network architecture efficiency solving sentiment analysis and punctuation error detection tasks. The biggest problems preventing the investigation of such tasks were the lack of structured datasets which were created as a part of this paper. Criteria were first established and the natural language processing tasks were identified before collecting the datasets. Two datasets were collected as a part of this research paper. The first dataset was online product review dataset for sentiment analysis task. Second dataset was collected from wikipedia lithuanian article corpus. It was later used to create a dataset for punctuation error detection. A brief overview was given on the latest achievements in the natural language space in transformer based neural networks like BERT. Comparison between the older Recurrent artificial neural network architecture and the newer transformer based neural networks is given. In the second part of the paper the analysis of sentiment analysis task is given, best neural network configuration and training methods are provided as well as comparison between recurrent neural network and transformer based neural network performance. The second analysis is given of punctuation error detection in lithuanian. The most accurate automatic punctuation error correction tool in lithuanian is created based on a mixed explicit programmed rule and transformer based neural network approach according to authors knowledge. The results of the work help advance natural language processing sphere in lithuanian language and commercial tools could be created by fine tuning the approaches presented in this research paper. The work could also help to solve natural language processing problems in other less popular languages like latvian, estonian, etc.

Įvadas

Problema

Natūralios kalbos analizė yra labai patobulėjusi per keletą paskutinių metų. Tobulėjančios dirbtinių neuroninių tinklų architektūros ženkliai pagerino rezultatus natūralios kalbos analizės uždaviniuose [Zhe18]. Šios architektūros dar nėra pritaikytos spręsti problemas lietuvių kalbos uždaviniams. Be to, iki šiol šiuos uždavinius buvo sudėtinga spręsti dėl struktūrizuotų duomenų rinkinių trūkumo. Taip pat pritaikant naujausius kalbos analizės uždavinių sprendimo būdus didėja galimybės automatizuoti įvairius su kalbos analize susijusius darbus, tokius kaip klaidų taisymas, vertimas, sentimentų analizė. Naujausių sprendimų pritaikymas taip pat leidžia pastebėti sprendimo būdų tikslumo skirtumus tarp skirtingų kalbų.

Tikslas

Įvertinti populiariausius bei tiksliausius kalbos analizės algoritmus pritaikant juos lietuvių kalboje. Sudaryti struktūrizuotus natūralios kalbos uždavinių duomenų rinkinius lietuvių kalba.

Uždaviniai

- Išanalizuoti esamą literatūrą bei išrinkti, kuriuos natūralios kalbos apdorojimo sprendimo būdus galima pritaikyti lietuvių kalboje.
- Pasirinkti norimus spręsti kalbos analizės uždavinius.
- Sukurti lietuviškus duomenų rinkinius dirbtiniams neuroniniams tinklams apmokyti pasirinktiems uždaviniams spręsti bei jų efektyvumui įvertinti.
- Atlikti tyrimą ir nustatyti efektyviausius būdus spręsti pasirinktus uždavinius lietuvių kalboje.

Darbo eiga

Norimų spręsti uždavinių pasirinkimas

Iš pradžių pasirenkami norimi spręsti keli iš populiariausių kalbos analizės uždavinių. Išanalizuojama literatūra populiariausiems kalbos analizės uždaviniams surasti. Pasirenkami kalbos uždaviniai pagal kuriuos vertinsime sprendimo būdus.

Duomenų rinkinių pagal pasirinktus spręsti uždavinius sudarymas

Norinti įvertinti uždavinių sprendimo būdus reikės sudaryti struktūrizuotus duomenų rinkinius. Sudaromi lietuvių kalbos duomenų rinkiniai tinkami spręsti pasirinktiems uždaviniams uždaviniams. Struktūrizuoti duomenų rinkiniai turėtų būti tinkami dirbtinių neuroninių tinklų apmokymui.

Sprendimo būdų pasirinkimas

Išanalizuojama literatūra ir pasirenkamos tinkamiausios šiems uždaviniams spręsti dirbtinių neuroninių tinklų architektūros. Pasirenkama pagal tai kokie sprendimo būdai šiuo metu yra lyderiai tam uždavinių tipui anglų kalboje.

Tyrimai

Kiekvienam duomenų rinkiniui atliekamas tyrimas keičiant dirbtinių neuroninių tinklų parametrus ir ieškoma optimalių rezultatų. Palyginami gauti rezultatai. Tyrimai atliekami kiekvienam duomenų rinkiniui atskirai. Keičiami tokie parametrai kaip duomenų kiekis, neuroninio tinklo dydis.

1. Uždavinių pasirinkimas

Kalbos analizės srityje yra populiarūs sentimentų analizės uždaviniai, teksto klasifikavimo uždaviniai, morfologinės teksto analizės uždaviniai, teksto kūrimo uždaviniai, klausimų atsakymo uždaviniai, teksto apibendrinimo bei vertimo uždaviniai [Kha19]. Kiekvienas iš uždavinių turi savo ypatumus bei skirtingus sprendimo būdus.

Sentimento analizės uždaviniai yra skirti žmonių nuomonei apie prekes, istorinius įvykius, politines partijas nustatyti. Jiem apmokyti dažnai naudojami forumų komentarai ar atsiliepimai apie prekes. Sentimento analizės uždaviniai sprendžiami kaip klasifikavimo uždaviniai ir dažnai rezultatas būna nustatyti ar aprašymas yra teigiamas ar neigiamas (binarinis klasifikavimo uždavinys) [Ram16].

Teksto klasifikavimo uždaviniai skirti suklasifikuoti duotus teksto segmentus į tam tikras kategorijas [KS15]. Pritaikymas gali būti suklasifikuoti forumo komentarus į įžeidžiančius, užgauliojančius, grasinančius bei gerus komentarus norint automatiškai uždrausti užgauliojimą internetinėje erdvėje.

Morfologinės analizės uždaviniuose sprendžiami tokie uždaviniai kaip kalbos dalių nustatymas, žodžio šakninės formos gavimas (lemavimas) [Loi18].

Teksto kūrimo uždaviniai dažnai būna uždaviniai skirti sugeneruoti tekstą [San18]. Tai gali būti tekstas skirtas atrašyti į žinutes (angl. chat bot). Taip pat egzistuoja įmonių, kurios generuoja finansinius straipsnius ar sporto naujienų straipsnius automatiškai.

Klausimų atsakymo uždaviniai būna tokie, jog dirbtiniam neuroniniam tinklui pateikiamas paragrafas teksto bei atskirai pateikiamas klausimas. Atsakymas į klausimą yra pateiktame paragrafe ir dirbtinis neuroninis tinklas turi nustatyti, kuri duoto paragrafo dalis yra atsakymas į klausimą. Paskutiniaisiais metais naujausios transformeriais grįstos dirbtinių neuroninių tinklų architektūros prilygsta žmonių atsakinėjimo lygiui [Pra16].

Teksto apibendrinimo uždaviniai yra uždaviniai, kurias siekiama apibendrinti ilgus tekstus kaip: legalius dokumentus, bylas, straipsnius [Ank19].

Teksto vertimo uždaviniai. Šie uždaviniai atliekami arba bandant išversti sakinius ar tekstus iš vienos kalbos į kitą, arba gali būti verčiama į kelias kalbas iš karto. Šių uždavinių sprendimo tikslumo pagerinimui taip pat prisidėjo transformeriais paremtos architektūros [Qia19].

Išnagrinėjus literatūrą pastebėta, jog lietuvių kalba jau yra sukurtas morfologinės teksto analizės įrankis lemuoklis [ZIN00]. Šis įrankis automatiškai nustato kokia yra žodžio dalis, linksnis, giminė. Taip pat yra ir daug morfologijos žodynų [And09] [Eri11]. Todėl šis uždavinių tipas atmes-tas. Vertimas iš bei į lietuvių kalbą yra patobulintas Google, nes pradėtas naudoti BERT dirbtinis neuroninis tinklas [Jac19]. Šis dirbtinis neuroninis tinklas, sukurtas Google kompanijoje dirbančių mokslininkų, leido pagerinti automatinio teksto vertimo tikslumą. Klausimų bei atsakymų uždaviniui reikalingi dideli duomenų rinkiniai, kuriuos reiktų surinkti rankomis, todėl atmetamas ir šis uždavinių tipas. Pasirinkta nagrinėti sentimentų analizės uždavinius bei teksto klasifikavimo uždavinius kadangi jiems galima lengviau surinkti duomenų rinkinius. Pavyzdžiui, galima surinkti atsiliepimų apie prekes rinkinį ir jis yra automatiškai anotuotas pagal žvaigždutinį įvertinimą. Teksto klasifikavimui taip pat galima surinkti duomenų rinkinį lengviau, kadangi tekstą anotuoti reikia

tik pagal galimas teksto klases. Taigi, pasirinkus uždavinius teksto surinkti duomenų rinkinius.

1.1. Sentimento analizė

Sentimento analizės uždaviniui naudosime viešai prieinamus atsiliepimus apie prekes iš elektroninių parduotuvių. Sentimento analizės uždaviniai dažnai būna dviejų klasių klasifikavimo uždaviniai (pvz. teigiamas aprašymas ar neigiamas). Taip pat, sentimento analizės uždaviniai gali būti klasifikuojami į daugiau skirtingų klasių uždavinius (pvz. Nuspėti aprašymo įvertinimą iš įrašyto teksto). Sentimento analizės uždaviniai yra skirti nustatyti asmens sentimentą ar nuomonę apie prekę, politinę partiją ar paslaugą.

