VILNIUS TECH
Elektroninių sistemų katedra

Daumantas TINTERIS

TEKSTO AUTORYSTĖS MODELIAVIMAS IR IDENTIFIKAVIMAS
TEXT AUTHORSHIP MODELING AND IDENTIFICATION

Baigiamasis magistro darbas

Informacinių elektroninių sistemų studijų programa, valstybinis kodas 6211BX015
Dirbtinio intelekto sistemų specializacija
Informatikos inžinerijos studijų kryptis

Vilnius, 2022

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
ELEKTRONIKOS FAKULTETAS
ELEKTRONINIŲ SISTEMŲ KATEDRA

TVIRTINU
Katedros vedėjas
_____
(parašas)
Prof. dr. A. Serackis
2022 m. 05 mėn. 25 d.

Daumantas TINTERIS

TEKSTO AUTORYSTĖS MODELIAVIMAS IR IDENTIFIKAVIMAS
TEXT AUTHORSHIP MODELING AND IDENTIFICATION

Baigiamasis magistro darbas

Informacinių elektroninių sistemų studijų programa, valstybinis kodas 6211BX015
Dirbtinio intelekto sistemų specializacija
Informatikos inžinerijos studijų kryptis

Vadovas	doc. dr. G. Tamulevičius (*pedag. vardas, moksl. laipsnis, vardas, pavardė*)	_____ (*parašas*) (*data*)
Lietuvių kalbos konsultantas	dr. A. Žemienė (pedag. vardas, moksl. laipsnis, vardas, pavardė)	_____ (parašas) (data)

Vilnius, 2022

Vilniaus Gedimino technikos universitetas
Elektronikos fakultetas
Elektroninių sistemų katedra

ISBN ISSN
Egz. sk. .....
Data .....-.....-.....

Antrosios pakopos studijų **Informacinių elektroninių sistemų** programos magistro baigiamasis darbas

Pavadinimas	**Teksto autorystės modeliavimas ir identifikavimas**
Autorius	**Daumantas Tinteris**
Vadovas	**Gintautas Tamulevičius**

Kalba: lietuvių

Anotacija

Baigiamajame magistro darbe nagrinėjama kalbos tekstų autorystės nustatymo tema. Tekstų lietuvių kalba autorystei nustatyti pasirinkti giliojo mokymosi tinklai: daugiasluoksnis perceptronas (MLP), konvoliuciniai neuronų tinklai (CNN), rekurentiniai neuronų tinklai (RNN), ilgos trumpalaikės atminties metodas (LSTM) ir autoenkoderiai. Tyrimo metu pasirinktieji metodai palyginti su kitais mašininio mokymosi metodais: atraminių vektorių klasifikatoriumi (SVM), k-artimiausių kaimynų algoritmu (KNN) ir Bajeso tikimybiniu klasifikatoriumi (Bayes). Kaip duomenys panaudojami lietuvių kalbos tekstai – 147 parlamentarų pasisakymai, kurių bendras skaičius siekė daugiau nei 110 tūkst. Metrikoms buvo pasirinktos n-gramos modelis. Tyrimo metu didžiausias gautasis lietuvių kalbos teksto autorystės nustatymo tikslumas siekė 74 %. Remiantis gautaisiais rezultatais, pateikiamos išvados ir rekomendacijos. Darbą sudaro: įvadas, tekstų autorystės identifikavimas, dirbtinio intelekto metodų analizė teksto autorystei identifikuoti, eksperimentinio tyrimo rezultatai, išvados, rekomendacijos ir literatūros sąrašas. Darbo apimtis – 46 p. teksto be priedų, 12 paveikslų, 5 lentelės, 37 bibliografiniai šaltiniai.

Prasminiai žodžiai: autorystės identifikavimas, dirbtinio intelekto metodai, n-gramos, analitinė tyrimų apžvalga, lietuvių kalbos tekstai.

Vilnius Gediminas Technical University
Faculty of Electronics
Department of Electronic Systems

ISBN ISSN
Copies No. .....
Date .....-.....-.....

Master Degree Studies **Information Electronics Systems** study programme Master Graduation Thesis

Title	**Text Authorship Modeling and Identification**
Author	**Daumantas Tinteris**
Academic supervisor	**Gintautas Tamulevičius**

Thesis language: Lithuanian

Annotation

The final master's thesis deals with the topic of authorship of language texts. The deep learning networks chosen for the authorship of English language texts are the Multilayer Perceptron (MLP), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) and autoencoders. The study compares the selected methods with other machine learning methods: support vector machine (SVM), k-nearest neighbours algorithm (CNN) and Bayesian probabilistic classifier (Bayes). The data used are Lithuanian language texts - 147 parliamentary speeches with a total number of more than 110,000. The n-gram model was chosen for the metrics. The highest accuracy obtained in the study was 74%. Based on the results, conclusions and recommendations are presented. The paper consists of: introduction, text authorship identification, analysis of artificial intelligence methods for text authorship identification, results of the experimental study, conclusions, recommendations and reference list. Thesis consists of 46 p. text without appendixes, 12 pictures, 5 tables, 37 bibliographical entries. Appendixes are included separately.

Keywords: authorship identification, artificial intelligence methods, n-grams, analytical research review, Lithuanian texts.

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Daumantas Tinteris, 20200074

(Studento vardas ir pavardė, studento pažymėjimo Nr.)

Elektronikos fakultetas

(Fakultetas)

Informacinės elektroninės sistemos, DISfm-20

(Studijų programa, akademinė grupė)

BAIGIAMOJO DARBO (PROJEKTO)

SĄŽININGUMO DEKLARACIJA

2022 m. gegužės 30 d.

Patvirtinu, kad mano baigiamasis darbas tema „Teksto autorystės modeliavimas ir identifikavimas“ yra savarankiškai parašytas. Šiame darbe pateikta medžiaga nėra plagijuota. Tiesiogiai ar netiesiogiai panaudotos kitų šaltinių citatos pažymėtos literatūros nuorodose.

Parenkant ir įvertinant medžiagą bei rengiant baigiamąjį darbą, mane konsultavo mokslininkai ir specialistai: docentas daktaras Gintautas Tamulevičius. Mano darbo vadovas docentas daktaras Gintautas Tamulevičius.

Kitų asmenų indėlio į parengtą baigiamąjį darbą nėra. Jokių įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs (-usi).

(Parašas)

Daumantas Tinteris

(Vardas ir pavardė)

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

ELEKTRONIKOS FAKULTETAS

ELEKTRONINIŲ SISTEMŲ KATEDRA

Technologijos mokslų sritis
Elektros ir elektronikos inžinerijos mokslo kryptis
Informatikos inžinerijos studijų kryptis
Informacinių elektroninių sistemų studijų programa, valst. kodas 6211BX015
Dirbtinio intelekto sistemų specializacija

TVIRTINU
Katedros vedėjas
_____
prof. dr. A. Serackis
2021 m. 05 mėn. 13 d.

MAGISTRO BAIGIAMOJO DARBO UŽDUOTIS

Studentui Daumantas TINTERIS

Baigiamojo darbo tema: Teksto autorystės modeliavimas ir identifikavimas

Text Authorship Modeling And Identification

Patvirtinta 2020 m. 11 mėn. 25 d. dekano potvarkiu Nr. 159 el.

Baigiamojo darbo užbaigimo terminas 2021 m. 05 mėn. ____ d.

6211BX01

Darbo tikslas – Išnagrinėti ir ištirti teksto autorystės identifikavimo metrikas ir metodus, įvertinti jų efektyvumą analizuojant lietuvių kalbos tekstus.

Darbo uždaviniai

1. Apžvelgti ir išnagrinėti teksto autorystės nustatymo uždavinį.
2. Išanalizuoti ir palyginti dažniausiai naudojamas teksto autorystės identifikavimo metrikas ir metodus.
3. Suformuluoti ir eksperimentiškai įvertinti sprendimus lietuvių kalbos teksto autorystei identifikuoti.

Aiškinamojo rašto turinys

Įvadas

Darbo aktualumas ir tikslas

Darbo uždaviniai

Naudoti tyrimo ir analizės metodai

Darbo naujumas ir praktinė nauda

Darbo struktūra

1. Tekstų autorystės identifikavimas

1.1. Autorystės identifikavimo uždavinys

1.2. Tekstų autorystės metrikos

1.3. Tekstų autorystės identifikavimo metodų apžvalga

2. Dirbtinio intelekto metodų analizė teksto autorystei identifikuoti

2.1. Dirbtinių neuronų tinklai

2.2. Giliojo mokymo tinklų taikymas

3. Lietuvių kalbos tekstų autorystės identifikavimo tyrimas

3.1. Eksperimento duomenys

3.2. Lietuvių kalbos tekstų identifikavimo metodika

3.3. Tyrimo planas

4. Eksperimentinio tyrimo rezultatai

Apibendrinimas. Išvados

Literatūra

Priedai

A priedas. Lietuvių kalbos tekstų identifikavimo eksperimentinio tyrimo rezultatai.

Baigiamojo darbo rengimo konsultantas: —

(pedag. vardas, moksl. laipsnis, vardas, pavardė)

Vadovas **Doc. dr. Gintautas Tamulevičius**

(parašas) *(pedag. vardas, moksl. laipsnis, vardas, pavardė)*

Užduotį gavau 2021 m. 05 mėn. 30 d.

(parašas)

TURINYS

ŽYMENYS IR SANTRUMPOS **11**

ĮVADAS **12**
Darbo aktualumas ir tikslas ... 12
Darbo uždaviniai ... 12
Naudoti tyrimo ir analizės metodai ... 13
Darbo naujumas ir praktinė nauda ... 13
Darbo struktūra ... 13

1. TEKSTŲ AUTORYSTĖS IDENTIFIKAVIMAS **14**
1.1. Autorystės identifikavimo uždavinys ... 14
1.2. Tekstų autorystės metrikos ... 17
1.3. Tekstų autorystės identifikavimo metodų apžvalga ... 18
1.4. Trumpas skyriaus apibendrinimas ... 18

2. DIRBTINIO INTELEKTO METODŲ ANALIZĖ TEKSTO AUTORYSTEI IDENTIFIKUOTI **19**
2.1. Dirbtinių neuronų tinklai ... 19
2.1.1. Daugiasluoksnis perceptronas ... 19
2.1.2. Konvoliuciniai neuronų tinklai ... 20
2.1.3. Rekurentiniai neuronų tinklai ... 21
2.1.4. Autoenkoderiai ... 22
2.2. Giliojo mokymo tinklų taikymas ... 23
2.2.1. Daugiasluoksnių perceptronų panaudojimas ... 23
2.2.2. Konvoliucinių neuronų tinklų panaudojimas ... 24
2.2.3. Rekurentinių neuronų tinklų panaudojimas ... 25
2.2.4. Autoenkoderių panaudojimas ... 26
2.3. Trumpas skyriaus apibendrinimas ... 27

3. LIETUVIŲ KALBOS TEKSTŲ AUTORYSTĖS IDENTIFIKAVIMO TYRIMAS **28**
3.1. Eksperimento duomenys ... 28
3.2. Lietuvių kalbos tekstų identifikavimo metodika ... 29
3.2.1. Įrankiai ... 30
3.2.2. Duomenų įkėlimas ... 30
3.2.3. Savybių išskyrimas ir paruošimas ... 30
3.2.4. Autorystės savybių klasifikavimas ... 31
3.2.5. Rezultatų išsaugojimas ... 32
3.3. Tyrimo planas ... 32
3.4. Trumpas skyriaus apibendrinimas ... 32

4. EKSPERIMENTINIO TYRIMO REZULTATAI **34**
4.1. Simbolių n-gramų tyrimas ... 34
4.1.1. Pradinio tikslumo vertinimas ... 34
4.1.2. Savybių aibės dydžio tyrimas ... 35
4.1.3. Savybių intervalo tyrimas ... 36
4.1.4. Gautų rezultatų įvertinimas ... 37
4.2. Žodžių n-gramų tyrimas ... 38
4.2.1. Pradinio tikslumo vertinimas ... 38
4.2.2. Savybių aibės dydžio tyrimas ... 39
4.2.3. Savybių intervalo tyrimas ... 39
4.2.4. Gautų rezultatų įvertinimas ... 40
4.3. Trumpas skyriaus apibendrinimas ... 42

APIBENDRINIMAS. IŠVADOS **43**

REKOMENDACIJOS **44**

LITERATŪRA **45**

PRIEDAI **49**

A priedas . Lietuvių kalbos tekstų identifikavimo eksperimentinio simbolių savybių tyrimo rezultatai. 49

B priedas . Lietuvių kalbos tekstų identifikavimo eksperimentinio žodžių savybių tyrimo rezultatai. 52

ŽYMENYS IR SANTRUMPOS

CNN (angl. *Convolutional neural network*)	Konvoliucinis neuronų tinklas
RNN (angl. R*ecurrent neural networks*)	Rekurentiniai neuronų tinklai
LSTM (angl. *Long Short-term memory method*)	Ilgalaikės–trumpalaikės atminties metodas
MLP (angl. *Multilayer perceptron*)	Daugiasluoksnis perceptronas
SVM (angl. *Support vector machine*)	Atraminių vektorių klasifikatorius
KNN (angl. *K-nearest neighbour*)	K-artimiausių kaimynų metodas
Bayes (angl. *Bayesian networks*)	Bajeso tikimybinis klasifikatorius,
bi_LSTM(angl. *Bidirectional long short term memory)*	Dvikryptė ilgos trumpalaikės atminties neuronų tinklas
ESN (angl. *Echo State Network*)	Rezervuarų skaičiavimo metodas
NLP (angl. *Natural language processing*)	Natūralus kalbos apdorojimas
VAE (angl. *Variational autoencoders*)	Variaciniai autoenkoderiai
DNN (angl. *Deep neural networks*)	Gilieji neuronų tinklai
SDAE (angl. *Stacked Denoising Autoencoder*)	Sukrautieji autoenkoderiai
RF (angl. *Random forest*)	Atsitiktinio miško algoritmas
NBM (angl. *Naive bayes multinomial*)	*Naive bayes* daugianaris klasifikatorius
R-CNN (angl. *Region based convolutional neural networks*)	Regioniniai konvoliuciniai neuronų tinklai
BWWV (angl. *best-worst weighted vote*)	Balsavimo metodas

ĮVADAS

Darbo aktualumas ir tikslas

Autoriaus kalbos stiliaus modeliavimas yra suvokiamas kaip asmens kalbos stiliaus priskyrimas, kuris yra viena iš svarbiausių natūralios kalbos apdorojimo užduočių. Stilius yra abstrakti sąvoka ir apima unikalius autorių rašymo įpročius, pavyzdžiui, pasižymėjimą savitu kalbėjimo stiliumi – žodynu, žodžių sekomis, sakinių konstrukcijomis, minčių nuoseklumu ir rišlumu. Nuo pat vaikystės susiformavę unikalūs kalbėjimo stiliai žmogui suteikia specifiškumo, nusakančio jo asmenybę. Apibūdinant asmens kalbos stiliaus modeliavimą ir analizę, ne mažiau svarbu yra paminėti ir autorystės identifikavimą, kuris yra laikomas neatsiejama modeliavimo ir analizės dalimi. Norint identifikuoti asmenį su konkrečiu kalbėjimo stiliumi palyginimo būdu, autorių tenka nustatyti iš kandidatų rinkinio. Turint pakankamai asmens kalbos informacijos (kalbos tekstų) galima nagrinėti asmens kalbos profilio uždavinį – asmens kalbos analizę ir atpažinimą (identifikavimą).

Kiekvieną sekundę pasaulyje sukuriama daugybė teksto fragmentų ir dokumentų, kurie nesuteikia tiek daug vertės, jei nėra atskiriamas autorius. Šioje vietoje asmens kalbos modeliavimas gali būti suvokiamas ir kaip unikalių kalbėjimo stilių suformavimas ir atskyrimas. Taip pat galima nustatyti kalbos modelio atskyrimą nuo kitų asmenų modelių, išsiskiriančių savitu sakinių konstravimu ar minčių logika ir nustatyti asmens kalbos identifikavimą kalbos sraute. Asmens kalbos stiliaus modeliavimas ir analizė yra plačiai panaudojami ir gali būti sietini su teismų kriminalistika, literatūros analize ir kitomis sritims, kurioms būtų ir yra naudingas autoriaus profiliavimas. Pavyzdžiui, aptariant taikymų sritį teismų kriminalistikos perspektyvoje, galima sakyti, kad tiriamo teksto, kuris nėra priskirtas ir identifikuotas, asmens kalbos profilio sudarymas ir nustatymas palengvintų visą teismų kriminalistikos darbą. O žvalgyboje, kuriai taip pat svarbi kriminalistikos praktika, autorių profiliavimas palengvintų informacijos apie galimą įtariamąjį gavimą. Be to, autorių profilio pagal jo kalbos stilių nustatymo galimybė aktuali ir įmonių rinkodaros specialistams, nes toks gebėjimas padėtų jiems gauti daugiau informacijos apie anoniminius vartotojus. Darbo tikslas – išnagrinėti ir ištirti teksto autorystės identifikavimo metrikas ir metodus, įvertinti jų efektyvumą analizuojant lietuvių kalbos tekstus. Viena iš pagrindinių autorystės modelio sudarymo – identifikavimo problemų, yra išgauti svarbiausius požymius, atspindinčius autoriaus savitą rašymo stilių: jo žodyną, logines žodžių sekas ar kalbos rišlumą.

Darbo uždaviniai

1. Apžvelgti ir išnagrinėti teksto autorystės nustatymo uždavinį.
2. Išanalizuoti ir palyginti dažniausiai naudojamas teksto autorystės identifikavimo metrikas ir metodus.

3. Suformuluoti ir eksperimentiškai įvertinti sprendimus lietuvių kalbos teksto autorystei identifikuoti.

Naudoti tyrimo ir analizės metodai

Šiame darbe naudota analitinė literatūros apžvalga, dirbtinio intelekto metodų analizė, duomenų analizė, kompiuteriniai natūralios kalbos apdorojimo metodai, duomenų gavybos metodai, dirbtinio intelekto metodai.

Darbo naujumas ir praktinė nauda

Atlikto darbo tema yra aktuali – dirbtinis intelektas su savo perspektyvomis vis stipriau veržiasi į priekį ir sugeba atlikti pačius įvairiausius uždavinius. Vis dažniau yra atliekami tyrimai, susiję su autorystės identifikavimu, tačiau neretu atveju tai atliekama ne su tekstine kalba. Susidomėjimas autorių nustatymu iš tekstų ateityje tik didės, kadangi šis uždavinys turi daug taikymo būdų.

Šis darbas yra atliktas su lietuvių kalbos tekstais pritaikant dirbtinio intelekto metodus. Savybių ir žodžių n-gramų panaudojimas lietuvių kalboje sukuria šio tyrimo vertę ir jo praktinę naudą, nes būtent n-gramų pasitelkimas padėjo pasiekti aukštų rezultatų tyrimo pabaigoje.

Darbo struktūra

Šį darbą sudaro įvadas (darbo uždaviniai, naudoti tyrimo ir analizės metodai, darbo naujumas ir praktinė nauda), pirmasis skyrius, kuriame pateikiama tekstų autorystės identifikavimo tyrimo apžvalga, sudaryta iš autorystės identifikavimo uždavinių, problemų apžvalgos, tekstų autorystės metrikų, tekstų autorystės identifikavimo metodų apžvalgos. Po to pristatoma dirbtinio intelekto metodų analizė teksto autorystei identifikuoti. Šį skyrių sudaro dirbtinių neuronų tikslų apžvalga ir apžvelgtų tinklų panaudojimas teksto autorytei nustatyti. Eksperimentinę darbo dalį sudaro lietuvių kalbos tekstų autorystės identifikavimo tyrimo skyrius, kuriame yra pateikiami ir išanalizuojami eksperimento duomenys, lietuvių kalbos tekstų identifikavimo metodika bei pateikiamas eksperimentinio tyrimo detalus planas. Toliau seka eksperimentinio tyrimo rezultatai. Šį skyrių sudaro 2 pagrindinės dalys: žodžių n-gramų modelių ir simbolių n-gramų modelių tyrimai: pradinio tikslumo vertinimai, savybių aibės dydžio tyrimai, savybių intervalo tyrimai bei gautųjų rezultatų įvertinimas. Paskutinieji darbo skyriai – išvados ir rekomendacijos bei naudotos literatūros sąrašas, darbo priedai.

