KAUNO TECHNOLOGIJOS UNIVERSITETAS

INFORMATIKOS FAKULTETAS

Arnas Karpavičius

ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMAS

Baigiamasis magistro darbas

Vadovas
Prof. dr. Algimantas Venčkauskas

KAUNAS, 2016

KAUNO TECHNOLOGIJOS UNIVERSITETAS
INFORMATIKOS FAKULTETAS
KOMPIUTERIŲ KATEDRA

ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMAS

Baigiamasis magistro darbas
Informacijos ir informacinių technologijų sauga (kodas 621E10003)

Vadovas
(parašas) Prof. dr. Algimantas Venčkauskas
(data)

Recenzentas
(parašas) Doc. dr. Romas Marcinkevičius
(data)

Projektą atliko
(parašas) Arnas Karpavičius
(data)

KAUNAS, 2016

KAUNO TECHNOLOGIJOS UNIVERSITETAS

Informatikos fakultetas

(Fakultetas)

Arnas Karpavičius

(Studento vardas, pavardė)

Informacijos ir informacinių technologijų sauga, 621E10003

(Studijų programos pavadinimas, kodas)

„Elektroninio diskurso autorystės nustatymas“

AKADEMINIO SĄŽININGUMO DEKLARACIJA

20 ___ m. _____ ____ d.

Kaunas

Patvirtinu, kad mano **Arno Karpavičiaus**baigiamasis projektas tema „Elektroninio diskurso autorystės nustatymas“ yra parašytas visiškai savarankiškai, o visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka)	*(parašas)*

Karpavičius, A.„Elektroninio diskurso autorystės nustatymas“. Magistro baigiamasis projektas / vadovas prof. dr. Algimantas Venčkauskas; Kauno technologijos universitetas, Informatikos fakultetas, Kompiuterių katedra.

Kaunas, 2016. 56 p.

SANTRAUKA

Dėl nuolat augančio, anonimiškai įvykdomų kibernetinių nusikaltimų skaičiaus, elektroninio diskurso autorystės nustatymas tampa aktualia šių dienų problema. Sėkmingas ir efektyvus nežinomo, elektronine forma parašyto teksto autoriaus išaiškinimas dažnai yra itin svarbus aspektas, tiriant nusikaltimus interneto erdvėje. Yra apibrėžta daugybė nuo kalbos nepriklausančių bendrųjų lingvistinių požymių bei charakteristikų, skirtų anoniminių tekstų, parašytų bet kokia kalba (su išimtimis), autorių tapatybei nustatyti. Taip pat egzistuoja daug metodų ir sistemos prototipų, skirtų anglų kalba parašytų tekstų autorystei nustatyti, naudojant pastaruosius požymius ir charakteristikas su atitinkamais papildymais anglų kalbai. Daugelyje kitų nacionalinių kalbų, ši sritis nėra taip gerai išvystyta, tačiau pastaruoju metu yra sparčiai tobulinama.

Šiame darbe nagrinėjami įvairūs metodai, priemonės bei lingvistinės charakteristikos, leidžiančios nustatyti anoniminio teksto autoriaus tapatybę pagal turimą identifikuotų autorių tekstų duomenų bazę. Aiškinamasis jų pritaikomumas ir efektyvumas lietuvių kalbai, siūlomi papildomi, nuo kalbos priklausantys požymiai, bei tiriamas jų efektyvumas. Taip pat darbe siūlomas inovatyvus metodas bei sistemos prototipas, skirtas lietuvių kalba parašytų anoniminių tekstų autorystei nustatyti, kuriame naudojami bendrieji bei papildomi, tik lietuvių kalbai būdingi, požymiai.

Siūlomos sistemos bandymams buvo naudojama 200 skirtingų autorių tekstų duomenų bazė. Tyrimai buvo atlikti naudojant tik bendruosius lingvistinius požymius bei bendruosius ir lietuvių kalbai būdingus požymius kartu. Atlikti eksperimentai parodė, kad papildomi, tik lietuvių kalbai būdingi, požymiai ženkliai pagerina autorystės nustatymo tikslumą nei vien tik naudojami bendrieji požymiai.

Reikšminiai žodžiai: autorystės nustatymas, teksto analizė, teksto gavyba, nacionalinės kalbos, kriminalistinė lingvistika, ekspertinė sistema.

Karpavičius, Arnas. *Electronic Discourse Authorship Verification*: *Master's* thesis / supervisor assoc. prof. Algimantas Venčkauskas. The Faculty of Informatics, Kaunas University of Technology.

Research area and field: Information and Information Technology Security

Key words: authorship identification, text analysis, text mining, national languages, forensic linguistics, expert system.

Kaunas, 2016. 56 p.

SUMMARY

Electronic discourse authorship identification becomes a significant nowadays problem due to an increasing number of anonymous cybercrimes. Successful and effective authorship revealing of an unknown text, written in electronic form, is often an especially important aspect for investigation process of internet cybercrimes. A lot of language independent general linguistic features and characteristics are already defined and intended for the authorship identification process of anonymous texts, written in any language (with exceptions). Also many methods and system prototypes, intended for the authorship identification of texts, written in English language, exist, which use these features and characteristics along with corresponding additions to English language. For the majority of other national languages, this area is not as well developed, but recently is being actively improved.

In this work, various methods, tools and linguistic characteristics, which allow to determine the authorship's identity of an anonymous text, according to a present database of identified authors' texts, are examined. Their applicability to Lithuanian language and the effectiveness are explored, as well asadditional, language dependent features are suggested and tested for their effectiveness. Also the innovative method and system prototype for the authorship identification of anonymous texts, written in Lithuanian language, are suggested, where general and additional, Lithuanian language specific, features are used.

For the experiments of the suggested system, a database with the texts of 200 different authors was used. The experiments were carried out using general linguistic features only, and general with Lithuanian language specific features together. The experiments proved that the additional, Lithuanian language specific, features significantly improve the accuracy of authorship identification rather than the general features used only.

Keywords: authorship identification, text analysis, text mining, national languages, forensic linguistics, expert system.

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

TERMINŲ IR SANTRUMPŲ ŽODYNAS

Terminas anglų k.	*Lietuviškas terminas*	*Paaiškinimas*
C4.5	C4.5 algoritmas	Klasifikavimo algoritmas apmokymo metu sukuriantis sprendimų medžius.
Emoticon	Jausmaženklis	Ženklų rinkinys emocijoms išreikšti.
Kolmogorov Complexity (KC)	Kolmogorovo sudėtingumas	Objekto aprašo (šioje ataskaitoje - teksto) suglaudinamumo matas, apibūdinantis aprašo sudėtingumą.
Latent Semantic Indexing (LSI)	Latentinis semantinis indeksavimas	Tekstinės informacijos klasifikavimo pagal koncepcijas metodas
n-gram		Šalia einančių n raidžių arba žodžių seka tam tikrame tekste.
Part-of-Speech (POS)	Morfologinė forma	Mažiausia reikšminė žodžio dalis.
Principal Component Analysis (PCA)	Principinių komponentų analizė	Duomenų dimensijų mažinimo metodas, naudojantis duomenų kovariacijos matricą.
Short Message Service (SMS)	Trumpoji žinutė	Daugelyje mobiliųjų telefonų esanti paslauga, skirta trumpųjų žinučių siuntimui tarp mobiliųjų telefonų arba kitų panašių įrenginių.
Stylometry	Stilometrija	Kompiuterinė statistinė tekstų stilistikos analizė.
Support Vector Machine (SVM)	Atraminių vektorių klasifikatorius	Algoritmas, kuris transformuoja pradinius duomenis į aukštesnę dimensiją, kur randama hiperplokštuma, skirianti dvi klases, su kiek galima didesniu atstumu tarp klasifikuojamų duomenų. Radus šią hipreplokštumą, duomenis galima suklasifikuoti į dvi atskiras klases.
Textese	Trumpųjų žinučių kalba	Trumposiose žinutėse ir kituose elektroniniuose tekstuose naudojama kalba, kuriai būdingos įvairios santrumpos ir žargonizmai.
Unified Modelling Language (UML)	Vieninga modeliavimo kalba	Modeliavimo ir specifikacijų kūrimo kalba, skirta specifikuoti, atvaizduoti ir konstruoti objektiškai orientuotų programų dokumentus.
Writeprint	Teksto antspaudas	Kai kurių autorių naudojamas terminas, reiškiantis konkrečiam autoriui būdingas stilometrines teksto charakteristikas.

ĮVADAS

Šiame magistro baigiamajame darbe nagrinėjami įvairūs elektroninio diskurso autorystės nustatymo metodai, skirti identifikuoti nežinomo autoriaus tekstą, parašytą tiek anglų, tiek kitomis nacionalinėmis kalbomis, taip pat nagrinėjami ir nuo kalbos nepriklausantys autorystės nustatymo metodai. Darbe siūlomas lietuvių kalbai pritaikytas metodas, kuriame naudojami ir bendrieji, ir lietuvių kalbai būdingi lingvistiniai požymiai bei kelių dimensiškumo sumažinimo ir mašinos mokymosi klasifikatorių architektūra.

Metodo realizacijoje sudaryta grafinė vartotojo sąsaja su galimybe pasirinkti keturis galimus klasifikavimo režimus, taip pat ir norimus lingvistinius požymius bei jų grupes.

Bandymu metu buvo naudojami 200 skirtingų autorių lietuviški tekstai, surinkti iš naujienų portalo komentarų. Atliekant eksperimentus, buvo naudojami bendrieji bei bendrieji ir lietuvių kalbai būdingi požymiai kartu, taip išsiaiškinant jų svarbą lietuviškų tekstų autorystės nustatymui. Taip pat buvo atskirai tiriami lietuvių kalbai būdingi požymiai ir jų grupės, taip nustatant kurie jų labiausiai įtakoja identifikavimo tikslumą. Vėliau buvo atlikti eksperimentai, aiškinantis tekstų ilgių bei jų padalinimo apmokymo bei tikrinimo procesams, įtaka autorystės nustatymui.

Darbo pabaigoje pateikiamos išvados bei idėjos tolimesniam šio darbo vystymui.

Tai Kauno Technologijos Universiteto, Informatikos fakulteto, Kompiuterių katedros, Informacijos ir informacinių technologijų saugos studijų programos studento Arno Karpavičiaus magistro baigiamasis darbas. Šis darbas buvo atliktas vykdant projektą „Inovatyvaus kibernetinių nusikaltimų daiktų internete tyrimo metodo sukūrimas ir tyrimas“.

Darbo problematika ir aktualumas

Internetas kibernetiniams nusikaltėliams tampa tinkama platforma anonimiškai vykdyti savo nusikalstamas veiklas, tokias kaip brukalo siuntinėjimas ir duomenų vagystės. Dėl šios priežasties pastaruoju metu anoniminių tekstų internete (pvz. el. laiškų, komentarų forumuose, žinučių socialiniame tinkle Twitter, trumpųjų žinučių) autorystės nustatymo analizė sulaukė kibernetinius nusikaltimus ir duomenų gavybą tiriančių bendruomenių dėmesio. Kriminalistinių tyrimų lingvistikosuždavinys yra teikti konsultacijas teisininkams, analizuojant kalba pagrįstus pėdsakus (tokius kaip anoniminiai ar tiriamieji nusikaltimams naudojami tekstai) per ikiteisminį tyrimą. Kriminalistinė tekstinių dokumentų internete analizė, skirta anonimiškumo problemai nagrinėti, vadinama autorystės analize. Autorystės tyrimas yra rašytinių dokumentų, kurių autoriai žinomi arba ne, kalbiniųir skaičiavimo ypatumų tyrimas. Jis apima rašymo stilių ar dokumento turinio stilometrines savybes.

Autorystės analize siekiama teisingai suklasifikuoti tekstus į klases pagal tekstų autorių pasirinktą stilistiką. Be autoriaus identifikacijos ir jo tapatybės patvirtinimo, kai tiriamas atskirų autorių stilius, autoriai taip pat gali būti skirstomi į klases pagal lytį, amžių, gimtąją kalbą ar asmenybės tipą.

Autorystės patvirtinimas padeda nustatyti, ar konkretus asmuo parašė nagrinėjamą dokumentą ,ar ne.

Rašymo stilius yra nesąmoningas asmens įprotis, kuris leidžia atskirti vieną autorių nuo kitų pagal tai, kokius komunikacijai skirtos kalbos žodžius, gramatiką ir kitus elementus autoriai naudoja. Patikimais autorystės rodikliais gali būti laikomi leksinės konfigūracijos ir žodyno vartosena (pvz. leksikos turtingumas) kompozicija ir rašymo forma, (pvz. išskirtiniai sintaksės ir struktūros išdėstymo bruožai),žodyno vartosenos modeliai, neįprasta kalbos vartosena, stiliaus ir substilių požymiai. Pagrindiniu iššūkiu, nustatant autorystę, tampa šių stiliaus požymių pakankamai tikslus atpažinimas ir analizė.

Šioje ataskaitoje pateikiama išsami apžvalga ir metodai autorystei nustatyti nacionalinės kalboselektroniniame diskurse.

Darbo tikslas ir uždaviniai

Šio darbo tikslas yra sudaryti ir ištirti lietuvių kalbai pritaikytą elektroninio diskurso autorystės nustatymo metodą.

Darbo uždaviniai:

- Atlikti išsamią esamų elektroninio diskurso autorystės nustatymo metodų bei sistemų analizę Lietuvos bei užsienio mastu.
- Suformuluoti reikalavimus kriminalistinėms autorystės nustatymo sistemoms.
- Sudaryti lietuvių kalbai pritaikytą elektroninio diskurso autorystės nustatymo metodą, naudojantį ir bendruosius lingvistinius požymius, ir tik lietuvių kalbai būdingus požymius.
- Ištirti sudaryto metodo efektyvumą, privalumus bei trūkumus.
- Suformuluoti išvadas bei pateikti idėjas tolimesniam šio darbo vystymui.

Darbo rezultatai ir jų svarba

Elektroninio diskurso autorystės nustatymo įrankis yra skirtas padėti kriminalistinės lingvistikos ekspertams identifikuoti nežinomo teksto autorių pagal turimų identifikuotų tekstų rinkinį. Tekstams tarpusavyje išskirti naudojami įvairūs lingvistiniai požymiai, pažymintys autoriaus rašymo stiliaus unikalumą. Įrankyje naudojami ne tik bendrieji, nuo kalbos nepriklausantys požymiai, bet ir specifiniai, tik lietuvių kalbai būdingi požymiai, todėl įrankis gali efektyviai identifikuoti lietuviškus tekstus. Programos nustatymuose yra galimybė pasirinkti, kurie požymiai ar požymių rinkiniai bus naudojami identifikavimo proceso metu.

Įrankyje yra naudojamas pažangus mašinos mokymosi klasifikatorius, leidžiantis efektyviai identifikuoti autorius pagal jų tekstuose esančius lingvistinius požymius. Taip pat įrankyje yra keturi galimi šio klasifikatoriaus klasifikavimo režimai, kurie skirtingai įtakoja identifikavimo tikslumą ir greitaveiką bei, kuriuos vartotojas gali naudoti pasirinktinai.