1.2. Skyrybos klaidų aptikimas

Teksto klasifikavimo uždaviniui pasirinkta kurti programą, kuri galėtų aptikti skyrybos klaidas lietuvių kalboje. Ši implementacija, pasiekus aukštą tikslumo lygį padėtų žmonėms išlaikyti taisyklingą lietuvių kalbos rašybą bei padėtų mokiniams bei studentams geriau išmokti taisyklingos rašybos ar patikrinti savo rašinius. Tai taip pat sutaupytų mokytojų laiką taisant rašinius. Šis uždavinys priskiriamas prie teksto klasifikavimo, nes iš esmės dirbtiniai neuroniniai tinklai tiesiog turi suklasifikuoti tarp kurių teksto dalių yra reikalingi skyrybos ženklai bei tarp kurių ne.

2. Duomenų rinkiniai

2.1. Kriterijai duomenų rinkinių sudarymui

Duomenų rinkiniai buvo surenkami atsižvelgiant į tai kokius uždavinius norima spręsti. Kadangi buvo pasirinkta spręsti sentimentų analizės ir skyrybos klaidų aptikimo uždaviniai, pagal tai ir buvo renkama kokių duomenų rinkinius sukurti. Vieną rinkinį nuspręsta surinkti, kuriame būtų išreikšta žmogaus nuomonė sentimentų analizės uždaviniams spręsti. Kitas rinkinys, kuris buvo renkamas buvo iš įvairių vikipedijos straipsnių gautas tekstas suskaidomas sakiniais ir surenkama, kad 50 procentų sakinių būtų be kablelių, o kiti 50 procentų sakinių būtų su kableliais.

2.2. Duomenų rinkinių surinkimui naudojami įrankiai

Duomenų surinkimui buvo naudojama python programavimo kalba bei selenium karkasas. Statiniams puslapiams buvo naudojama BeautifulSoup4 python biblioteka. Dinaminiam interneto puslapiams (parašytiems su javascript) buvo naudojamas Selenium karkasas. Parašytos programos, kurios surenka duomenis iš internetinių puslapių. Tuomet duomenys sutvarkomi į duomenų rinkinius, kurie tinka dirbtinių neuroninių tinklų apmokymui. Standartizuojama įvestis bei išvestis.

Taip pat darbe buvo renkama morfeminė informacija apie žodžius. Tai informacija koks žodis priklauso atitinkamai kalbos daliai, kokio linksnio yra žodis ir panašiai. Morfologinei informacijai surinkti patogiu json formatu buvo naudojama PDFMiner python biblioteka. Ištraukti reikalingą informaciją iš teksto pdf formatu buvo ganėtinai nelengva, tačiau pavyko išgauti visą naudingą informaciją iš žodynų. Iš žodynų gauta informacija: žodžio dalis, giminė, skiemenys, linksnis, laipsnis (jei būdvardis).

Taip pat transformerinių dirbtinių neuroninių tinklų apmokymui naudota Google Cloud platforma. Tai debesų kompiuterijos platforma, kuri suteikia galimybę nusipirkti įvairius galingus procesorius specifiskai pritaikytus dirbtinių neuroninių tinklų mokymui.

2.3. Duomenų rinkinių sąrašas

2.3.1. Prekių aprašymų rinkinys

Buvo pasirinkta sudaryti prekių aprašymų rinkinį iš pradžių automatiškai ropojančių per eko-mercijos tinklapius bei surenkant prekės įvertinimo aprašymą bei įvertinimą. Duomenų rinkinys puikiai atitinka realius prekių aprašymus, kurie kartais turi gramatinių bei rašybos klaidų.

```
{ "rating": "5", "description": "Tik peilis labai jau minkstas. Nuo pirmo kurmerausio sulinko. :-)  
Ireko nusipirkti kietesny.  
O pats aparatas tai super." },  
{ "rating": "4", "description": "Gera vejapjove nera sunki, tai lengva valdyt ir as galiu" }
```

1 pav. Prekių aprašymų duomenų rinkinio pavyzdiniai įrašai.

Duomenų rinkinyje yra 4000 atsiliepimų apie prekes.

2.3.2. Lietuvių kalbos skyrybos klaidų rinkinys

Buvo surinktas lietuvių kalbos skyrybos duomenų rinkinys, kuriame surinkta 50 procentų sakinių su kableliais, o kiti 50 procentų sakiniai be kablelių. Duomenų rinkinys suskirstytas taip, jog dirbtiniai neuroniniai tinklai galėtų būti apmokomi atskirti sakinius su taisyklinga skyryba. Šiam rinkiniui buvo parsisiųsti visi lietuviški wikipedia straipsniai. Wikipedija suteikia galimybę parsisiųsti visus straipsnius viena kalba. Buvo pasirinkta parsisiųsti lietuvių kalbos wikipedijos straipsnius ir suskaidyti visą duomenų rinkinį į atskirus sakinius. Wikipedia yra puikus duomenų šaltinis, dėl to, jog straipsniai yra labai įvairių temų ir dėl to ten yra vartojamas įvairesnis žodynas, apmokomi dirbtiniai neuroniniai tinklai susiduria su įvairesniais sakiniai ir tai leidžia dirbtiniams neuroniniams tinklams pasiekti geresnį rezultatą. Duomenų rinkinys apdorotas taip, jog kiekvienas sakinytis būtų atskiroje eilutėje. Tai padaryta todėl, kad wikipedijos žodynas yra ganėtinai platus. Rašoma apie daugybę įvairių temų. Tai leidžia geriau ištestuoti dirbtinių neuroninių tinklų sugebėjimus įvairesniam tekstui. Duomenų rinkinyje yra 200 000 sakinių be kablelių bei 200 000 sakinių turinčių kablelius.

3. Sprendimo būdų pasirinkimas

3.1. Santrauka

Išanalizavus literatūrą buvo pasirinkti du sprendimo būdai rekurentiniai dirbtiniai neuroniniai tinklai, transformeriais grįsti dirbtiniai neuroniniai tinklai. Transformeriais grįsti neuroniniai tinklai buvo pasirinkti dėl to, kad literatūros duomenimis šiuo metu jie duoda geriausius rezultatus [Ash17]. Rekurentiniai neuroniniai tinklai buvo pasirinkti dėl to, jog jie taip pat duoda plačiai aprašomi literatūroje bei duoda puikius rezultatus kalbos analizės uždaviniuose [Pro01].

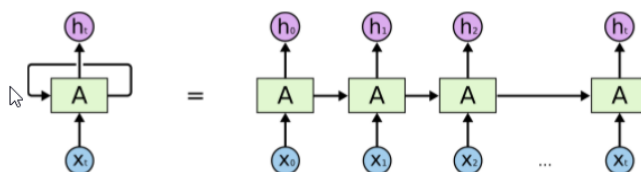
3.2. Natūralios kalbos apdorojimo algoritmų raidos apžvalga

Maždaug nuo 1980-ųjų metų pradėjo įsigalėti statistiniai modeliai, vietoje griežtomis taisyklėmis bei sprendimų medžiais paremtų modelių. Buvo pereita prie mašininio mokymosi bei statistinių algoritmų. Šiais algoritmais ir yra paremti dabartiniai natūralios kalbos apdorojimo modeliai [Hri11]. 2006 metais atliktas tiriamasis darbas apie “word2vec” sistemą, kuri transformuoja žodžius į vektorius ir poto atlieka operacijas su jais [Dea06]. Šis straipsnis sukelia didelį susidomėjimą bei išpopuliarina natūralios kalbos apdorojimo tyrimą pasitelkiant panašius modelius. Nuo 2006 iki 2016 metų yra atliekami pagrinde tik inkrementalūs pagerinimai panašūs į “word2vec” sistemą pasitelkiant rekurentinius dirbtinius neuroninius tinklus. Išpopuliarėja LSTM (Long Short Term Memory) Rekurentiniai dirbtiniai neuroniniai tinklai [Pro01]. 2017 metais buvo parašytas straipsnis “Attention is all you need”, kuriame buvo pristatyta transformeriais paremta natūralios kalbos apdorojimo sistema BERT. Ši architektūra padarė šuolį atliekant įvairias kalbos užduotis (žodžių nuspėjimas, klausimų atsakinėjimas, teksto klasifikavimas) [Ash17]. Nuo 2017 iki 2019 metų “Attention is all you need” straipsnis sukėlė didžiulį susidomėjimą tarp mokslininkų natūralios kalbos apdorojimo srityje. Sukuriami patobulinti modeliai (XLM-Roberta) [Zhe18]. Transformerių architektūrą vis labiau linkstama pripažinti kaip efektyvesnę spręsti natūralios kalbos apdorojimo problemas nei rekurentinius dirbtinius neuroninius tinklus. Automatizuotos sistemos pradeda priartėti arba pranoksta žmonių sugebėjimus tam tikrose nišinėse srityse [Phu18]. BERT - (angl. Bidirectional Encoder Representations from Transformers) yra transformeriais paremta natūralios kalbos apdorojimo sistema, kuri padeda lengviau paruošti dirbtinius neuroninius tinklus treniravimui. Ši sistema pakeitė mokslininkų suvokimą apie tai kas yra iš tikrųjų įmanoma natūralios kalbos apdorojimo srityje. Taigi, matome, jog pastaraisiais metais susidomėjimo ir atradimų natūralios kalbos analizės srityje buvo daug. Dėl to šiame darbe pasirinkta naudoti būtent anksčiau geriausius rezultatus teikusios architektūros (RNN) ir dabartinės geriausius rezultatus teikiančios architektūros (BERT).