1. TEKSTŲ AUTORYSTĖS IDENTIFIKAVIMAS

Atsižvelgiant į šių dienų mokslo pasaulio greitį, dirbtinis intelektas yra vienas sparčiausiai besivystančių mokslo sričių, vis didesniu pagreičiu įsiliejanti į dabartinį gyvenimą. Naudojant dirbtinio intelekto metodus galima automatizuoti daug monotoniško rankų darbo reikalaujančių užduočių, kaip pavyzdžiui, autorių nustatymą iš tekstų. Šiame skyriuje pristatoma analitinė publikacijų apžvalga, skirta identifikuoti pagrindinius tekstų autorystės modeliavimo ir identifikavimo uždavinius. Be to, aptariamos problemos, įvairios metrikos ir metodai, su kuriais susiduria tyrėjai, spręsdami panašius uždavinius.

1.1. Autorystės identifikavimo uždavinys

Žemiau pateikiami pagrindiniai argumentai, liudijantys tekstų autorystės atpažinimo uždavinio realistiškumą ir aktualumą.

a) Kitų autorių tyrimuose yra bandomi spręsti labai panašūs uždaviniai. Viename tokių, yra sakoma, kad autorių profilį sudaro daugybė elementų, o kai kurie iš jų yra stipriai susiję su pačiu autoriumi: asmens lytis ir amžius. Atliktame tyrime buvo sudaromi internetinių tekstinių laikmenų autorių profiliai, kurie profilyje nustato autoriaus lytį bei amžių. Buvo pasiektas 83 % tikslumas nuspėjant lytį [1].

b) Kitas darbas puikiai atspindi šio darbo uždavinį – autorystės atpažinimą vertinant teksto kalbą. Darbe teigiama, kad pagrindinis aspektas, norint priskirti autorių ir jį atpažinti, yra kalbos autorių unikalus stilius, kuriam aprašyti ir įvertinti naudojama taip vadinamoji stilometrija. Sprendžiant pagrindinį uždavinį, autoriai pasiekė apie 90 % efektyvumą [26].

c) Ankstesniuose darbuose, kuriuose jau buvo bandoma spręsti panašaus pobūdžio uždaviniai, buvo atskleista, kad autoriaus požymius yra įmanoma suskirstyti bent į 4 grupes: leksines, sintaksines, struktūrines ir turinio ypatybes. Tyrėjai teigia, kad įgyvendinant užduotį geriausias rezultatas buvo pasiektas sujungus šešias funkcijų kategorijas [12].

d) Autorystės identifikavimo uždaviniui išspręsti galima naudoti sintaksines asmens kalbėjimo savybes: sakinių struktūras, sakinių sekas, žodžių eiliškumą, kurios atneša didesnį potencialą autoriaus atpažinimui. Atliktame tyrime pateikiamas skirtingas – (nuo 64 iki 88 %) tikslumas, naudojant skirtingas kalbos savybes [14].

e) Tokią pačią užduotį išsprendė ir kiti tyrėjai. Viename tyrime pažymima, kad atliekant darbą buvo konstruojamos teksto savybės, po to jos klasifikuojamos. Eksperimente buvo pasiektas net 95 % tikslumas [17].

f) Tyrėjai, naudodami giliojo mokymosi bei tradicinius mašininio mokymosi metodus, savo darbe sprendžia lietuvių autorių profiliavimo uždavinį, nustatydami autoriaus profilio du

aspektus: metus ir lytį. Be to, autoriai lygina ir tiria mokymo duomenų rinkinio dydžio įtaką autoriaus profiliavimo tikslumui. Geriausi rezultatai buvo pasiekti naudojant didžiausius duomenų rinkinius su 5000 egzempliorių kiekvienoje klasėje. Pasiektas nuo 17 % iki 45 % tikslumas autoriaus metų atpažinime bei nuo 47 % iki 75 % tikslumas nustatant lytį. Šiame tyrime teigiama, kad gilusis mokymasis nėra geriausias autorių profiliavimo užduoties sprendimas [4].

g) Aiškinantis, kaip galima spręsti autoriaus asmenybės tipo nustatymą iš teksto, paaiškėjo, kad tai galima atlikti nustatant penkių psichologinių profilio savybių buvimą arba nebuvimą tekste. Kiekvienam iš penkių bruožų yra parengiamas atskiras dvejetainis klasifikatorius, identiška architektūra, pagrįsta nauja dokumentų modeliavimo technika. Klasifikatorius yra paremtas giliojo mokymusi metodu. Pirmieji tinklo sluoksniai kiekvieną teksto sakinį traktuoja atskirai, tada sakiniai sujungiami į dokumento vektorių. Filtruojant emociškai neutralius įvesties sakinius yra pagerinamas našumas. Šio tyrimo autoriams pavyko pasiekti 50 % – 63 % tikslumą atpažįstant autoriaus asmenybės tipą iš teksto [7].

h) Sprendžiant uždavinį, galima suformuluoti klausimą, ar tik labai specializuoti ir gerai įrengti subjektai galėtų identifikuoti autorių pasinaudodami tekstu. Siekiant atsakyti į šį klausimą, yra siūlomas labai paprastas ir skaičiavimo resursų nereikalaujantis autorių identifikavimo metodas, pagrįstas glaudinimo be nuostolių algoritmu. Teigiama, jog naudojant paprastą metodą galima sėkmingai identifikuoti autorius iš tekstų [11].

i) Tyrėjai taip pat sprendžia kiek kitokį autoriaus analizės uždavinį – atskleisti paslėptas autorių savybes iš tekstinių duomenų. Remiantis rašymo stiliais yra nustatoma autoriaus tapatybė ir sociolingvistinės savybės. Tyrimo autoriai, norėdami imituoti žmogaus sakinių komponavimo procesą, naudojasi neuronų tinklo pagalba. Visų pirma, siūlomi modeliai išgauna kiekvieno dokumento aktualius leksinius, sintaksinius ir simbolių lygio vektorius kaip stilistiką, kuri yra pritaikoma autoriaus analizės uždaviniui spręsti naudojant „Twitter“, tinklaraščio, apžvalgų, romanų ir esė duomenų rinkinius. Ši analizė buvo pritaikyta skirtingoms kalboms, tikslumas siekė 88–99 % [13].

j) Uždavinio sprendimui yra siūlomas autorių identifikavimo metodas, pagrįstas automatine funkcijų inžinerija, naudojant žodžių įterpimus, atsižvelgiant į kiekvieno autoriaus rašymo stilių. Siūloma sistema apima du mokymosi etapus. Pirmuoju etapu siekiama sugeneruoti kiekvieno autoriaus semantinį vaizdą, kad modelį būtų galima mokyti ir būtų išgautos svarbiausios neapdoroto dokumento charakteristikos. Antrasis etapas yra klasifikatoriaus pritaikymas, kad būtų nustatytos klasifikavimo taisyklės. Atlikus eksperimentą teigiama, kad pavyko pasiekti 95,83 % tikslumą autoriaus tapatybės nustatyme iš tekstų [16].

k) Pagrindinis šio tyrimo uždavinys yra apibrėžti tekstinės kalbos struktūrą, pagal kurią būtų galima nustatyti autoriaus amžių ir lytį. Atlikti tyrimai, susiję su autoriaus lytimi ir amžiumi pagal rašytinį tekstą daugeliu kalbų. Tyrime pasiūlytas modelis, kuris nustato pagrindinius autoriaus parametrus. Šie parametrai matuojami lyčiai ir trims amžiaus kategorijoms [18].

l) Panašus uždavinys yra bandomas spręsti kitų tyrėjų atliktame darbe, kuriame jis išdėstomas kaip uždavinys priskirti autorystę. Svarbu paminėti ir tai, kad autorių darbe buvo naudojamas būdas, patenkinantis naujų autorystės priskyrimo užduočių poreikiams tenkinti. Autoriai sako, jog patikrinę tekstą, kuriam reikia priskirti autorių duomenų rinkinyje nustato, kad tekstų, kurių pradinis ilgis yra maždaug 200 žodžių, skaičius sudaro daugiau nei 70 procentų viso teksto. Tai reiškia, kad geriausius rezultatus autorystės priskyrime galima pasiekti tada, kai teksto dydis neviršija 200 žodžių ribos [22].

m) Natūralios kalbos apdorojimas (NLP), be kurio nebūtų įsivaizduojamas ir autorystės priskyrimas, yra sprendžiamas pasitelkiant sintaksinius n-gramų modelius, dar vadinamus kaip sn-gramos, kurie yra taikomi autorystės priskyrimui. Pradiniams duomenims naudoti tradiciniai žodžių n-gramų modeliai, kalbos žymės ir simboliai. Naudojant sn-gramų modelius galima pasiekti gerų rezultatų, o kai kuriais atvejais efektyvumas siekia net 100 %, tačiau autorystė priskiriama tik mažam skaičiui autorių [20].

n) Dar vieni autoriai sprendė artimą nagrinėjamai temai uždavinį – autorystės patvirtinimą, arba kitaip tariant, autorystės verifikavimą. Darbe apibūdinama tai, kad tyrėjams pavyko išspręsti pateiktą uždavinį, nes jie sužinojo geriausius žodžių vaizdus automatiniam autorystės patvirtinimui. Rezultatai atsispindi, kad atliekant skirtingas autorystės patvirtinimo funkcijas buvo gautas 88–98 % efektyvumas [24].

o) Yra sakoma, kad tradicinis autorystės priskyrimas apibūdina uždarojo scenarijaus klasifikavimo užduotį. Atsižvelgiant į ribotą autorių kandidatų rinkinį ir atitinkamai pažymėtus tekstus, tikslas – nustatyti, kuris iš autorių parašė anoniminių ar ginčytinų tekstų rinkinį. Tam pasiekti yra atlikta tikimybinė automatinio kodavimo sistema, skirta šiai prižiūrimai klasifikavimo užduočiai spręsti. Variaciniai autoenkoderiai (VAE) yra sėkmingi mokantis latentinių vaizdų, tačiau esamus VAE vis dar riboja apribojimai, kuriuos lemia numanomas pagrindinio tikimybių pasiskirstymo latentinėje erdvėje Gausas. Šioje vietoje yra pratęsiamas VAE su įterptu Gauso mišinio modeliu iki studento-*testo* mišinio modelio, kuris leidžia nepriklausomai kontroliuoti numanomų tikimybių tankių atitinkamų svorių „sunkumą“. Eksperimentai per „Amazon“ peržiūros duomenų rinkinį rodo puikų siūlomo metodo našumą [25].

Sprendžiant pagrindinį darbo uždavinį – asmens kalbos stiliaus atpažinimą – tenka susidurti su įvairiomis problemomis.

a) Kiekvienas asmuo turi savitą kalbėjimo stilių, kalbą ir tematikos lemiamą žodyną. Būtent dėl to yra susiduriama su problemomis norint atpažinti konkrečios kalbos autorių. Tyrėjų problema buvo išspręsta pritaikant skirtingus stilius kaip savybes, pagal kurias ir buvo bandoma identifikuoti teksto autorius [14].
b) Taip pat galima susidurti ir su žymėtų duomenų, kuriuos naudoja mašininiai mokymosi modeliai, trūkumu. Autoriai šią problemą bandė spręsti pasinaudodami įvairių iššūkių konkursų, kitais viešai prieinamais duomenų rinkiniais [1].
c) Nepaisant to, galima numanyti, kad asmuo, rašantis skirtingus paskirties tekstus (pvz., dalykinį, asmeninį), gali vartoti visiškai skirtingą rašymo stilių, kitaip dėlioti mintis. Šis bruožas neleidžia surinkti tinkamų ir panašių duomenų apie asmenį ir jo savitas rašymo savybes, nes kintant tekstų stiliui bei temai, skiriasi ir asmens rašymo stilius. Su šita problema susidūrė ir kiti tyrėjai, ją bandė spręsti derindami skirtingus metodus ir leisdami neuronų tinklams rasti tam tikrus kiekvieno vartotojo rašymo stiliaus modelius [12].
d) Be to, neigiamą įtaką gali turėti teksto apdorojimo būklė: redaguotas jis, ar ne. Redaguotame straipsnyje neišvengiamai atsispindės redaktoriaus mąstymo ir kalbos stilius, kuris didina identifikavimo klaidos tikimybę.
e) Gilinantis į mokytų modelių sąveiką su galimais uždaviniais, galima teigti, kad apmokyti modeliai aprašo tam tikrus (modelių apmokymo) duomenis, kurie gali stipriai pasikeisti, kai modelis yra pritaikomas naujai užduočiai atlikti.
f) Didelis teksto fragmentų skaičius gali kelti problemų. Pavyzdžiui, kai kurie fragmentai yra per trumpi, todėl pasirinktasis identifikavimo metodas negali gauti pakankamai informacijos, o viso modelio teksto vaizdavimas sunkiai atspindi semantinę informaciją [22].

1.2. Tekstų autorystės metrikos

Nagrinėjant dažniausiai naudojamas autorių identifikavimo metrikas, buvo atlikta panašių tyrimų analizė. Kaip stiliaus požymiai (metrikos) naudojami lingvistiniai požymiai, simbolių lygmens požymiai (pvz., simbolių n-gramos, simbolių pasirodymo dažniai), žodžių lygmens požymiai (pvz., žodžių n-gramos, santykiniai žodynų dydžiai, žodžių pasirodymo dažniai, *hapax legomena* – vienąkart tekste pasirodančių žodžių skaičius, *hapax dislegomena* – dukart tekste pasirodančių žodžių skaičius), sintaksės požymiai (pvz., skyrybos ženklų kiekis, klausiamųjų sakinių kiekis), kalbos struktūros požymiai (pvz., vidutinis sakinio ilgis žodžiais, simboliais ir/arba tarpais, vidutinis sakinių

skaičius viename pasisakyme, ilgiausias trumpiausias sakinio ilgis), taip pat funkciniais žodžiais grįsti požymiai (pvz., jungtuko „ir“ kiekis tekste).

Buvo naudojamos ir tokios tekstų autorystės metrikos kaip: „word2wec“, „žodžių maišas“, „žodžių įterpiniai“, bei paprastos teksto savybės kaip: žodžių ilgis, sakinių ilgis, skirtingų žodžių tekste santykis [13, 16, 20, 25]. Išanalizuotos įvairios publikacijos, kuriose, kaip paaiškėjo, buvo pasitelkiamos n-gramų modelio metrikos. Šie minimi požymiai buvo pasirinkti atsižvelgiant į tyrėjų gautus rezultatus. Autoriai teigia, kad uždaviniui spręsti tinkamiausios metrikos yra išskiriamos būtent n-gramos modelis dėl to, nes jos identifikuojamos kaip efektyviausias būdas nustatyti tam tikrą autorių iš daugelio kitų autorių tekstų. N-gramos modelis yra apibūdinama kaip metrika, kuri yra skirta tekstų atvaizdavimui stilistiniais tikslais, nes ji turi gebėjimą užfiksuoti ne tik leksinio ir sintaksinio, bet ir struktūrinio lygmens niuansus. Apibūdinant n-gramų metrikas, svarbu paminėti, jog jas gali sudaryti: žodžiai, įvairūs simboliai, jų intervalas. Nustatant n-gramų intervalą yra nustatomas jo ilgis, kuris reiškia, kokiu dažniu minėti požymiai bus sugeneruojami [3, 15, 17, 20].

1.3. Tekstų autorystės identifikavimo metodų apžvalga

Nustatant konkretų autorių iš daugelio tekstų yra naudojami dirbtinio intelekto metodai. Jie yra skirstomi į neuronų tinklus, kurie, pasižymi puikiais rezultatais sprendžiant įvairaus tipo uždavinius, tačiau geriausiai jų efektyvumas atsispindi būtent natūralios kalbos apdorojimo užduotyse [29]. Tokiems tinklams priskiriami: daugiasluoksnis perceptronas (MLP), konvoliuciniai neuronų tinklai (CNN), rekurentiniai neuronų tinklai (RNN), ilgos – trumpalaikės atminties metodas (LSTM) ir autoenkoderiai. Kiti dirbtinio intelekto tinklai, kurie neretai naudojami kaip lyginamieji metodai tekstų autorystės nustatymui, yra tradiciniai mašininio mokymosi tinklai. Šiems tinklams yra priskiriama: atraminių vektorių klasifikatorius (SVM), k-artimiausių kaimynų algoritmas (KNN) ir Bajeso tinklas (Bayes). Platesnė minimų metodų analizė bus pateikta sekančiame skyriuje.

1.4. Trumpas skyriaus apibendrinimas

Naudojantis įvairiose publikacijose įvardijamais uždaviniais, kurie yra susiję su autorių identifikavimu iš tekstų, galima sakyti, kad toks pats autorių nustatymas iš lietuviškų tekstų yra įmanomas įgyvendinti. Remiantis problemomis, su kuriomis susidūrė panašius uždavinius sprendę tyrėjai, yra suformuluotos problemos, su kuriomis galima susidurti atliekant eksperimentinį šio darbo tyrimą. Po pirminės panašių darbų analizės, galima teigti, kad autoriaus nustatymas ar modeliavimas yra įmanomas pasitelkus išmaniuosius sprendimus, pagrįstus mašininio mokymosi ir dirbtinio intelekto metodais.

2. DIRBTINIO INTELEKTO METODŲ ANALIZĖ TEKSTO AUTORYSTEI IDENTIFIKUOTI

Apžvelgus dalį atliktų tyrimų, galima teigti, kad naudojami metodai gali būti pritaikyti asmens kalbos modeliavimo ir identifikavimo uždaviniams. Po metodų apžvalgos nustatyti tinkamiausi metodai, nes juos naudojant buvo pasiekti geriausi autorystės modeliavimo ir nustatymo iš tekstų rezultatai.

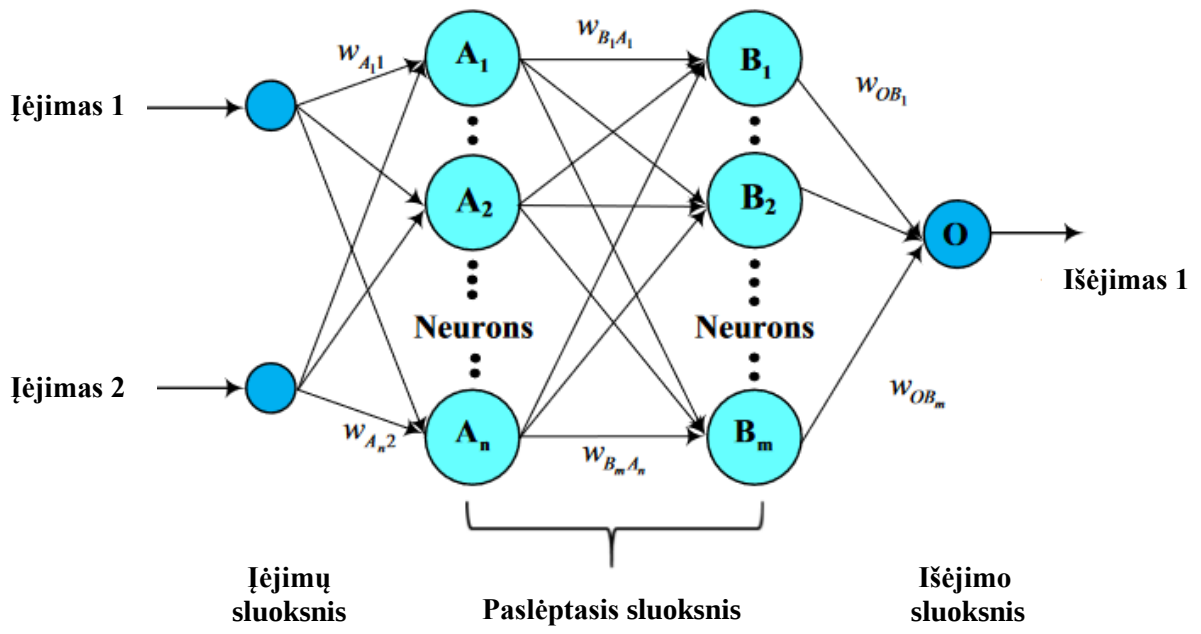
2.1. Dirbtinių neuronų tinklai

Dirbtinių neuronų tinklai (DNN) yra sukurti žmogaus nervų sistemos sudedamosios dalies – neurono – dirbtinės versijos pagrindu. Tokie tinklai geba apdoroti ir analizuoti didelės apimties ir sudėtinius (kompleksinius) duomenis apie fizinį reiškinį ar sprendimo procesą. Išskirtinė DNN savybė yra jų gebėjimas įvertinti ryšius tarp nepriklausomų ir priklausomų duomenų požymių ir iš reprezentatyvių duomenų rinkinių išgauti nematomą informaciją, formuoti sudėtingas priklausomybes. Ryšiai tarp nepriklausomų ir priklausomų kintamųjų gali būti nustatyti be jokių prielaidų apie reiškinių fizikinę prigimtį, jų matematinį vaizdavimą. DNN modeliai suteikia pranašumų, lyginant su regresija pagrįstais modeliais, įskaitant jos gebėjimą tvarkyti triukšmingus duomenis [27].