Įrankyje atskirai patalpinamas vieno nežinomo autoriaus tekstas, bei daug identifikuotų autorių tekstų. Rezultatuose įrankis parodo nežinomo autoriaus tekstui panašiausių identifikuotų autorių eilę, surikiuotą pagal panašumą. Taip pat šioje eilėje prie kiekvieno autoriaus yra parodomas panašumo įvertis, leidžiantis geriau įvertinti panašumą tarp skirtingų autorių.

Darbo struktūra

Šį dokumentą sudaro penki pagrindiniai skyriai:

- Elektroninio diskurso autorystės nustatymo metodų analizė
- Elektroninio diskurso autorystės nustatymo metodo sudarymas
- Elektroninio diskurso autorystės nustatymo metodo realizavimas
- Elektroninio diskurso autorystės nustatymo metodo bandymai
- Galutiniai rezultatai ir išvados

1. ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMO METODŲ ANALIZĖ

1.1. Analizės tikslas

Šios analizės tikslas yra atlikti išsamią tiek bendrosios paskirties, tiek atskiroms nacionalinėms kalboms skirtų elektroninio diskurso autorystės nustatymo metodų apžvalgą bei efektyvumo įvertinimą, nustatyti jų privalumus bei trūkumus ir padėti priimti tinkamus sprendimus tolimesnei šio darbo eigai.

1.2. Elektroninio diskurso autorystės nustatymo iššūkiai

Šiame skyriuje pateikiama trumpa autorystės nustatymo elektroniniame diskurse apžvalga, įskaitantelektroninių pranešimų lingvistinės savybes, atskirų kalbos dalių atpažinimą, tarpkalbinį autorystėsatpažinimą, didelį duomenų dimensiškumą, lingvistines elektroninių tekstų charakteristikas, kalbosdalių atpažinimą, tarpkalbinį autoriaus atpažinimą, duomenų daugiadimensiškumą ir galimusbandymus nuslėpti autorystę.

1.2.1. Elektroninių pranešimų lingvistinės savybės

Trumpi internete aptinkami tekstai pasižymi keliomis savybėmis, dėl kurių autorystės nustatymo problema tampa iššūkiu, lyginant su ilgesniais, formaliais tekstiniais dokumentais, tokiais kaip įvardijami literatūriniai kūriniai. Pagrindinės tokių tekstų savybės yra šios:

1. Internetiniai tekstai įprastai yra trumpi; tai rodo, kad tam tikra kalba grįsta metrika, priklausanti nuo žodžių tekste skaičiaus, gali būti netinkama.
2. Internetiniuose tekstuose gali būti daug gramatinių klaidų
3. Gali būti vartojamas specifinis kalbinis sociolektas (pvz. trumpųjų tekstinių žinučių kalba [1]), kuris gali apriboti mūsų taikomus žodyno ir leksinės analizės metodus.
4. Autoriaus naudojamas rašymo stilius gali kisti priklausomai nuo numatyto adresato.
5. Tam tikrose autorių bendruomenėse gali būti vartojami specifiniai terminai (pvz. technologijoms pritaikyti žodžiai)

Kai skaitmeniniuose kriminalistiniuose tyrimuose tradiciniai metodai yra taikomi trumposioms tekstinėmis žinutėmis, kyla įvairių problemų. Visų pirma, taikant pagrindinius standartinius metodus, požymiams išskirti ir suklasifikuoti reikia didelės apimties teksto. Nepaisant to, atliekant mobiliojo telefono kriminalistinius tyrimus, duomenys apie trumpąsias žinutes ir naudotojui išsiųstų žinučių skaičius dažniausiai būna riboti. Taip pat galimų autorių grupė gali būti didelė, dėl ko tampa sudėtinga klasifikuoti tekstus pagal jų autorystę. Be to, trumpųjų tekstinių žinučių struktūra neatitinka standartinės kalbai būdingos sintaksinės struktūros ir yra skirtinga skirtingiems vartotojams. Dėl šios priežasties sintaksinės savybės tampa neveiksmingos ir sumažėja klasifikacijai tinkančių požymių rinkinys. Galiausiai, atliekant trumpųjų tekstinių žinučių kriminalistinius tyrimus, reikia aukštos duomenų apdorojimo spartos. Nagrinėjant atvejus, tyrėjų laikas yra ribotas ir jie negali skirtijo labai daug įtariamam autoriui nustatyti.

Y.H.Segerstad [2] išskiria šias kalbines trumpųjų tekstinių žinučių savybes: nenaudojami skyrybos ženklai, netradicinė skyryba, vengtini tarpai tarp žodžių, šnekamąją kalbą primenanti rašyba, priebalsinėrašyba, tradiciniai ir netradiciniai trumpiniai, visos didžiosios arba mažosios raidės, ilgų žodžių pakeitimas trumpesniais, jausmaženkliai, asteriskai, žodžių, skyrybos ženklų keitimas simboliais,-C. Thurlow [3] nurodo šiuos: 1) trumpiniai (t.y. praleistos paskutinės raidės), kontrakcijos (t.y.trūkstamos raidės žodžio viduryje), bei sutrauktiniai žodžiai (t.y. paskutinio skiemens trūkumas), 2) akronimai ir iš pirmų žodžio raidžių sudaryti trumpiniai, 3) raidiniai/skaitiniai homofonai, 4)”netaisyklinga rašyba” ir tipografinės klaidos, 5) nestandartinė rašyba, 6) kirčių stilizacija.

Maclead ir Grantas [4] išskiria trumpųjų tekstinių žinučių kalbai būdingas savybes, tokias kaip žodžiai, parašyti klaidingai (bet koks žodis, neaptinkamas anglų kalbos žodyne), kirčių stilizacija (fonetinė rašyba naudojama išskirtiniam tarimui reikšti), prozodiniai kirčiai (rašyba išreiškiamas išskirtinis tarimas), viso žodžio pakeitimas raidiniu homofonu (viso žodžio pakeitimas viena raide), viso žodžio pakeitimas skaitiniu homofonu (viso žodžio pakeitimas skaičiumi), žodžio skiemenų keitimas skaitiniais trumpiniais (bendriniai žodžiai sutrumpinami iki kelių pirmų žodžio raidžių), jausmaženkliai (ženklųrinkinys emocijoms išreikšti) ir t.t.

1.2.2. Atskirų kalbos dalių atpažinimas

Morfologinių formų ir kalbos dalių (angl. POS) žymėjimas kiekvieną žodį tekste priskiria morfosintaksinei klasei, atsižvelgiant į to žodžio savybes ir kontekstą, kuriame yra naudojamas. POS žymėjimas dar gali būti vadinamas morfologiniu, žodžio klasės ar leksiniu žymėjimu [5]. Morfologinių formų žymėjimo metodai gali būti klasifikuojami įvairiai; tačiau visuotinė klasifikacija skirsto metodus į [6]:

1) statistinius metodus ir

2) taisyklėmis grindžiamus metodus.

Pagal faktinį ženklinimo procesą, POS žymėjimas gali būti skiriamas į šias kategorijas [6]:

1. Išgryninimo metodas (angl. *bold approach*) - visa informacija, kurią žymintysis naudoja, yrasujungiama ir pasirenkamas geriausiai atitinkantis sprendimas.

2. Apdairiu pasirinkimu grįstas metodas (angl. *cautious approach*) - visų pirma visi žodžiai pažymimi visomis įmanomomis POS žymenomis, tuomet netinkamos žymės pašalinamos dėl kontekstinių apribojimų.

3. Nenuspėjamumu paremtas metodas (angl. *whimsical approach*) - visų pirma, kiekvienam žodžiui yra priskiriama viena žymena, kuri, taikant transformavimo taisykles, pakeičiama, remiantis kontekstinėmis sąlygomis.

POS žymėjimas taip pat gali būti skirstomas į apačioje išvardintas kategorijas, atsižvelgiant į lingvistinės informacijos gavimo būdą [6]:

1. Mašininis mokymas - reikiama informacija gaunama automatiškai. Dauguma morfologinių formų žymėjimo metodų taiko save mokančias sistemas. Dauguma POS žymėjimo metodųtaiko save apsimokančias sistemas (pvz. atmintimi grįstą žymėjimą, tikimybinį žymėjimą, transformacinį žymėjimą, neuroninių tinklų žymėjimą). Mokymasis gali būti skirstomas į prižiūrimą mokymąsi (mokymuisi naudojami identifikuoti duomenys) ir neprižiūrimą mokymąsi (mokymuisi naudojami neapdoroti duomenys), nors geresni rezultatai gaunami taikant prižiūrimą mokymąsi.

2. Apdorojimas rankiniu būdu - POS žymėjimo sistemai keliami apdorojimo rankiniu būdureikalavimai.

Bendrą morfologinių formų architektūrą sudaro šie etapai [5]:

- Teksto skaidymas į žodžių/sakinių seką - tekstas suskaidomas į reikalingus segmentus (pvz.žodžiai, skyrybos ženklai, žodžių junginiai).
- Informacijos apie dviprasmybes paieška - leksikonas ir asmenys padeda ieškoti leksikone leksemų, kurios į leksikoną nėra įtrauktos (leksikonas gali būti suvokiamas kaip žodynas, kurį sudaro žodžių formos ir jas atitinkančios kalbos dalys ar baigtiniai modeliai). Parinkėjas analizuoja likusias leksemas, panaudodami informaciją apie leksikoną, pvz. kokio tipo duomenų leksikone yra ir kokių ne - jis taip pat gali padėti išsirinkti tam tikros rūšies veiklas. Leksikonas ir parinkėjas, naudojami kartu su kompiliatoriumi ar interpretatoriumi, sudaro morfologinį/leksinį analizatorių, kuris priskiria visas tinkamas morfologinių formų kategorijas kiekvienai leksemai. Taip pat sudurtinės gairės ir lemos gali būti naudojamos kaip žymėjimo kategorijos.
- Dviprasmybių raiška/dviprasmybių sprendimas - pritaikoma apie pačius žodžius turima informacija (pvz. tam tikra morfologinė žodžio forma yra dažnesnė nei kita,jei tažodžio forma priklauso kartotinėms morfologinėms formoms). Šiam etapui taip pat reikalingakontekstinė informacija arba informacija apie leksemų ir jų gairių sekas (pvz. norint

suteikti pirmenybę nuo kitų besiskiriančių morfologinių formų analizei, priklausančiai morfologinei formai, jei prieš tai buvusi leksemos gairė daug dažniau naudojama su leksema, kuri priklausotai morfologinei formai).

1.2.3. Tarpkalbinis autorystės atpažinimas

Tarpkalbinis autorystės priskyrimas išplečia įprastinį autorystės pasiskirstymą, siekiant išspręsti šioje srityje kylančias daugiakalbiškumo problemas [7]. Taip siekiama atpažinti viena kalba parašyto dokumento autorių, nors tekstas, priskiriamas tam pačiam autoriui, yra parašytas kita kalba.

Tarpkalbinis autorystės priskyrimas yra svarbus uždavinys, kadangi autoriai gali rašyti skirtingomis kalbomis, tokiu atveju paversdami įprastinį autorystės priskyrimą neveiksmingu. Tai naujas iki šiol dar nenagrinėtas uždavinys. Buvo atlikti keli tyrimai [8-10], nagrinėjantys tarpkalbinio plagiato aptiktį.

Tarpkalbinėje perspektyvoje, priskiriant autorystę, svarbu išskirti požymius, kurie nepriklausytų nuo kalbos ir išskirtų autoriaus rašymo stilių. Šaltinyje [7] aprašomame tyrime buvo pritaikyti nuo kalbos nepriklausomi stilometriniai požymiai, kurie išlieka net ir po mašininio vertimo. Jie buvo suskirstyti įšias kategorijas:

- Jausminiai požymiai (teigiamų / neigiamų žodžių dažnumas);
- Emociniai požymiai (bendrų emocijų - džiaugsmo, pykčio, baimės, liūdesio, pasišlykštėjimo, nuostabos - pasikartojimas);
- Morfologinių formų požymiai (daiktavardžių, veiksmažodžių, būdvardžių, prieveiksmių pasikartojimas);
- Percepciniai požymiai (garsinių, vaizdinių ir kinestetinių percepcinių ženklų dažnumas);
- Vidutinis sakinio ilgis.

Kaip teigia šaltinio [7] autoriai, žmonių nuotaikų raiška nepriklauso nuo jų gimtosios kalbos, taip pat tyrimas, kurį atliko [11] šaltinio autoriai, patvirtino, kad žmonių nuotaikų raiška yra skirtinga.Nepaisant to, kad morfologinių formų požymiai priklauso nuo kalbos, jie nekinta ir po vertimo. Taippat šaltinyje [12] atliktas tyrimas rodo, kad dauguma žmonių gali būti suskirstyti į audialus, vizualus ir kinestetikus. Mokymosi stilius atsiskleidžia autoriaus užrašuose, taip pat pateikdamas naudingosinformacijos, reikalingos autorystei priskirti tarpkalbiniu aspektu.

1.2.4. Duomenų daugiadimensiškumo problema

Yra problemų, kurias labai sunku išspręsti dėl požymių, kurių kiekis dažnai viršija turimus skaičiavimų resursus. Šis fenomenas yra vadinamas didelio matavimo prakeiksmu, o originalus terminas pirmiausiabuvo paminėtas [13]. Kalbant apie informacijos paiešką, terminas buvo pirmiausiai paminėtas [14].

Didelio matavimo vektorinėje erdvėje, informacijos yra nedaug ir norint gauti pakankamai tikslius rezultatus, reikia daug mėginių [15]. Ši problema taip pat galioja tekstų klasifikavimo sritims [16].Siekiant ją išspręsti, buvo sukurti sudėtiniai didelio matavimo sumažinimo būdai (pvz. principinių komponentų analizė (angl., *Principal Component Analysis*, PCA), matricos skaidymas atskiromis reikšmėmis (angl. *Singular Value Decomposition*, SVD)), daugelį jų pritaikant specialiai tekstiniamsdokumentams.

1.2.5. Autoriaus tapatybės klastojimas

Atvejų, kada autoriaus tapatybę mėginama suklastoti, analizė yra išskirtinis autorystės priskyrimo uždavinys, itin susijęs su kibernetiniais kriminalistiniais tyrimais. Tikslas surasti suklastotą autorystę gali būti laikomas specifiniu autorystės priskyrimo uždaviniu [17].

Brennan ir kt. [18] analizuoja tris specifinius metodus ir jų atsparumą dviejų tipų priešiškoms atakoms.Pirmoji yra klaidinimo ataka (angl. *obfuscation attack*), kai autorius stengiasi parašyti dokumentą taip, kad asmeninis rašymo stilius būtų neatpažintas. Antroji yra ataka, kada autorius mėgina parašyti tekstą taip, kad jis primintų kito atitinkamo autoriaus rašymo stilių. Klaidinimo atveju, subjektas mėgina nuslėpti savo tapatybę. Imitavimo atveju, subjektas mėgina susitapatinti su kitu subjektu, mėgdžiodamas pastarojo rašymo stilių. Verčiant pradinės frazės

suklastojamos, naudojantis mašininio vertimo paslaugomis. Tokie atvejai gali smarkiai sumažinti tradicinių autorystės nustatymo metodų efektyvumą.