3.3. Dirbtinių neuroninių tinklų architektūrų įvertinimas

3.3.1. Rekurentiniai dirbtiniai neuroniniai tinklai

Rekurentiniai dirbtiniai neuroniniai tinklai - tai tokie dirbtiniai neuroniniai tinklai, kurie sugeba “įsiminti” prieš tai buvusias detales. Kaip apačioje esančiame paveiksle (paimta iš “Understanding LSTM Neural networks”). Taigi rekurentiniai dirbtiniai neuroniniai tinklai sugeba perduoti informaciją iš vieno žingsnio į kitą. Tai įvyksta vartų esančių kiekviename dirbtinio neuroninio tinklo neurone dėka. Pirmieji vartai nusprendžia ar kokia informacija priima tai yra įvestis vartai. Antrieji vartai esantys neurone išfiltruoja, kuri informacija bus perduodama, šie vartai vadinami užmaršties vartais. Tretieji vartai tai yra išvesties vartai, kurie dar kartą transformuoja išėigą. Natūralios kalbos apdorojime tai reikštų, jog dirbtinis neuroninis tinklas sugeba, bent teoriškai, prisiminti faktus. Realybėje šis kontekstas yra ganėtinai trumpas (neuroninis tinklas nesugeba prisiminti faktų pavyzdžiui, iš praito paragrafo). Neuroniniai tinklai yra naudojami būtent tokiose dalyse. Rekurentinių dirbtinių neuroninių tinklų panaudojimas išsiplėtė dėl specialios šių dirbtinių neuroninių tinklų architektūros LSTM (angl. Long short term memory). LSTM leido iš dalies išspręsti tolimo konteksto problemą bei pagerinti rekurentinių dirbtinių neuroninių tinklų teikiamus rezultatus. LSTM neuronai leidžia rekurentiniam dirbtiniam neuroniniam tinklui atsiminti daugiau nei prieš jį ėjusio neurono būseną. Pavyzdžiui, tai gali leisti nuspėti, jog tas pats garsas pasikartoja kas 5 sekundes dainoje, nes dirbtinis neuroninis tinklas prisimena prieš tai ėjusius ciklus.



2 pav. RNN architektūra [3].

Kur panaudojami RNN

Rekurentiniai dirbtiniai neuroniniai tinklai yra panaudojami įvairiose srityse, kuriose analizuojami eiliškumą turintys duomenys. Tos sritys yra:

- teksto analizė
- video analizė / garsų analizė
- finansinių rinkų analizė

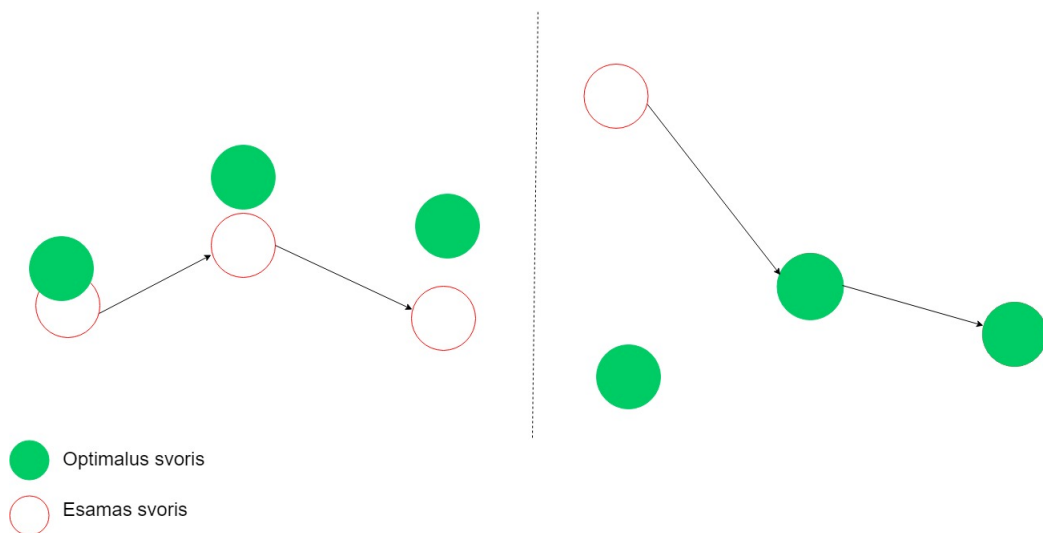
Šiose srityse duomenys turi tam tikrą eiliškumą, tekstas priklauso nuo prieš tai ėjusių žodžių, melodija taip pat priklauso nuo prieš tai ėjusių natų, akcijos kainos priklauso nuo istorinių akcijų kainų.

3.3.2. Rekurentinių neuroninių tinklų pranašumai

Rekurentiniai dirbtiniai neuroniniai tinklai yra pranašesni už kitus būdus, tokius kaip žodžių maišo modelis, ar kaip žodžių vektorizavimo modelis "word2vec". Kadangi jie gauna informaciją iš prieš tai apdorotų duomenų, taigi jų architektūra iš karto yra pritaikyta apdoroti eilišką informaciją. Iš esmės tekstas yra eiliška informacija (sekantys sakiniai konstruojami iš jau turimos informacijos). Dar vienas rekurentinių neuroninių tinklų pranašumas yra tai, jog jie sugeba prisiminti anksčiau pateiktus faktus, pavyzdžiui prieš kelis sakinius ir teoriškai prieš kelis paragrafus parašytus teiginius.

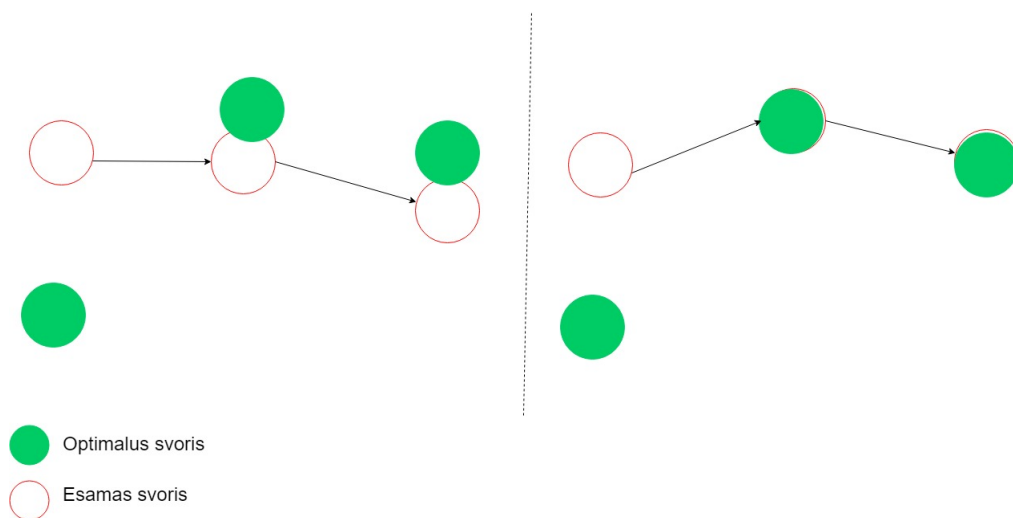
3.3.3. Rekurentinių dirbtinių neuroninių tinklų trūkumai

Rekurentiniai neuroniniai tinklų problemos bei sunkumai juos apmokant: nykstantys nuolydžiai bei sprogstantys nuolydžiai. Kai kalbama apie nuolydžius turima omenyje pradžios funkcijos (angl. loss function) nuolydžiai [Kar18]. Sprogstantys nuolydžiai įvyksta tada kai rekurentinio neuroninio tinklo apmokymo metu gradientas išauga nepaprastai didelis ir dėl to pirminiai sluoksniai pradeda per daug kisti taip mažėjant neuroninio tinklo tikslumui.



3 pav. Sprogstantis nuolydis

Nykstantys nuolydžiai įvyksta atvirkščiai, kai gradientas sumažėja tiek, kad rekurentiniam dirbtiniam neuroniniam tinklui nebepavyksta išmokyti asociacijų tarp tolimų duomenų. Nykstantys nuolydis reiškia, jog pirminiai rekurentinio neuroninio tinklo sluoksniai keisis labai nežymiai. Tai reiškia, jog rekurentinis tinklas negalės efektyviai mokytis. Dėl to bus gauti prastesni rezultatai.