Giliuoju mokymu grįsti algoritmai yra mašininio mokymosi metodų šeimos dalis, pagrįsta didelės apimties dirbtiniais neuronų tinklais. Gilusis mokymasis leidžia skaičiavimo modeliams, sudarytiems iš keleto, neretai itin didelės apimties apdorojimo sluoksnių, vaizduoti nagrinėjamus duomenis skirtingos abstrakcijos lygiais. Giliojo mokymo metodai žymiai pagerina pasiekiamus rezultatus kalbos atpažinimo, vizualių objektų atpažinimo, objektų aptikimo ir daugelyje kitų sričių, pavyzdžiui, vaistų sukūrime ar genomikoje. Giliuoju mokymu grįsti algoritmai atranda sudėtingą struktūrą dideliuose duomenų rinkiniuose, naudodami atbulinio sklidimo algoritmą [28].

2.1.1. Daugiasluoksnis perceptronas

MLP architektūra dėl savo paprastumo yra dažniausiai naudojamas neuronų tinklas. Tai grįžtamojo ryšio neuronų tinklo struktūra. Perceptronas tinkle priima įvestį, atlieka skaičiavimus ir sukuria išvestį. Išvestis tikrinama naudojant įvestį. Svoriai generuojami atsitiktinai ir pridedami prie neuronų skaičiavimo metu, kol skirtumas tarp įvesties ir išvesties yra ribinis arba nulinis. Kai perceptronų tinklas turi bent 1 paslėptą sluoksnį, jis yra vadinamas daugiasluoksniu perceptronu. Toliau pateikiama daugiasluoksnio perceptrono tinklo struktūra ir šio tinklo pritaikomumas autoriaus atpažinimo uždaviniui [30]. Toliau pateikiama schema, apibūdinanti daugiasluoksnio perceptrono struktūrą (žiūrėti 2.1 pav.).



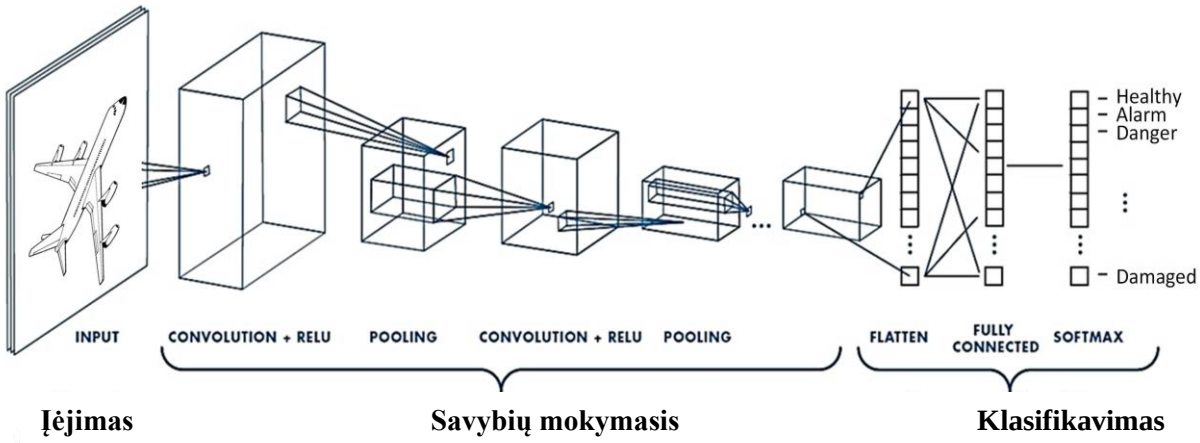
2.1 pav. Daugiasluoksio perceptrono tinklo struktūra [30]

Schemoje pateikta MLP struktūra, sudaryta iš įvesties sluoksnio, į kurį pateikiami duomenys. Šie duomenys vaizduoja uždavinio kintamuosius. Tinklas taip pat sudarytas iš vieno ar daugiau paslėptų sluoksnių, kuriuose yra keli ar daugiau neuronų. Šie paslėptų sluoksnių neuronai gauna duomenis iš įvesties sluoksnio ir siunčia juos į kitą sluoksnį. Paskutinis sluoksnis – išvesties, kuris su neuronais vaizduoja modeliuojamus priklausomus kintamuosius.

2.1.2. Konvoliuciniai neuronų tinklai

CNN tinklai suteikia galimybes spręsti tokioms užduotims kaip: vaizdo klasifikavimas ir atpažinimas, natūralios kalbos apdorojimas ir kt. Dauguma konvoliucinių neuronų tinklų labiausiai yra skirti dirbti su vaizdo duomenimis, tačiau šiuos tinklus galima naudoti ir su visų kitų tipų duomenimis. Paprasčiausias tokio tinklo duomenų pavyzdys yra dvimatis vaizdas. Šis vaizdas yra paimamas ir jam priskiriama svarba įvairiems vaizdo aspektams ar objektams. Tokie tinklai gali sėkmingai užfiksuoti erdvines ir laiko priklausomybes vaizde, taikydamas atitinkamus filtrus. Svarbiausia konvoliucinių neuronų tinklų charakteristika yra konvoliucijos operacija. Dėl šios priežasties konvoliuciniai neuronų tinklai apibrėžiami kaip tinklai, kurie naudoja konvoliucinę operaciją bent viename sluoksnyje, nors dauguma konvoliucinių neuronų tinklų naudoja šią operaciją keliuose sluoksniuose. Konvoliuciniuose neuronų tinkluose būsenos kiekviename sluoksnyje yra išdėstytos pagal erdvinio tinklo struktūrą. Šie erdviniai ryšiai yra paveldimi iš vieno sluoksnio į kitą, nes kiekviena ypatybės reikšmė yra pagrįsta mažu vietiniu erdviniu regionu ankstesniame sluoksnyje. Svarbu išlaikyti šiuos erdvinius ryšius tarp tinklo ląstelių, nes konvoliucijos operacija ir

transformacija į kitą sluoksnį labai priklauso nuo šių ryšių. Konvoliucinio tinklo schema pateikta 2.2 pav.

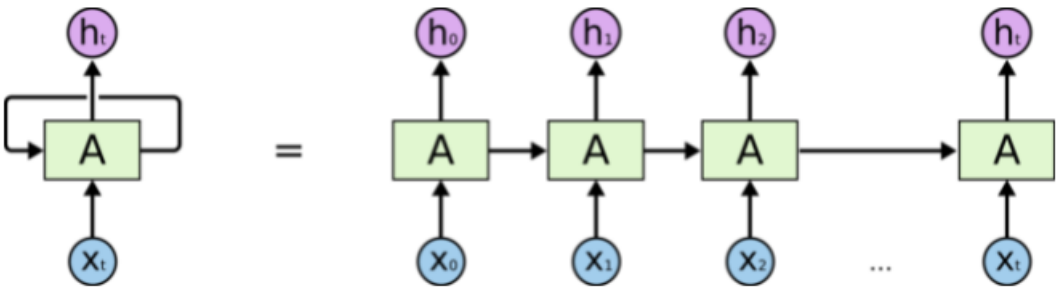


2.2 pav. CNN tinklo pavyzdys [32]

Kiekvienas konvoliucinio tinklo sluoksnis yra 3 matmenų tinklo struktūra, kurią sudaro aukštis, plotis ir gylis. Konvoliucinis neuronų tinklas yra sudarytas iš kelių sluoksnių, vieni tokių: konvoliucija, jungiamasis–telkiamasis (angl. pooling), aktyvavimo funkcija ReLU ir pilno sujungimo sluoksnis [33].

2.1.3. Rekurentiniai neuronų tinklai

Žmonės, kaip visiškai atskiri individai, turi savitą ir unikalią mąstymą, logiką, kuri neretai pasąmoningai susieja žmogaus gyvenimo įvykius. Naudojantis šia logika galima paaiškinti ir žmogaus supratimą apie atskirus žodžius. Pavyzdžiui, žmogus, skaitydamas žodžius, interpretuoja pagal anksčiau žinomus jų atitikmenis, kurie yra siejami su ankstesnėmis situacijomis gyvenime. Kalbant apie neuronų tinklus – tradicinį neuronų tinklą, jis negali nuspėti iš ankstesnių interpretacijų ar samprotavimų, kas buvo manoma, t.y. toks tinklas nenaudoja praeities duomenų. Rekurentiniai neuronų tinklai išsprendžia šią problemą. Šiuos tinklus galima įvardinti kaip tokius, kurie turi grįžtamojo ryšio elementus, leidžiančius išlaikyti praeities informaciją.

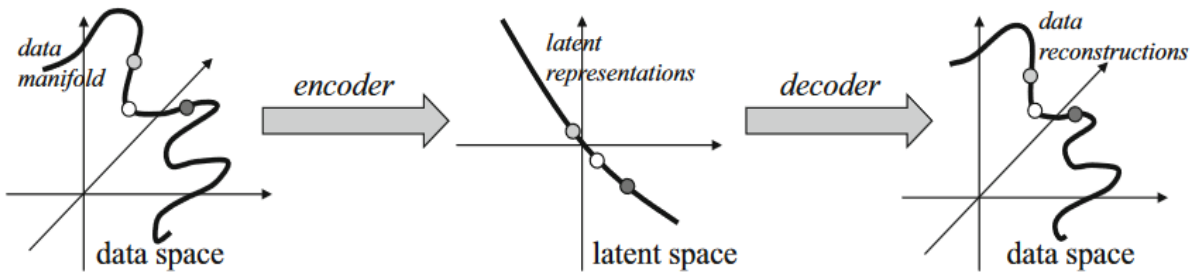


2.3 pav. Išskleisto rekurentinio tinklo pavyzdys [34]

2.3 pav. pateiktame paveiksle matoma neuronų tinklo A dalis, į kurios struktūrą įvedama tam tikra įvestis x_t ir išvedama reikšmė h_t. Ciklas leidžia perduoti informaciją iš vieno sluoksnio į kitą. Rekurentinis neuronų tinklas gali būti laikomas keliomis to paties tinklo kopijomis, o kiekviena iš jų perduoda pranešimą vėlesniam. Šis grįžtamojo ryšio pobūdis atskleidžia, kad rekurentiniai neuronų tinklai yra glaudžiai susiję su apdorotomis kintamo ilgio įvesties sekomis. Taip pat jie yra natūrali neuronų tinklo architektūra, naudojama tokiems duomenims. Norint pasiekti rezultatą, būtinas LSTM tinklo naudojimas. Jį galima nusakyti kaip rekurentinių neuronų tinklo tipą, kuris daugeliui užduočių veikia daug geriau nei standartinė versija. Su jais pasiekiami beveik visi geri rezultatai, pagrįsti rekurentiniais neuronų tinklais [34].

2.1.4. Autoenkoderiai

Autoenkoderiai yra giliųjų dirbtinių tinklų architektūra, susidedanti iš dviejų dalių: kodavimo tinklo, kuris susieja kiekvieną įvesties duomenų tašką su tašku kodinėje erdvėje ir dekoderio tinklo, kuris latentinės erdvės taškus atvaizduoja atgal į duomenų erdvę. Abu tinklai yra mokomi kartu, be priežiūros, kad jų sudėtis maždaug išsaugotų taškus iš tam tikro mokymo duomenų rinkinio. Automatinio kodavimo mokymosi tikslas yra surasti tam tikrą latentinį mokymo duomenų rinkinio taškų atvaizdą, kuris išsaugotų duomenų taškuose esančią informaciją, tuo pačiu supaprastindamas tuos duomenų taškus. Automatinio kodavimo atveju duomenų taškų atvaizdavimas iki latentinių jų atvaizdų yra suskirstomas pasitelkiant giliojo perdavimo (kodavimo) tinklą. Mokymosi metu taip pat siekiama atkurti apytikslį atvirkštinį kodavimo įrenginio atvaizdavimą, kuris yra apibrėžiamas kaip gilusis tinklas (dekoderis). Kodavimo ir dekoderio mokymas lygiagrečiai užtikrina, kad aptiktas latentinis duomenų atvaizdavimas išsaugo didžiąją dalį duomenyse esančios informacijos. Norint atskleisti latentinę (paslėptą) duomenų rinkinio (arba pagrindinio paskirstymo) struktūrą, kodavimo įrenginio ir (arba) dekoderio architektūrai paprastai taikomi tam tikri apribojimai. Latentiniai duomenų taškų atvaizdai gali būti naudojami kaip įvairių mašininio mokymosi užduočių savybės.



2.4 pav. Autoenkoderio tinklo diagrama [35]

Autoenkoderio architektūra 2.4 pav. apima kodavimo įrenginį, susiejantį duomenis su latentine erdve, ir dekoderio, kuris yra apmokytas lygiagrečiai su dekoderiu, kad apytiksliai apverstų juos

duomenų taškuose. Mokymosi tikslas – gauti latentinį duomenų atvaizdą, kuris tam tikru būdu būtų paprastesnis arba labiau tinkamas tolesniam apdorojimui nei pirminis duomenų atvaizdavimas [35].

2.2. Giliojo mokymo tinklų taikymas

Po pateiktų giliojo mokymosi tinklų apibūdinimų, toliau yra kalbama apie jų taikymą. Kitaip tariant, apžvelgiami prie giliojo mokymo priskiriami metodai.

2.2.1. Daugiasluoksnių perceptronų panaudojimas

Iš teksto atpažinti jo autorių yra gana nelengva. Tačiau būtent tokią užduotį pasirinko spręsti vieni iš tyrėjų savo atliktame darbe. Jie, panaudodami neuronų tinklus su suderintomis kalbos stiliaus savybėmis, tyrimo metu siekė atpažinti tekstinio darbo autorystę. Kaip pagrindinį autorystės atpažinimo iššūkį tyrėjai įvardijo savybių identifikavimą. Jie surinko 138 savybes, suskirstydami jas į 6 grupes: vidutiniai žodžio ilgiai ir sakinių ilgiai, prielinksnių, jungtukų, jungtinių prieveiksmių, skyrybos ženklų, bei 50 dažniausiai vartojamų žodžių naudojimas. Šios savybės sudarė kiekvieno atskiro autoriaus stilių. Pasinaudojant pateiktomis savybėmis buvo gaunami rezultatai keletu klasifikavimo modelių: daugiasluoksnio perceptrono, atsitiktinio miško algoritmu (RF), SVM, KNN. Gauti rezultatai rodo, kad, lyginant su kitais keturiais gerai žinomais klasifikatoriais, daugiasluoksnio perceptrono (MLP) klasifikatorius leido pasiekti geriausius autorių identifikavimo rezultatus (92–98 %) [12].

Tam, kad nekiltų abejonių dėl daugiasluoksnio perceptono rezultatų sprendžiant autorystės nustatymo iš tekstų uždavinį, galima pateikti pavyzdį, įrodantį jo veiksmingumą. Tyrėjai siekė nustatyti anoniminį tekstų autorių remdamiesi jo rašymo pavyzdžių rinkiniu. Atliktame tyrėjų darbe siūlomas autorių identifikavimo metodas, pagrįstas automatiniu savybių išgavimu, naudojant „word2vec“ žodžių įterpimo funkciją. Ši funkcija atsižvelgia į kiekvieno autoriaus unikalų rašymo stilių. Autoriai darbe sugeneravo kiekvieno autoriaus semantinį vaizdą naudodami „word2vec“. Tuomet taikė daugiasluoksnį perceptrono klasifikatorių klasifikavimo taisyklėms nustatyti. Naudojant MLP klasifikatorių su „word2vec“ modeliu pasiektas net 95,83 % tikslumas. Taigi, naudojant „word2vec“ žodžių įterpimo modelį akivaizdžiai galima pagerinti autorystės identifikavimo rezultatus, lyginant su kitais klasikiniais modeliais, pavyzdžiui, n-gramų intervalai ir „žodžių maišas” (angl. *bag of words*). Galima daryti prielaidą, kad naudojant daugiasluoksnio perceptono klasifikatorių galima apdoroti natūraliąją tekstų kalbą ir nustatyti tekstų autorystę. Be abejo, tokiam uždaviniui atlikti reikalingi ir duomenys, kurie neretai būna pateikti kaip tam tikro autoriaus parašyto teksto pavyzdys [16].

2.2.2. Konvoliucinių neuronų tinklų panaudojimas

Lietuvių tyrėjų atliktame darbe tinklas buvo pritaikytas norint sudaryti lietuvių kalbos autorių profilius iš lietuviškų tekstų pagal jų amžių bei lytį. Rezultatai buvo pasiekti panaudojant CNN bei LSTM metodus, kurie buvo palyginti su tradiciniais mašininio mokymosi metodais NBM bei SVM. Pasitelkus anksčiau minėtas žinias apie CNN metodą, kuris yra priskiriamas konvoliuciniams neuronų tinklams, akivaizdu, kad šis metodas yra orientuotas, pirmiausia, į modelių atpažinimą vaizduose, tačiau autoriai jį pritaikė norėdami atpažinti lietuvių autorius iš naudojamų lietuviškų tekstų. Tai reiškia, kad tyrėjai orientuojasi labiau ne į kalbos stilistiką, o į asmens žodyną ir jo unikalų turtingumą. Abu giliojo mokymosi metodai buvo taikomi lietuvių kalbos žodžių įterpiniams. Žodžių įterpimas yra vienas iš naujesnių natūralios kalbos apdorojimo metodų, leidžiantis mokytis vektorinių žodžių vaizdų iš neapdorotų tekstų, turinčių mažą matmenį, ir užfiksuoti sintaksinius bei semantinius ryšius. Autoriai teigia, kad pats žodynas tampa vienu stipriausių autorystės įrodymų, todėl jie mano, kad žodžių įterpimas turėtų būti tinkamas jų uždavinio sprendimo požymis. Atlikus tyrimą paaiškėjo, jog didžiausią tikslumą, remiantis vidurkiu, pasiekė CNN beveik 32 % nustatant amžių bei beveik 61 % lyties nustatyme (tyrime buvo nagrinėjama 147 autorių tekstai). Geriausi tyrimo rezultatai buvo pasiekti naudojant didžiausius duomenų rinkinius, kuriuose yra 5000 egzempliorių kiekvienoje klasėje. Iš atlikto autorių darbo rezultatų yra matoma, kad daugiausiai problemų sukėlė amžiaus nustatymas. Apžvelgdami atlikto tyrimo rezultatus autoriai atkreipė dėmesį į tai, jog giliojo mokymosi metodai nebuvo tinkamiausias sprendimas lietuvių kalbos autorių profilių identifikavimui [4].