1.3. Autorystės atpažinimo metodai anglų kalbos tekstuose

Šiame skyriuje pateikiama anglų kalbos tekstų autorystės atpažinimo metodų apžvalga, įskaitantpožymių atranką, tekstų skaitmeninimą, požymių skaičiaus su sumažinimą, klasifikavimo metodais grįstą autorystės nustatymą, panašumo (atstumo) metrikomis grįstus metodus, vizualizacija grįstus autorystės nustatymo metodus, tekstų modeliavimą tinklais ir tekstų topologiją.

1.3.1. Požymių atranka

Autorystei atpažinti išskiriami šie požymiai: grafemų ypatybės, leksiniai požymiai, sintaksiniai požymiai, semantiniai požymiai ir atitinkamam taikymui būdingi požymiai. Nė vienas iš jų nėra reikšmingesnis nei kiti.

Grafemų ypatybės rodo tekstą esant simbolių seka, o įvairūs vertinimo būdai yra: ženklų dažnumas, skaitmenų, didžiųjų ir mažųjų raidžių simbolių dažnumas, skyrybos ženklų dažnumas ir t.t. [19] Leksiniai požymiai yra charakterizuojami, skirstant tekstą į ženklų (žodžių) seką, kurie grupuojami į sakinius, ir požymiai yra apibūdinami kaip žodžių ilgumas, žodžio/ simbolio dažnumas, žodžio/ simbolion-gramos, sakinių ilgumas ir žodyno turtingumas [20]. Leksiniai požymiai nepriklauso nuo kalbos [21].

Žodžių ilgumą įprastai nurodo simbolių skaičius [3], o sakinių ilgumą atitinkamai - žodžių ir/arba simbolių skaičius. Žodyno turtingumo požymiai nustato teksto žodyno įvairovę. Kitas variantas yra tik kartą tekste aptinkamas žodžių skaičius (hapakslegomena). Tačiau žodžiais pagrįsti požymiai, tokie kaip žodyno turtingumas, nėra itin efektyvūs, kai taikomi trumpiems tekstams, dėl jų apimties, palyginus su literatūros kūriniais, ir tokie požymiai įprastai priklauso nuo konteksto ir autorius gali sąmoningai jais manipuliuoti [22]. N-gramos yra tęstinės teksto žodžių, simbolių ar baitų sekos. Simboliųn-gramos pateikia tekstą kaip baitų sekas ir apima visus spausdinamus ir nespausdinamus simbolius[23]. Žodinės n-gramos (arba kolokacijos) pateikia kontekstinę informaciją [24]. Simbolių n-gramospateikia stilistinę, leksinę ir tam tikrą kontekstinę informaciją bei atskleidžia autoriaus įpročius, susijusius su didžiųjų raidžių ir skyrybos vartojimu tekste. Simbolių n-gramos yra atsparios triukšmui[25], nes tekste padaryta klaida įprastai paveikia tik mažą skaičių simbolių n-gramų.

Sintaksiniai požymiai charakterizuoja tekstą pagal tam tikrų sintaksinių struktūrų buvimą ir dažnumą, tokių kaip veiksnio, tarinio,papildinio ir t.t. dažnumas, gramatinės klaidos ir neoficialusstilius (pvz. sakinių rašymas didžiosiomis raidėmis) [26]. Sintaksiniai požymiai nurodo, kaip autorius sudaro sakinius, bei atskleidžia jų rašymo stilių. Sintaksiniai požymiai gali būti klasifikuojami į fukcinių žodžių, skyrybos ženklų, santrumpų dažnumą, POS (morfologinį) žymėjimą (kalbos dalių statistika), paviršinę sintaksinę analizę ir sintaksines klaidas.

Funkciniai žodžiai, tokie kaip prieveiksmiai, prielinksniai, jungtukai ar jaustukai, yra žodžiai, kurie patysturi nedaug arba iš viso neturi semantinio turinio, tačiau jie lemia, kaip formuojami sakiniai. Įprastai jie nurodo gramatinį santykį arba bendrą savybę. Dėl didelio funkcinių žodžių dažnumo ir jų svarbaus vaidmens gramatikoje autorius įprastai sąmoningai nekontroliuoja jų vartojimo konkrečiame tekste.

Taigi, funkciniai žodžiai yra geri autoriaus stiliaus rodikliai [27]. Tipišką angliškų funkcinių žodžių rinkinį, naudojamą autorystės tyrimo užduotims, įprastai sudaro keli šimtai žodžių. Dėl dažno vartojimo kalboje ir aukšto gramatikalizavimo lygmens, mažiau tikėtina, kad funkciniai žodžiai bus apgalvotai autoriaus kontroliuojami, nors jų vartojimas smarkiai skiriasi tarp įvairių autorių, tokiu būdu tai tampa geru kriterijumi, atliekant autorystės tyrimo užduotis [28]. Nepaisant fakto, kad kai kurios kalbosstilistiškai yra panašios, jose dažnai aptinkamos skirtingos charakteristikos, tokios kaip žodžių skaidymo savybės ar funkciniai žodžiai [21]. Tam tikrų skyrybos ženklų vartojimas gali suteikti informaciją apie sakinio struktūrą, ypač kai tai yra sintaksiškai klasifikuota skyryba (kur

skyrybosženklai vertinami pagal ribos ar pabaigos, kurią jie žymi, tipą [29]), ir tai yra efektyviau, palyginus su tradiciniais [30] [31] skyrybos dažnumo matavimais.

Struktūriniai požymiai pateikia informaciją, kuri nurodo, kaip autorius organizuoja rašymo išdėstymą.Struktūriniai bruožai yra pastraipų ilgis, įtraukų statistika, cituojamo turinio statistika, parašoįterpimas, tarpų tarp pastraipų vartojimas, pasisveikinimo/ atsisveikinimo frazių vartojimas.Struktūriniai požymiai nurodo pastovius autoriaus rašymo modelius [22] ir yra labai svarbūs internetinių žinučių identifikavimo užduotims [19].

Su turiniu susiję požymiai nurodo autoriaus susidomėjimą tam tikromis sritimis. Šie požymiai įprastaiyra su turiniu susijusių raktinių žodžių, frazių ar simbolių dažnumas. Pavyzdžiui, vienas autorius gali rašyti tekstus santykinai mažu skaičiumi temų, o įvairūs autoriai gali rašyti tekstus skirtingomis temomis. Martindale ir McKenzie [32] įrodė, kad šie požymiai yra labai svarbūs ir gali net viršyti leksinių požymių svarbą bei padidinti autorystės identifikavimo užduoties efektyvumą. Tokių požymių parinkimas priklauso nuo konkrečių taikymo sričių.

Kai kurie požymių tipai gali priklausyti kelioms kategorijoms, nes jie pateikia informaciją, kuri gali būti naudojama skirtingiems tikslams, atliekant autorystės analizės užduotis (pvz. informacija apie skyrybos dažnumą yra naudinga autoriaus skyrybos vartojimo (leksikos kategorija) bei sakinių struktūros niuansų nustatymui (sintaksinė kategorija)) [22].

1.1 lentelė. Anglų kalboje dažniausiai naudojami lingvistiniai požymiai ir jų rinkiniai

Nr.	Aprašymas
1	Žodžių skaičius
2	Eilučių skaičius
3	Didžiųjų raidžių santykis
4	Skaičių dažnis
5	Tuščios erdvės simbolių dažnis
6	Raidžių dažnis
7	Trumpų (< 4 simbolių ilgio) žodžių santykis
8	Vidutinis žodžio ilgis
9	Sakinių skaičius
10	Vidutinis sakinio ilgis
11	Unikalių žodžių santykis
12	Daugiausiai pasirodančio žodžio dažnis
13	Skyriklių santykis
14	Paragrafų skaičius
15	Vidutinis eilutės ilgis
16	Žodžių galūnių dažniai
17	Bigramų dažniai
18	Unikalių bigramų santykis
19	Sutrumpinimų santykis
20	Jausmaženklių santykis

1.3.2. Tekstų konvertavimas į skaitinius požymius

Tekstinių žinučių konvertavimas į skaitines sekas nėra gerai ištyrinėta sritis. Tradiciškai, dokumentai yra pateikiami kaip požymių vektoriai, kurie neišlaiko nuoseklios informacijos esančios tekstinėseežinutėse. Pavyzdžiui, bendrieji požymiai, tokie kaip žodžių skaičius, tenkinantis tam tikras lingvistines ypatybes (pvz. žodis baigiasi galūne „ing"), nepasikeis, net jeigu originali žodžių seka tekste bus pakeista. Dėl to bus prarandama svarbi stilistinė ir substilistinė informacija. Kitos dokumentų savybėstokios, kaip n-gramų dažnis, išlaiko tik kai kurią lokalią nuoseklią informaciją, esančią greta n-gramosvietos, bet nesuteikia galimybės rekonstruoti originalų tekstą iš jo n-gramų dažnio.

Tekstui būdingų savybių išskyrimas, pasitelkiant panašumą išsaugančias maišos funkcijas, yra naudojamas plagiato aptikimui. Šis metodas generuoja unikalius skaitmeninius šablonus iš dokumento ar teksto segmento. Tada, šie šablonai yra lyginami su dokumentų rinkiniu, kad būtų rastos sutampančios kopijos [33].

Toa ir kiti [34] tyrinėja, ar teksto ir laiko eilučių susiejimas yra įmanomas taip, kad aktualios duomenų gavybos problemos tekste gali turėti savo atitikmenis laiko eilutėse (ir atvirkščiai). Jie pasiūlė T3 (tekstas į laiko eilutes) karkasą, kuris naudoja skirtingas detalumo (pvz. raidžių ar žodžių lygmuo) ir n-gramų (pvz. unigramų ar bigramų) kombinacijas. Kad kiekvienai raidei būtų pasirenkama tinkama skaitmeninė reikšmė, T3 karkasas pasirenka skirtingas ploto užpildymo kreives (pvz. tiesines, Hilberto, Z eilės), priklausomai nuo klaviatūros išdėstymo.

Prieš apdorojant tekstą, paprastai reikalinga transformacija, pervedanti jį į skaitinių požymių rinkinį, priimtiną algoritmams. Paprastai dokumentas pateikiamas kaip tam tikrų teksto elementų svarbą atspindinčių svorių vektorius. Dažniausiai naudojami du teksto elementų parinkimo ir reprezentavimo būdai: žodžių krepšelis (angl. *bag of words*) ir fiksuoto ilgio vektorius. Pirmuoju atveju dokumentą reprezentuoja visų jo žodžių rinkinys, neretai ignoruojant jų tvarką. Sudėtingesnes struktūras naudojantys pateikimo būdai dažnai neduoda apčiuopiamai geresnių rezultatų, tačiau smarkiai padidina skaičiavimų apimtis. Antruoju atveju fiksavus teksto terminų (teksto fragmentai, nebūtinai pavieniai žodžiai) žodyną, dokumentas pateikiamas kaip skaičių vektorius.

1.3.3. Požymių skaičiaus mažinimas

Yra išskiriami dviejų tipų matmenų skaičiaus mažinimo metodai [35]: požymių atranka ir požymių išskyrimas. Požymių atrankos atveju, iš visos požymių aibės išrenkamas požymių poaibis, o likę požymiai skaičiavimo procese nenaudojami. Požymių išskyrimo atveju, mažinimas yra vykdomasnaujoje vektorių erdvėje (su specialiomis charakteristikomis), kuri buvo transformuota iš pradinės vektorių erdvės.

Principinių komponenčių analizė (PCA) - populiariausias metodas, naudojamas žemesnių dimensijų daugiamačių duomenų pateiktims apskaičiuoti. Šis metodas iteratyviai skaičiuoja didžiausios dispersijos kryptį, kuri atitinka vertikalios hiperplokštumos projekciją, taip nustatomos kelios statmenos kryptys, kurios atitinka didžiausias dispersijas duomenyse, šitaip sumažinama dimensijų erdvė [36].

1.3.4. Klasifikavimu grįstas autorystės nustatymas

Automatizuoti autorystės priskyrimo metodai gali būti grupuojami į klasifikavimu grįstus (mašininis mokymasis) ir panašumu grįstus metodus [26]. Klasifikavimu grįstų metodų atveju, anoniminių raštųklasifikavimui naudojamas klasifikatorius, kuris yra sudaromas naudojant tekstus, sukurtus kiekvienožinomo autoriaus kandidato (naudojamas apmokymo rinkinys, sudarytas iš atskirų dokumentų) [4].Patys populiariausi klasifikavimo metodai yra šie.

Atraminių vektorių klasifikavimo algoritmas (SVM) - vienas iš populiariausių autorystės nustatymo metodų. Šiame klasifikatoriuje naudojamas netiesinis susiejimas, kuris transformuoja apmokymo duomenis į aukštesnę dimensiją. Jis ieško optimalios tiesinės hiperplokštumos, kuri skiria dvi klases ir tarnauja kaip etalonas sprendimo priėmimui.

Bajeso klasifikatorius - tai paprastas tikimybinis klasifikatorius, kuris remiasi Bajeso teoremakartu su stipriomis nepriklausomumo prielaidomis. Nepaisant fakto, kad šis metodas veikia susupaprastintomis prielaidomis, paprastai jis teikia gerus rezultatus. Bajeso klasifikatorius veikiasu nedideliais apmokymo duomenų kiekiais ir efektyviai naudoja skaičiavimo resursus. Teksto klasifikavimo srityje, šis metodas naudoja požymius (pvz. žodžius) ir maksimalią a posteriori (MAP) sprendimo taisyklę.

C4.5 - tai sprendimų medžiais grįstas algoritmas. Jis konstruoja sprendimo medžius, naudodamas apmokymo duomenis. Kiekvienam sprendimo mazgui medyje yra parenkamas vienas duomenų požymis, kuris efektyviausiai skiria jo imties aibę į poaibius, pridedamus į vieną ar į kitą

klasę.Sprendimai yra išrenkami iš požymių, kurie turi didžiausią normalizuotą informacijos kiekį, bendrąvisam poaibiui.

1.3.5. Panašumu grįsti metodai

Panašumu grįsti metodai taiko įvairias metrikas skirtumams tarp dviejų dokumentų įvertinti ir priskiria duotą dokumentą konkrečiam autoriui, kurio žinomas tekstų rinkinys (traktuojamas kaip vienas dokumentas) yra labiausiai panašus į tą dokumentą. Šiuos metodus rekomenduojama pasirinkti tada, kai reikia dirbti su dideliu kiekiu autorių kandidatų [37]. Labiausiai priimtinas priskyrimas yrapasiekiamas, kai dokumentas yra pateikiamas daugiamatėje erdvėje ir atliekant duoto dokumento priskyrimą autoriui, naudojamos tinkamos atskyrimo metrikos [38]. Panašumu grįsti metodai taip pat leidžia patvirtinti, ar autorius parašė duotą dokumentą, jeigu panašumas tarp žinomų autoriaus tekstų ir pateikto dokumento peržengia tam tikrą slenksčio reikšmę. Labiausiai paplitę metodai yra:

- Kosinuso panašumo metrika - viena iš populiariausių metrikų, vertinančių panašumą tarp tekstinių dokumentų. Panašumas išreiškiamas koreliacija tarp dokumentų terminų vektorių.Kosinusinio panašumo atveju, skaičiuojama kosinuso reikšmė tarp vektorių kampų [39].
- Euklido atstumas yra standartinė geometrinė metrika. Ji parodo vidutinį atstumą tarp dviejųtaškų. Ji taip pat plačiai naudojama tekstų klasterizavime [39].
- Žakardo (Jaccard) koeficientas lygina suminę terminų, kurie yra pateikiami abiejuose dokumentuose, reikšmę su sumine terminų, kurie skiriasi ir nėra bendri abiejuose dokumentuose, reikšme.
- Pirsono (Pearson) koreliacijos koeficientas matuoja vektorių susietumo dydį.