4 pav. Nykstantis nuolydis

3.3.4. transformeriais grįstų dirbtinių neuroninių tinklų architektūros

BERT taip pat yra transformeriais paremta natūralios kalbos apdorojimo sistema, kuriai reikia daug resursų norint pasiekti aukštus rezultatus įvairiuose užduočių rinkiniuose. Daug resursų reikia nes pats dirbtinis neuroninis tinklas susideda iš daug sluoksnių bei pačiuose sluoksniuose yra daug neuronų, kadangi sąryšiai tarp žodžių būna ganėtinai sudėtingi ir pats bet kurios kalbos žodynas yra labai didelis. Pats transformeris susideda iš užkodavimo bei dekodavimo dalių, kurios apmokomos norint gauti reikiamą rezultatą. Patys užkodavimo ir dekodavimo sluoksniai susideda iš dėmesio filtrų bei dirbtinių neuroninių tinklų. Duomenų rinkiniai gali būti tokie kaip (SQuAD 1.1, SQuAD 2.0 ir RACE). Šie užduočių rinkiniai yra standartai natūralios kalbos apdorojimo srityje. SQuAD bei RACE duomenų rinkiniai yra klausimais bei atsakymais iš duoto teksto paragrafo paremti duomenų rinkiniai. BERT išpradžių būna apmokomas pilnai ant didelio teksto kiekio. Gaunamas bendrinis teksto analizės dirbtinis neuroninis tinklas. Gautą dirbtinį neuroninį tinklą galima pritaikyti specifinėms užduotims.

3.3.5. Kiti transformeriais grįsti dirbtiniai neuroniniai tinklai

BERT yra transformerių šeimos dirbtinis neuroninis tinklas. Paskelbus BERT atsirado daug mokslininkų, kurie apmokė panašius dirbtinius neuroninius tinklus. Keičiant mokymosi parametrus, dirbtinio neuroninio tinklo dydį, mokymosi žingsnius bei keičiant duomenis su kuriais šie dirbtiniai neuroniniai tinklai yra apmokomi. Pasiekiami vis geresni rezultatai. Tokių dirbtinių neuroninių tinklų pavyzdys: XLM-Roberta. Tyrimai bus atliekami ir su šiais dirbtiniais neuroniniais tinklais atsižvelgiant į tai kokioms užduotims šie dirbtiniai neuroniniai tinklai duoda pačius geriausius rezultatus.

3.3.6. Kaip veikia BERT

Dirbtinių neuroninių tinklų modeliai prieš BERT bandydavo suvokti žodžius vien pasitelkdamie žodžius esančius prieš užmaskuotą žodį, neįvertinant pilno konteksto. pvz:

„Žmogus ėjo namo ir sutiko juodą katę.“

Šiuo atveju būtų bandoma atspėti žodį juodą pasitelkiant tik „Žmogus ėjo namo ir sutiko“. Tuo tarpu BERT modelis tiesiog maskuoja norimą atspėti žodį ir tada bando jį atspėti pasitelkdamas pilną kontekstą: „Žmogus ėjo namo ir sutiko ... katę“.

Tai leidžia išgauti daugiau informacijos iš teksto ir veikia labiau kaip žmogaus smegenys, nes žmonės teksto nesuvokia vien iš kairės į dešinę. Žmonės perskaito visą sakinį ar paragrafą ir iš to konstruoja atskirus faktus. Taigi, BERT yra algoritmas su kuriuo galima atlikti neprižiūrimąjį mokymąsi pasitelkiant didelius duomenų kiekius (tūkstančius knygų ar šimtus tūkstančių straipsnių). Užtenka tiesiog užmaskuoti dalį žodžių (dažniausiai 15%) visame tekste ir tada apmokant sistemą jai atspėjant žodžius. Taip pat, apmokant BERT yra naudojamas ir sekančio sakiniu nuspėjimo apmokymas, paimant sakinį ir bandant nuspėti koks sakinyš eis po to. Tai padeda BERT neuroniniam tinklui suvokti tokį dalyką kaip kontekstas, kad sakiniai turi tam tikrą loginę seką [Ash17]. Apmokant sekančio sakinio nuspėjimo uždaviniu dirbtinis neuroninis tinklas įgauną platesnį temos ir paragrafo suvokimą, nes dirbama ne tik sakinio lygyje, bet ir paragrafo lygmenyje tarp sakinių. Taip pat reiktų paminėti, jog transformerių tipo dirbtiniai neuroniniai tinklai naudoja dėmesio principą, kuris nustato į kuriuos žodžius ar žodžio dalis reiktų kreipti dėmesį

3.3.7. BERT pritaikymas specifinėms užduotims

Google tyrėjai paviešino BERT apmokytus neuroninių tinklų modelius, taigi bet kas gali juos panaudoti specifinėms užduotims. Norint panaudoti juos specifinėms užduotims, tereikia turėti treniravimo bei įvertinamuosius duomenis ir tiesiog apmokyti neuroninį tinklą tai užduočiai. Didysis BERT privalumas ir yra tai, jog jie yra bendrieji kalbos modeliai, kuriuos galima pritaikyti kitoms užduotims [Jac19]. Neuroninių tinklų modelių perpanaudojamumas sukėlė tikrą natūralios kalbos apdirbimo revoliuciją, kadangi bet kas gali pritaikyti modelį savo reikmėms. BERT buvo pritakytas atsakinėti į klausimus bei nustatyti sekantį žodį įvairiems etaloniniams duomenų rinkiniams ir leido pasiekti rekordinius rezultatus. Kaip pavyzdžius galime paimti Glue benchmark bei Squad benchmark (Populiarūs natūralios kalbos analizės tyrimams naudojami duomenų rinkiniai). BERT Glue benchmark pasiekė 86.7 procento tikslumą kas tuo metu buvo geriausi rezultatai. Squad klausimų atsakinėjimo užduotyje BERT pasiekė 80.411% tikslumą ir priartėjo prie žmonių lygio (86%) [Ash17]. Glue tai yra duomenų rinkinys susidedantis iš 9 kalbos analizės užduočių [Ale19]. Užduotys vyrauja nuo klausimų atsakinėjimo iki sakinių panašumo nustatymo uždavinių. Glue rezultatas leidžia nustatyti kaip neuroninis tinklas bendrai gali atlikti labai skirtingas užduotis.

3.3.8. BERT pranašumai

- Lengviau susiejama informacija, kuri išsidėsčiusi per kelis sakinius/paragrafus pasitelkiant naujai atrastą dėmesio metodą. Šis metodas priskiria koeficientus kiekvienam žodžiui ir pagal tai tiek dėmesio skiria tam žodžiui atlikdamas užduotį su duotais duomenimis. [Ash17] Tai teoriškai turėtų leisti neuroniniam tinklui pastebėti tokius dalykus kaip ironija bei sarkazmas.

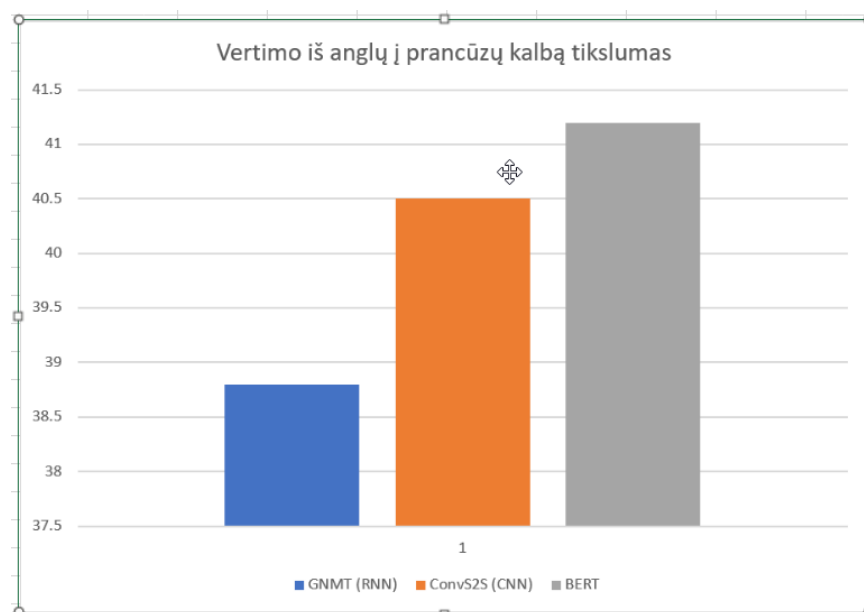
- Kadangi BERT neturi nykstančių bei sprogstančių nuolydžių problemos.[Kar18] Nykstančių bei sprogstančių gradientų problema yra kai dėl matematinių veiksmų atliekamų bandant optimizuoti kainos funkciją, pirminiai sluoksniai pradeda nekisti arba kinta pernelyg greitai. Tai yra dėl giles rekurentinių dirbtinių neuroninių tinklų struktūros.
- BERT yra bendrasis kalbos modelis, taigi tarkim apmokius BERT anglų kalbos modelį, galima lengvai jį pritaikyti skirtingoms užduotims, padarius mažai pakeitimų [Che19].

3.3.9. BERT trūkumai

Allyson Ettinger teigia, jog BERT tiesiog sugeba gerai išmokti tam tikrus pasikartojimus tekste ir iš jų sudaryti atsakymus ir neturi suvokimo [Ett19]. Tam tikroms užduotims, kaip sakinių generavimas, BERT nėra labai geresnis nei rekurentiniai neuroniniai tinklai.

3.4. Sprendimo būdų palyginimas

Viena iš sričių, kuri labai patobulėjo sukūrus transformerių architektūrą yra vertimas iš vienos kalbos į kitą. Kaip pavyzdį galime paimti iš anglų į prancūzų kalbos vertimą:



5 pav. Vertimo kokybės palyginimas [3].