Aptariamas metodas buvo panaudotas kitame tyrime, nustatant žmogaus asmenybės tipą iš teksto. Jame nustatyta, kuriame iš 5 asmenybės tipų, tokių kaip: ekstraversija, neurotizmas, geranoriškumas, sąžiningumas, atvirumas, žmogus priklauso. Asmenybės tipo priklausomumas nustatytas iš penkių psichologinių profilio bruožų savybių buvimo arba nebuvimo tekste. Autorių naudotas metodas apima išankstinį įvesties duomenų apdorojimą ir filtravimą, požymių ištraukimą bei klasifikavimą. Darbe naudoti dviejų tipų savybės: dokumento lygio stilistinės ir kiekvieno žodžio semantinės savybės. Jos sujungiamos į kintamo ilgio įvesties teksto atvaizdą. Šis kintamo ilgio vaizdavimas įvedamas į CNN, kur jis apdorojamas hierarchiniu būdu, sujungiant žodžius į n-gramus, n-gramus – į sakinius, sakinius – į visą dokumentą. Taip gautos reikšmės sujungiamos su dokumento lygio stilistinėmis ypatybėmis, kad būtų suformuotas dokumento atvaizdas, naudojamas galutinei klasifikacijai. Tyrime buvo naudojami penki skirtingi CNN tinklai su ta pačia architektūra, taip pat SVM, MLP bei keletas kitų klasifikatorių. Buvo lyginami pasiekti rezultatai, pritaikyti kiekvieno iš asmenybės bruožams nustatyti. Pirmieji tinklo sluoksniai kiekvieną teksto sakinį traktuoja atskirai, tada sakiniai sujungiami į dokumento vektorių. Kiekvienam iš penkių bruožų autoriai parengė atskirą dvejetainį klasifikatorių, kuris nustatė, ar bruožas bus teigiamas, ar neigiamas. Filtruojant emociškai

neutralius įvesties sakinius yra pagerinamas našumas. Tyrėjams su CNN tinklu pavyko pasiekti 52–63 % tikslumą atpažįstant autoriaus asmenybės tipą iš teksto [7].

CNN metodą taip pat pasitelkė kiti autoriai, kurie minėtą metodą naudojo savo darbe kartu su SVM metodu. Naudodami tekstus internete, tyrėjai sudarė žmogaus profilį pagal metus ir amžių. Darbe naudojamas standartinis CNN neuronų tinklas. Uždaviniui spręsti autoriai išskyrė tokius parametrus kaip: žodžių ilgis, specialių simbolių naudojimo dažnumas, funkciniai žodžiai, emocinių jaustukų naudojimas ir žymėti žodžiai. Šie bruožai buvo pritaikomi klasifikuojant PAN14-sm, PAN15-twitter, Google-blogs, PAN14-blogs duomenų rinkinius, kurie buvo sudaryti iš įvairaus amžiaus bei lyties socialinių tinklų įrašų. Rezultatai atskleidžia internetinių įrašų klasifikavimą, kuomet buvo pasiektais 55–83 % tikslumas nustatant lytį. Identifikuojant amžių tyrimų tikslumas siekė 23–44 %. Tyrime teigiama, kad SVM modelis parodė geresnius rezultatus, taip pat kaip ir (Kapočiūtė-Dzikienė ir Damaševičius, 2018) darbe yra matoma, kad nustatyti žmogaus amžių sekėsi daug sunkiau nei žmogaus lytį [1].

Kitame tyrime siūloma nauja mašininio mokymosi modelio schema, kuri pagrįsta trejomis skirtingomis ir sugretintomis architektūromis: CNN, R-CNN ir SVM. Tyrimui atlikti naudojamas PAN 2015 duomenų rinkinys. Jis buvo sukurtas su 100 autorių, įskaitant 100 žinomų dokumentų ir 100 nežinomų dokumentų anglų kalba. Teksto analizei siūloma „word2vec“ analizė, pagrįsta žodžių įterpiniais. Galutinis sistemos sprendimas gaunamas derinant trijų modelių rezultatus ir naudojant balsavimo metodą (BWWV). Šis būdas yra pasitelkiamas trijų modelių rezultatams sujungti ir pasinaudoti šių trijų klasifikatorių papildomumu, kad būtų gautas bendras rezultatas. Rezultatai rodo, kad autoriaus atpažinime iš tekstų buvo pasiektas 88–98 % tikslumas [24].

2.2.3. Rekurentinių neuronų tinklų panaudojimas

Aptariant dirbtinių neuronų tinklų pobūdį apie kalbos taikymą ir jos apdorojimą, galima pastebėti, kad daugelis autorystės nustatymo metodų grindžiami tradiciniais mašininio mokymosi metodais. Nors šie metodai yra veiksmingi, žmogui sunku pasirinkti efektyvius požymius, todėl publikacijos autoriai pasiūlė naują autorystės priskyrimo modelį, naudojant rekurentinius neuronų tinklus (RNN). Jie taip pat naudojo dėmesio mechanizmą, kad užfiksuotų svarbiausią informaciją sekoje. Be viso to, tyrėjai įdiegė naują duomenų apdorojimo metodą, siekdami patenkinti naujų autorystės priskyrimo užduočių poreikius. Šiame darbe teigiama, kad pagrindinė priežastis yra ta, kad teksto fragmentų šaltinis ir ilgis daro įtaką tekstinės informacijos kaupimui. RNN tinklas yra naudojamas tam, kad apdorotų žodžių seką, jis įtraukia konteksto informaciją, o dėmesio mechanizmas yra naudingas norint rasti svarbų žodį, kad būtų galima padidinti šių svarbių žodžių indėlį į teksto atvaizdavimą. RNN problema yra ta, jog šis tinklas negali gerai susitvarkyti su labai

ilgais tekstais. Taigi, tekstams, ilgesniems nei 200 žodžių, RNN veikimas gali būti netinkama išeitis. Galiausiai su RNAN-200 duomenų rinkiniu buvo pasiektas 91–92 % rezultatas [22].

Dar vienas pavyzdys, susijęs su aptariamais rekurentiniais neuronų tinklais, identifikuojant tekstų autorystę, yra pristatomas [23] tyrimas. Autorius norėjo ištirti Atgarsių būsenos (angl. *Reservoir Computing*) tinklu pagrįstas rezervuarų skaičiavimo (ESN) modeliu su papildomais įterpimo sluoksniais, galintį klasifikuoti „Reuters C50“ duomenų rinkinio tekstinius dokumentus pagal autorystę. Ištirtos įvairios pateiktys, tokios kaip: žodžių ir simbolių įterpimas ir giliųjų funkcijų išėmimas. Pastarosiomis pasinaudodamas tyrėjas išskyrė požymius autorystei nustatyti. Atlikti eksperimentai parodė, kad ESN modeliai pasiekia gerus rezultatus ir yra konkurencingi lyginant su tradiciniais modeliais, tokiais kaip atraminių vektorių klasifikatorius (SVM). Darbe nurodoma, kad modelis analizuoja dokumentus kaip duomenų srautus, aptikdamas įvykius ir segmentuodamas tekstą. Geriausius rezultatus pasiekė ESN tinklai, turintys didelį 1500 neuronų kiekį su žodžių vektoriais – 81,5–92,1 % tikslumu [23].

Turinio atskyrimas nuo stiliaus yra pagrindinė autorystės priskyrimo problema – siekiama, kad būtų atstovaujamas nuo temos nepriklausomas asmeninis autorių stilius. Tyrėjai [26], atlikdami savo darbą, siūlo atskirti temą ir stilių pasinaudojant kelių užduočių mokymosi metodą. Jie vadovavosi tuo, kad būtų galima išmokyti atskirų temų ir stilių reprezentacijas. Šiame darbe pristatytas kelių užduočių mokymosi metodas, skirtas temų stiliaus atskyrimui, siekiant priskirti kelių temų autorystę. Temos ir stiliaus atskyrimui autoriai pristatė konkurencinio dėmesio mechanizmą ir atskyrimo–rekonstrukcijos apribojimą. Užduoties atlikimui buvo naudojami RNN, LSTM bei kiti tradiciniai tinklai. Taip pat naudotas dėmesiu grįstas LSTM tam, kad būtų gautos teksto reprezentacijos, priskiriamos dviem užduotims ir tiesiogiai susietos su konkrečiais užduočių išvesties sluoksniais. Rezultatai parodė, kad siūlomas kelių užduočių mokymosi metodas yra perspektyvus, ypač atsižvelgiant į įvairias temas. Autoriai nustatė, kad temų derinimas gali padėti užfiksuoti aktualų turinį, o konkurencinis dėmesys gali būti naudingas norint temą atskirti. Svarbu tai, jog modelis atskiria temą ir stilių tikimybiniu būdu ir nereikalauja žmogaus įsikišimo. Buvo pasiektas 71–91 % tikslumas [26].

2.2.4. Autoenkoderių panaudojimas

Norint dar geriau įvertinti dirbtinių neuronų tinklų efektyvumą ir svarbumą tekstų autorystės modeliavime ir identifikavime, galima pastebėti, kad uždavinys – nustatyti autorių, gali kelti problemų, kadangi autorių reikia atpažinti ne tik individualiai pagal jo išskirtinius bruožus, kalbos, ar rašymo stilių, bet ir atskirti jį iš daugelio kitų grupių ar žmonių. Šia problema susidomėję tyrėjai paskyrė sau nelengvą užduotį – atpažinti duoto teksto autorių iš tam tikros grupės kitų [17]. Tyrimo autoriai pagrindinius požymius pasirinko remdamiesi kitų studijų tyrimais: simbolių n-gramos ir

dažniausiai pasirodantys žodžiai yra vieni iš plačiausiai naudojamų bei veiksmingiausių savybių autorystės nustatymo užduotyje. Taip pat buvo pasitelktas savybių normalizavimas (mastelio keitimas), standartizuojantis savybių diapazoną bei funkcijų pasirinkimo metodus. Jie buvo panaudoti tam, kad būtų pašalintos perteklinės savybės ir išsaugotos tik informatyviausiosios. Tuomet gilusis mokymasis buvo naudojamas dar aukštesnio lygio požymiams išgauti – buvo panaudotas sudėtinis nutriukšminamasis autoenkoderis (SDAE). Klasifikavimui buvo pasitelktas atraminių vektorių mašinų (SVM) klasifikatorius. Autoriai teigia, jog pavyko pasiekti 95,1 % klasifikavimo tikslumą su normalizacija bei 92,4 % tikslumą be normalizacijos [17].

Autoriai [25] atliko tyrimą, kuriame yra praplečiami variaciniai autoenkoderiai (VAE) su įterptu Gauso mišinio modeliu į student-*testo* modelį, kuris leidžia nepriklausomai kontroliuoti numanomų tikimybės tankių „sunkumą“. VAE įrodė savo naudą atliekant užduotis, nes jie apjungia neuronų tinklo mokymosi stipriąsias puses su neapibrėžtumo metrikos, gaunamos iš statistinių modelių, galia. Jie gali išmokti mažo matmens kolektorių, kad apibendrintų svarbiausias duomenų charakteristikas ir pateiktų natūralų, statistinį aiškinimą tiek latentinėje, tiek stebėjimo erdvėje. Siekdami tikslo, autoriai darbe pasiūlė ir įvertino VAE, kuriame yra įmontuotas student-*testo* modelis. Rezultatas gaunamas bandant vienu metu išspręsti abi užduotis: kartu išmokti netiesinį atvaizdavimą, kad pateiktas dominantis duomenų rinkinys būtų transformuotas į (žemesnės dimensijos) poerdvę ir suskirstant latentinius vaizdus į kategorijas. Šiam tikslui pasiekti buvo sukurtas variacinis mokymosi algoritmas ir parodyta jo nauda mokantis latentinių reprezentacijų, panaudojant tiek egzistuojančius duomenis, tiek realaus laiko. Autorių siūlomas modelis suteikė galimybę gauti geresnius rezultatus nei SVM pagrįstas klasifikatorius, taip pat standartinis Gauso VAE [25].

2.3. Trumpas skyriaus apibendrinimas

Išanalizavus dirbtinių neuronų tinklų taikymą tekstų autorystės identifikavimui, galima teigti, jog įvairiuose šaltiniuose pateiktų metodų pritaikymas yra skirtingas. Geriausius rezultatus parodė MLP, LSTM bei CNN metodai, taip pat palyginamieji tinklai, tokie kaip: SVM ir KNN. Juos būtų galima pritaikyti asmens kalboje, nustatant jos stilių, kurį po to būtų galima analizuoti ar modeliuoti.

3. LIETUVIŲ KALBOS TEKSTŲ AUTORYSTĖS IDENTIFIKAVIMO TYRIMAS

Išanalizavus dirbtinius neuronų tinklus, jų veikimo principus, atlikus dirbtinio intelekto metodų analizę ir jų taikymo perspektyvas, buvo išsirinkti metodai, kurie gali būti naudingi ir tinkami atliekant lietuvių kalbos tekstų autorystės identifikavimo tyrimą. Vienas iš potencialių metodų, kuris prisideda prie autorystės nustatymo lietuvių kalbos tekstuose – dirbtiniai neuronų tinklai. Toliau bus kalbama apie visą nuoseklią tyrimo eigą: pateikiami eksperimento duomenys, lietuvių kalbos tekstų identifikavimo metodika ir atliekamo tyrimo planas.

3.1. Eksperimento duomenys

Norint atlikti tyrimo, kuris orientuotas į rašytinę kalbą, šiuo atveju populiarumo atžvilgiu mažai vartojamą kalbą – lietuvių, svarbu tikslingai pasirinkti tyrimo medžiagą, su kuria bus dirbama. Šiam darbui pasitelkti diskusinio pobūdžio tekstai – parlamentarų kalbos Seime. Pasisakymai vyko 1990 m. kovo mėn. – 2013 m. gruodžio mėn. (žiūrėti 3.1 lentelė).

3.1 lentelė. Informacija apie eksperimento duomenis [36]

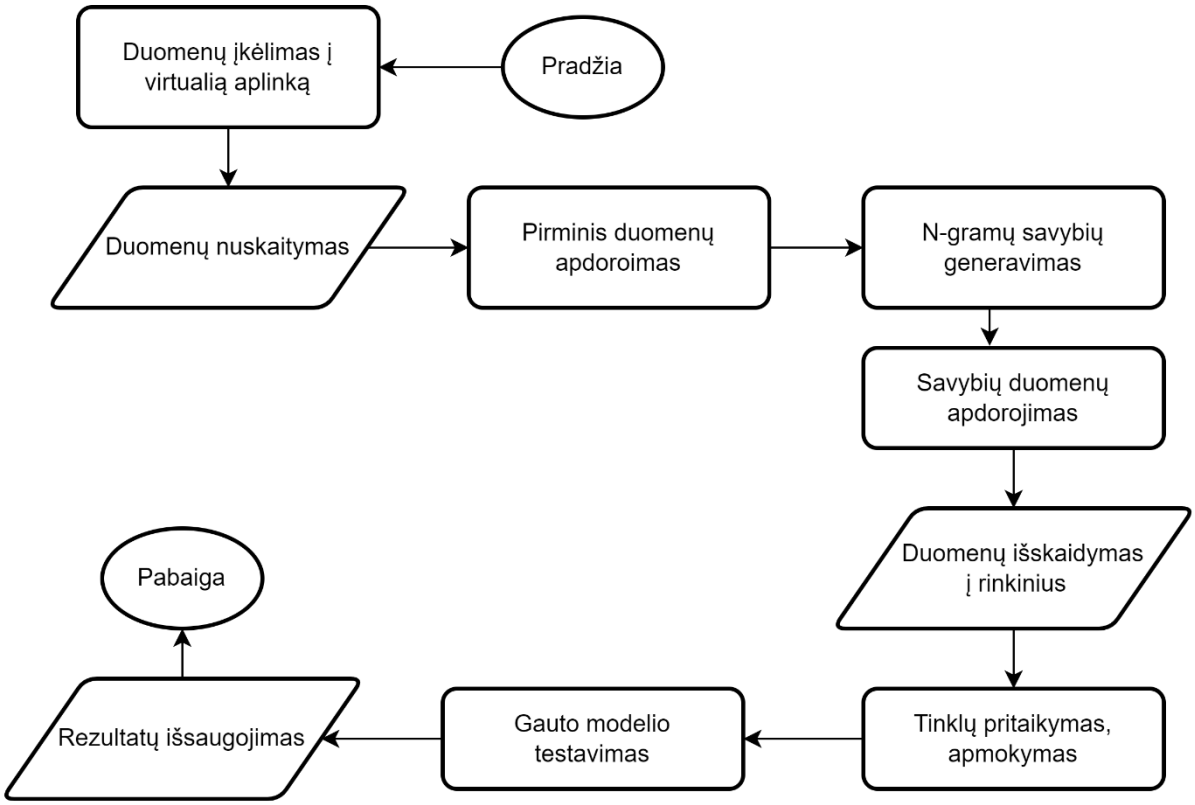
Autorius	Tekstų kiekis	Žodžių kiekis	Vid. teksto ilgis žodžiais
1. JURŠĖNAS Č.	7 779	1 454 762	187,0
2. KUBILIUS A.	5 093	1 076 084	211,3
3. ANDRIUKAITIS V. P.	3 932	926 247	235,6
4. VESELKA J.	3 018	672 644	222,9
5. VIDŽIŪNAS A.	2 654	455 756	171,7
6. STEPONAVIČIUS G.	2 539	456 598	179,8
7. DEGUTIENĖ I.	2 491	484 498	194,5
8. PEČELIŪNAS S.	2 430	458 652	188,7
9. MUNTIANAS V.	2 241	384 398	171,5
10. SYSAS A.	2 130	451 444	211,9
11. SAKALAS A.	2 024	400 853	198,0
12. DAGYS R. J.	1 934	406 769	210,3
13. PAULAUSKAS A.	1 808	361 556	200,0
14. LANDSBERGIS V.	1 732	595 325	343,7
15. RAZMA J.	1 719	316 059	183,9
...	...	...	...
VISO: 147	**110 908**	**23 908 302**	**215,6**

Iš viso panaudota 147 autorių tekstai, išsakyti Seimo salėje. Atkreipiant dėmesį į parlamentarų pasisakymų laikotarpių ribas, galima išskirti, jog pateiktuose pavyzdžiuose gali būti pastebimas lietuvių kalbos žodžių naujumas, skirtingi kalbos stiliai, sakinių struktūra, ar net žodžių vartojamumo dažnis. Eksperimento duomenys atitinka asmenų, kurių lietuvių kalba ir viešojo kalbėjimo įgūdžiai privalo būti geri arba nepriekaištingi, pasisakymai, atitinkantys tam tikrą specifiką ir stilių. Būtent dėl autorių įvairovės, kuri pasireiškia unikaliomis žodžių sekomis, sakinių nuoseklumu ar žodžių

vartojamumu, galima išskirti daugiau savybių, padedančių identifikuoti autorių. Atsižvelgiant į tai, jog kiekvieną autorių leidžiantis atpažinti tekstų rinkinys turi būti išsamus, būtent dėl to eksperimento duomenims pasitelkti autoriai, kurie pasisakė ne mažiau nei 200 kartų. Daugiausiai vieno autoriaus pasisakymų skaičius – 7779. Kalbant apie žodžių kiekį tekste, minimalus jo skaičius – 100, pagrindinė to priežastis yra ta, jog trumpas tekstas neatspindi autoriaus kalbėjimo stiliaus ir neparodo tikslių duomenų, kurių būtų galima pasiekti su didesniu žodžių kiekiu tekste. Svarbu paminėti, jog atliekant tyrimą aktualus ir visų naudotų žodžių skaičius, kuris siekia apie 24 mln. Taip pat, eksperimento duomenys gali būti įvardijami kaip 110908 kalbos, kuriomis pasidalino 147 parlamentarai [36].