1.3.6. Kompresija grįsti metodai

Kompresija grįsti metodai remiasi Kolgomorovo kompleksiškumo teorija, kuri apskaičiuojama, taikant duomenų glaudinimo įrankius. Ši teroija teigia, kad panašūs duomenys yra suspaudžiami geriau negu skirtingi. Pasak šaltinio [67] autorių, teksto kategorijų modeliai yra sudaromi, taikant PPM (angl.*Prediction by Partial Matching*) teksto glaudinimo algoritmą, remiantis ženklais grįstu kontekstu.Tuomet nežinomas tekstas yra kategorizuojamas, remiantis kategorijomis, atsirandančiomis dėl dokumento entropijos (vidutinis bitų koduotam simboliui skaičius), atsižvelgiant į kategorijos modelį, ir dokumento entropijos skirtumą tarp kategorijos ir kategorijos modelių komplementų.

Šaltinio [68] autoriai teigia, kad nežinomo autoriaus parašytas tekstas yra pridedamas prie tekstų, kurių autoriai nustatyti, ir ekspertinės paieškos tekstų, kuriuose skirtumas tarp suglaudinto originalo ir pridėtų tekstų ilgių yra mažiausias. Nepaisant to, kompresiją grįsti metodai reikalauja daug tekstų kiekvienam autoriui, kad rezultatas būtų reikšmingas.

1.3.7. Vizualizacija grįstas autorystės nustatymas

Autorystės nustatymas, grįstas vizualizacija, yra nagrinėjamas ribotame darbų kiekyje. Literatūrosšaltinyje [40] rašto šablonų vizualizavimui yra naudojama pagrindinių komponenčių analizė kartu su kosinuso panašumo metrika. Literatūros šaltinyje [41] autorystės, išreikštos vektoriais daugiamatėje erdvėje, vizualizavimui buvo naudojamas latentinis semantinis indeksavimas (LSI).

Antspaudai (angl. *Writeprints*) yra autorystės vizualizavimo metodas pasiūlytas [69]. Šis metodas naudoja PCA algoritmą. Vizualizavimo proceso metu sukuriami grafiniai šablonai, kuriuos galimalengvai palyginti ir identifikuoti vizualiai. Tačiau metodas netinkamas naudoti su trumpais (mažiau nei 30-40 žodžių) tekstais.

1.3.8. Funkciniai žodžiai

Funkciniai žodžiai, tokie kaip prielinksniai, jungtukai ir įvardžiai, patys turi nedidelę kontekstinę reikšmę, bet vietoj to apibrėžia gramatines priklausomybes tarp žodžių. Funkcinių žodžių tyrinėjimas yra naudingas, nes jie teikia informaciją labiau apie sintaksę nei kontekstą. Kadangi funkciniai žodžiai pateikiami kaip priemonė stilometrinių savybių išskyrimui, buvo pasiūlyta daug metodų, skirtųi šių žodžių dažniui, skirtingų autorių parašytuose tekstuose ,analizuoti. Taip pat buvo atkreiptas dėmesys į požymių analizę, susijusią ne tik su dažnai pasikartojančių žodžių pasirodymu. Šių požymių pavyzdžiai: žodyno turtingumas, žodžio stabilumas (dydis, nurodantis galimybę žodį pakeisti jo ekvivalentu) ar sintaksiniai požymiai, tokie kaip POS požymiai.

Burrows [42] pasiūlė žinomiausią dažniausiai naudojamų žodžių metodą, pavadintą Delta metodu.Delta metodas dirba tokiu būdu. Pirmiausiai, metodas skaičiuoja funkcinių žodžių aibės z skirstinį (paprastai naudojama 150 dažniausiai naudojamų žodžių). Tada, kiekvienam tekstui iš z reikšmės suskaičiuojamas kiekvieno žodžio dažnio nuokrypis nuo normos, tokiu būdu apytikriai įvertinama, arpasikartojimas yra didesnis (teigiama z reikšmė), ar mažesnis (neigiama z reikšmė) už vidurkį. Galiausiai, Delta metrika yra absoliutaus skirtumo tarp visos funkcinių žodžių aibės, esančios to paties autoriaus parašytų apmokymo tekstų aibėje, z reikšmės ir nežinomų tekstų atitinkamos z reikšmės vidurkis, parodantis skirtumą tarp apmokymo tekstų ir nežinomų tekstų Kuo mažesnė Delta metrika, tuo didesnis stilistinis panašumas tarp nežinomo teksto ir pretenduojančiojo į autorius.

Taip pat Mosteller ir Wallace [43] įrodė, kad funkcinių žodžių (tokių kaip „ir“, „joks“, „kada nors“, „arba“, „iki“, ir „su“) dažnis gali būti naudojamas autoriaus stiliaus kiekybei įvertinti.

1.3.9. Rašybos klaidos

Rašybos klaidos yra dažnos elektroniniame diskurse. Tokios klaidos gali suteikti papildomosinformacijos apie autoriaus išsilavinimą, tautybę, lytį ir profilį. Rašybos klaidų pavyzdžiai yra raidžių pakartojimai, trūkstamos raidės, greta esančių raidžių inversija, perteklinės raidės, pakeistos raidės iržodžių sujungimas. Analizuojant tekstą, rašybos klaidas galima nustatyti ir ištaisyti naudojant eilučių atstumo (pvz. Levenshtein) algoritmą. Klaidų dažniai pagal atskiras klaidų kategorijas ir bendras klaidų skaičius gali būti panaudoti kaip teksto požymiai [65].

1.3.10. Tekstų kaip tinklų modeliavimas

Antiqueira ir kt. [44] demonstruoja, kaip tekstas gali būti modeliuojamas žodžių tinklu ir taip pat pateikia kai kurias tinklo metrikas, kurios gali būti naudojamos autorystės charakterizavimui. Analizė rodo, kad yra įmanoma grupuoti kai kurias tris metrikų kombinacijas (pvz. braižant grupavimo koeficientą su išeinančiais ryšiais arba grupavimo koeficientą su laipsnio koreliacija).

Segarra ir kiti [45] pateikia autorystės nustatymo metodą, grįstą teksto modeliavimo normalizuotu gretimų žodžių tinklu. Šie tinklai yra aiškinami kaip Markovo grandinės ir gali būti lyginami pagal entropijos laipsnius. Jie naudojami su funkciniais žodžiais, bet vietoj to, kad naudotų dažnių pasiskirstymą autorystei nustatyti, jie siūlo naudoti ryšiais susietą funkcinių žodžių struktūrą. Kad būtų galima klasifikuoti tekstą pagal autorystę, naudojamas asimetrinis tinklas, sudarytas iš funkcinių žodžių gretimumo įverčių, parodančių, kaip tikėtina rasti konkretų funkcinį žodį kituose ,toliau esančiuosežodžiuose, su sąlyga, kad bus aptiktas kitas duotas žodis. Rezultatų matricos gali būti interpretuojamos kaip Markovo grandinės pereinamosios tikimybės. Skirtingų tekstų panašumas yra įvertinamas pagalšių pereinamųjų tikimybių santykinę entropiją.

Žodžių kolokacijos tinklai [46] yra žodžių, randamų dokumente ar dokumentų rinkinyje, tinklai, kur kiekvienas tinklo mazgas atitinka unikalų žodžio tipą, o briaunos atitinka žodžių rinkinius.

Supaprastintu atveju, kiekviena briauna atitinka unikalią originalaus dokumento bigramą. Pavyzdžiui, jeigu žodžiai A ir B dokumente pasirodo kartu kaip bigrama AB, tada to konkretaus dokumento žodžių kolokacijos tinklas turės briauną A → B. Briaunos gali būti su svoriais (su bigramos AB dažniu) arba be svorių. Globalūs tinklo parametrai, tokie kaip diametras, globalus

grupavimo koeficientas,susitraukimo eksponentė [47] ir vienoje vietoje sukoncentruoti ryšiai tarp grafų [48] gali būtinaudojami autorystei identifikuoti.

Kokybinis tinklas [49] parodo, kaip dažnai žodis yra aukščiau nei kiti žodžiai sakinyje, atskirtame bet kokiu žodžių skaičiumi. Kad būtų normalizuoti dažnių skaičiavimai, jie yra dalinami iš dažnio, pagal kurį du žodžiai pasirodo toje pačioje kalboje bet kokia tvarka.

Pasikartojimai tinkle yra jungūs ir kryptingi subgrafai, daug dažniau pasitaikantys sudėtinguose tinkluose, nei atsitiktiniuose tinkluose. Rizvic ir kiti [50] naudoja tinklų šablonus struktūriniam panašumui aptikti tarp kryptingų tinklų, sudarytų iš skirtingų tekstų, pateikiamų kryptingais junglumo tinklais. Toks panašumo matas gali būti naudojamas tekstų autorystei klasifikuoti.

1.3.11. Kompiuterinė tekstų topologija

Gaunant didesnės dimensijos tekstinių dokumentų topologinę struktūrą, galima suvokti duomenų sąvoką [51]. Patrauklesnis modelis paaiškintų kai kurias erdvinių priklausomybių funkcines formastokias, kaip "patrauklumas ir atstūmimas" [52]. Lingvistiniuose modeliuose žodžių patrauklumo ir atstūmimo efektas yra visada svarbus fenomenas [53]. Svarbu pabrėžti, kad buvo nustatyta, jog knygų autorių galima atpažinti pagal tam tikrų žodžių grandines.

1.4. Autorystės nustatymo metodai nacionalinėse kalbose

Šiame skyriuje apžvelgiami nacionalinėmis kalbomis parašytų tekstų autorystės nustatymo metodai.Apžvalga atlikta su Europoje labiausiai paplitusiomis kalbomis (romanų, germanų, slavų, baltų), naudojančiomis lotyniškąją abėcėlę.

1.4.1. Romanų kalbos

Autorių [54] atliktame tyrime, n-gramų modeliai buvo naudojami italų kalbos tekstyno autorystės priskyrimui, manant, kad tekstai yra tik simbolių sekos, ignoruojant gramatiką ir teksto turinį (laikantraides, skyrybos ženklus, tarpus tik abstrakčiais simboliais). Gauti rezultatai rodo, kad eksperimentas buvo itin sėkmingas.

Prancūzų kalba parašytiems tekstams analizuoti buvo svarbūs: žodžių, sakinių ilgis, daiktavardžių, prieveiksmių, būdvardžių, veiksmažodžių, neskaidomos/sudėtinės daiktavardinės frazės [55].

Varela ir kt.. [56] pasiūlė portugalų kalba parašytų tekstų autorių atpažinimui naudoti įvardžius ir dažniausius brazilų kalbos veiksmažodžius.

Pavelec ir kt.. [57], kalbėdami apie autoriaus tapatybės nustatymo problemą, jai spręsti pasiūlė stilometrinių požymių rinkinį, paremtą portugalų kalbos jungtukais ir prieveiksmiais (77 jungtukai ir 94 prieveiksmiai).

1.4.2. Germanų kalbos

Vokiečių kalba pasižymi turtinga morfologija, taiga tokio tipo požymiai gali suteikti naudingosinformacijos apie tekstą. Autorių [58] atliktame tyrime, kartu su daugeliu bendrų savybių, vokiečių kalbos derivacinės priesagos (ant, arium, ast, at, ator, atur, ei, er, ent, enz, eur, heit, ist, ion, ismus, ität, keit, ling, nis, shaft, tum, ung, ur, werk, wesen) buvo naudojamos išvestiniams daiktavardžiams vertinti, įvertinant daiktavardžių santykį.

Eksperimente, kuriam buvo panaudoti vokiečių kalba parašyti tekstai, be bendrų savybių, reikalingų priskiriant autorystę, svarbūs: žodžių ir sakinių ilgis, žodyno turtingumas / įvairovė, nuo konteksto nepriklausantys funkciniai žodžiai, sintaksinės žodžių klasės, žodžiai, kurie tekstyne sutinkami tik vieną kartą, ir t.t., mažosiomis raidėmis parašytų žodžių formų skaičius buvo panaudotas kaip atraminiovektoriaus klasifikatoriaus (SVM) įvesties duomenys [59].

1.4.3. Slavų kalbos

Rusų kalbai buvo efektyviai pritaikyti daugelis požymių, tokių kaip:

1. Sakinių ilgis (vidutinis žodžių sakinyje skaičius).
2. Žodžių ilgis (vidutinis žodžio skiemenų skaičius).
3. Funkcinių žodžių - prielinksnių, jungtukų, dalelyčių bendras dažnumas.
4. Daiktavardžių dažnumas.
5. Veiksmažodžių dažnumas.
6. Būdvardžių dažnumas.
7. Funkcinių žodžių skaičius sakinyje (vidutinis jungtukų, prielinksnių ir dalelyčių skaičius).

Autorių [60] atliktame tyrime, buvo pasiūlyti šie požymiai:

1. Raidžių porų sekos, t.y., žodžiuose, kur jos atskirtos brūkšniais.
2. Raidžių porų sekos pagrindinėse žodžių formose. Rusų kalboje ši redukcija yra daug svarbesnė, kadangi rusiški žodžiai turi daug formų.
3. Labiausiai apibendrintų gramatinių žodžių klasių ir morfologinių formų (POS) sakiniuose porų sekos.Rusų kalboje išskiriama 14 kalbos dalių.
4. Mažiau apibendrintų žodžių gramatinių klasių poros.

Reicher ir kt. [27] išskiria, kad autorystės priskyrimo problema, kai susiduriame su morfologiškai sudėtinėmis kalbomis, tokiomis kaip kroatų kalba, gali būti sėkmingai išspręsta, sujungiant kelis palyginti paprastus požymius, tokius kaip funkcinių žodžių, skyrybos ženklų, žodžių ir sakinių ilgumopasikartojimas.

Nagrinėdami serbų kalbą, Zecevic ir Utvic [61] pasiūlė suskaidyti tekstą į žodžius, o žodžius - į skiemenis. Tuomet autoriaus profilis gali būti sudarytas iš skiemeniniu skirtymu grįstų požymių, tokių kaip vidutinis žodžio skiemenų skaičius, žodžio ilgumo dažnumo pasiskirstymas skiemenyse (vienskiemenių žodžių skaičius, dviskiemenių žodžių skaičius ir t.t.) ir vidutinis atstumo tarp nskiemeniųžodžių bei skiemenų pasikartojimas. Norint palyginti skirtingų autorių profilius, tarp požymių vektorių matuojamas atstumas.