Matome, jog transformerių architektūra pagerino šios užduoties efektyvumą net dviem procentais [Ash17]. Tai yra puikus rezultatas įvertinant ir tai, jog tranformerių architektūra yra pakankamai nauja ir galima tikėtis dar daug naujovių bei tobulinimų (angl. fine tuning) šiai architektūrai lyginant su rekurentiniais neuroniniais tinklais, kurie jau yra ilgai tobulinami. Taip pat dar vienas transformerių architektūros pranašumas yra tai, jog galima atvaizduoti į kokius žodžius transformėris kreipia dėmesį pasirinkdamas atitinkamą kitos kalbos žodį [Ash17]. Atsižvelgus į išanalizuotą literatūrą, pastebėta, jog BERT veikia panašiau kaip mūsų smegenys ir atsižvelgia į žodžio kontekstą sakinyje. Rekurentiniai neuroniniai tinklai taip pat yra labiau ištirta mokslo sritis [Zhe18]. Dėl to

galima būtų tikėtis daugiau atradimų naujose architektūrose kaip BERT. Taip pat, BERT pranašumas kad galima naudoti neprižiūrijamą mokymąsi sukurti bendriniam kalbos modeliui turėtų būti labai naudingas turint mažai struktūrizuotų duomenų. Taip pat vienas iš skirtumų yra apmokymo laikas, rekurentiniai dirbtiniai neuroniniai tinklai gali būti apmokomi greičiau nei transformerių tipo dirbtiniai neuroniniai tinklai, nes jie būna mažesni. Beto, vienas didžiausių pranašumų dėl ko transformerių tipo dirbtiniai neuroniniai tinklai turėtų būti pranašesni lietuvių kalboje yra sąlyginai maži struktūrizuotų duomenų rinkiniai. Transformerių tipo dirbtiniai neuroniniai tinklai taip pat naudoja daugiau RAM atminties apmokymo metu, į tai turėtų būti atsižvelgiama, jei nėra prieigos prie daug RAM atminties turinčių kompiuterių ar serverių. Pavyzdžiui, BERT neuroninis tinklas buvo apmokomas 4 dienas naudojant angliškos vikipedijos duomenis [Ash17].

4. Sentimento analizės lietuvių kalba tyrimas

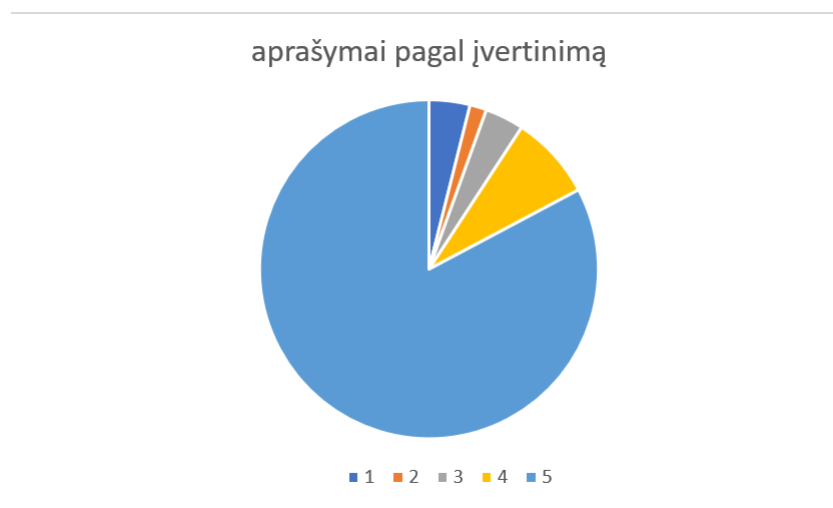
4.1. Sentimento analizė pasitelkiant rekurentinius dirbtinius neuroninius tinklus

Sentimento analizės uždaviniui spręsti nutarta keisti duomenų kiekį, neuronų skaičių bei sluoksnių skaičių dirbtiniuose neuroniniuose tinkluose norint rasti konfigūraciją, kuri tiksliai nustatytų sentimentą.

duomenys	neuronų skaičius	sluoksnių skaičius	tikslumas
3200	200	1	91.8 %
3200	300	1	92.4 %
3200	400	1	90.6 %
3200	300	2	91.8 %
3200	300	3	91.4 %

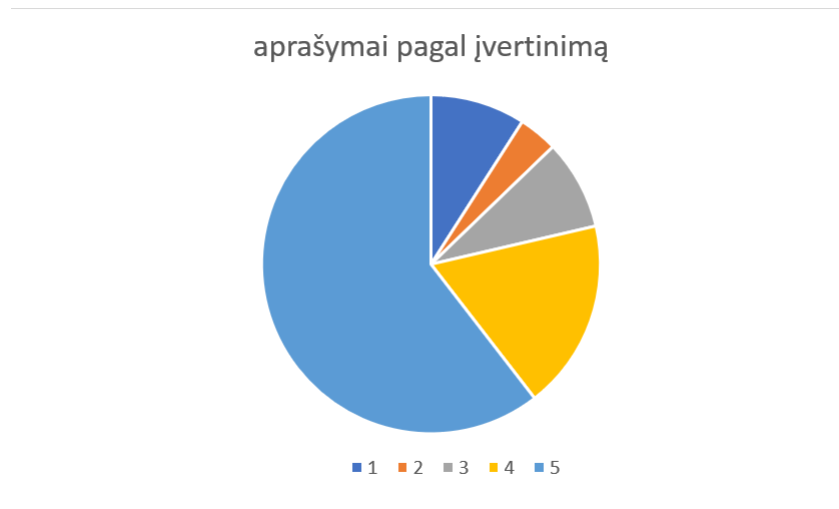
Sentimento analizei naudotas surinktas prekių aprašymų duomenų rinkinys. Sentimento analizę galima išskirstyti į du uždavinius. Galima suskirstyti prekių aprašymus į dvi klases pozityvius bei negatyvius (lengvas uždavinys). Taip pat galima palikti duomenų rinkinį koks jis yra. Tai būtų penkios klasės pagal įvertinimą nuo vieno iki penkių. Apmokymui naudota 3200 aprašymų, testavimo daliai palikta 800 aprašymų.

Sprendžiant pirmąjį uždavinį pasitelkiant rekurentinius dirbtinius neuroninius tinklus gautas 92.4 % procentų tikslumas. Geriausiai pasirodė 300 neuronų 1 sluoksniu dirbtinis neuroninis tinklas. Norint pagerinti tikslumą visada gera idėja peržvelgti apmokomuosius duomenis. Įdomiausia yra atsižvelgti į duomenų pasiskirstymą.



6 pav. Originalus duomenų pasiskirstymas prekių aprašymo duomenų rinkinyje

Matoma, jog 5 žvaigždučių aprašymai labai duominuoja turimuose duomenyse. Galima sumažinti dominavimą išimant 40 % 5 žvaigždučių aprašymų ir gaunant tokį aprašymų pagal įvertinimus pasiskirstymą:



7 pav. Modifikuoto prekių aprašymų rinkinio duomenų pasiskirstymas

Sumažinus 5 žvaigždučių duomenų kiekį, dirbtinis neuroninis tinklas gali apsimokyti tiksliau nustatyti kitų įvertinimų aprašymus, nes penkių žvaigždučių aprašymai nebeužgožia kitų duomenų.

Atlikus tyrimą gavome tokius rezultatus:

duomenys	neuronų skaičius	sluoksnių skaičius	tikslumas
2700	300	1	93.3 %

Pastebėta, jog rezultatas geresnis, dėl prieš tai jau minėtų užduočių. Užgožiančių duomenų dažnio sumažinimas yra standartinė praktika duomenų moksle. Gautas 93.3 procentų rezultatas yra ganėtinai geras šiam uždaviniui, palyginimui, anglų kalboje panašaus uždavinio (IMDB filmų aprašymai ir jų įvertinimai) geriausias rezultatas yra 97.9 procento.

Toliau tęsiamas sprendimas su sudėtingesniu įvertinimo nuspėjimo uždaviniu. Sprendžiant šį uždavinį įvestis bus aprašymo tekstas o išvestis bus skaičius nuo vieno iki penkių. Dirbtinis neuroninis tinklas bandys nuspėti įvertinimą tik turint tekstą.

duomenys	neuronų skaičius	sluoksnių skaičius	tikslumas
3200	200	1	74.4 %
3200	300	1	82.3 %
3200	300	2	82 %
3200	300	3	80.6 %

Matome, jog rezultatai gerokai prastesni nei binariniam klasifikavimo uždavinyje, nes uždavinys yra sunkesnis bei duomenų rinkinys yra ganėtinai mažas. Dar labiau sumažinus rinkinį nesitikima geresnio rezultato, taigi tas atvejis nėra tiriamas. Taip yra todėl, nes dar labiau sumažinus duomenų kiekį duomenų bus per mažai penkioms klasėms identifikuoti.