3.2. Lietuvių kalbos tekstų identifikavimo metodika

Pasitelkus anksčiau apibūdintus duomenis, kurie yra naudojami tyrimo eigoje kaip pagrindinis įrankis, norint atpažinti lietuvių kalbos tekstų autorius pasitelkiant gilųjį mokymąsi, bus pradedamas lietuvių kalbos tekstų autorių identifikavimo proceso vykdymas. Šiame darbe autorystės nustatymo tyrimas vykdomas remiantis žemiau pateikta schema (žiūrėti 3.1 pav.).



3.1 pav. Tyrimo eigos schema

Schemoje pateikiama pilna tyrimo eiga, kurioje atsispindi veiksmai nuo pradžios ir duomenų įkėlimo į virtualią aplinką, iki pat gautų rezultatų testavimo, kurie ir užbaigia eksperimentinį tyrimą.

3.2.1. Įrankiai

Šis darbas atliktas naudojantis „Google Colaboratory“ aplinka. Sutrumpintai ši aplinka vadinama „Colab“ ir jai apibūdinti yra vartojamas toks apibrėžimas – „Google Research“ produktas, leidžiantis bet kam rašyti ir vykdyti „python“ kodą per naršyklę. Ši aplinka ypač tinka mašininiam mokymuisi bei duomenų analizei. „Colab“ veikia „Jupyter notebook“ pagrindu. Ji nereikalauja specifinės konfigūracijos lyginant su „Jupyter notebook“ [37]. Duomenų laikyti ir saugoti bus naudojamas „Google Drive“ debesis.

3.2.2. Duomenų įkėlimas

Šis procesas bus pakankamai intensyvus, kadangi duomenų kiekis viršija 110 tūkst. atskirų failų, kuriuos reikia išarchyvuoti. Tam bus panaudojamas suspaustas failų archyvas, kuris laikomas saugykloje. Eksperimento kode bus nurodomas duomenų rinkinio archyvo failo „ID“. Jį paleidus, failas bus atsiunčiamas į darbinę virtualią aplinką. Todėl, virtualioje aplinkoje bus sukuriami nauji aplankai, skirti atsiųstų failų saugojimui. Duomenų rinkinys išarchyvuojamas ir sugrupuojamas pagal autoriaus pavardę.

Tyrimui naudojamų 110 tūkst. failų bus išarchyvuota į virtualią aplinką. Svarbu pabrėžti, jog archyve kiekvienam autoriui yra sudarytas aplankas, į kurį yra sukelti visi to autoriaus pasisakymų failai. Atitinkamai yra 147 aplankai, kurie reiškia, kad tiek pat yra ir autorių. Aplankų pavadinimai atitinka autorių vardą ir pavardę, kiekviename aplanke yra mažiausiai po 200 autoriaus kalbos tekstų.

3.2.3. Savybių išskyrimas ir paruošimas

Išarchyvuoti failai virtualioje aplinkoje yra laikomi diske. Norint šiuos failus panaudoti tolimesniuose tyrimuose, reikalinga juos nusiskaityti į atmintį tam, kad būtų galima išvengti, kad skaičiavimai vyktų sparčiau. Toliau iš asmens kalbos autorių tekstų failų bus ištraukiamas tekstas. Šis tekstas atskirai bus išsaugomas į vieną bendrą failą su kitų autorių tekstais.

Sekantis tyrimo žingsnis – pirminių duomenų apdorojimas. Šiam etapui atlikti galima pasitelkti ir pritaikyti įvairius metodus. Analizuotuose tyrėjų darbuose buvo išbandyta daugybė būdų norint tinkamai atlikti duomenų apdorojimą. Apžvelgus įvairių autorių darbus paaiškėjo, kad dažniausiai naudojami apdorojimo metodai, kuriais siekiama pašalinanti simbolius, didžiąsias raides pakeisti mažosiomis ir panaikinti skaičius, darbuose padėjo pasiekti geresnių rezultatų sprendžiant autoriaus identifikavimo iš rašytinių tekstų uždavinį. Toliau apžvelgiant lietuvių kalbos tekstų apdorojimo metodikas, verta paminėti, kad eksperimentiniame tyrime autorių vardų inicialai ir pavardės bus užkoduoti atitinkamais skaičiais nuo 1 iki 147 tam, kad tolimesniuose žingsniuose juos būtų lengviau priskirti dirbtinio neuronų tinklo mokymosi metu kaip klasių (autorių) etiketes.

Kitas eksperimento etapas yra n-gramų savybių generavimas. Tai reiškia, kad bus sugeneruojamos savybės iš autoriaus pasisakymų, kurie yra laikomi tekstais. Šios savybės bus sugeneruojamos nurodant tikslų norimų savybių skaičių ir atitinkamą n-gramų intervalą. Savybių generavimo procesas bus vykdomas dviem būdais panaudojant tiek žodžių n-gramų skirtingus intervalus, tiek ir simbolių n-gramų skirtingus intervalus. Žodžiams bus naudojamos vienos eilės n-gramos (pvz., 1-os eilės) ir mišrių eilių n-gramos (pvz., visos n-gramos nuo 1-os iki 3-os eilės), o simboliams bus naudojamos ne tik tos pačios vienos eilės n-gramos, bet ir mišrių eilių n-gramos (pvz., visos n-gramos nuo 1-os iki 5-os eilės). Šios sugeneruotos savybės bus laikomos žodžių ar simbolių dažniais tekste. Sudarytos savybės bus saugomos atskirame faile tam, jog būtų galima užtikrinti, kad mašininio mokymosi tinklams bus nurodomi tik skaičių masyvai. Sudarytos savybės bus dar kartą apdorojamos – jos bus normalizuojamos. Atlikta panašių tyrimų analizė atskleidė, kad šis žingsnis kai kuriuose tyrėjų tyrimuose dar labiau pagerino gaunamus rezultatus, kadangi modeliai geriau apdoroja normalizuotus duomenis. Toliau duomenys bus išskaidomi į 3 dalis, santykiu 70:20:10. Mašininio mokymosi tinklams apmokyti bus paskirta 70 % duomenų, 20 % bus skirti modelio validacijai ir paskutinieji 10 % bus palikti apmokyto mašininio mokymosi modelio testavimui atlikti.

3.2.4. Autorystės savybių klasifikavimas

Klasifikuojant autorystės savybes, į jų procesą svarbu įtraukti ir tokius esminius veiksmus kaip: tinklų pritaikymas, apmokymas, ir gautų rezultatų testavimas. Sekantis etapas yra prieš tai paruoštų savybių panaudojimas skirtingų mašininio mokymosi tinklų apmokymui, validacijai bei testavimui. Pasirinkti klasifikavimo metodai: MLP, CNN, bi-LSTM, SVM, KNN ir Bayes klasifikatorius. Tinklams buvo parinktos tokios topologijos:

- MLP: įėjimų skaičius, priklausantis nuo savybių skaičiaus (1000–11000), 2 sluoksniai, pirmame sluoksnyje 350, antrame 750 neuronų, taip pat po paskutiniu sluoksniu yra panaudota bachnormalization bei dropout sluoksniai, neuronų, išėjimų skaičius yra tolygus 147 autorių skaičiui.
- CNN: įėjimų skaičius (1000–11000), 3 sluoksniai, pirmame sluoksnyje 64, antrame 64, o trečiame 300 neuronų, tinkle panaudoti averagepooling, bei flatten sluoksniai, išėjimų skaičius, lygus autorių skaičiui 147.
- Bi-LSTM: įėjimų skaičius (1000–11000), 3 sluoksniai, pirmame sluoksnyje 64, antrame 64, o trečiame 32 neuronų, tinkle panaudoti bachnormalization, bei dropout sluoksniai, išėjimų skaičius yra lygus autorių skaičiui 147.

Kaip palyginamieji metodai pasirinkti SVM, KNN ir Bayes. Jie visi grįsti skirtingais klasifikavimo principais, todėl leis atlikti pakankamai išsamų palyginimą su tyrime naudojamais dirbtinių neuronų tinklų pagrindu grįstais klasifikatoriais. Pagrindiniam uždaviniui spręsti –

autorystės nustatymui iš lietuviškų politinio pobūdžio rašytinių tekstų bus naudojami pagrindiniai tinklai MLP, CNN, bi-LSTM. Pasirinkti šie metodai, nes atliktos įvairių autorių tyrimų analizių išvados parodė, kad būtent šie metodai siekia geriausius rezultatus autorių identifikavimo uždavinyje. Tuo remiantis galima manyti, jog gerus rezultatus metodai pasieks ir su lietuviškais rašytiniais tekstais.

Toliau seka apmokytų modelių patikrinimo procesas naudojant testavimo duomenis. Visi prieš tai minėti tinklai bus dar kartą paleidžiami naudojant tuos pačius testavimo duomenis ir iš jų bus nustatoma, koks bus gautas galutinis rezultatas sprendžiant lietuvių kalbos autorių nustatymo uždavinį.

3.2.5. Rezultatų išsaugojimas

Rezultatuose bus išsaugojami paskutiniųjų modelių apmokymų epochų rezultatai, kurie bus gaunami naudojant simbolių bei žodžių n-gramų duomenų rinkinius. Kiekvienas rezultatas turės skirtingą intervalą ir savybių skaičių. Prie jų taip pat bus pateikiami ir testavimo proceso rezultatai. Iš visų rezultatų rinkinių bus sprendžiama, kurie tinklai pasiekė geriausius rezultatus, kurie savybių rinkiniai veikė efektyviausiai bei bus nustatoma, koks savybių skaičius parodė geriausius rezultatus sprendžiant šį uždavinį.

3.3. Tyrimo planas

Tyrimai bus vykdomi, rezultatai bus pristatomi tokiu nuoseklumu:

1. Pradinio autorystės tikslumo įvertinimas – atliekama su 1000 savybių dydžio aibe, apsiribojant 1-gramos modeliu.
2. Autorystės savybių (požymių) aibės dydžio tyrimas – kaip nustatymo tikslumą įtakoja aibės dydis.
3. Autorystės savybių tyrimas – kurios eilės n-gramos modeliai leidžia pasiekti didžiausią tikslumą.
4. Gautojo rezultato įvertinimas – palyginimas su kitais autoriais, įvertinamas tikslumo padidėjimas lyginant su pradiniu (1 punktu). Rezultatų apibendrinimas.

Pagrindinis eksperimentų rezultatų vertinimo kriterijus – autoriaus identifikavimo (atpažinimo) tikslumas, nurodantis, kuri autorių dalis buvo identifikuota teisingai.

Tyrimai bus vykdomi dviem atvejais: žodžių ir simbolių n-gramomis.

3.4. Trumpas skyriaus apibendrinimas

Lietuvių kalbos tekstų autorystės identifikavimo tyrimui buvo pasirinkti parlamentarų pasisakymų tekstai. Eksperimentinio tyrimo atlikimui išskirtos tokios dalys: lietuvių kalbos tekstų

identifikavimo metodika, pasitelkti įrankiai, duomenų įkėlimas, savybių išskyrimas ir paruošimas, autorystės savybių klasifikavimas ir rezultatų išsaugojimas. Taip pat pateiktas tyrimo planas su informacija apie kiekvieną plano dalį.

4. EKSPERIMENTINIO TYRIMO REZULTATAI

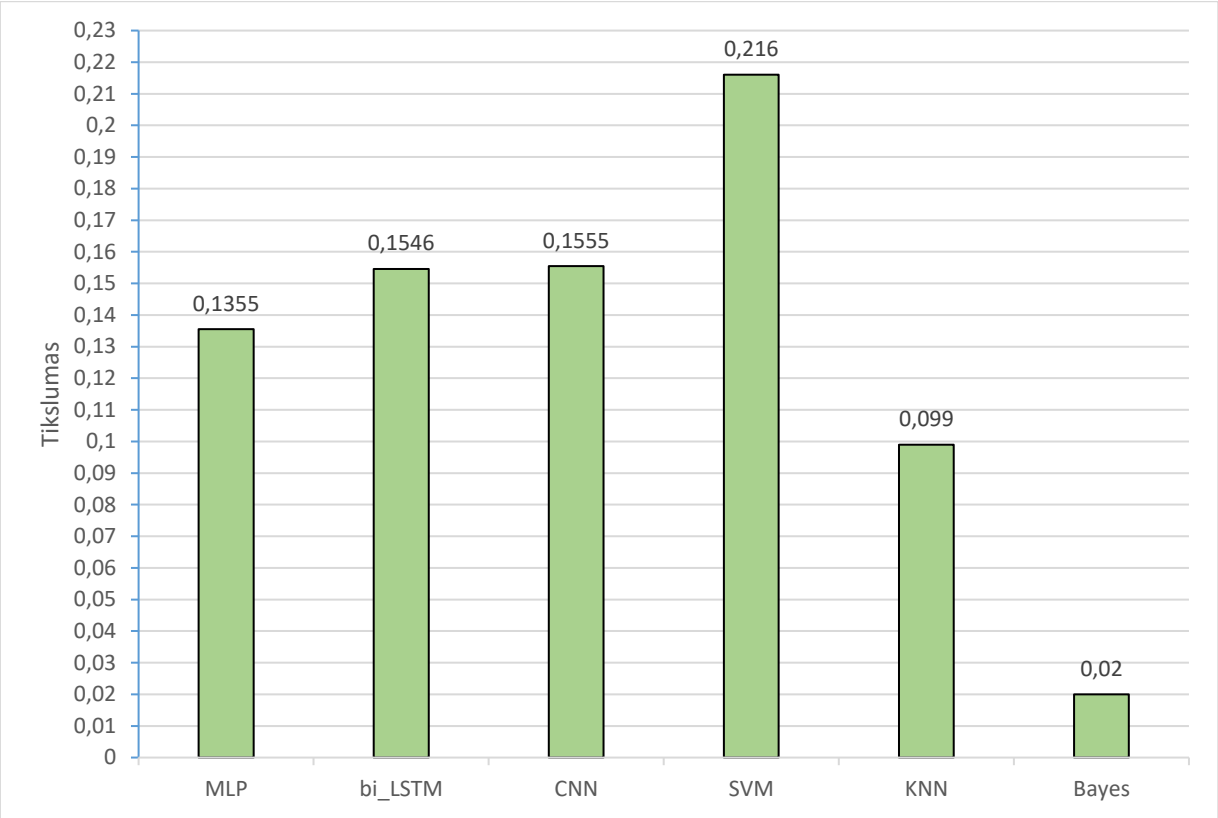
Atlikus lietuvių kalbos tekstų autorių tekstų identifikavimo tyrimą, išanalizavus ir pateikus tyrimui naudotus duomenis bei visą atliekamo eksperimento planą, toliau pateikiami lietuvių kalbos tekstų autorių modeliavimo ir identifikavimo rezultatai. Jie gauti eksperimentuojant su skirtingomis simbolių, žodžių n-gramomis, giliojo mokymosi modeliais bei lyginant gautuosius rezultatus su tradicinio mašininio mokymosi modeliais tyrimo metu.

4.1. Simbolių n-gramų tyrimas

Pirmoje tyrimo dalyje atliktas eksperimentinis tyrimas su simbolių n-gramomis. Poskyryje pristatomi tyrimo rezultatai, kuriuos sudarys išskaidyta rezultatų pateikimo eiga. Ją sudaro: pradinio tikslumo vertinimas, savybių aibės dydžio tyrimas, savybių intervalo tyrimas ir gautų rezultatų įvertinimas.

4.1.1. Pradinio tikslumo vertinimas

Eksperimentinis tyrimas su skirtingais metodais parodė skirtingus rezultatus. Paruošiamųjų eksperimentų metu, buvo atliekami tyrimai su 7 skirtingais giliojo mokymosi ir tradiciniais mašininio mokymosi modeliais: daugiasluoksniu perceptronu (MLP), ilgalaikės–trumpalaikės atminties neuronų tinklais (LSTM), dvikrypčiu ilgos trumpalaikės atminties neuronų tinklais (bi_LSTM), konvoliuciniais neuronų tinklais (CNN), atraminių vektorių mašina (SVM), *K* artimiausių kaimynų klasifikatoriumi (KNN) ir Bejeso klasifikatoriumi (Bayes). Atsižvelgiant į pradinius gautų rezultatų duomenis, buvo pastebėta tai, kad LSTM bei bi-LSTM rezultatai pakankamai stipriai skyrėsi. Kadangi bi_LSTM veikia tokiu pačiu pagrindu kaip LSTM, buvo nuspręsta LSTM metodo atsisakyti. Pirmasis eksperimentas atliktas naudojant prieš tai minėtus tinklus ir simbolių savybių n-gramas, naudojant modelio eilės intervalą 1–1 bei 1000 savybių (4.1 pav.).



4.1 pav. Tinklų su simbolių 1–1-gramomis rezultatų palyginimas

Pateiktoje diagramoje yra matomas visų naudotų tinklų rezultatas. Iš eksperimentų metu gautų rezultatų galima pastebėti, kad didžiausias tikslumas buvo pasiektas naudojant SVM klasifikatorių 21,6 %, antras pagal gerumą yra CNN su daugiau nei 15 %, tačiau tikslumas nėra labai aukštas. Likę klasikiniai mašininio mokymosi metodai, KNN bei Bayes, parodė žemiausius rezultatus, siekiančius 2–9% tikslumą. Tam įtakos galėjo turėti ganėtinai mažas savybių skaičius bei žema simbolių n-gramų modelio eilė. Kadangi autorių kalbos tekstų yra daugiau nei 110 tūkstančių, galima teigti, kad panaudotas savybių skaičius bei simbolių n-gramų aibės dydis yra per žemas tokios apimties uždaviniui spręsti.

4.1.2. Savybių aibės dydžio tyrimas

Toliau yra pateikiamas gautųjų rezultatų palyginimas ir priklausomybė nuo didesnio savybių skaičiaus naudojant simbolių n-gramų intervale 1–3, nes būtent šis intervalas pasiekė geriausius rezultatus (žiūrėti 4.1 lentelė).