1.4.4. Baltų kalbos

Lietuvių kalbos elektroniniame diskurse (ypatingai el. laiškuose), dažnai aptinkami šie kalbos vienetai[62]: vulgarizmai, skoliniai, trumpiniai, ženklai, perteklinė kalba, jausmaženkliai. Išsamiausia leksinė analizė reikalinga idiolektui atpažinti. Tame pačiame tyrime teigiama, kad teksto vienetų pasikartojimo sąrašai sudaro svarbią tyrimo metodų dalį. Labai svarbus kriterijus autorystei nagrinėti yra atitinkama autoriaus skyrybos ženklų vartosena, o ypatingai vyraujančių skyrybos ženklų aibė ir ženklas, kuriam suteikta pirmenybė. Tyrime pabrėžiama, kad skyrybos vartojimas skiriasi priklausomai nuo autoriaus.

Sudėtinga lingvistinė lietuvių kalbos struktūra, įskaitant tokius elementus kaip diakritiniai ženklai,linksniai, turtinga kalbos morfologija ir žodynas, gali kelti problemų analizuojant tekstą. Dėl didelio galimų galūnių, prielinksnių ir priesagų skaičiaus gali būti sudėtinga korektiškai identifikuoti atskiraižodžio dalis (morfemas). Dėl turtingo žodyno, kai kurie kitų kalbų tekstams naudojami metodai (pvz.,žodžių krepšelio metodas) nepasiteisina. Lietuvių kalba turi 9 raides su diakritiniais ženklais: ą, č, ę, ė,į, š, ų, ū, ž. Tačiau elektroniniame diskurse šios raidės dažniai pakeičiamos analogiškomis lotyniškosabecėlėmis raidėmis (pvz., ę įe) arba panašiai tariama anglų kalbos abecėlės raidžių pora (pvz., š [ʃ] įsh) [71].

1.4.5. Nacionalinių kalbų lingvistinių požymių palyginimas

1.2 lentelė. Nacionalinių kalbų lingvistinių požymių lyginamoji lentelė

Lingvistiniai požymiai		
Bendri požymiai		
Nr.	**Požymiai**	
1	**Bendras simbolių skaičius (M)**	
2	Alfabetinių simbolių santykis su M	
3	Didžiųjų raidžių ir simbolių santykis su M	
4	Skaitmenų santykis su M	
5	Tuščios erdvės(white spaces) santykis su M	
6	Tabuliatorių santykis su M	
7	Alfabetinių simbolių dažnių santykis su M (didžiosios ir mažosios raidės)	
8	Specialiųjų simbolių dažnio santykis su M	
9	**Bendras simbolių skaičius subjekte (S)**	
10	Alfabetinių simbolių subjekte santykis su S	
11	Didžiųjų raidžių ir simbolių subjekte santykis su S	
12	Skaitmenų subjekte santykis su S	
13	Tuščių erdvių(white spaces) subjekte santykis su S	
14	Tabuliatorių subjekte santykis su M	
15	Alfabetinių simbolių(dižiosios ir mažosios raidės) dažnio santykis su S	
16	Specialiųjų simbolių dažnio santykis su S	
17	Trumpų žodžių (mažiau negu 4 simboliai) santykis su W	
18	Vidutinis žodžio ilgis	
19	Vidutinis simbolių skaičius sakiniuose	
20	Vidutinis žodžių skaičius sakiniuose	
21	Skirtingų žodžių santykis su W	
22	Hapax legomena(žodžiai pasirodantys tik vieną kartą) santykis su W	
23	Hapax dislegomena(žodžiai pasirodantys tik du kartus) santykis su W	
24	Žodžio dažnio(daugiausiai pasirodantis žodis) santykis su W	
25	Žodžio dažnio(antras daugiausiai pasirodantis žodis) santykis su W	
26	Žodžio ilgio distribucija(1-20 simbolių žodžių dažnis) santykis su W	
27	Bendras žodžių skaičius subjekte	
28	Vidutinis žodžių ilgis subjekte	
29	Funkcinių žodžių dažnio santykis su W	
30	Funkcinių žodžių santykis su W	
31	Stabdančių žodžių dažnių santykis su W(tiktai tie,kurie nepaminėti funkcinių žodžių sąraše)	
32	Stabdančių žodžių santykis su W(tiktai tie,kurie nepaminėti funkcinių žodžių sąraše)	
33	Punktuacijos dažnių santykis su M	
34	Punktuacijų santykis su M	
35	Bendras punktuacijų skaičius subjekte	
36	Bendras sakinių skaičius	
37	Bendras paragrafų skaičius	
38	Bendras eilučių skaičius	
39	Žymeklis, tuščių eilučių tarp paragrafų pažymėjimui	
40	Vidutinis simbolių skaičius paragrafe	
41	Vidutinis žodžių skaičius paragrafe	
42	Vidutinis sakinių skaičius paragrafe	
Požymiai, priskiriami skirtingoms kalboms		
Nr.	**Požymis**	**Kalbos**

		Anglų k.	Lenkų k.	Rusų k.	Serbų-kroatų k.	Lietuvių k.
43	Žodžių ,turinčių galūnę “able” santykis su W	+				
44	Žodžių,turinčių galūnę”al” santykis su W	+				
45	Žodžių,turinčių galūnę”ful” santykis su W	+				
46	Žodžių,turinčių galūnę”ible” santykis su W	+				
47	Žodžių,turinčių galūnę”ic” santykis su W	+				
48	Žodžių,turinčių galūnę”ive” santykis su W	+				
49	Žodžių,turinčių galūnę”less” santykis su W	+				
50	Žodžių,turinčių galūnę”ly” santykis su W	+				
51	Žodžių,turinčių galūnę”ous” santykis su W	+				
52	Didžiųjų ir mažųjų raidžių kombinacijos žodžiuose		+			
53	Afikso –anie dažnis		+			
54	Afikso –arka dažnis		+			
55	Afikso –ca dažnis		+			
56	Afikso –acz dažnis		+			
57	Afikso –cie dažnis		+			
58	Afikso –ista dažnis		+			
59	Afikso –enie dažnis		+			
60	Afikso –izna dažnis		+			
61	Afikso –ość dažnis		+			
62	Afikso –ek dažnis		+			
63	Afikso –ik dažnis		+			
64	Afikso –ec dažnis		+			
65	Afikso –acja dažnis		+			
66	Afikso –owiec dažnis		+			
67	Afikso –arz dažnis		+			
68	Afikso –ówka dažnis		+			
69	Afikso –izm dažnis		+			
70	Afikso –ak dažnis		+			
71	Afikso –nik dažnis		+			
72	Afikso –nica dažnis		+			
73	Afikso –stwo dažnis		+			
74	Afikso –arnia dažnis		+			
75	Afikso –ka dažnis		+			
76	Kalbos struktūrų ilgis		+			
77	Žodžių porų vartojimo dažnumas		+			
78	Jungimo dalių skaičius			+		
79	Atitinkamo ilgio sakinių skaičius			+		
80	Jungtukų,dalelyčių ir prielinksnių skaičius			+		
81	Raidžių poros,natūraliai einančios tekste ir tarpai tarp jų			+		

82	Raidžių poros duotame tekste bei sakinių formos			+		
83	Labiausiai apibendrintų gramatinių klasių žodžių poros(nepilnieji sakiniai). Tai daiktavardžiai,veiksmažodžiai ir kt. kl. dalys bei jų eiliškumas sakinių tekste.Iš viso 14 gramatinių klasių bei 4 santykinės kategorijos ir "neaiški "klasė,nes ne visi žodžiai aptinkami automatiniu būdu			+		
84	Mažiau apibendrintų(pilnų) gramatinių klasių žodžiai (pvz. rusų kalboje gyvieji daiktavardžiai)			+		
85	Daiktavardžių naudojimo dažnumas(jų procentinis kiekis)			+		
86	Veiksmažodžių naudojimo dažnumas(jų procentinis kiekis)			+		
87	Prieveiksmių naudojimo dažnumas proc.			+		
88	Prielinksnio"в" procentinis dažnumas			+		
89	Dalelytės "не" proc.dažnumas			+		
90	Tarnybinių žodelių kiekis sakinyje			+		
91	Atitinkamas kalbos dalių santykis			+		
92	Kai kurių "konstrukcinių" teksto elementų santykinis dažnumas			+		
93	Lingvistinė sakinių struktūra			+		
94	Vidutinis skiemenų kiekis žodyje			+		
95	Sakinių skaičius				+	
96	"ne" dažnumas				+	
97	"I" dažnis				+	
98	"ili" dažnis				+	
99	"je" dažnis				+	
100	"se" dažnis				+	
101	"pa" dažnis				+	
102	"da" dažnis				+	
103	"kao poput" dažnis				+	
104	Priešdėlio "ne" dažnumas					+
105	Vulgarizmai					+
106	Kitų kalbų intarpai					+
107	Sutrumpinimai					+
108	Kalbinis perteklius					+
109	Jausmaženkliai					+
110	Lietuviškų/nelietuviškų raidžių naudojimas					+
111	Nenorminė leksika					+
112	Atskirų kalbos dalių žodžių vartosena					+
113	Liepiamosios nuosakos dažnis					+
114	Tariamosios nuosakos dažnis					+
115	Palyginimai					+
116	Vulgarumo / mandagumo žodžiai					+

117	Emocinių bruožų žodžiai					+

1.5. Reikalavimai kriminalistinėms autorystės nustatymo sistemoms

Šiame skyriuje formuluojami reikalavimai kriminalistinėms autorystės nustatymo sistemoms.Atlikus autorystės nustatymo metodų, taikomų elektroniniame, o ypač valstybiniame elektroniniamediskurse, apžvalgą ir analizę, pristatomos bendros rekomendacijos kriminalistinių metodų ir įrankių, skirtų autorystei nustatyti, kūrėjams.

Autorystės nustatymo sistemos turi tenkinti bendrus ekspertinėms, kriminalistinėms sistemoms keliamus reikalavimus bei specifinius autorystės nustatymo reikalavimus.Autorystės nustatymo sistema, kaip tam tikras ekspertinės sistemos tipas ,turi veikti greičiau nei perprideramą laiką atsakymą pateikiantys jos žmogiškieji kolegos, turi atpažinti klaidas ir veikti, naudodama neapibrėžtus duomenis ir vartotojams jos sąsaja turėtų būti patogi ir naudinga [63].

Iš kriminalistinių tyrimų perspektyvos, autorystės nustatymo sistema privalo gebėti užfiksuoti duomenis apie nusikaltimo vietą ir suformuluoti galimas hipotezes apie nusikaltimą, pritaikant struktūrizuotus įrodymus. Kiekviena hipotezė patvirtina, jog yra susijusi su aibe faktų, reikalingų jaiįrodyti arba paneigti. Faktai parodo įkalčius, kurie turi būti surinkti iš informacijos šaltinių, aptinkamųnusikaltimo vietoje.

Galiausiai turi būti patenkinti autorystės nustatymo sistemai keliami specifiniai reikalavimai. Jie turi gebėti efektyviai atskirti, kokios įrodinėjamos savybės patvirtinamos, bei galėti atsilaikyti prieš į autoriaus nustatymą nukreiptas atakas. Pasak Granto [64], norint sėkmingai pateikti autorystės nustatymo rezultatus kaip įrodymus, reikia parodyti pastovumo ir išskirtinumo darną: autoriui būdingas pastovumas reikalauja, kad savybės, išskirtos iš tekstų, kurių autorius žinomas, vertės nekistų kiekvienam galimam autoriui, kas leistų išgauti pagrįstą lyginamąją analizę pagal žanrus ir kituslingvistinius kintamuosius. Tarpautorinis savitumas reikalauja, kad savybių, pasirinktų autorystei nustatyti, vertės skirtųsi visoms autorių poroms.

Reikalavimai "idealiai" kriminalistinės lingvistikos sistemai buvo suformuluoti [66]: nuosavos požymių aibės sudarymo lankstumas, galimybė atlikti automatinę stiliaus požymių paiešką įvairiuose lygmenyse(fonologiniame, morfologiniame, leksikiniame-semantiniame), pagalba pažymint reikšmingus požymius tekste, duomenų kiekybinis ir statistinis įvertinimas, galimybė atlikti statistinį įvertinimą(pvz., t-testą, dispersijos analizę), pateikti autorystės nustatymo patikimumo įvertį (pvz., požymio vertės to paties autoriaus ir kitų autorių tekstuose).

1.6. Išvados

1. Kiekvienam autoriui yra būdingos tam tikros lingvistinės savybės, leidžiančios jį išskirti iš kitų autorių.
2. Elektroninio diskurso autorystės nustatymo sistemose autoriams aprašyti yra naudojami įvairių lingvistinių požymių skaitinių reikšmių rinkiniai (vektoriai), dar vadinami teksto antspaudais.
3. Elektroninio diskurso autorystės nustatymo sistemose autoriams atpažinti (suklasifikuoti) yra naudojami mašinos mokymosi klasifikatoriai arba panašumo įvertinimo metrikos.
4. Trumpų (kriminalistinėje lingvistikoje dažniausiai pasitaikančių) ir ilgų tekstų autorystės nustatymo sudėtingumas ir metodika skiriasi.
5. Kriminalistinėje lingvistikoje elektroninio diskurso autorystės nustatymas turi būti atliekamas per priimtiną laiko tarpą, todėl svarbu naudoti tinkamus metodus, optimalų požymių rinkinį, bei esant reikalui naudoti dimensiškumo sumažinimo algoritmus.
6. Daugeliui kalbų, elektroninio diskurso autorystės nustatymą galima atlikti naudojant tik bendruosius, nuo kalbos nepriklausančius lingvistinius požymius, tačiau tyrimai rodo, kad papildomi, tam tikrai kalbai būdingi požymiai dažniausiai pagerina autorystės nustatymo tikslumą.

7. Elektroninio diskurso autorystės nustatymo sistema yra pagalbinė priemonė, leidžianti paspartinti tyrimo procesą, tačiau galutinį sprendimą priima tyrėjas.

2. ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMO METODO SUDARYMAS

2.1. Metodo aprašymas

Pagrindinis šio siūlomo metodo tikslas yra kuo efektyviau atlikti lietuvių kalba parašyto elektroninio diskurso autorystės nustatymo uždavinį bei ištirti lietuvių kalbos lingvistinių požymių įtaką autorystės nustatymo procesui bei jo rezultatams.

Skirtingai nei kituose sprendimuose, šiame metode yra naudojami keli skirtingi dimensiškumo sumažinimo algoritmai bei mašinos mokymosi klasifikatoriai. Tokiu būdu galima rasti bei naudoti efektyviausiai veikiantį jų derinį. Dimensiškumo sumažinimo algoritmų dėka galima sumažinti požymių dimensijų erdvę ir tokiu būdu paspartinti sistemos greitaveiką.