4.2. Sentimento analizė pasitelkiant transformerių tipo dirbtinius neuroninius tinklus

Sprendžiamasis pirmasis uždavinys panašiai kaip ir rekurentinių dirbtinių neuroninių tinklų skiltyje.

duomenys	tinklo tipas	tikslumas
3200	BERT	89.7 %
3200	XLM-ROBERTA	90.9 %
2700	BERT	83.6 %
2700	XLM-ROBERTA	88.0 %

Matome, jog rezultatai yra prastesni nei rekurentinių dirbtinių neuroninių tinklų, gali būti, jog tai yra dėl mažo turimų duomenų kiekio, matomas mažesnis tikslumas net sumažinus penkių žvaigždučių įvertinimų kiekį. Po to sprendžiamas 5 klasių uždavinys.

duomenys	tinklo tipas	tikslumas
3200	BERT	80.4 %
3200	XLM-ROBERTA	82.9 %
2700	BERT	71.3 %
2700	XLM-ROBERTA	72.8 %

4.3. Dirbtinių neuroninių tinklų sprendimo būdų palyginimas su esamais sprendimo būdais sentimentų analizei nustatyti

Yra ir kitų būdų spręsti sentimentų analizės uždavinius. Vienas iš jų yra žodynu paremtas sprendimo modelis. Šio modelio trūkumas yra, jog žodžiai dažniausiai turi būti surenkami rankomis ir kiekvienam iš jų priskiriamas sentimentų koeficientas, ar žodis teigiamas ar neigiamas [Tam18]. Tam reikia skirti labai daug laiko kai tuo tarpu apmokytą dirbtinį neuroninį tinklą yra lengviau, bet reikalauja turėti daugiau struktūrizuotų duomenų. Tuo tarpu žodynu paremtas sprendimo būdas nereikalauja turėti daug struktūrizuotų duomenų ir gali būti naudojamas turint mažą duomenų kiekį. Žodynų paremtų algoritmų tikslumas svyruoja tarp 70 - 90 % [Tam18] ir tai priklauso nuo žodyno dydžio ir žodžių nesančių žodyne toleravimo kiekio. Taigi, renkantis sprendimo būdą labiausiai reiktų atsižvelgti į turimą duomenų kiekį, jei jis yra mažas - geriau rinktis žodynu paremtą metodą, kitu atveju - dirbtinius neuroninius tinklus.

4.4. Rezultatų apibendrinimas

Pastebėta, jog tiksliausiai binarinį sentimentą nustatė rekurentiniai dirbtiniai neuroniniai tinklai. Geriausiai tiko 1 sluoksnio 300 neuronų konfigūracija. Buvo pasiektas 93.3 % tikslumas binariame uždavinyje. XLM-Roberta geriausiai iš visų dirbtinių neuroninių tinklų atliko įvertinimo iš teksto atspėjimo uždavinį ir pasiekė 82.9 % tikslumą. Anglų kalboje sentimentų analizės uždaviniuose rekurentiniai dirbtiniai neuroniniai tinklai turi panašius rezultatus kaip ir transformeriais grįsti dirbtiniai neuroniniai tinklai. Lietuvių kalboje matoma ta pati tendencija.

5. Skyrybos klaidų aptikimo tyrimas

5.1. Tyrimo apibendrinimas

Tyrimo metu buvo analizuojama, kaip rekurentinių dirbtiniams neuroniniams tinklams bei transformeriais grįstiems dirbtiniams neuroniniams tinklams pavyks identifikuoti skyrybos klaidas lietuvių kalboje. Siekiama, kad pavyzdžiui įrašius "Tomas norėjo kad šuo pasilikėtų jo namuose." neuroniniai tinklai sugebėtų nusakyti, jog trūksta kablelio prieš žodį "kad". Neuroniniai tinklai buvo mokomi pasitelkiant lietuviškos vikipedijos duomenis.

5.2. Duomenys

Tyrimui naudoti lietuviškos vikipedijos duomenų rinkinys. Iš pradžių jis buvo apdorotas išskaidant kiekvieną sakinį į atskirą eilutę. Poto sakiniai buvo filtruojami pagal jo ilgį, palikti sakiniai nuo 5 iki 12 žodžių. Tai leido išvengti nereikalingų duomenų kur sakiniai labai trumpi ir neturi skiriamųjų ženklų ir labai ilgų sakinių, kurių būtų nepakankamai dirbtiniam neuroniniam tinklui apmokyti. Dirbtiniams neuroniniams tinklams būtų sudėtinga surasti daugybę skiriamųjų ženklų ilgame sakinyje [Val17]. Po filtravimo duomenys suklasifikuojami į skyrybos ženklus turinčius (šiuo atveju tai būtų kableliai) bei papildomų skyrybos ženklų neturinčius sakinius. Duomenų rinkinys sudarytas taip, jog proporcingai jame būtų 50 procentų sakinių be papildomų skyrybos ženklų bei likę 50 procentų su papildomais skyrybos ženklais.

5.3. Duomenų paruošimas neuroniniams tinklams

Norint apmokyti dirbtinius neuroninius tinklus reikalingas duomenų apdorojimas ir teksto pavertimas atitinkamais vektoriais. Sakinys yra paverčiamas į skaičių seką, tai padaroma pasitelkiant žodyną. Žodyne būna surašytos žodžių dalys ir tą žodžio dalį atitinkanti skaičius. Pavyzdžiui žodis "Jonas" gali būti išskirstytas į tokias dalis ["Jo", "nas"]. Šios dalys tada paverčiamos į jas atitinkančius skaičius [1, 2] ir duomenys tokiu formatu perduodami dirbtiniam neuroniniam tinklui.

Duomenys buvo paruošti naudojant tensorflow "tokenizavimo" biblioteką. Ši biblioteka sudaro žodyną, kur tam tikras skaičius atitinka vieną žodį. Žodžiai buvo gauti atskiriant žodžiu pagal tarpo simbolį. Skiriamieji ženklai esantys prie žodžių buvo pašalinti. Sudarytas žodynas turintis skaičių atitikmenis kiekvienam žodžiui.

Prieš paverčiant sakinius į skaičių vektorius reikėjo sugalvoti kokius išvestinius duomenis mums duos dirbtinis neuroninis tinklas nustatant ar sakinyje esančias skyrybos klaidas. Buvo nuspręsta, jog paskutiniame dirbtinio neuroninio tinklo sluoksnyje bus tiek neuronų kiek maximum gali būti tarpų sakinyje (šiuo atveju 49). Jei toje vietoje dirbtinis neuroninis tinklas nuspręs, jog yra reikalingas skiriamasis ženklas, tikimės gauti vienetą toje vietoje, o kitu atveju 0. Pavyzdžiui: įvesties sakinyje bus "Tomas norėjo, kad šuo pasilikėtų pas jį". Jis bus paverčiamas į skaičių vektorius [1, 2, 3, 4, 5, 6, 7] ir tikimasi gauti rezultatą [0, 1, 0, 0, 0, 0] tai reikšų, jog antrojo tarpo vietoje yra reikalingas kablelis. Aišku pavyzdyje dar nepavaizduota tai, jog gale įvesties bei išvesties pridedami

nuliai tam, kad išlaikyti vienodas dimensijas skirtingo ilgio sakiniams bei kad vieną žodis gali būti suskaidytas į kelias dalis ir jį gali reprezentuoti keli skaitmenys.

5.4. Tyrimo būdai

Šiam uždaviniui spręsti buvo pasirinkta taip pat naudoti rekurentinius dirbtinius neuroninius tinklus dėl to, kad jie yra naudojami panašioms gramatikos klaidų aptikimo uždaviniams atlikti anglų kalboje. Taip pat buvo pasirinkta iš transformeriais grįstų dirbtinių neuroninių tinklų naudoti XML-ROBERTA kadangi tinklas turi jau apmokytus daugiakalbius modelius, kurie buvo mokomi įskaitant ir lietuvių kalbą. Tai duoda aukštesnius tikslumo rezultatus, nes dirbtinis neuroninis tinklas jau būna apmokytas atspėti žodžius sakinyje.

Išanalizavus literatūrą, buvo rastas kinų kalboje atliktas panašus uždavinys, ir ten buvo nuspręsta uždavinį spręsti ne tik naudojant dirbtinius neuroninius tinklus, bet uždavinį spręsti mišriu būdu [Yaj16]. Šiuo atveju pasvarsčius apie lietuvių kalboje esančias taisykles, buvo nuspręsta, jog kai kurias taisykles galima lengvai įgyvendinti suprogramavus paprastą žodžių tikrinimo logiką bei žinant žodžių morfologinę informaciją (ar žodis yra daiktavardis, būdvardis, veiksmažodis ir t.t.). Taigi, buvo pasirinkta atlikti uždavinius pasitelkiant tik suprogramuotas taisykles, pasitelkiant tik dirbtinius neuroninius tinklus, bei pasitelkiant mišrų sprendimo būdą, kuris pasitelks suprogramuotas taisykles ir dirbtinius neuroninius tinklus.