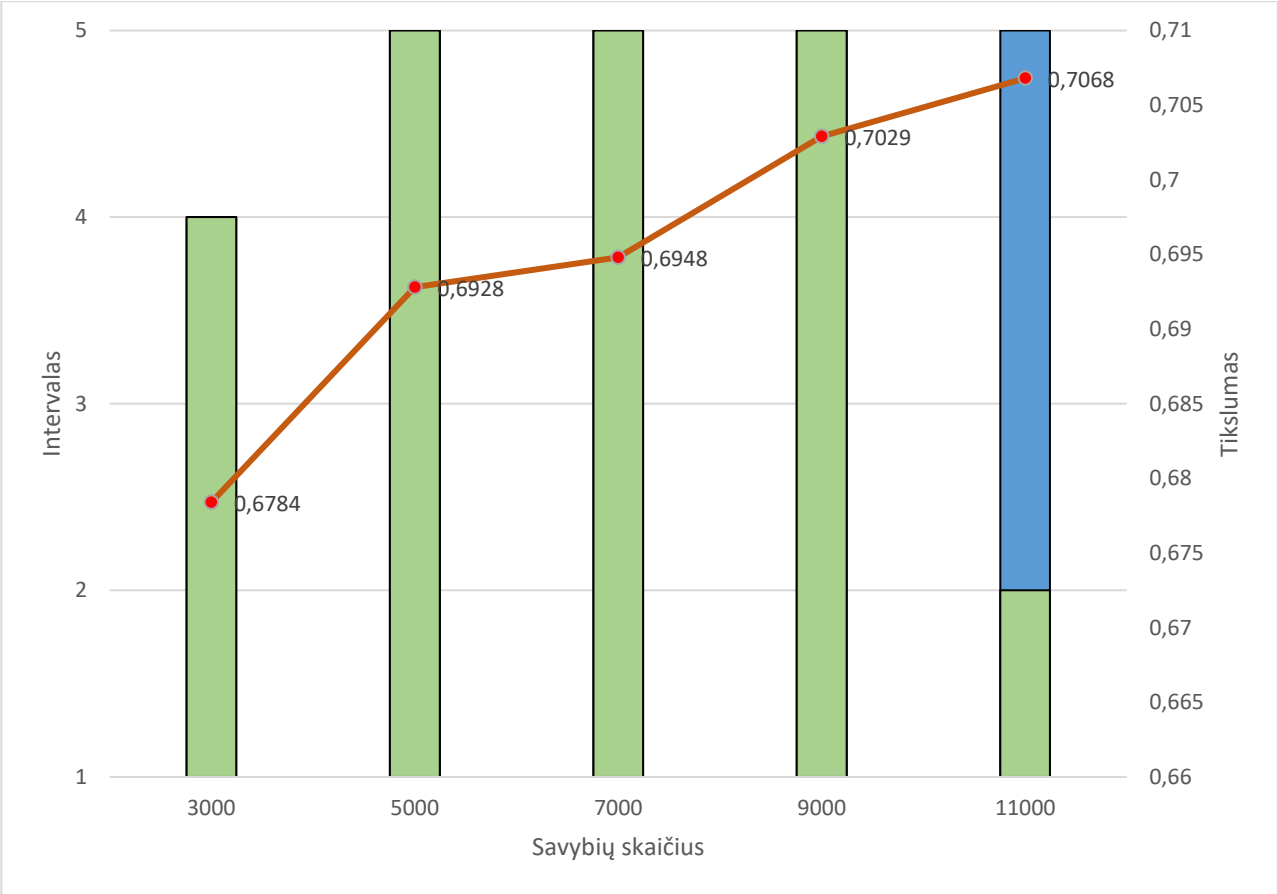
4.1 lentelė. Tinklų rezultatų priklausomybė nuo simbolių požymių skaičiaus

Tinklai	3000	5000	7000	9000	11000
MLP	0,6477	0,6524	0,6757	0,6754	0,6820
bi_LSTM	0,5827	0,5907	0,5947	0,6006	0,6060
CNN	0,6553	0,6587	0,6740	0,6882	0,7068

Lentelėje pateikti 3 tinklų tikslumo priklausomybė nuo požymių skaičiaus, stulpeliuose – požymių skaičius, o eilutėse – tinklų pavadinimai. Gauti eksperimento rezultatai rodo, jog pasitelkiant didesnį savybių skaičių, lyginant su praeitu bandymu, tikslumas padidėjo daugiau nei 50 %. Didesnis savybių skaičius palaipsniui didina sprendimų tikslumą, ir lentelėje matyti, jog tinklas CNN pasiekė geriausius rezultatus. Nustatyta, kad tinklo rezultatai padidėjo nuo 15,5 % iki 65,5 % pakėlus naudojamų savybių skaičių iki 3000. Palaipsniui didinant savybių skaičių kas 2000, galima matyti, jog rezultatai kyla beveik 1 %, o su 11000 savybių buvo pasiektas 70,5 % tikslumas. Remiantis gautais rezultatais, galima teigti, kad šiam uždaviniui spręsti didesnis savybių skaičius lemia geresnius rezultatus.

4.1.3. Savybių intervalo tyrimas

Toliau pateikiama didžiausią tikslumą demonstruojančio CNN tinklo geriausi rezultatai bei juos leidžiantis pasiekti požymių skaičius, n-gramos modelio eilės intervalas (žiūrėti 4.2 pav.).



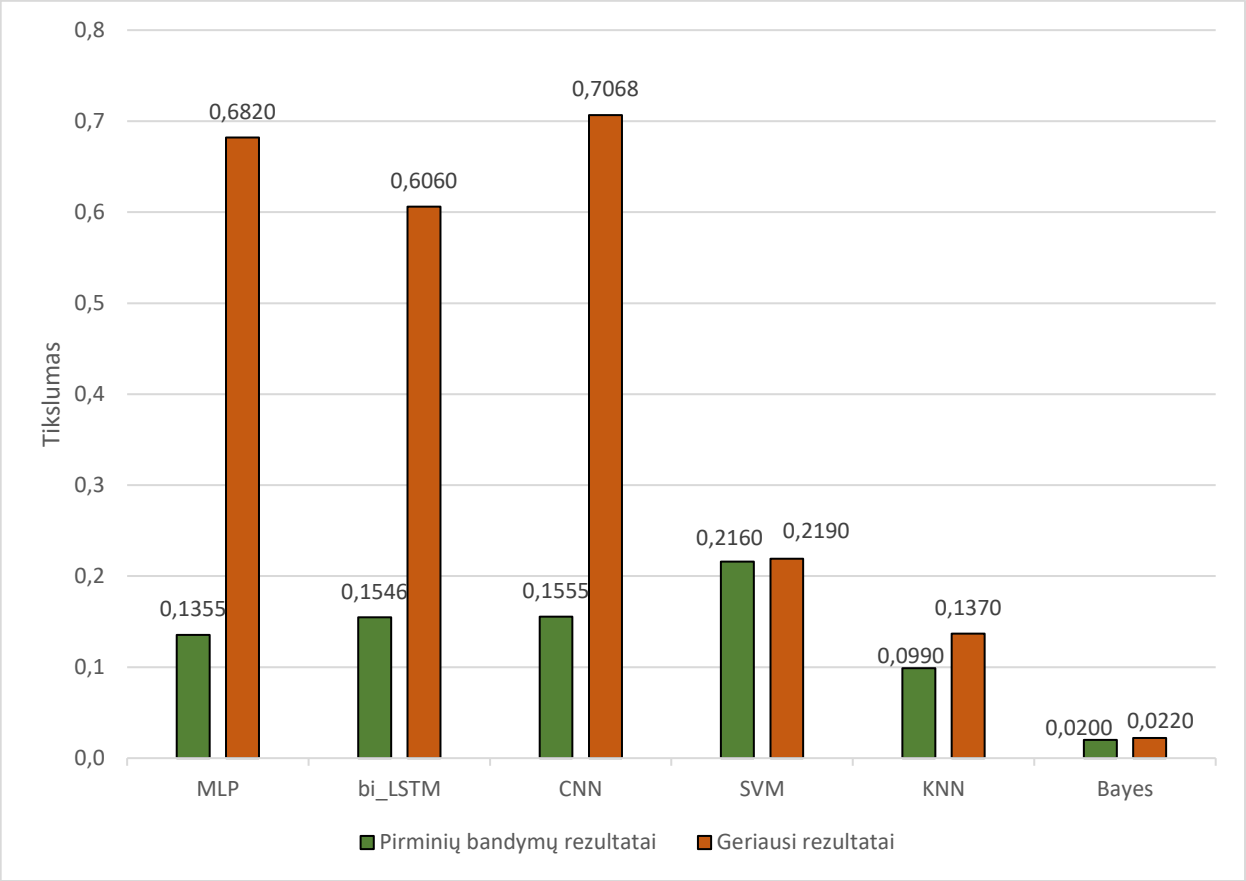
4.2 pav. Tinklo rezultatų priklausomybė nuo požymių skaičiaus geriausius rezultatus gaunančių simbolių n-gramų intervaluose

Iš paveikslo puikiai matyti, kad didėjant savybių skaičiui, identifikavimo rezultatai taip pat gerėja. Eksperimentų metu pastebėta, kad norint pasiekti geriausią rezultatą, tai galima atlikti naudojantis savybių intervalu 2–5. Šis intervalas ir 11000 simbolių savybių pasiekė daugiau nei

70,5 % tikslumo rezultatą. Didesnis n-gramos modelio intervalas analizuoja platesnę simbolių seką, išgautų iš autorių tekstų, diapazoną. Būtent dėl to galima daryti prielaidą, kad autorystės nustatymo uždavinyje tai suteikia daugiau informacijos apie autoriaus unikalų stilių.

4.1.4. Gautų rezultatų įvertinimas

Toliau pateikiamas geriausių rezultatų palyginimas, kur pirmojo bandymo (su 1000 savybių) atvejis lyginamas su geriausiais gautais rezultatais (su 11000 savybių) (žiūrėti 4.3 pav.).



4.3 pav. Geriausių rezultatų palyginimas su pirminiais bandymų rezultatais

Pateiktoje diagramoje matomi pirminio bandymo rezultatai ir geriausi rezultatai apskritai. Matoma, kad CNN tinklas pasiekė geriausius rezultatus, daugiau nei 70,5 % tikslumą, lyginant su pradiniu bandymu gautųjų 15,5–55 % padidėjimą. Nagrinėjant kitais neuronų tinklais gautus rezultatus, juose taip pat pastebimas žymus tikslumo padidėjimas. MLP tinklo pradinių bandymų rezultatai buvo pagerinti 54 %, o bi_LSTM – beveik 45 %. Klasikinių metodų rezultatai pakito mažiausiai. Galima teigti, kad didesnis savybių skaičius tolimesniuose eksperimentuose turėjo įtakos gaunamų rezultatų didėjimui neuronų tinkluose. Lyginant gautą rezultatą su tyrėjų, naudojančių tuos pačius duomenis eksperimentams ir sprendžiančius panašų autoriaus identifikavimo uždavinį,

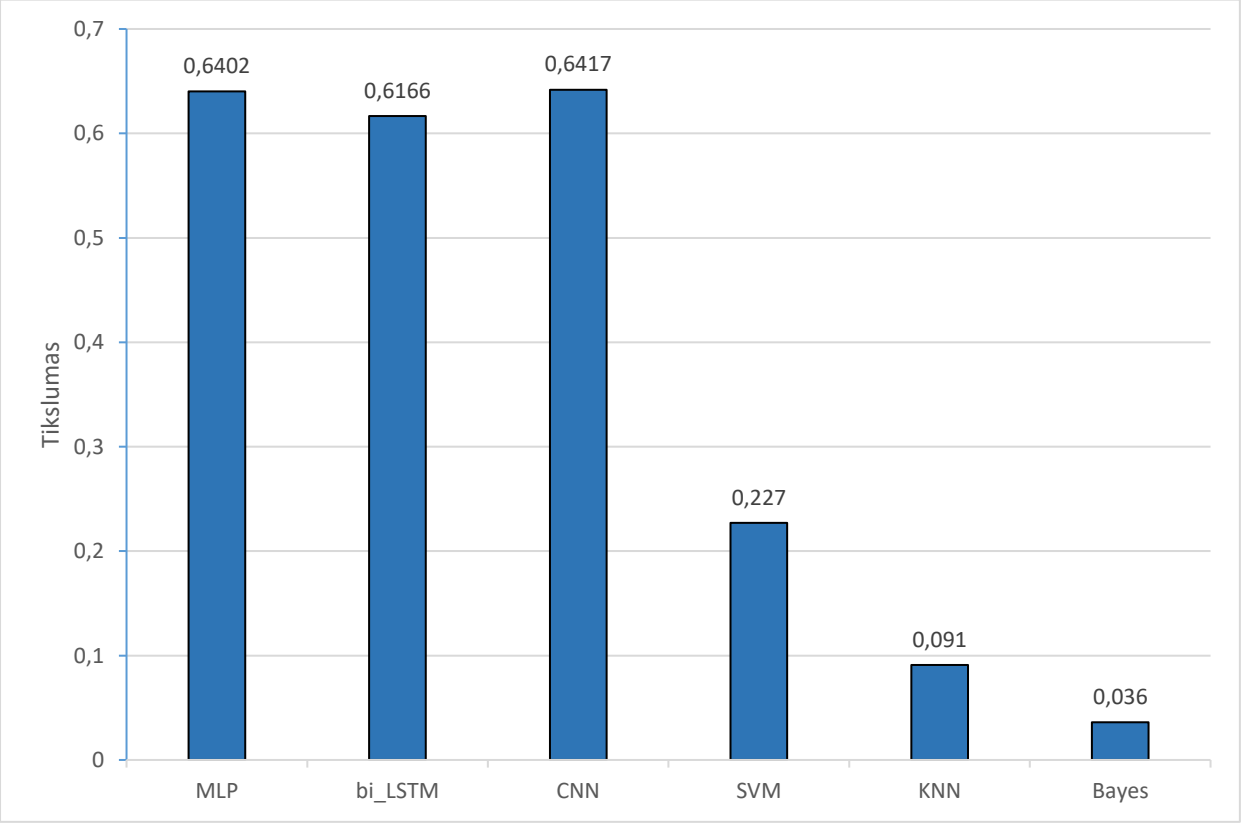
rezultatas buvo pagerintas daugiau nei 7,5 % [36]. Pilni tyrimo rezultatai yra pateikti (žiūrėti A priedas).

4.2. Žodžių n-gramų tyrimas

Toliau pateikiama antrosios tyrimų dalies – žodžių n-gramų rezultatai. Ją sudaro: pradinio tikslumo vertinimas, savybių aibės dydžių tyrumas, savybių intervalo tyrimas ir gautų rezultatų įvertinimas.

4.2.1. Pradinio tikslumo vertinimas

Kita eksperimento dalis yra atlikta naudojant žodžių savybių n-gramų rinkinį. Pirmieji eksperimentai yra atlikti naudojant tokius pačius tinklus, kurie buvo pasitelkiami ir pirmiesiems bandymams su simbolių savybių n-gramų rinkiniu. Toliau pateikiamas grafikas naudojant 1–1 intervalą ir 1000 savybių (žiūrėti 4.4 pav.).



4.4 pav. Tinklų su žodžių 1–1-gramomis rezultatų palyginimas

Didžiausias tikslumas buvo pasiektas naudojant CNN bei MLP tinklus – maždaug 64 %. Gauti rezultatai nedideliu skirtumu lenkia ir bi_LSTM tinklą, kuris pasiekė daugiau nei 61,5 % tikslumą. Norint palyginti neuronų tinklais gautą tikslumą su ankstesniu simbolių n-gramų savybių rinkiniu, galima paminėti tai, kad rezultatui įtakos galėjo turėti autoriaus tekstų žodžių ir žodyno unikalumas, pagal kurį tinklams daug geriau sekasi atpažinti autorių negu su simbolių rinkiniu. Kiek prastesnius

rezultatus, sprendžiant autorystės identifikavimo uždavinį, parodė klasikiniai mašininio mokymosi metodai, pasiekę 3–22,5 % tikslumą. Taigi, neuronų tinklų ir žodžių n-gramų savybių efektyvumas žymiai lenkia klasikinių metodų efektyvumą.

4.2.2. Savybių aibės dydžio tyrimas

Pateiktas gautų rezultatų palyginimas, jų priklausomybė nuo savybių skaičiaus naudojant žodžių n-gramos modelio eilių intervalą 1–3, nes būtent šis intervalas pasiekė geriausius rezultatus (žiūrėti 4.2 lentelė).

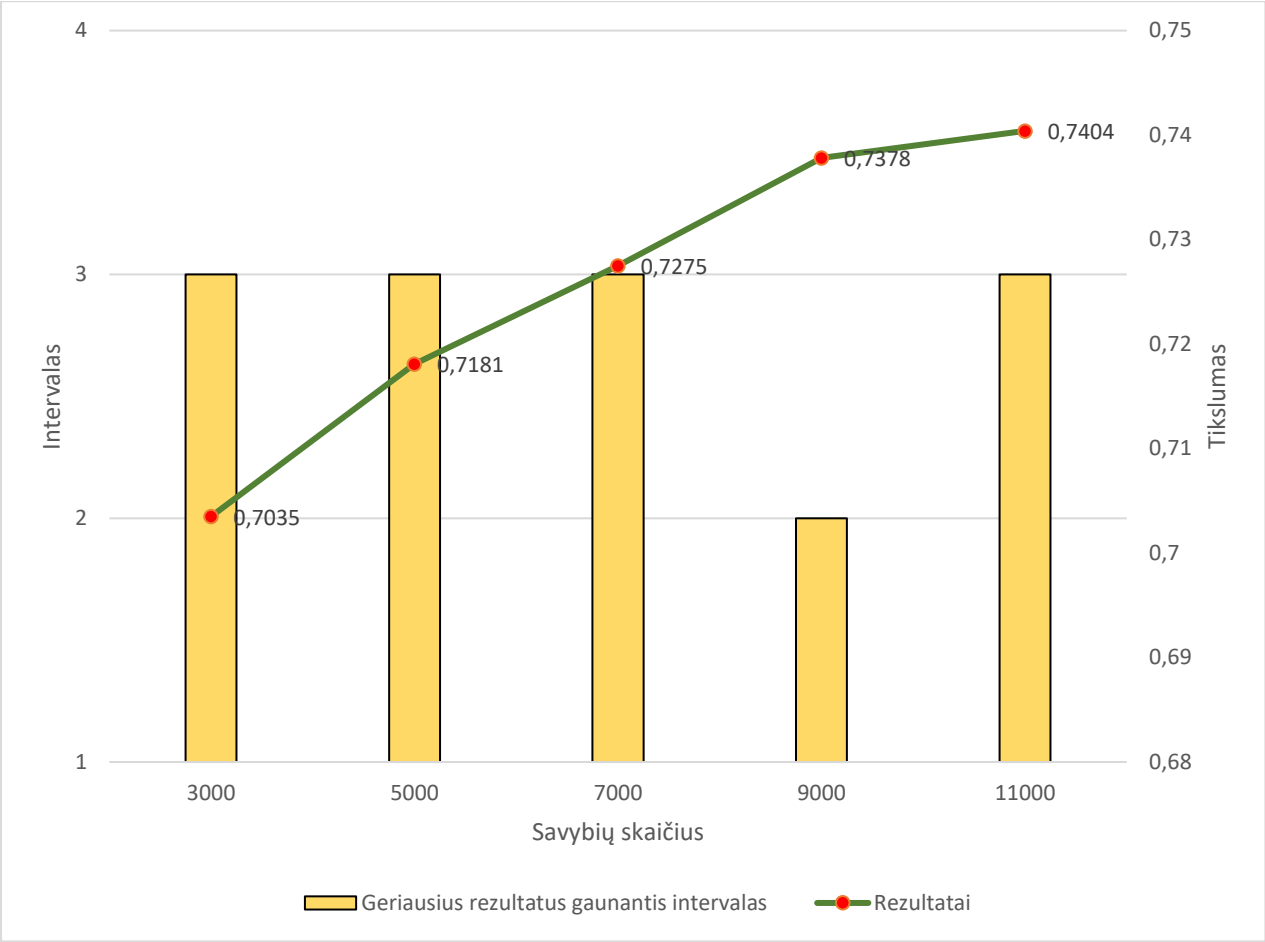
4.2 lentelė. Tinklų rezultatų priklausomybė nuo žodžių požymių skaičiaus

Tinklai	3000	5000	7000	9000	11000
MLP	0,6640	0,6830	0,6818	0,6797	0,6799
bi_LSTM	0,6572	0,6622	0,6666	0,6661	0,6539
CNN	0,7035	0,7181	0,7275	0,7280	0,7404

Gauti eksperimento rezultatai rodo, jog pasitelkiant didesnį savybių skaičių, rezultatai padidėjo 2–5 % Didinant savybių skaičių palaipsniui didėja ir identifikavimo tikslumas. Taip pat pastebėta, kad nėra tokio didelio tikslumo šuolio (buvusio su simbolių savybėmis, kuriame rezultatai padidėjo daugiau nei 50 %), tačiau vis dėl to, būtent žodžių savybių n-gramų rinkiniai leidžia didžiausią tikslumą . Lentelėje matyti, jog CNN tinklas pasiekė aukščiausią, daugiau nei 74 % tikslumą. Taip pat matyti, kad tinklo rezultatai palaipsniui didėjo nuo 64 % iki 74 % didinant naudojamų savybių skaičių – rezultatai kilo po maždaug 1% savybių rinkinio dydžiui padidėjus 2000 savybių. Kiti tinklai, kurių tikslumas neaugo daugiau 66–68 %, aukštesnio rezultato, didinant savybių skaičių, nepasiekė. Atsižvelgiant į gautus rezultatus, galima teigti, kad šio uždavinio sprendimui didesnis savybių skaičius atneša geresnius rezultatus tik CNN tinklui.

4.2.3. Savybių intervalo tyrimas

Toliau nagrinėjama, kas lemia didžiausią tikslumą CNN tinkle. 4.5 paveiksle pateikti geriausi CNN pademonstruoti autorių identifikavimo rezultatai ir juos atitinkantys savybių rinkinių dydžiai, naudotų n-gramos modelio eilės intervalai (žiūrėti 4.5 pav.).