2.2. Lingvistinių požymių rinkiniai

2.1 lentelė. Metode naudojamų lietuvių kalbai būdingų lingvistinių požymių grupės

Nr.	Aprašymas
1	Funkcinių žodžių santykiniai dažniai
2	Funkcinių žodžių ir visų žodžių santykis
3	Stabdančių žodžių santykiniai dažniai
4	Stabdančių žodžių ir visų žodžių santykis
5	Žodžių su nurodytomis galūnėmis santykiniai dažniai
6	Lietuvių kalbai nebūdingų simbolių bigramų santykiniai dažniai
7	Lietuvių kalbai nebūdingų bigramų ir visų simbolių bigramų santykis
8	Priešdėlio „ne“ santykinis dažnis
9	Lietuviškų raidžių santykiniai dažniai
10	Lietuviškų raidžių ir visų raidžių santykis
11	Sutrumpinimų santykiniai dažniai
12	Sutrumpinimų ir visų žodžių santykis
13	Palyginimų santykiniai dažniai
14	Palyginimų ir visų žodžių santykis

2.2 lentelė. Metode naudojami lietuvių kalbai būdingi lingvistiniai požymiai

Nr.		Aprašymas
1	Funkciniai žodžiai	ir, kur, į, ant, kad, nėra, iš, jį, su, todėl, buvo, prieš, o, jeigu, yra, daug, kaip, vis, tai, nei, ar, galima, tik, bus, ne, mano, bet, jog, savo, man, lietuvos, tiek, jis, kurie, taip, jei, nuo, reikia, m, pagal, apie, gal, jo, daugiau, jau, vienas, kai, tuo, dar, jam, dėl, žemės, jų, v, aš, t, po, darbo, už, nr, per, metu, jos, j, kas, tą, bei, valstybės, tačiau, tas, a, tarp, ji, seimo, to, visi, labia, ją, prie, kuri, čia, kiek, būti, žmogaus, gali, tu, arba, p, pat, žmonių, jie, pirmininkas, turi, pats, d, lietuvių, mūsų, kitų, metų, tam, iki, vėl, nors, tada, dabar, žmogus,

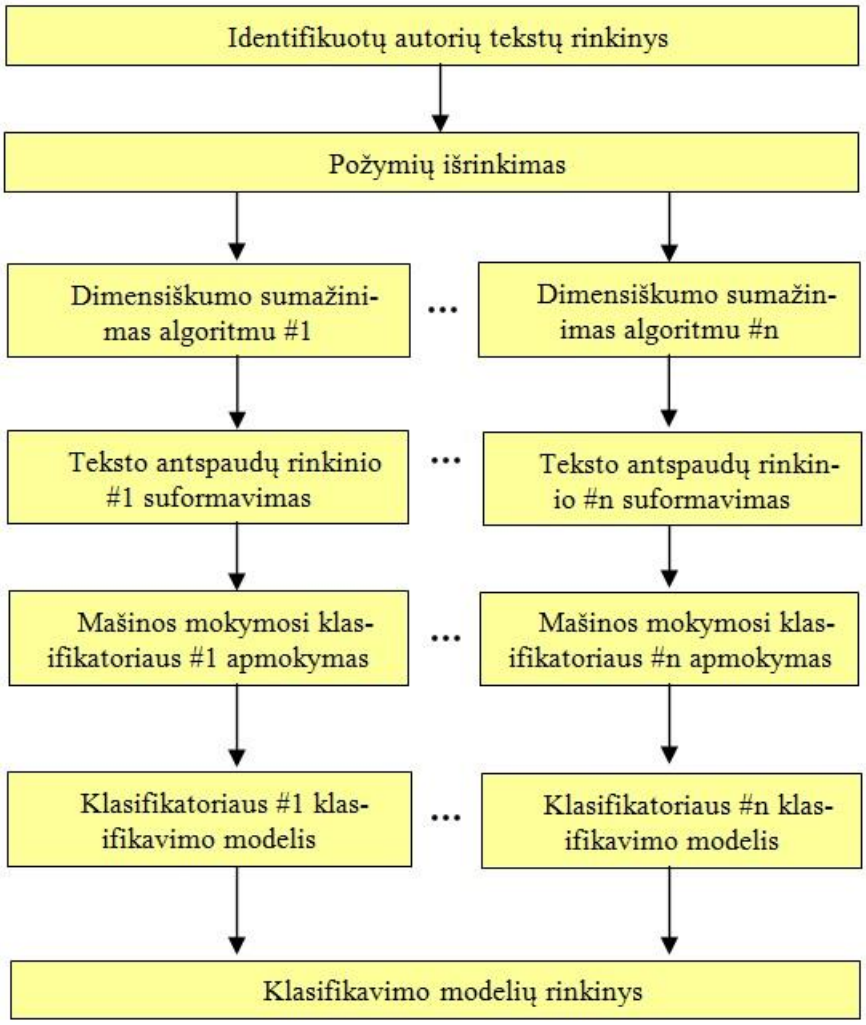
		respublikos, juos, be, ten, nes, kurios, net, vyriausybės, ką, vieną, mes, g, būtų, lietuvoje
2	Stabdantys žodžiai	į, įkypai, įstrižai, šįjį, šalia, še, šiąją, šiąja, šiąsias, šiųjų, šiųjų, ši, šiaisiais, šiajai, šiajam, šiajame, šiapus, šiedvi, šieji, šiesiems, šioji, šiojo, šiojoje, šiokia, šioks, šiosiomis, šiosioms, šiosios, šiosios, šiosiose, šis, šisai, šit, šita, šitas, šitiedvi, šitokia, šitoks, šituodu, šiuodu, šiuoju, šiuosiuose, šiuosius, štai, žemiau, aš, abi, abidvi, abiejų, abiejose, abiejuose, abiem, abigaliai, abipus, abu, abudu, ai, anąją, anąjį, anąja, anąsias, anųjų, anųjų, ana, ana, anaiptol, anaisiais, anajai, anajam, anajame, anapus, anas, anasai, anasis, anei, aniedvi, anieji, aniesiems, anoji, anojo, anojoje, anokia, anoks, anosiomis, anosioms, anosios, anosios, anosiose, anot, ant, antai, anuodu, anuoju, anuosiuose, anuosius, apie, aplink, ar, ar, arba, argi, arti, aukščiau, be, be, bei, beje, bemaž, bent, bet, betgi, beveik, dėka, dėl, dėlei, dėlto, dar, dargi, daugmaž, deja, ech, et, gal, galbūt, galgi, gan, gana, gi, greta, iš, išilgai, išvis, idant, iki, iki, ir, irgi, it, itin, jąją, jąja, jąsias, jįjį, jųjų, jųjų, jūsų, jūs, jūsiškė, jūsiškis, jaisiais, jajai, jajam, jajame, jei, jeigu, ji, jiedu, jiedvi, jieji, jiesiems, jinai, jis, jisai, jog, joji, jojo, jojoje, jokia, joks, josiomis, josioms, josios, josios, josiose, judu, judvi, juk, jumis, jums, jumyse, juodu, juoju, juosiuose, juosius, jus, kažin, kažkas, kažkatra, kažkatras, kažkokia, kažkoks, kažkuri, kažkuris, kad, kada, kadangi, kai, kaip, kaip, kaipgi, kas, katra, katras, katriedvi, katruodu, kiaurai, kiek, kiekvienas, kieno, kita, kitas, kitokia, kitoks, kodėl, kokia, koks, kol, kolei, kone, kuomet, kur, kurgi, kuri, kuriedvi, kuris, kuriuodu, lai, lig, ligi, link, lyg, mūsų, mūsiškė, mūsiškis, maždaug, mažne, manąją, manąjį, manąja, manąsias, manęs, manųjų, manųjų, man, manaisiais, manajai, manajam, manajame, manas, manasai, manasis, mane, maniškė, maniškis, manieji, maniesiems, manim, manimi, mano, manoji, manojo, manojoje, manosiomis, manosioms, manosios, manosios, manosiose, manuoju, manuosiuose, manuosius, manyje, mat , mes, mudu, mudvi, mumis, mums, mumyse, mus, nė, na, nagi, ne, nebe, nebent, nebent, negi, negu, nei, nei, nejau, nejaugi, nekaip, nelyginant, nes, net, netgi, netoli, neva, nors, nors, nuo, o, ogi, ogi, oi, paėių, paėiais, paėiam, paėiame, paėiu, paėiuose, paėius, paeiliui, pagal, pakeliui, palaipsniui, palei, pas, pasak, paskos, paskui, paskum, patį, pat, pati, patiems, paties, pats, patys, per, per, pernelyg, pirm, pirma, pirmiau, po, prieš, priešais, prie, pro,

		pusiau, rasi, rodos, sau, savąją, savąjį, savąja, savąsias, savęs, savųjų, savųjų, savaisiais, savajai, savajam, savajame, savas, savasai, savasis, save, saviškė, saviškis, savieji, saviesiems, savimi, savo, savoji, savojo, savojoje, savosiomis, savosioms, savosios, savosios, savosiose, savuoju, savuosiuose, savuosius, savyje, skersai, skradžiai, staėiai, su, sulig, tąją, tąjį, tąja, tąsias, tųjų, tųjų, tūlas, tačiau, ta, tad, tai, tai, taigi, taigi, taip, taipogi, taisiais
3	Galūnės	a, ai, ajam, ame, ams, ant, as, asis, au, ą, ąjį, e, ė, ei, ei, ėj, ėje, ėmis, es, ėse, ęs, i, ia, iai, iai, iai, iais, iams, ias, iau, ią, iąsias, ieji, int, io, ioje, iomis, ioms, ios, iose, iosiomis, iosios, iosiose, is, iuos, iuose, ius, ių, iųjų, į, y, yje, ys, ysim, k, o, oje, oji, ojo, os, s, š, t, ti, ui, um, uoju, uose, uosius, us, us, ūs, ų, ųjų
4	Nebūdingos bigramos	qu, sh, zh, ch, ux, xu, wa, we, wi, wu, wo, aw, ew, iw, uw, ow, aj, ej, xe, 2x, 2k, 2c, 2h, 2r, 2s, 4e, sq, šq, sw, wx, zd, rw, rc, rč, zn, tc, gh, žn, kz, kž, jk, ee
5	Sutrumpinimai	a., a. k., a. s., adv., akad., aklg., akt., al., apyg., aps., apskr., asist., asmv., avd., atsak., aut., avd., asmv., biol., b. k., bkl., bot., bt., buv., chem., d., dail., dek., dėst., dir., dirig., doc., dr., drp., dš., e. (el) p., egz., eil., ekon., el. (e.), etc., ež., fak., faks., filol., filos., g., G., gen., geol., gerb., gim., gyd., gv., įl., Įn., insp., inž., ir pan., ir t. t., istor., J. E., J. Em., k., K., k. a., kand., kat., kyš., kl., kln., kn., koresp., kpt., kr., kt., kun., l. e. p., ltn., m., m., mst., m. e., m. m., mat., med., mgr., mgnt., mjr., mln., mlrd., mok., mokyt., moksl., mot., mst., m., mstl., N., nkt., ntk., Nr., p., p. d., p. m. e., pav., pavad., pav., pirm., pl., plg., plk., pr., pr. Kr., prof., prok., prot., pss., pšt., pvz., r., red., rš., s., sąs., sav., saviv., sekr., sen., sk., skg., skyr., sk., skv., sp., spec., sr., st., str., stud., š. m., šnek., t., t. y., t. p., techn., tel., teol., tir., tūkst., up., upl., V., vad., ved., vet., virš., vyr., vyresn., vlsč., vs., Vt., vtv., vv., vv., vtv., zool., žml., žr., ž. ū.
6	Palyginimai	pvz, pvz., pavyzdžiui, kaip antai, lyg, kaip, tarkim, sakykim, kaip koks, kaip kokia, kaipkokie, kaip kokios,kaip kokiuose, kaip kokiose, tarsi, būtent kaip, panašiai kaip,tarytum, tarytum koks, tarytum kokia, tarytum, kaip

2.3. Siūlomo metodo architektūra

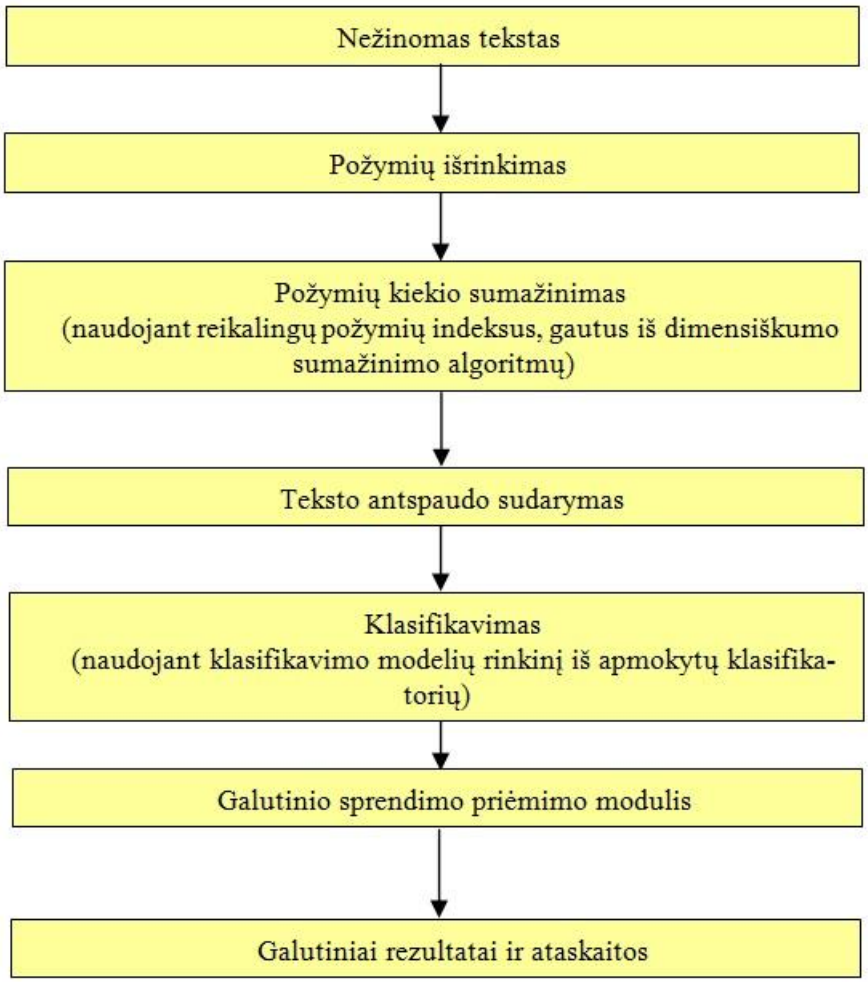
Šiame skyriuje pateikiama autorystės identifikavimo sistemos architektūra. Ją sudaro du pagrindiniai procesai: sistemos apmokymo, ir autorystės tikrinimo. Pirmasis, sistemos apmokymo procesas yra pavaizduotas žemiau esančioje koncepcinėje schemoje. Šio proceso metu, iš turimų,

identifikuotų autorių tekstų yra išrenkami numatyti lingvistiniai požymiai, kurių skaitines reikšmes apdoroja keli dimensiškumo sumažinimo algoritmai, paliekantys tik reikšmingų požymių skaitinių reikšmių rinkinius. Tuomet šie rinkiniai yra suvedami į reikiamo formato duomenų struktūras, ir taip gaunami taip vadinami teksto antspaudai. Naudojant šiuos požymių skaitinių reikšmių rinkinius yra apmokomi sistemos mašinos mokymosi klasifikatoriai. Šio subproceso metu yra gaunamas klasifikavimo modelių rinkinys, pagal kurį vėliau atliekamas autorystės tikrinimo procesas.