5.5. Skyrybos klaidų aptikimo tyrimas pasitelkiant suprogramuotas taisykles

Atsižvelgus į lietuvių kalbos skyrybos taisykles, buvo nuspręsta, jog pakankamai lengvai galima suprogramuoti logiką taisykle "kablelis dedamas prieš žodžius o, jog, tačiau, nes, kad". Bei į taisyklę, jog vienaarūšės sakinio dalys yra atskiriamos kableliais. Pirmoji taisyklė buvo suprogramuota tiesiog surašant specialius žodžius prieš kuriuos reikalingas kablelis. Antroji taisyklė suprogramuota pasitelkiant morfeminę lietuvių kalbos žodyną [Eri11]. Žodžius bandoma surasti žodyne ir jei pavyksta rasti abu šalia vienas kito sakinyje esančius žodžius žodyne, tikrinama ar žodžiai yra vienaarūšės sakinio dalys. Kaip duomenys naudojami 40 000 sakinių. 20 000 be kablelių, o kiti 20 000 su kableliu. Testavimui naudojami 4 000 sakinių, kur 2000 sakinių be kablelių, o kiti 2000 sakinių su kableliais, nebent tyrime parašyta kitaip. Sprendžiamas pilno sakinio skyrybos klaidų aptikimo uždavinys:

duomenys	tikslumas
40000	53.8 %

Matome, jog tikslumas yra ganėtinai mažas, nes suprogramuotos tik kelios taisyklės.

5.6. Tyrimas pasitelkiant rekurentinius dirbtinius neuroninius tinklus

Pasitelkinat rekurentinius dirbtinius neuroninius tinklus skyrybos klaidoms aptikti buvo sudarytos dvi užduotys. Pirmoji užduotis buvo, tokia, jog kaip įvestis buvo pateikiamas visas sakinyš,

jei jis turėjo skiriamuosius ženklus, jie ištrinami. Dirbtinis neuroninis tinklas turi nuspėti ar kableliai buvo ištrinti ar ne ties kiekviena vieta tarp žodžių. Užduotis rekurentiniams dirbtiniams neuroniniams tinklas buvo pernelyg sudėtinga. Tikslumas maksimaliai pasiekė 7 procentus.

Taip pat buvo atliktas ir antrasis uždavinys. Paimami keturi žodžiai ir rekurentinis dirbtinis neuroninis tinklas bando nustatyti ar tarp antro bei trečio žodžių yra reikalingas kablelis. Šis uždavinys dirbtiniam neuroniniam tinklui suteikia daugiau konteksto ir iš to lengviau galima nustatyti tokius kalbos darinius kaip įterpinys.

Trūkstatų kablelių nustatymas visame sakinyje:

duomenys	neuronų skaičius	sluoksnių skaičius	tikslumas
30000	200	3	3.1 %
30000	300	4	4.7 %
30000	600	4	6.4 %
40000	600	4	7 %

Pavyko pasiekti tik 7 % tikslumą, didesni duomenų kiekiai nebuvo tiriami dėl ilgo rekurentinių dirbtinių neuroninių tinklų apmokymo laiko.

Trūkstatų kablelių nustatymas tarp antrojo bei trečiojo žodžio turint keturis žodžius:

duomenys	neuronų skaičius	sluoksnių skaičius	tikslumas
30000	300	4	70.2 %
30000	600	4	74.6 %
40000	600	4	78.5 %

Rezultatas daug geresnis sprendžiant lengvesnį uždavinį.

5.7. Tyrimas pasitelkiant transformeriu grįstus dirbtinius neuroninius tinklus

Pasitelkiant XLM-Roberta dirbtinį neuroninį tinklą taip pat buvo atliktos dvi tos pačios užduotys. Su tais pačiais duomenų rinkiniais.

Trūkstatų kablelių nustatymas visame sakinyje:

duomenys	neuroninio tinklo modelis	tikslumas
30000	XLM-ROBERTA	19.3 %
40000	XLM-ROBERTA	20.6 %

Trūkstatų kablelių nustatymas tarp antrojo bei trečiojo žodžio turint keturis žodžius:

duomenys	neuroninio tinklo modelis	tikslumas
30000	XLM-ROBERTA	90.2 %
40000	XLM-ROBERTA	94.3 %

5.8. Skyrybos tyrimas pasitelkiant mišrų metodą

Kaip jau minėta, atsižvelgus į lietuvių kalbos gramatikos taisykles [Dru]. Buvo nuspręsta, jog būtų labai sudėtinga parašyti programą, kuri galėtų teisingai sudėti kablelius lietuvių kalboje. Dauguma taisyklių priklauso nuo konteksto. Viena taisyklė, kuriai būtų lengva sudėti kablelius yra ta, kad prieš žodžius (o, bet, jog, tačiau) turi būti kablelis. Taip pat pakankamai lengva nustatyti vienasarūšės sakinio dalis. Taigi buvo pasirinkta atlikti užduotį kombinuojant suprogramuotas taisykles bei dirbtinius neuroninius tinklus. Parašyta programa veikia tokiais žingsniais:

1. Paima du žodžius ir tikrina ar antrasis žodis yra o, bet, tačiau, kad, nes, jei taip, tarp žodžių dedamas kablelis. 2. Tikrina ar žodžiai yra vienasarūšės sakinio dalys, jei taip, dedamas kablelis. 3. Jei prieš praitus du žingsnius vis dar nėra padėtas kablelis, imamas keturių žodžių kontekstas kaip įvestis dirbtiniam neuroniniam tinklui (šiuo atveju XLM-ROBERTA) ir dirbtinis neuroninis tinklas nusprendžia ar tarp žodžių visgi turi būti kablelis.

Naudojant šį algoritmą gautas 65 % tikslumas bandant įvertinti individualius sakinius. Naudojant tik dirbtinius neuroninius tinklus (3 žingsnis) 61.2 %.

5.9. Tyrimo metu patirti sunkumai

Tyrimo metu buvo patirta daugybė sunkumų. Tarp jų - dirbtinių neuroninių tinklų apmokymo laikas. Apmokyti dirbtinius neuroninius tinklus dažnai prireikdavo nuo 4 iki 24 valandų laiko, pareikalauta daug resursų. Taip pat transformerių tipo dirbtinių neuroninių tinklų apmokymas reikalavo daug RAM atminties. Šioms problemoms išspręsti buvo pasitelkta Google Cloud platforma, kuri suteikė galimybę naudoti TPU (angl. Tensor Processing Unit) procesorius, kad pagreitinti apmokymo laiką bei buvo galima naudoti skirtingas virtualias mašinas turinčias iki 64GB RAM atminties.

5.10. Tyrimo rezultatai

Vien tik programuojant skyrybos taisykles lietuvių kalbai buvo pasiektas mažas 7 % tikslumas, nes daugumai lietuvių kalbos taisyklių sunku suprogramuoti taisykles. Pasitelkiant rekurentinius dirbtinius neuroninius tinklus rezultatai taip pat ne tokie tikslūs kaip pasitelkiant transformerių tipo dirbtinius neuroninius tinklus. Transformerių tipo dirbtinis neuroninis tinklas rodė labai gerus rezultatus bandant nuspėti ar turint keturis žodžius tarp antrojo ir trečiojo žodžio turėtų būti kablelis, pasiektas net 94.3 % tikslumas. Geriausias rezultatas (65 % tikslumas) aptinkant skyrybos klaidas visame sakinyje gautas naudojant mišrų sprendimo būdą pasitelkiant suprogramuotas taisykles bei dirbtinius neuroninius tinklus.

5.11. Skyrybos klaidų aptikimo palyginimas su esamais įrankiais

Buvo rasta tik viena viešai prieinama programa, kuri taiso lietuvių kalbos skyrybą. Tai būtų www.semantika.lt svetainė. Palyginus programų tikslumą su įvairiais sakiniiais, pastebėta, jog www.semantika.lt esanti programa veikia tik kelių procentų tikslumu, sudeda kablelius tiesiog prieš tam

tikrą žodžių sąrašą (o, kad, bet, tačiau, nes), nesudeda kablelių ties įterpiniais ir t.t. Taigi šiame darbe sukurta programa yra keletą kartų tikslesnė už rastus alternatyvius įrankius.

5.12. Plėtros galimybės

Šiame darbe sukurta programa nėra pritaikyta komerciniams tikslams. Egzistuoja galimybė sukurti skyrybos taisymo programą lietuvių kalbai ir pritaikyti viešajam sektoriui ar privačiuose sektoriuose. Taip pat darbe aprašytu būdu galima sukurti skyrybos tikrinimo programas būtų galima sukurti panašioms kalboms latvių bei estų, nes iš dalies naudojamas neprižiūrimojo mokymosi algoritmas, kur nuo sakinių automatiškai yra nuimami kableliai ir dirbtinis neuroninis tinklas bando nustatyti, kurioje sakinio vietoje turėtų būti kablelis. Taip pat programos rezultatus yra galimybė pagerinti naudojant daugiau duomenų. Darbe buvo naudota duomenų kiek buvo įmanoma su turimais įrankiais bei finansais.

Rezultatai

- Paruoštas prekių aprašymų struktūrizuotas duomenų rinkinys su 4000 duomenų egzempliorių.
- Paruoštas iš wikipedia.lt straipsnių gautas struktūrizuotas duomenų rinkinys skyrybos klaidoms aptikti. Jis sudarytas iš 400 000 egzempliorių.
- Ištirtas sentimentų analizės uždavinio sprendimo tikslumas naudojant prekių aprašymų duomenų rinkinį pasitelkiant rekurentinių dirbtinių neuroninių tinklus bei transformeriais grįstus neuroninius tinklus. Gautas 93.3 % tikslumas.
- Ištirtas skyrybos klaidų aptikimo uždavinio sprendimas pasitelkiant rekurentinius dirbtinius neuroninius tinklus bei transformeriais grįstus neuroninius tinklus. Pasiūlytas mišrus problemos sprendimo algoritmas. Gautas 65 % tikslumas.
- Sukurtas ganėtinai tikslus du trečdalius skyrybos klaidų aptinkantis įrankis.
- Sukurtas lietuviškų žodžių bei juos atitinkančių kalbos dalių sąrašas. Sąrašas sukurtas lengvai programoms perskaitomu json formatu.