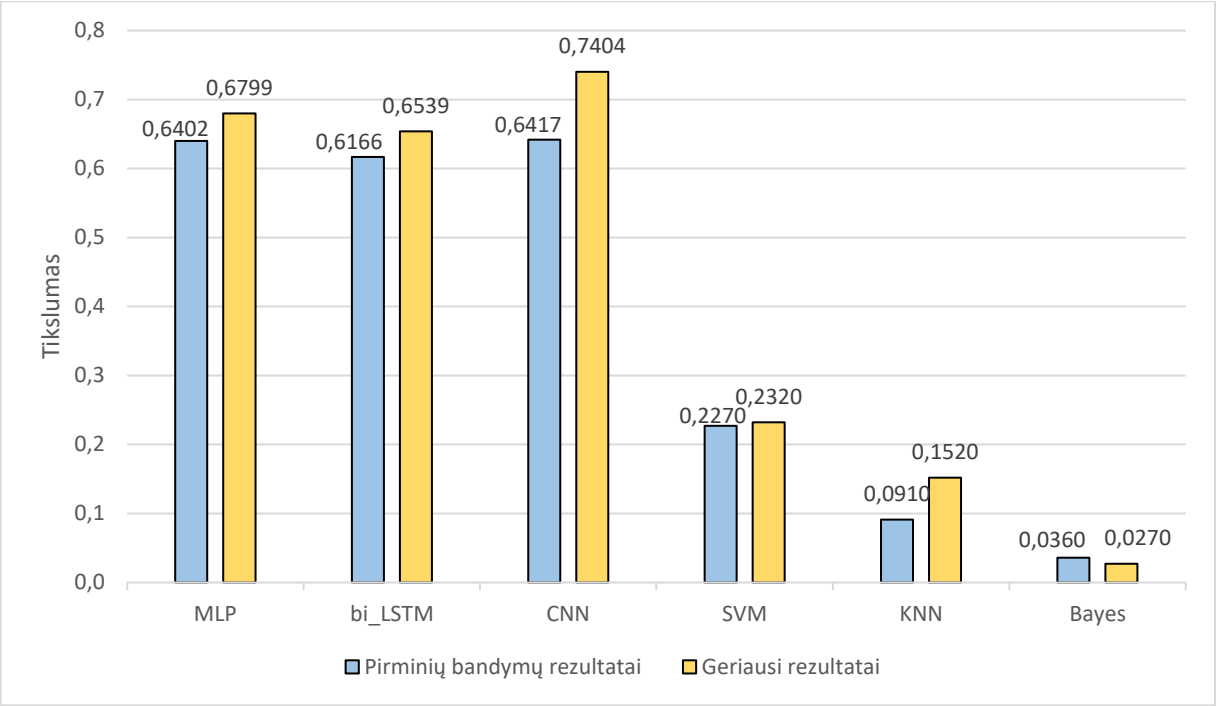


4.5 pav. Tinklo rezultatų priklausomybė nuo požymių skaičiaus geriausius rezultatus gaunančių žodžių n-gramų intervaluose

Vėl matoma, kad kuo didesnis savybių skaičius, tuo identifikavimo rezultatai yra geresni. Taip pat eksperimentų metu pastebėta, kad didžiausi rezultatai pasiekiami naudojantis 1–3 eilių n-gramų savybes. Šis intervalas ir 11000 simbolių savybių leido pasiekti daugiau nei 74 % tikslumą. Jį palyginus su didžiausiu tikslumu simbolių n-gramų savybių atveju (70,5 %), galima teigti, kad rezultatai pagerėjo daugiau nei 3,5 %. Didesnis savybių intervalas analizuoja platesnį žodžių, išgautų iš autorių tekstų, skirtumą, dėl to daroma prielaida, kad autorystės nustatymo uždavinyje tai suteikia daugiau informacijos apie autoriaus unikalų stilių.

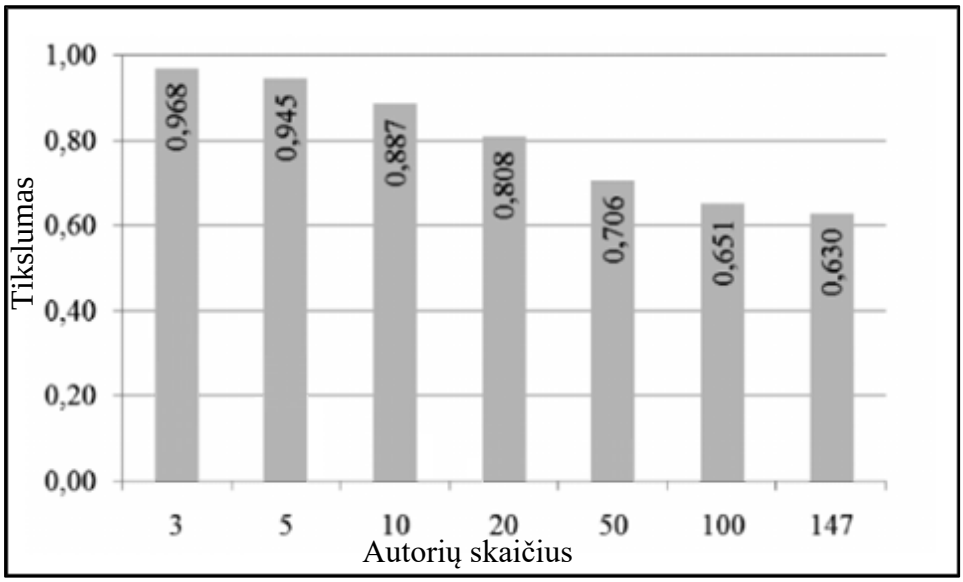
4.2.4. Gautų rezultatų įvertinimas

Toliau bus pateikiama tarpusavio tinklų geriausių rezultatų palyginimas, pirmojo bandymo su 1000 savybių rezultatų palyginimas su geriausiais gautais rezultatais naudojančius 11000 savybių (žiūrėti 4.6 pav.).



4.6 pav. Tinklų geriausių rezultatų palyginimas su pirminiais bandymų rezultatais

Matoma, kad CNN tinklas vėl parodė geriausią rezultatą – daugiau nei 74 % tikslumą. Lyginant su pradiniais 64 %, gautas beveik 10 % pagerėjimas. Kitų tinklų rezultatai ne taip stipriai pagerėjo: MLP ir bi_LSTM tinko pradinių bandymų rezultatai buvo pagerinti beveik 4 %. Panašiai padidėjo klasikinių metodų rezultatai, tačiau lyginant su neuronų tinklais jie pasiekė blogiausius rezultatus, o Bayes rezultatai, šiuo atveju, net sumažėjo beveik per 1 %. Remiantis šiais rezultatais, didesnis savybių skaičius tolimesniuose eksperimentuose labiausiai pagerino tik CNN tinklo gaunamus rezultatus. Lyginant su simbolių n-gramų rinkinio eksperimento rezultatais, labiausiai pasižymėjo CNN tinklas, jis pagerino rezultatą 4 % ir siekė net 74 %. MPL tinklo rezultatai beveik nesiskyrė, abiem atvejais pasiektas maždaug 68 % tikslumas, o bi_LSTM tinklas geriau veikė su žodžių savybėmis, kadangi tikslumas padidėjo beveik 5 % ir siekė daugiau nei 65 %. Lyginant su kitais tyrimais, naudojančiais tuos pačius duomenis, rezultatas buvo pagerintas daugiau nei 11 %. 4.7 pav. pateikiamas palyginimas su kitų tyrėjų rezultatais.



4.7 pav. Tyrėjų, dirbančių su tokiais pačiais duomenimis, gauti rezultatai [36]

Paveiksle matoma, autorių identifikavimo tikslumas siekia beveik 97 %, kai yra bandoma identifikuoti tik 3 autorius, tačiau, kai bandoma identifikuoti 147 asmenis iš visų 110 tūkstančių tekstų, rezultatas siekė tik 63 %. Lyginant su šiuo rezultatu, mūsų tyrime rezultatas buvo pagerintas 7,5 % iki 70,5 % identifikavimo tikslumo, o žodžių savybių atveju tikslumas buvo pagerintas 11 % iki daugiau nei 74 % (žiūrėti B priedas).

4.3. Trumpas skyriaus apibendrinimas

Šiame skyriuje pateikiamas dviejų skirtingų n-gramos modelių (žodžių ir simbolių lygmens) savybių rinkinių tyrimas bei jo rezultatai. Abiem atvejais palyginimo tikslais neuronų tinklais (MLP, LSTM, CNN) gautieji rezultatai yra palyginami su klasikiniais mašininio mokymo algoritmais (SVM, KNN, Bayes) gautaisiais rezultatais. Nustatyta autoriaus identifikavimo tikslumo priklausomybė nuo simbolių ar žodžių požymių skaičiaus. Didžiausio tikslumo rezultatai gauti CNN tinklo atveju – 74 %. Gautieji rezultatai palyginti su analogiškų tyrimų (su tais pačiais duomenimis) rezultatais.

APIBENDRINIMAS. IŠVADOS

Iš analitinės panašių darbų apžvalgos rezultatų matoma, kad tekstų autorystės identifikavimo uždaviniui būdingi šie aspektai:

1. Didžioji dalis autorių savo tyrimuose naudoja automatiškai sugeneruotas savybes iš teksto, pavyzdžiui: įvairaus ilgio intervalų ir dydžio žodžių ar simbolių n-gramas, „žodžių maišus“, „word2wec“, žodžių dažnio vektorius bei žodžių įterpinius. Dažniausiai naudojami giliojo mokymosi metodai yra šie: daugiasluoksnis perceptronas, konvoliuciniai ir rekurentiniai neuronų tinklai, autoenkoderiai, ilgalaikės–trumpalaikės atminties neuronų tinklai.
2. Esminiai problemų šaltiniai: per didelis arba per mažas teksto fragmentų skaičius, kalbėjimo stilių įvairovė, kalbos tematika, to paties autoriaus kalbos stiliaus kitimas. Labiausiai išsiskyrė dažnio ir ilgio simbolių arba žodžių n-gramais metodai, nes juos naudojant pasiekti geriausi tyrimų rezultatai išskiriant tiek žodžius, tiek ir simbolius iš teksto. Dėl to būtent šie metodai buvo naudojami atliekant šį eksperimentinį tyrimą.

Iš atlikto teksto autorystės modeliavimo ir identifikavimo eksperimentinio tyrimo rezultatų matoma, naudojant ankščiau minėtus metodus ir būdus galima identifikuoti ir lietuviškų tekstų autorystę. Remiantis eksperimentiniais rezultatais suformuluotos tokios išvados:

1. Neuronų tinklais grįsta autorystės nustatymas leido pasiekti 40–72 % didesnį tikslumą nei naudojant klasikinius mašininio mokymo metodus.
2. Žodžių n-gramomis grįstos savybės tyrimuose rodė nedidelį, keleto procentų eilės pranašumą prieš kitus tyrimus su simbolių n-gramų grįstomis savybėmis. Maksimalus pasiektas identifikavimo tikslumas siekė daugiau nei 74 % pirmuoju atveju ir 70,5 % antruoju. Galima teigti, jog žodžių sekos yra šiek tiek unikalesnė kalbančiojo savybė.
3. Naudojamų savybių skaičiaus didinimas leido padidinti identifikavimo tikslumą. Savybių skaičių padidinus nuo 1000 iki 11000, tikslumas didėjo 11–14 %. Tai rodo, jog autorystės identifikavimas – itin sudėtingas uždavinys, reikalaujantis atlikti savybių, su didesniais eilių intervalais, analizę.
4. Tiek žodžių, tiek simbolių atveju didžiausias tikslumas gautas naudojant kompleksines įvairių n-gramos eilių modelių savybes. Žodžių atveju geriausi rezultatai pasiekti su 1–3 eilės n-gramos modeliais, simbolių atveju – su 2–5 eilės modelių savybėmis. Galima tokios situacijos priežastis lietuvių kalbos sudėtingumas, lemiantis sudėtingas simbolių ir žodžių sekas.

REKOMENDACIJOS

1. Siekiant efektyviai spręsti autorystės identifikavimo uždavinį, svarbu atlikti išankstinį autorių tekstų apdorojimą, kurio metu yra pašalinami tokie teksto elementai kaip: simboliai, didžiosios raidės bei skaičiai. Papildomas tekstų apdorojimas ypač pradiniuose eksperimentuose galėtų pagerinti tyrimų rezultatus, nes šis etapas padeda išgryninti tekstus ir juos paruošti sekantiems etapams, kuriems tikslingai apdoroti tekstai turi tik didesnės pridėtinės vertės.
2. Kadangi eksperimento metu su turimais resursais maksimalus automatinis savybių generavimo skaičius siekė 11000, o atliekant tyrimą buvo pastebėta tendencija, nuo didesnio skaičiaus savybių priklauso ir geresnių rezultatų pasiekimas, tad turint didesnius resursus, būtų galima išbandyti dar didesnį skaičių savybių su minėtais dažniais naudojant mašininio mokymosi modelį. Šis resursų kiekio pasitelkimas galėtų padidinti rezultatus dar keletu procentų.
3. Siekiant pagerinti eksperimentų rezultatus, tai būtų galima padaryti parenkant ir koreguojanti neuronų tinklų hyperparametrus.

LITERATŪRA

1. Kaati, L., Lundeqvist, E., Shrestha, A., Svensson, M. (2017). Author Profiling in the Wild. *European Intelligence and Security Informatics Conference (EISIC).* 155-158. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8240786
2. Pokorny, F. B., Graf, F., Pernkopf, F., Schuller B. W. (2015). Detection of Negative Emotions in Speech Signals Using Bags-of-Audio-Words**.** *International Conference on Affective Computing and Intelligent Interaction (ACII)*. 879-884.a. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7344678
3. Beke, A. (2018). Forensic speaker profiling in a Hungarian speech corpus. *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)***.** 379-384. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8639932
4. Kapočiūtė-Dzikicnė, J., Damaševičius, R. (2018). Lithuanian Author Profiling with the Deep Learning. *Federated Conference on Computer Science and Information Systems (FedCSIS).* 169-172. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8511159
5. Yalamanchili, B., Kota S. N., Abbaraju, M. S., Nadella, S. S., Alluri, S. V. (2020). Real-time Acoustic based Depression Detection using Machine Learning Techniques. *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE).* 1-6. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9077698
6. Asni, M., Shapiro. D., Bolic, M., Mathew, T., Grebler, L. (2018). Speaker Differentiation Using A Convolutional Autoencoder. *IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE).* 1-5. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8646307
7. Majumder, N., Poria, S., Gelbukh, A., Cambria, E. (2017). Deep Learning-Based Document Modeling for Personality Detection from Text. *IEEE Intelligent Systems.* Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7887639
8. Tsung-Hsien Yang, Chung-Hsien Wu, Kun-Yi Huang, Ming-Hsiang Su. (2016). Detection of Mood Disorder Using Speech Emotion Profiles and LSTM**.** *International Symposium on Chinese Spoken Language Processing (ISCSLP).* 1-5. Prieiga per internetą: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7918439
9. Wang, B., Sun, Y. (2020). SSM-Seq2Seq: A Novel Speaking Style Neural Conversation Model. *Journal of Physics: Conference Series*. 1-11. Prieiga per internetą: https://iopscience.iop.org/article/10.1088/1742-6596/1576/1/012001/meta
10. Joge, D. S., Shirsat, A. S. (2016). Different language recognition model. *International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT).* 1096 – 1101. Prieiga per internetą: https://ieeexplore.ieee.org/document/7877756

11. Meluch, L., Tokarova, I., Farkaš, P., Schlinder, F. (2016). Simple Method Based on Complexity for Authorship Detection of Text. 2016 *39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO).* Prieiga per internetą: https://ieeexplore.ieee.org/document/7522349
12. Hurtado, J., Taweewitchakreeya, N., Zhu, X. (2014). Who wrote this paper? Learning for authorship de-identification using stylometric featuress. *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/7051981
13. Ding, S., Fung, M., Iqbal, F., Cheung, W. (2017). Learning Stylometric Representations for Authorship Analysis. *IEEE Transactions on Cybernetic*s. 107-1*21*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8116753
14. Zhao, C., Song, W., Liu, X., Liu, L., Zhao, X. (2017). Research on Author Identification Based on Deep Syntactic Features. 2017 *10th International Symposium on Computational Intelligence and Design (ISCID)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8275769
15. Houvardas, J., Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. *Lecture Notes in Computer Science.* 77-86. Prieiga per internetą: https://www.researchgate.net/publication/221655968_N-Gram_Feature_Selection_for_Authorship_Identification
16. Benzebouchi, N. E., Azizi N., Hammami, N. E., Schwab, D., Khelaifia, M. C. E., Aldwairi, M. (2019). Authors' Writing Styles Based Authorship Identification System Using the Text Representation Vector. *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8894872
17. Mohsen, A. M., El-Makky, N. M., Ghanem, N. (2016). Author Identification Using Deep Learning. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/7838265
18. Baseer, F., Jaafar, J., Habib, A. (2019). Gender and Age Identification Through Romanized Urdu Dataset. 2019 *1st International Conference on Artificial Intelligence and Data Sciences (AiDAS)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8971016
19. Jain, A., Kulkarni, G., Shah, S. (2018). Natural Language Processing. *International journal of computer sciences and engineering.* 161-167. Prieiga per internetą: https://www.myecole.it/biblio/wp-content/uploads/2020/11/7semt_2018-Natural_Language_Processing.pdf
20. Sidorov, I., Velaques, F., Stamatatos, E., Gelbukh, A., Chanona-Hernández, L. (2014). Syntactic N-grams as machine learning features for natural language processing. Exper*ts Systems with Applications. 853*-860. Prieiga per internetą: https://www.sciencedirect.com/science/article/abs/pii/S0957417413006271
21. Thanaki, J. (2017). *Python Natural Language Processing*: vadovėlis. Packt, Birmingham-Mumbai. 221-228 p.

22. Zhao, C., Song, W., Liu, X., Liu, L., Zhao, X. (2018). Research on Authorship Attribution of Article Fragments via RNNs. *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8663814
23. Schaetti, N. (2019). Behaviors of Reservoir Computing Models for Textual Documents Classification. *2019 International Joint Conference on Neutral Networks (IJCNN)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8852304
24. Benzebouchi, N. E., Azizi, N., Aldwairi, M., Farah, N. (2018). Multi-classifier system for authorship verification task using word embeddings. *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*. Prieiga per internetą: https://ieeexplore.ieee.org/document/8374391
25. Boenninghoff, B., Zeiler, S., Nickel, M. R., Kolossa, D. (2020). Variational Autoencoder with Embedded Student-t Mixture Model for Authorship Attribution. *Proceedings of the 28th International Conference on Computational Linguistics.* 519–529. Prieiga per internetą: https://www.aclweb.org/anthology/2020.coling-main.45.pdf
26. Song, W., Zhao, Ch., Liu, L. (2019). Multi-Task Learning for Authorship Attribution via Topic Approximation and Competitive Attention. *IEEE Access*. 177114-177121. Prieiga per internetą: https://ieeexplore.ieee.org/document/8930008
27. Sadiq, R., Rodriquez, J. R., Mian, R. H. (2019). Empirical Models to Predict Disinfection By-Products (DBPs) in Drinking Water: An Updated Review. SinceDirect. *Encyclopedia of Environmental Health. (Second Edition)* 324-338. Prieiga per internetą: https://www.sciencedirect.com/science/article/pii/B9780124095489111935
28. LeCun, Y., Bengio, Y., Hilton, G. (2015). Deep Learning. Nature. Prieiga per internetą: https://www.nature.com/articles/nature14539
29. Song, W., Zhao, Ch., Liu, L. (2019). Multi-Task Learning for Authorship Attribution via Topic Approximation and Competitive Attention. *IEEE Access*. 177114-177121. Prieiga per internetą: https://ieeexplore.ieee.org/abstract/document/9194070
30. Venugopal, A., Madanan, M., Venugopal, A. (2020). Multilayer Perceptron Approach for Character Classification. *European Journal of Molecular & Clinical Medicine*, 1283-1289p. Prieiga per internetą: https://ejmcm.com/article_4964.html
31. Hamedi, S., Kordrostami, Z. (2019). Artificial neural network approaches for modeling absorption spectrum of nanowire solar cells. *Neural Computing and Applications*. 8985–8995. Prieiga per internetą: https://www.researchgate.net/publication/335059192
32. Tabian, I., Fu, H., Khaodaei, S. Z. (2019). A Convolutional Neural Network for Impact Detection and Characterization of Complex Composite Structures. *Internet of Things for Structural Health Monitoring*. Prieiga per internetą: https://www.mdpi.com/1424-8220/19/22/49333
33. Aggarwal, C. (2018). *Neural Networks and Deep Learning*: vadovėlis. Springer Cham. 54-56 p.
34. *Aprašymas* [interaktyvus]. (2015). Understanding LSTM [žiūrėta 2021 m. gruodžio 11 d.]. Prieiga per internetą: http://colah.github.io/posts/2015-08-Understanding-LSTMs/
35. Ikeuchi, K. (2021). *Computer Vision*: vadovėlis. Springer Nature, Switzerland. 111-116 p.