2.1 pav. Sistemos apmokymo proceso koncepcinė schema

Žemiau esančioje koncepcinėje schemoje pavaizduotas autorystės tikrinimo procesas. Šis procesas pradedamas tų pačių lingvistinių požymių, kaip ir sistemos apmokyme, išrinkimu iš tiriamo nežinomo autoriaus teksto. Tuomet, pagal apmokymo proceso metu įvykdytų dimensiškumo sumažinimo algoritmų gautus reikšmingų požymių indeksų rinkinius, paliekami tik atitinkamai tie patys reikšmingi požymiai tiriamojo teksto požymių rinkiniuose (gaunami keli rinkiniai, kadangi pradinis tiriamojo teksto požymių rinkinys yra atskirai naudojamaskiekvienam dimensiškumo sumažinimo algoritmų reikšmingų požymių indeksų rinkiniui). Tuomet šių požymių skaitinės reikšmės yra suvedamos į atitinkamai tokias pačias duomenų struktūras, kurias naudojant kartu su apmokymo procese gautais klasifikavimo modeliais, atliekamas klasifikavimo subprocesas, naudojant kelis mašinos mokymosi klasifikatorius. Šių klasifikatorių rezultatai yra įvertinami ir apibendrinami sistemos galutinio sprendimo priėmimo modulyje, kuris pateikia galutinius rezultatus bei sugeneruoja ataskaitas.

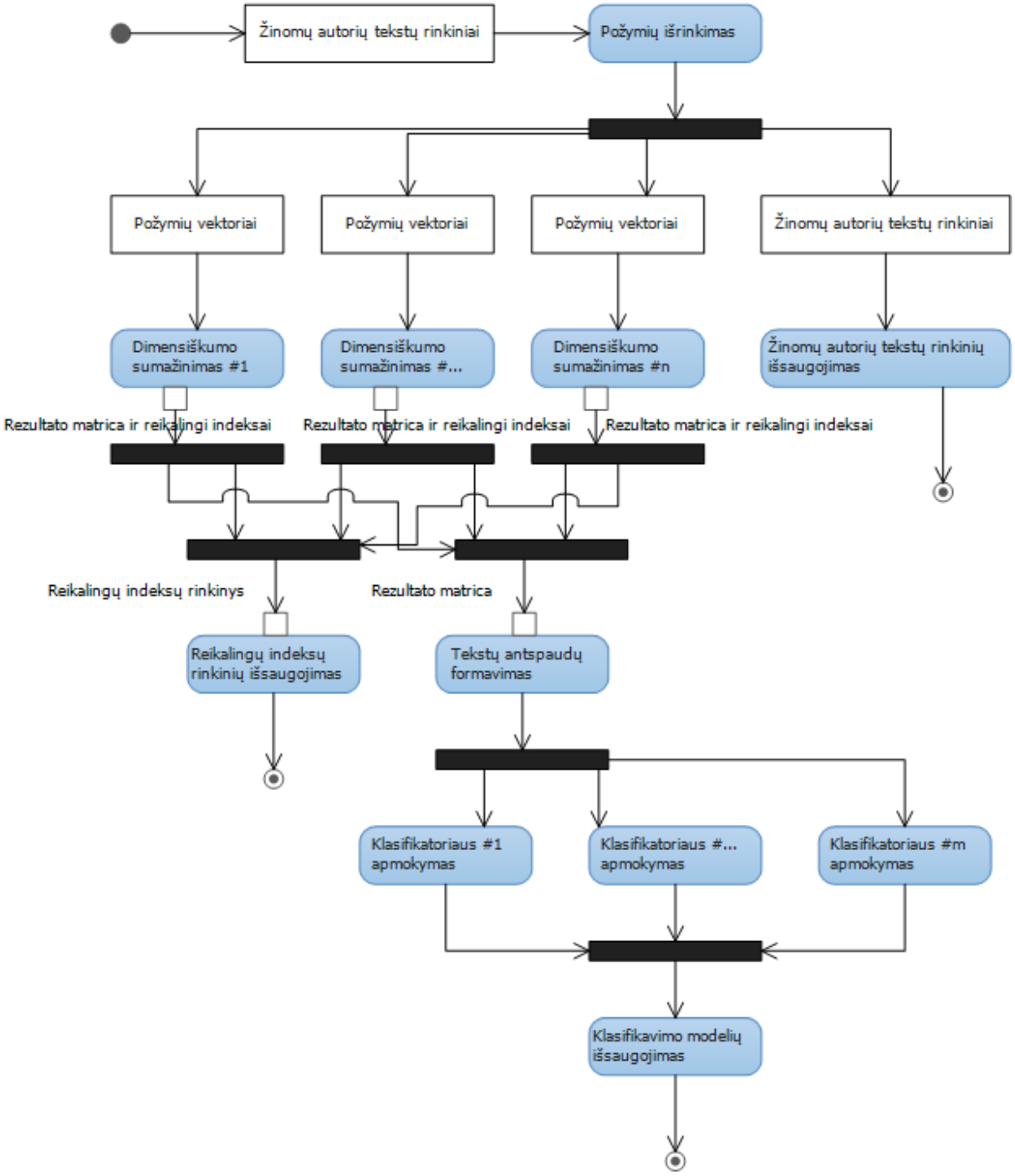


2.2 pav. Autorystės tikrinimo proceso koncepcinė schema

Šie procesai ir subprocesai detaliau aprašomi tolesniuose skyreliuose.

2.3.1. Klasifikatorių apmokymas

2.3 pav. pavaizduoti apmokymo proceso duomenų srautai. Šio proceso eigoje surenkami iš žinomų tekstų sugeneruoti pavyzdininiai tekstai. Apmokymo rezultatai yra klasifikacinių modelių rinkinys, galintis padėti suklasifikuoti autorius, atsižvelgiant į pateikto teksto autorystės tikimybę.



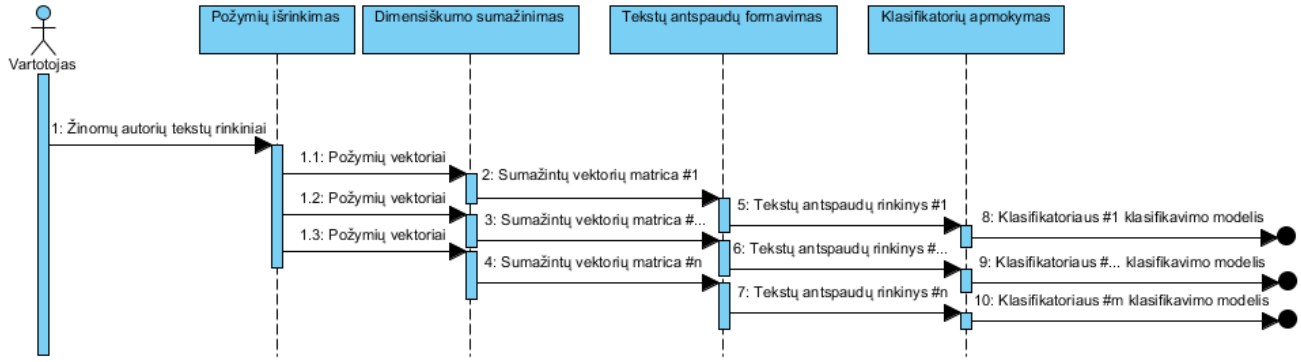
2.3 pav. Apmokymo duomenų srautų diagrama

Klasifikacinių modelių konstravimą sudaro keturi žingsniai :

1. Požymių išskyrimas - suskaičiuojamos prieduose įvardintų požymių savybės ;
2. Dimensijos mažinimas - požymiai yra suklasifikuojami pagal jų galėjimą atskirti tekstų autorius. Gali būti pritaikyti keli dimensijos mažinimo metodai : PCA (tiesioginių komponenčių analizė), LDA (tiesioginė gradientinė analizė) ir t.t. Dimensijos metodai specifiniams požymiams tirti bus apibrėžti po eksperimentinio vertinimo.

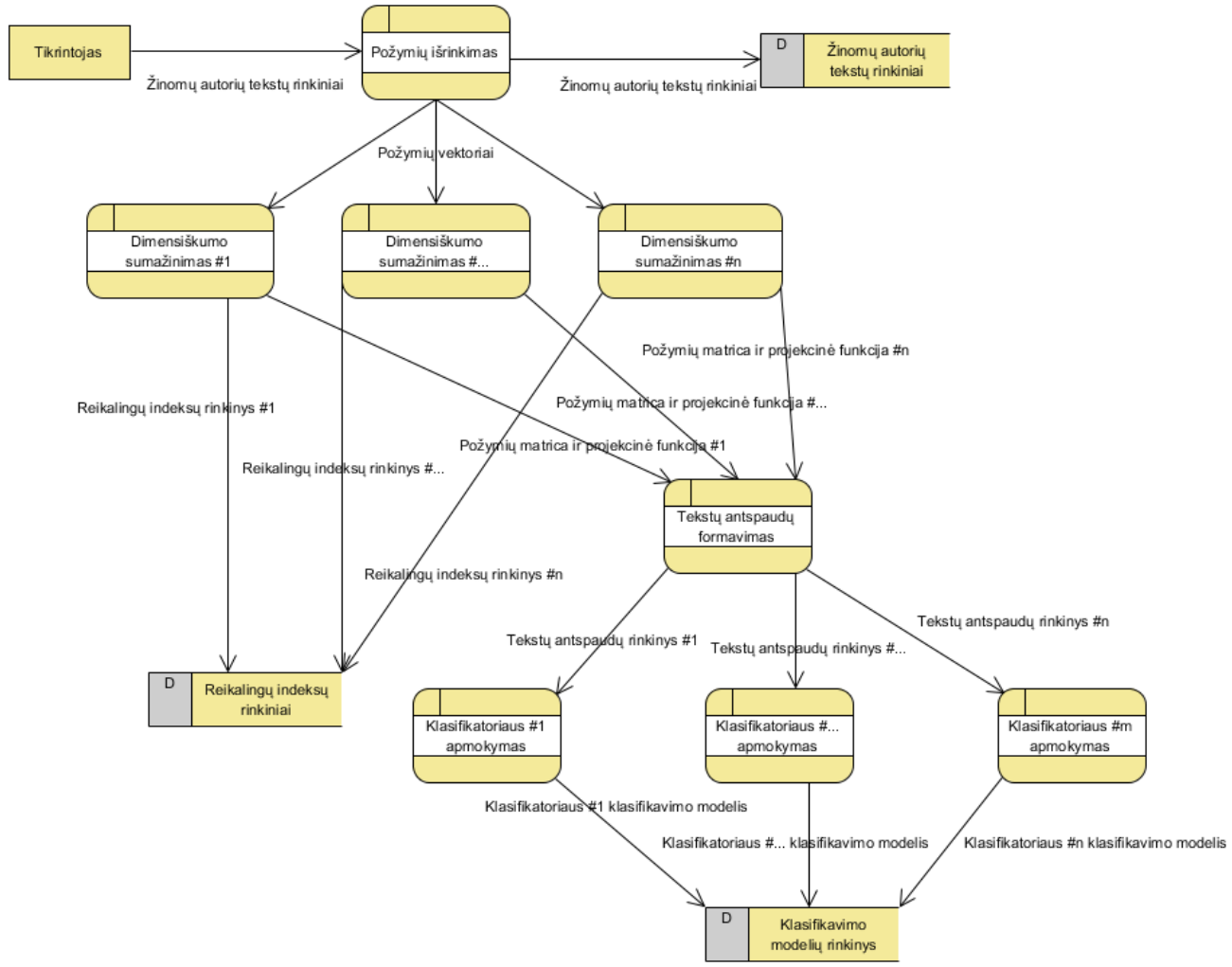
3. Teksto antspaudai (angl. *Writeprints*) - esminiai požymiai bus išskirti ir tik šie požymiai bus naudojami apmokymų klasifikatoriams, norint išvengti dimensionalumo prakeiksmo.Eksperimentiniu būdu bus apibrėžta konkreti ribinė vertė požymiams atrinkti.
4. Mokymo klasifikatoriai - klasifikatorių rinkinys bus parengtas, naudojant skirtingus klasifikavimo modelius, tiesiogiai susijusius su autorystės nustatymo sritimi. Konkretus klasifikatorių, kurie turi būti įtraukti į sistemą, rinkinys bus apibrėžtas po atliktų išsamių bandymų. Bus nagrinėjami šie metodai : atraminių vektorių metodas (SVM), ekstremalausmokymosi metodas (ELM), k-Artimiausio kaimyno (angl. *k-Nearest Neighbours*) ir kiti.

2.4 pav. vaizduojama apmokymo proceso sekos diagrama. Ji rodo, kaip susidaro apmokytoklasifikatoriaus modeliai, kaip sąveikos, išsidėstančios laiko juostoje. Kai tik įmanoma, požymių išskyrimas, dimensionalumo mažinimas, teksto antspaudavimas ir klasifikatorių apmokymas bus atliekami paraleliai, kad kompiuterio ištekliai būtų panaudojami efektyviau, o rezultatai gauti pertrumpesnį laiką.



2.4 pav. Apmokymo sekų diagrama

2.5 pav. vaizduoja apmokymo proceso sąveikos diagramą. Ji rodo, kaip sudaromi apmokyto klasifikatoriaus modeliai, kaip sudėtinis veiklų, kurias reikia atlikti, planas. Apmokymo rinkinių rezultatai yra sudaryti svarbių požymių rinkiniai (vaizduojami kaip indeksų vektorius) ir apmokytų klasifikatorių modelių rinkinys.



2.5 pav. Apmokymo sąveikos diagrama

2.3.2. Klasifikatorių apmokymas naudojant atraminių vektorių klasifikatorių

Šiame skyrelyje aprašoma Atraminių vektorių klasifikatoriaus (angl. santrumpa *SVM*) realizacija. SVMyra klasifikavimo algoritmas, pagrįstas struktūrinės rizikos minimizavimu. Pirmiausiai, SVM atvaizduoja klasifikatoriaus modelio apmokymui skirtus duomenis į aukštesnio dimensiškumo požymių erdvę. Tada požymių erdvėje sukonstruojama skiriančioji hiperplokštuma, kuri atskiria duomenų klases ir maksimizuoja atstumą tarp hiperplokštumos ir jai artimiausių taškų. Ši hiperplokštuma naudojama nežinomų duomenų klasifikavimui į klases.
SVM yra galingas algoritmas, kurį naudojant galima pasiekti labai gerų rezultatų. Tačiau šie rezultatai labai priklauso nuo apmokymui naudojamo duomenų rinkinio, duomenų atvaizdavimo į požymių erdvę, probleminei sričiai tinkamos branduolio funkcijos parinkimo, branduolio funkcijos parametrų (jei tokie yra) ir klasifikatoriaus apmokymo parametrų. SVM algoritmo įvesties duomenys yra realių skaičių vektorius požymių erdvėje. Tinkamai parinkus požymius galima pasiekti aukštesnius klasifikavimo rezultatus. Taip pat klasifikavimo tikslumąįtakoja SVM algoritmo branduolio funkcijos parinkimas. Dažniausiai SVM algoritme naudojamostiesinė, polinominė, radialinė ir sigmoidinė funkcijos. Sistemoje bus realizuotos visos šios funkcijos, o eksperimentinių tyrimų metu bus nustatyta branduolio funkcija, leidžianti pasiekti geriausius rezultatus. Taip pat apmokymą įtakoja ir kiti specifiniai SVM algoritmo parametrai, tokie kaip SVMapmokymo parametrai. Sistemoje bus realizuota galimybė ekspertui parinkti šių parametrų reikšmes.