Išvados

- Dirbtiniai neuroniniai tinklai pasiekia labai aukštą tikslumą sprendžiant sentimentų analizės uždavinius lietuvių kalba. Rezultatai yra panašūs kaip ir anglų kalboje.
- Dirbtiniams neuroniniams tinklams pavyko pasiekti 65 procentų tikslumą sprendžiant skyrybos klaidų aptikimo uždavinius. Tai įrodo, jog galima sukurti programą automatizuotam lietuvių kalbos skyrybos taisymui.
- Transformeriais grįsti neuroniniai tinklai lietuvių kalboje pasiekia didesnę tikslumą sprendžiant skyrybos klaidų aptikimo uždavinį. Tuo tarpu rekurentiniai dirbtiniai neuroniniai tinklai pasiekė geresnius rezultatus sprendžiant sentimentų analizės uždavinį.

Literatūra

- [Ale19] Julian Michael Alex Wang Amanpreet Singh. *GLUE: A MULTI-TASK BENCHMARK AND ANALYSIS PLATFORM FOR NATURAL LANGUAGE UNDERSTANDING*, 2019.
- [And09] Vidas Valskys Andrius Utkė Erika Rimkutė. Dažninis rašytinės lietuvių kalbos žodynas, Žodynas, 2009.
- [Ank19] Ming Xu Ankit Kumar Zixin Luo. *Text Summarization using Natural Language Processing*, 2019.
- [Ash17] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani Noam Shazeer. *Attention Is All You Need*, 2017.
- [Che19] Alexander J. Smola Chenguang Wang Mu Li. *Language Models with Transformers*, 2019.
- [Dea06] Tomas Mikolov Ilya Sutskever Kai Chen Greg Corrado Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2006.
- [Dru] Albinas Drukteinis. *Lietuvių kalbos gramatikos taisyklės*. tinklapis <http://www.vlkk.lt/vlkk-nutarimai/nutarimai/nutarimo-del-lietuviu-kalbos-taisykles>.
- [Eri11] Gailius Raškinis Erika Rimkutė Asta Kazlauskienė. *Abėcėlinis lietuvių kalbos morfemikos žodynas*. Žodynas, 2011.
- [Ett19] Allyson Ettinger. *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*, 2019.
- [Hri11] Florentina Hristea. *Statistical Natural Language Processing*, 2011-01, p. 1452–1453.
- [Yaj16] Liyan Zhuo Yajun Liu Yingjie Han. *Automatic Grammatical Error Detection for Chinese based on Conditional Random Field*, 2016.
- [Jac19] Chang Kenton Lee Kristina Toutanova Jacob Devlin Ming-Wei. *Deep contextualized word representations*, 2019.
- [Kar18] Andrej Karpathy. Understanding lstm networks, 2018.
- [Kha19] Oleksii Kharkovyna. *Natural Language Processing (NLP): Top 10 Applications*. Straipsnis, 2019.
- [KS15] Jasleen Kaur ir Jatinderkumar Saini. A study of text classification natural language processing algorithms for indian languages. *VNSGU Journal of Science and Technology*, 4:162–167, 2015-07.
- [Loi18] Jurgita Dzikienė Loïc Boizou. *Deeper Error Analysis of Lithuanian Morphological Analyzers*, 2018.
- [Phu18] Kyunghyun Cho Phu Mon Htut Samuel R. Bowman. *Training a Ranking Function for Open-Domain Question Answering*, 2018.
- [Pra16] Konstantin Lopyrev Pranav Rajpurkar Jian Zhang. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, 2016.

- [Pro01] Prof. Paolo Frasconi Dr. MER Martin Rajman Prof. R. Hersch Prof. Wulfram Gerstner Dr. habil. Jurgen Schmidhuber. *Long Short-Term Memory in Recurrent Neural Networks*, 2001.
- [Qia19] Tong Xiao Qiang Wang Bei Li. *Learning Deep Transformer Models for Machine Translation*, 2019.
- [Ram16] Apoorv Agarwal Boyi Xie Ilia Vovsha Owen Rambow. *Sentiment Analysis of Twitter Data*, 2016.
- [San18] Alessandro Sordoni Sandeep Subramanian Sai Rajeswar. *Towards Text Generation with Adversarially Learned Neural Outlines*, 2018.
- [Tam18] Justas Tamašauskas. *Socialiniuose tinkluose naudojamų emotikonų ir sentimentų analizės vertinimo tyrimas*. Magistro baigiamasis darbas, 2018.
- [Val17] Hojjat Salehinejad Sharan Sankar Joseph Barfett Errol Colak Shahrokh Valaee. *Recent Advances in Recurrent Neural Networks*, 2017.
- [Zhe18] Sebastian Goodman Kevin Gimpel Piyush Sharma Radu Soricut Zhenzhong Lan Mingda Chen. *ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS*, 2018.
- [ZIN00] Vytautas ZINKEVIČIUS. *Lemuoklis morfologinei analizei*. Straipsnis, 2000.

Sąvokų žodynas

BERT - Bidirectional Encoder Representations from Transformers

RNN - Rekurentinis neuroniniai tinklai

TPU - tensorinis apdorojimo įrenginys (procesorius skirtas dirbtinių neuroninių tinklo mokymui).

Priedas nr. 1

Kalbos dalių žodyno pavyzdys

```
{
  "word": "abatijos", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "kilm."},
  "word": "abécèle", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "inag."},
  "word": "abécèlè", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abécèlèms", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "dgs.", "linksnis": "naud."},
  "word": "abécèlès", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "kilm."},
  "word": "abejingumą", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "gal."},
  "word": "abejingumas", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abejingumo", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "kilm."},
  "word": "abejingumu", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "inag."},
  "word": "abejojimas", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abejone", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "inag."},
  "word": "abejonė", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abejonėmis", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "dgs.", "linksnis": "inag."},
  "word": "abejonėms", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "dgs.", "linksnis": "naud."},
  "word": "abejones", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "dgs.", "linksnis": "gal."},
  "word": "abejonės", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "vns.", "linksnis": "kilm."},
  "word": "abejonių", "kalbos_dalis": "dkt.", "gimine": "mot.", "skaita": "dgs.", "linksnis": "kilm."},
  "word": "abipusiškumą", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "gal."},
  "word": "abipusiškumas", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abipusiškumo", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "kilm."},
  "word": "abipusiškumu", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "inag."},
  "word": "abipusiškumui", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "naud."},
  "word": "abitu", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "inag."},
  "word": "abitų", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "dgs.", "linksnis": "kilm."},
  "word": "abiturientas", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "vard."},
  "word": "abolicionistų", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "dgs.", "linksnis": "kilm."},
  "word": "abolicionizmu", "kalbos_dalis": "dkt.", "gimine": "vyr.", "skaita": "vns.", "linksnis": "inag."}
```

Priedas nr. 2

Prekių aprašymų duomenų rinkinio pavyzdys

```
[{"rating": "4", "description": "Pagal kainą viskas normalu, puiki vaza"}, {"rating": "5", "description": "Labai graži skaidraus stiklo vaza, gan didelė."}, {"rating": "5", "description": "Efektyvus dezodorantas, malonus kvapas, ilgai užtenka, perku jau nebe pirmą kartą."}, {"rating": "5", "description": "Nebe pirmas šios prekės pirkinys. Viskas ok"}, {"rating": "5", "description": "Geras produktas su švelniu kvapu ir efektyvus."}, {"rating": "4", "description": "Kaina atitinka kokybę"}, {"rating": "5", "description": "Pirkau, zinodamas ką perku. Problemos nebuvo."}, {"rating": "2", "description": "puodeliai gražūs, bet vienas po pirmo plovimo skilo"}, {"rating": "5", "description": "Siurbliui Sencor SVC 1030 CAT & DOG filtras tiko"}, {"rating": "5", "description": "Tinka puikiai siurbliui Sensor SVC 1030"}, {"rating": "5", "description": "nes reikia atsarginio filtro"}, {"rating": "3", "description": "Labai maži tik ekspres kavai"}, {"rating": "5", "description": "....."}, {"rating": "4", "description": "TIKO MANO SIURBLIUI. NORS JUSU DARBUOTOJAI NEGALEJO PATVIRTINTI AR TIKS"}, {"rating": "5", "description": "Viskas puikiai."}, {"rating": "5", "description": "dar nepakeičiau, nes pirkau pakeitimui į turimą dulkių siurbly. bet šiaip tai gera prekė kadangi galima plauti visą"}]
```

Skrybos klaidų aptikimo duomenų rinkinio pavyzdys

34

Priedas nr. 4

Duomenų rinkiniai ir kodo pavyzdžiai

Duomenų rinkinius bei kodo pavyzdžius galima rasti čia: <https://github.com/mindaugasbarc/BakalauroDarbas>