36. Kapočiūtė-Dzikicnė, J., Utka, U., Šarkutė, L. (2016). Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. *Kalbotyra*. 1392-1517p. Prieiga per internetą: https://www.researchgate.net/publication/303846166

37. *Aprašymas* [interaktyvus]. (2021). What is Colaboratory? [žiūrėta 2021 m. gruodžio 21 d.]. Prieiga per internetą: https://research.google.com/colaboratory/faq.html

PRIEDAI

A priedas. Lietuvių kalbos tekstų identifikavimo eksperimentinio simbolių savybių tyrimo rezultatai.

features N	n_gra0ms	mpl	bidirectional_LSTM	CNN
1000	1_1	loss: 4.0063 - accuracy: 0.1355	loss: 3.8103 - accuracy: 0.1546	loss: 3.8441 - accuracy: 0.1555
	1_2	loss: 3.0777 - accuracy: 0.3130	loss: 2.4313 - accuracy: 0.4255	loss: 2.1387 - accuracy: 0.5018
	1_3	loss: 1.7990 - accuracy: 0.5622	loss: 1.8999 - accuracy: 0.5419	loss: 1.7532 - accuracy: 0.5848
	1_4	loss: 2.4636 - accuracy: 0.4234	loss: 1.9955 - accuracy: 0.5211	loss: 1.8597 - accuracy: 0.5686
	1_5	loss: 1.8004 - accuracy: 0.5601	loss: 2.0634 - accuracy: 0.4995	loss: 1.9109 - accuracy: 0.5408
	2_2	loss: 2.2307 - accuracy: 0.4656	loss: 2.1846 - accuracy: 0.4782	loss: 2.0767 - accuracy: 0.5053
	2_3	loss: 1.6609 - accuracy: 0.5873	loss: 1.8681 - accuracy: 0.5467	loss: 1.7693 - accuracy: 0.5810
	2_4	loss: 1.7181 - accuracy: 0.5744	loss: 2.0024 - accuracy: 0.5131	loss: 1.8476 - accuracy: 0.5633
	2_5	loss: 1.8278 - accuracy: 0.5496	loss: 2.0240 - accuracy: 0.5115	loss: 1.8829 - accuracy: 0.5650
	3_3	loss: 1.5569 - accuracy: 0.6154	loss: 1.7972 - accuracy: 0.5718	loss: 1.7498 - accuracy: 0.5873
	3_4	loss: 1.6998 - accuracy: 0.5745	loss: 1.9245 - accuracy: 0.5430	loss: 1.8157 - accuracy: 0.5698
	3_5	loss: 1.9862 - accuracy: 0.5182	loss: 2.0216 - accuracy: 0.5041	loss: 1.8750 - accuracy: 0.5665
	4_4	loss: 1.5925 - accuracy: 0.6087	loss: 1.7865 - accuracy: 0.5742	loss: 1.6484 - accuracy: 0.6017
	4_5	loss: 1.7213 - accuracy: 0.5771	loss: 1.9382 - accuracy: 0.5319	loss: 1.7977 - accuracy: 0.5639
	5_5	loss: 1.5991 - accuracy: 0.6096	loss: 1.7334 - accuracy: 0.5732	loss: 1.6490 - accuracy: 0.6007
3000	1_2	loss: 2.2120 - accuracy: 0.4714	loss: 2.2053 - accuracy: 0.4818	loss: 2.1038 - accuracy: 0.5100
	1_3	loss: 2.3626 - accuracy: 0.4605	loss: 1.6051 - accuracy: 0.6127	loss: 1.4804 - accuracy: 0.6518
	1_4	loss: 1.6129 - accuracy: 0.6029	loss: 1.7360 - accuracy: 0.5748	loss: 1.5803 - accuracy: 0.6496
	1_5	loss: 1.3831 - accuracy: 0.6539	loss: 1.7000 - accuracy: 0.5895	loss: 1.5136 - accuracy: 0.6375
	2_2	loss: 2.4353 - accuracy: 0.4086	loss: 2.1100 - accuracy: 0.5006	loss: 2.0697 - accuracy: 0.5188
	2_3	loss: 1.3793 - accuracy: 0.6522	loss: 1.6797 - accuracy: 0.5961	loss: 1.4671 - accuracy: 0.6586

features N	n_gra 0ms	mpl	bidirectional_LSTM	CNN
	2_4	loss: 1.4096 - accuracy: 0.6463	loss: 1.6543 - accuracy: 0.5952	loss: 1.4592 - accuracy: 0.6489
	2_5	loss: 1.4061 - accuracy: 0.6477	loss: 1.7262 - accuracy: 0.5827	loss: 1.4538 - accuracy: 0.6553
	3_3	loss: 1.4240 - accuracy: 0.6619	loss: 1.5868 - accuracy: 0.6255	loss: 1.5164 - accuracy: 0.6511
	3_4	loss: 1.3672 - accuracy: 0.6568	loss: 1.6229 - accuracy: 0.6028	loss: 1.5125 - accuracy: 0.6591
	3_5	loss: 1.3984 - accuracy: 0.6522	loss: 1.7249 - accuracy: 0.5831	loss: 1.4545 - accuracy: 0.6494
	4_4	loss: 1.5629 - accuracy: 0.6165	loss: 1.5151 - accuracy: 0.6361	loss: 1.3598 - accuracy: 0.6784
	4_5	loss: 1.3873 - accuracy: 0.6510	loss: 1.6650 - accuracy: 0.6059	loss: 1.4220 - accuracy: 0.6566
	5_5	loss: 1.3500 - accuracy: 0.6714	loss: 1.5514 - accuracy: 0.6346	loss: 1.3557 - accuracy: 0.6737
5000	1_3	loss: 1.9119 - accuracy: 0.5525	loss: 1.6049 - accuracy: 0.6226	loss: 1.5131 - accuracy: 0.6630
	1_4	loss: 1.4632 - accuracy: 0.6398	loss: 1.5893 - accuracy: 0.6212	loss: 1.4564 - accuracy: 0.6683
	1_5	loss: 1.9678 - accuracy: 0.5466	loss: 1.8359 - accuracy: 0.5918	loss: 1.4684 - accuracy: 0.6646
	2_3	loss: 2.4981 - accuracy: 0.4321	loss: 1.6421 - accuracy: 0.6126	loss: 1.4656 - accuracy: 0.6579
	2_4	loss: 1.3895 - accuracy: 0.6658	loss: 1.6662 - accuracy: 0.6021	loss: 1.4534 - accuracy: 0.6702
	2_5	loss: 1.4142 - accuracy: 0.6524	loss: 1.7294 - accuracy: 0.5907	loss: 1.4817 - accuracy: 0.6587
	3_3	loss: 1.4791 - accuracy: 0.6521	loss: 1.6700 - accuracy: 0.6183	loss: 1.3984 - accuracy: 0.6661
	3_4	loss: 1.4026 - accuracy: 0.6660	loss: 1.6462 - accuracy: 0.6162	loss: 1.4232 - accuracy: 0.6768
	3_5	loss: 1.4484 - accuracy: 0.6430	loss: 1.7165 - accuracy: 0.5947	loss: 1.4777 - accuracy: 0.6569
	4_4	loss: 1.3683 - accuracy: 0.6746	loss: 1.4936 - accuracy: 0.6384	loss: 1.2778 - accuracy: 0.6901
	4_5	loss :1.3932 - accuracy: 0.6622	loss: 1.6805 - accuracy: 0.6055	loss: 1.4168 - accuracy: 0.6730
	5_5	loss: 1.4298 - accuracy: 0.6686	loss: 1.5222 - accuracy: 0.6363	loss: 1.2751 - accuracy: 0.6928
7000	1_3	loss: 1.5137 - accuracy: 0.6390	loss: 1.7208 - accuracy: 0.5974	loss: 1.6686 - accuracy: 0.6395
	1_4	loss: 1.3342 - accuracy: 0.6815	loss: 1.6400 - accuracy: 0.6218	loss: 1.4795 - accuracy: 0.6775
	1_5	loss: 1.3623 - accuracy: 0.6678	loss: 1.7376 - accuracy: 0.5980	loss: 1.4637 - accuracy: 0.6742
	2_3	loss: 1.5442 - accuracy: 0.6225	loss: 1.7141 - accuracy: 0.6104	loss: 1.6342 - accuracy: 0.6454

features N	n_gra0ms	mpl	bidirectional_LSTM	CNN
	2_4	loss: 1.3733 - accuracy: 0.6734	loss: 1.6329 - accuracy: 0.6251	loss: 1.3142 - accuracy: 0.6805
	2_5	loss: 1.3591 - accuracy: 0.6757	loss: 1.7307 - accuracy: 0.5947	loss: 1.4209 - accuracy: 0.6740
	3_3	loss: 1.4915 - accuracy: 0.6467	loss: 1.6972 - accuracy: 0.6127	loss: 1.3994 - accuracy: 0.6657
	3_4	loss: 1.3868 - accuracy: 0.6704	loss: 1.6890 - accuracy: 0.6221	loss: 1.4468 - accuracy: 0.6771
	3_5	loss: 1.3441 - accuracy: 0.6711	loss: 1.6965 - accuracy: 0.6033	loss: 1.4300 - accuracy: 0.6799
	4_4	loss: 1.4633 - accuracy: 0.6691	loss: 1.4959 - accuracy: 0.6445	loss: 1.2926 - accuracy: 0.6924
	4_5	loss: 1.3614 - accuracy: 0.6733	loss: 1.6251 - accuracy: 0.6150	loss: 1.4066 - accuracy: 0.6900
	5_5	loss: 1.3767 - accuracy: 0.6725	loss: 1.5066 - accuracy: 0.6417	loss: 1.2654 - accuracy: 0.6948
9000	1_3	loss: 1.5112 - accuracy: 0.6414	loss: 1.7909 - accuracy: 0.5970	loss: 1.5052 - accuracy: 0.6536
	1_4	loss: 2.4950 - accuracy: 0.4540	loss: 1.5976 - accuracy: 0.6195	loss: 1.3147 - accuracy: 0.6943
	1_5	loss: 1.4216 - accuracy: 0.6557	loss: 1.7015 - accuracy: 0.6059	loss: 1.3116 - accuracy: 0.6915
	2_3	loss: 1.6339 - accuracy: 0.6108	loss: 1.6972 - accuracy: 0.6080	loss: 1.4556 - accuracy: 0.6558
	2_4	loss: 1.3835 - accuracy: 0.6639	loss: 1.6783 - accuracy: 0.6192	loss: 1.4478 - accuracy: 0.6832
	2_5	loss: 1.3890 - accuracy: 0.6754	loss: 1.6478 - accuracy: 0.6006	loss: 1.2761 - accuracy: 0.6882
	3_3	loss: 1.5352 - accuracy: 0.6375	loss: 1.7076 - accuracy: 0.6176	loss: 1.3802 - accuracy: 0.6638
	3_4	loss: 1.3948 - accuracy: 0.6691	loss: 1.5576 - accuracy: 0.6315	loss: 1.2733 - accuracy: 0.6915
	3_5	loss: 1.3547 - accuracy: 0.6695	loss: 1.6432 - accuracy: 0.6085	loss: 1.2656 - accuracy: 0.6961
	4_4	loss: 1.3503 - accuracy: 0.6776	loss: 1.5435 - accuracy: 0.6421	loss: 1.2510 - accuracy: 0.7002
	4_5	loss: 1.3657 - accuracy: 0.6814	loss: 1.6229 - accuracy: 0.6210	loss: 1.2503 - accuracy: 0.6993
	5_5	loss: 1.3481 - accuracy: 0.6778	loss: 1.5416 - accuracy: 0.6466	loss: 1.2670 - accuracy: 0.7029
11000	1_3	loss: 1.6519 - accuracy: 0.6073	loss: 1.7981 - accuracy: 0.5937	loss: 1.4951 - accuracy: 0.6510
	1_4	loss: 1.9174 - accuracy: 0.5721	loss: 1.6512 - accuracy: 0.6131	loss: 1.3758 - accuracy: 0.6872
	1_5	loss: 1.4427 - accuracy: 0.6533	loss: 1.7240 - accuracy: 0.6149	loss: 1.2662 - accuracy: 0.7019
	2_3	loss: 1.5702 - accuracy: 0.6403	loss: 1.7385 - accuracy: 0.6002	loss: 1.4848 - accuracy: 0.6551

features N	n_grams	mpl	bidirectional_LSTM	CNN
	2_4	loss: 1.3738 - accuracy: 0.6704	loss: 1.7284 - accuracy: 0.6251	loss: 1.3087 - accuracy: 0.6916
	2_5	loss: 1.3335 - accuracy: 0.6820	loss: 1.7462 - accuracy: 0.6060	loss: 1.1932 - accuracy: 0.7068
	3_3	loss: 1.5145 - accuracy: 0.6503	loss: 1.7362 - accuracy: 0.6092	loss: 1.3919 - accuracy: 0.6641
	3_4	loss: 1.3980 - accuracy: 0.6757	loss: 1.6296 - accuracy: 0.6251	loss: 1.2807 - accuracy: 0.6946
	3_5	loss: 1.3629 - accuracy: 0.6795	loss: 1.6109 - accuracy: 0.6185	loss: 1.2926 - accuracy: 0.6903
	4_4	loss: 1.3928 - accuracy: 0.6745	loss: 1.5906 - accuracy: 0.6361	loss: 1.3398 - accuracy: 0.6897
	4_5	loss: 1.3787 - accuracy: 0.6796	loss: 1.6263 - accuracy: 0.6165	loss: 1.2372 - accuracy: 0.6996
	5_5	loss: 1.3951 - accuracy: 0.6768	loss: 1.6055 - accuracy: 0.6391	loss: 1.2900 - accuracy: 0.7004

B priedas. Lietuvių kalbos tekstų identifikavimo eksperimentinio žodžių savybių tyrimo rezultatai.

features N	n_grams	mpl	bidirectional_LSTM	CNN
1000	1_1	loss: 1.4827 - accuracy: 0.6402	loss: 1.5793 - accuracy: 0.6166	loss: 1.4674 - accuracy: 0.6417
	1_2	loss: 1.5312 - accuracy: 0.6352	loss: 1.5689 - accuracy: 0.6157	loss: 1.4653 - accuracy: 0.6396
	1_3	loss: 1.4922 - accuracy: 0.6319	loss: 1.5926 - accuracy: 0.6070	loss: 1.4611 - accuracy: 0.6369
	2_2	loss: 2.2074 - accuracy: 0.4870	loss: 2.2232 - accuracy: 0.4759	loss: 2.1725 - accuracy: 0.4894
	3_3	loss: 3.1095 - accuracy: 0.3088	loss: 3.0589 - accuracy: 0.3213	loss: 3.0908 - accuracy: 0.3217
3000	1_1	loss: 1.3896 - accuracy: 0.6727	loss: 1.4984 - accuracy: 0.6449	loss: 1.2844 - accuracy: 0.6916
	1_2	loss: 1.3523 - accuracy: 0.6776	loss: 1.4306 - accuracy: 0.6566	loss: 1.4306 - accuracy: 0.6566
	1_3	loss: 1.3939 - accuracy: 0.6640	loss: 1.4402 - accuracy: 0.6572	loss: 1.1879 - accuracy: 0.7035
	2_2	loss: 2.0112- accuracy: 0.5510	loss: 2.7852- accuracy: 0.3702	loss: 1.8595- accuracy: 0.5620
	3_3	loss: 2.8805- accuracy: 0.3630	loss: 2.7852- accuracy: 0.3702	loss: 2.8764 - accuracy: 0.3802
5000	1_1	loss: 1.5321 - accuracy: 0.6621	loss: 1.5542 - accuracy: 0.6401	loss: 1.2521 - accuracy: 0.6932
	1_2	loss: 1.3262 - accuracy: 0.6773	loss: 1.5379 - accuracy: 0.6522	loss: 1.1803 - accuracy: 0.7083
	1_3	loss: 1.3838 - accuracy: 0.6830	loss: 1.4986 - accuracy: 0.6622	loss: 1.1680 - accuracy: 0.7181

features N	n_grams	mpl	bidirectional_LSTM	CNN
	2_2	loss: 2.1120- accuracy: 0.5493	loss: 2.0623- accuracy: 0.5396	loss: 1.8213- accuracy: 0.5820
	3_3	loss: 2.8952- accuracy: 0.3812	loss: 2.8232- accuracy: 0.3826	loss: 2.8621- accuracy: 0.4085
7000	1_1	loss: 1.4023- accuracy: 0.6625	loss: 1.5941- accuracy: 0.6225	loss: 1.2892- accuracy: 0.6921
	1_2	loss: 1.4311 - accuracy: 0.6834	loss: 1.4538 - accuracy: 0.6571	loss: 1.2041 - accuracy: 0.7144
	1_3	loss: 1.4453 - accuracy: 0.6818	loss: 1.4376 - accuracy: 0.6666	loss: 1.1399 - accuracy: 0.7275
	2_2	loss: 1.8767 - accuracy: 0.5703	loss: 2.0121 - accuracy: 0.5395	loss: 1.8201 - accuracy: 0.6007
	3_3	loss: 3.0423 - accuracy: 0.3798	loss: 2.8142 - accuracy: 0.3901	loss: 2.8462 - accuracy: 0.4186
9000	1_1	loss: 1.4353 - accuracy: 0.6598	loss: 1.6324 - accuracy: 0.6125	loss: 1.2094- accuracy: 0.7027
	1_2	loss: 1.3105 - accuracy: 0.6877	loss: 1.4572 - accuracy: 0.6686	loss: 1.0952 - accuracy: 0.7378
	1_3	loss: 1.3600 - accuracy: 0.6797	loss: 1.4522 - accuracy: 0.6661	loss: 1.1616 - accuracy: 0.7280
	2_2	loss: 1.9077 - accuracy: 0.5693	loss: 2.0698 - accuracy: 0.5328	loss: 1.7838 - accuracy: 0.6083
	3_3	loss: 2.7125 - accuracy: 0.4027	loss: 2.8652 - accuracy: 0.3968	loss: 2.8562 - accuracy: 0.4189
11000	1_1	loss: 1.4138 - accuracy: 0.6623	loss: 1.6392 - accuracy: 0.6238	loss: 1.2387 - accuracy: 0.7103
	1_2	loss: 1.3156 - accuracy: 0.6893	loss: 1.5324 - accuracy: 0.6578	loss: 1.1067 - accuracy: 0.7397
	1_3	loss: 1.3686 - accuracy: 0.6799	loss: 1.5351 - accuracy: 0.6539	loss: 1.0666 - accuracy: 0.7404
	2_2	loss: 1.9051 - accuracy: 0.5692	loss: 2.1056 - accuracy: 0.5436	loss: 1.8390 - accuracy: 0.6044
	3_3	loss: 2.7437 - accuracy: 0.4103	loss: 2.7918 - accuracy: 0.3923	loss: 2.8610 - accuracy: 0.4247