2.3.3. Vienos klasės SVM klasifikatorius

Sistemoje naudojamas vienos klasės atraminių vektorių mašinos (SVM) klasifikatorius. Ši klasifikatoriaus versija naudoja požymius be žymų ir stengiasi nustatyti, kurios požymių erdvės dalys saugo informaciją įprastai, o kurios ne. Paprastai ši priemonė naudojama ieškant išsiskiriančių objektų ar identifikuojant neįprastus duomenų pavyzdžius.
Požymiams saugoti naudojami stulpeliniai vektoriai. Vektoriai sudaromi iš požymių numerių ir požymių skaitinių reikšmių arba tik požymių skaitinių reikšmių (t. y. naudojami dvimačiai arba vienmačiai taškai), priklausomai nuo pasirinkto klasifikavimo režimo.

Sistemoje yra galimi keturi klasifikavimo režimai, kurie skirtingai įtakoja sistemos greitaveiką ir efektyvumą. Šiems režimams pasirinktinai naudojamos dvi branduolio funkcijos, iš kurių kiekvieną galima naudoti tik su požymių skaitinėmis reikšmėmis arba su požymių numeriais ir požymių skaitinėmis reikšmėmis. Galimos branduolio funkcijos yra tiesinės (linear kernel) arba radialinės (radial basis kernel).

Galimi keturi klasifikavimo režimai:

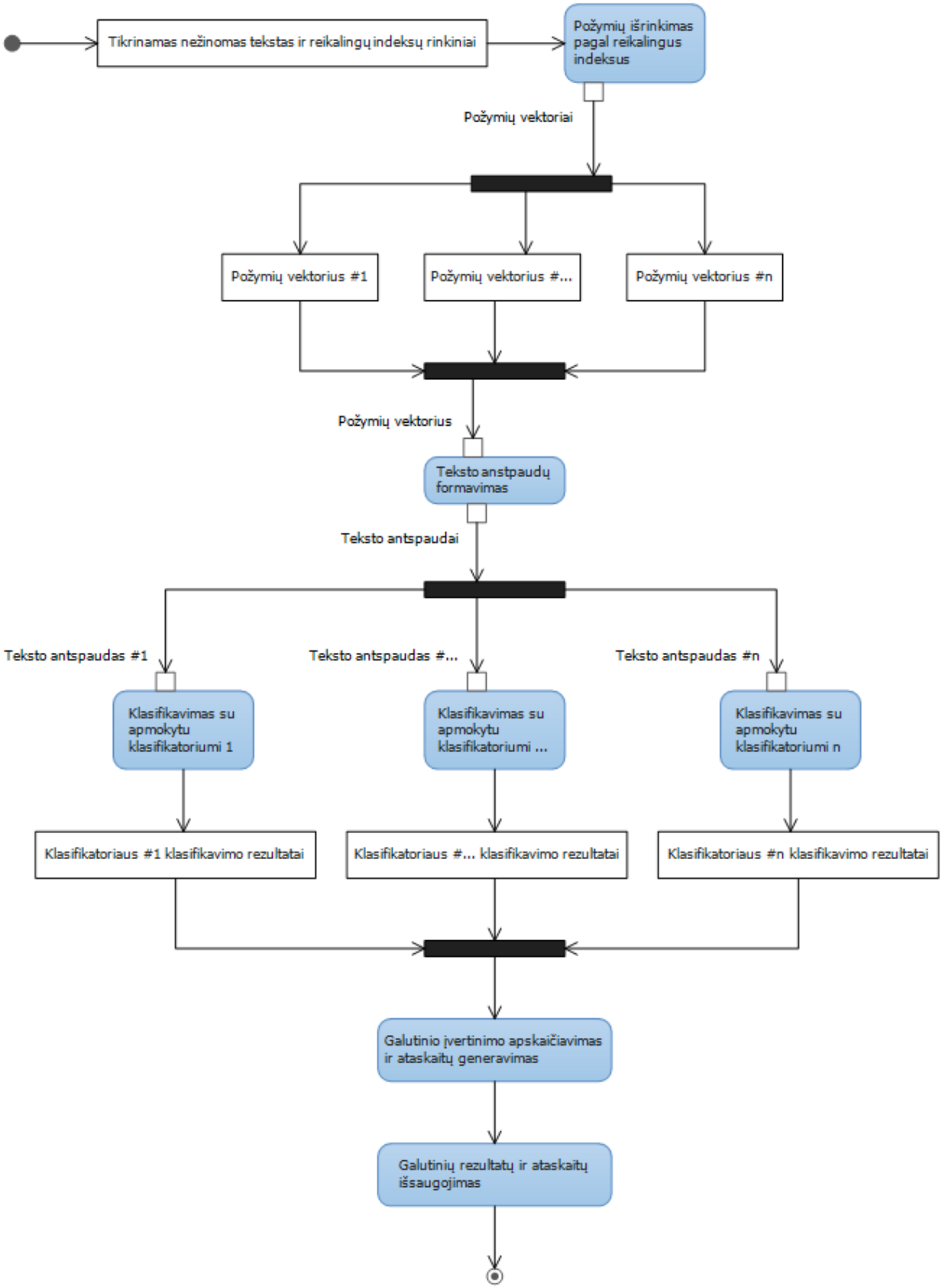
1. Radial basis kernel (naudojamos tik požymių skaitinės reikšmės)
2. Radial basis kernel (naudojami požymių numeriai ir požymių skaitinės reikšmės)
3. Linear kernel (naudojamos tik požymių skaitinės reikšmės)
4. Linear kernel (naudojami požymių numeriai ir požymių skaitinės reikšmės)

Naudojant radialinę branduolio funkciją (radial basis kernel), požymių erdvė yra atvaizduojama į didesnio dimensiškumo požymių erdvę. Tai leidžia geriau atskirti duomenų taškus, kadangi mažo dimensiškumo erdvėje juos gali būti sunku išskirti. Atvaizdavus požymių erdvę į didesnio dimensiškumo erdvę, yra ieškoma optimali hiperplokštuma, tiesiogiai neanalizuojant pačios požymių erdvės. Ši branduolio funkcija leidžia atlikti didelės erdvės padalinimą, o SVM atlieka generalizavimą, naudojant maksimalios ribos kriterijų, ir taip išvengiant perpildymo.

Tiesinė branduolio funkcija (linear kernel) tinkama naudoti su tiesiškai gerai atskiriamais duomenimis. Taip pat ši funkcija pasižymi didele greitaveika.

2.3.4. Klasifikavimas ir ataskaitos generavimas

2.6 pav. vaizduojami klasifikavimo ir ataskaitos rengimo procesai, nustatant teksto autorystę.



2.6 pav. Klasifikavimo ir ataskaitos generavimo duomenų srautų diagrama

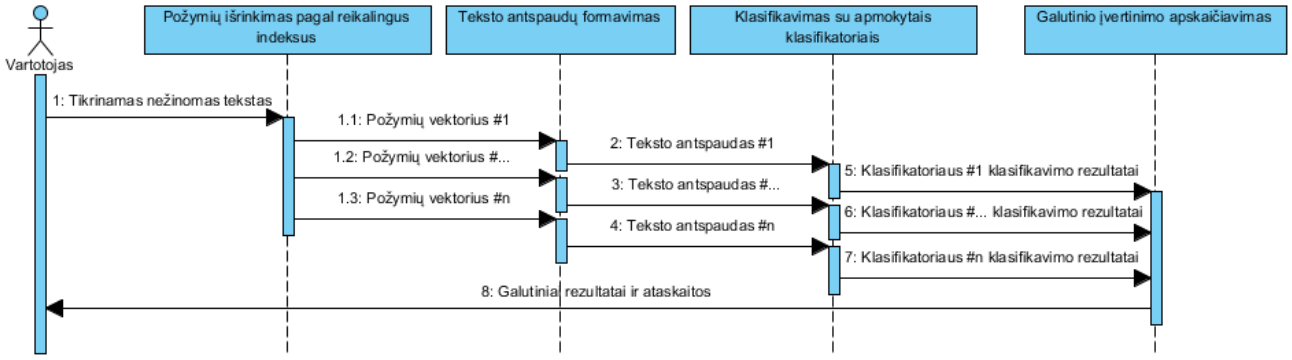
Pav. 2.7. Klasifikavimo ir ataskaitos generavimo duomenų srautų diagrama

Šį procesą sudaro tokie žingsniai:

1. Požymių išskyrimas - tik požymių rinkiniai yra išskiriami iš nenustatyto autoriaus parašyto teksto, norint sumažinti požymių išskyrimo laiką. Kiekvieno išskirtopožymio vektorius žymi poaibį reikšmingų požymių, išskirtų, taikant skirtingus požymių dimensijos mažinimo metodus.
2. Klasifikuojama, naudojant išankstinio apmokymo klasifikacijos modelių rinkinį.Klasifikavimo rezultatai bus išreikšti kaip sąrašas n-labiausiai tikėtinų analizuojamo teksto autorių.
3. Galiausiai skirtingus klasifikatorius taikant gauti rezultatai bus sugrupuoti, taikant sistemossprendimo palaikymo modulį. Modulis sudarys galutinį tikėtinų autorių sąrašą, sujungiant sąrašus, sugeneruotus skirtingų klasifikatorių modelių, taikant svertinio balsavimo schemą, kur kiekvieno klasifikatoriaus reikšmė bus apibrėžta atlikus eksperimentus, pagrįstus prieinamų tekstų duomenų rinkiniu. Bus parengta autoriaus nustatymo ataskaita, kuri tik patvirtins sistemų sprendimą, remiantis statistiniais pačių svarbiausių požymių skirtumais, išskiriant svarbiausius tarpautorinius požymius, kurie tuo pačiu metu nekinta to patiesautoriaus tekstuose.

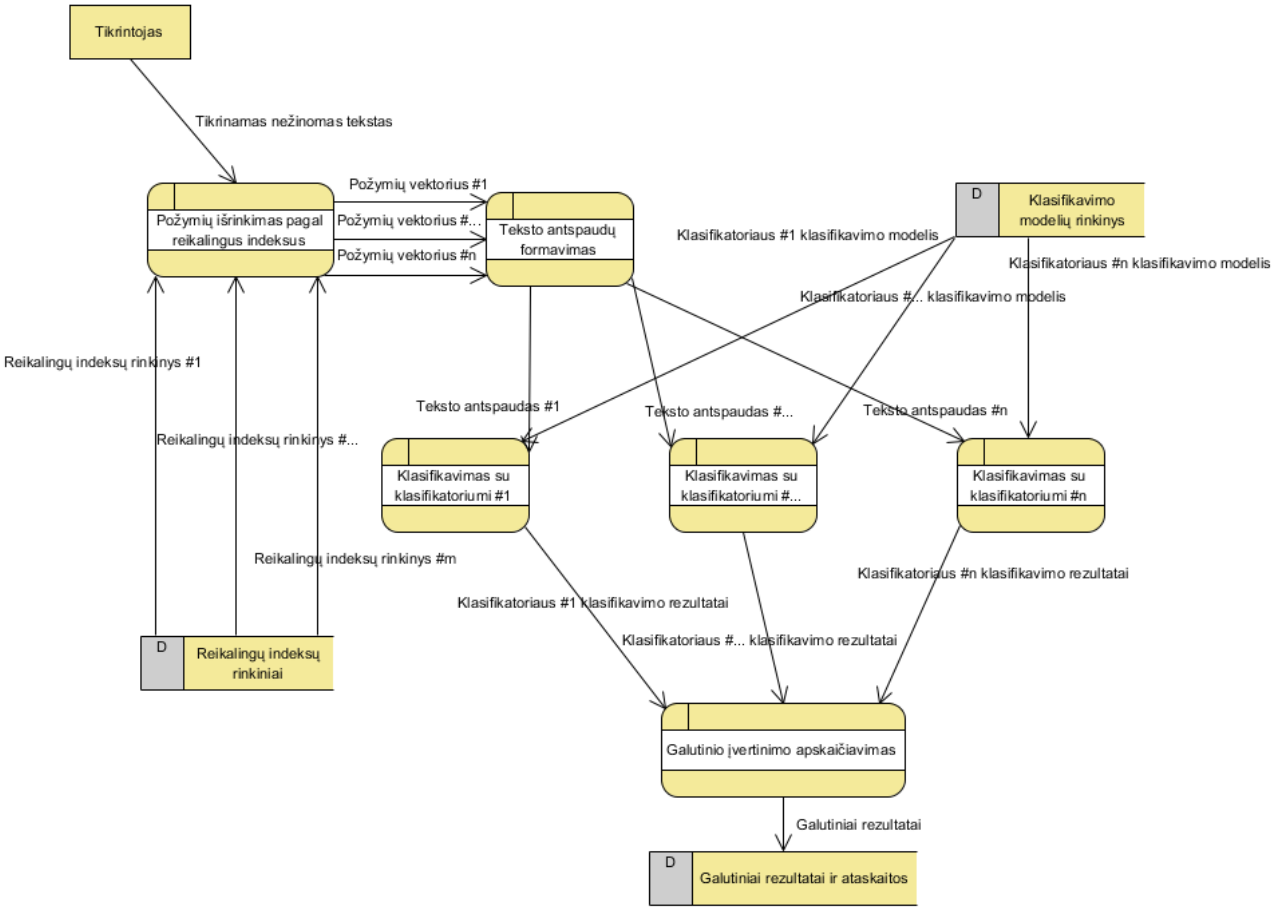
Naudojant SVM klasifikatorių, dvejetainio klasifikavimo atveju rezultato klasė vadinama teigiama arba neigiama klase. Klasifikuojant nežinomus duomenis SVM klasifikatoriaus grąžinama reikšmė yraduomenų vektoriaus klasė. Mūsų realizuojamoje autorystės nustatymo sistemoje, bus realizuotas modifikuotas SVM algoritmo rezultatas, kuris grąžins duomenų vektoriaus kaip taško požymių erdvėje atstumą iki klases skiriančios hiperplokštumos. Ši atstumo reikšmė bus naudojama įvertinti tikimybei, kad duomenų vektorius priklauso tam tikrai konkrečiai klasei. Eksperimentų pagalbą bus nustatyta statiškai reikšminga atstumo reikšmė, leidžianti patikimai priskirti duomenų vektorių tam tikraiklasei. Tai leis naudoti dvejetainį klasifikatorių kaip daugelio klasių klasifikatorių, kuris grąžins klasių sąrašą, kurioms galima priskirti analizuojamą duomenų vektorių (t.y., analizuojamam tekstui bus grąžinamas įtariamų autorių sąrašas). Klasifikavimo įvertinimui bus naudojama daugelio klasių klasifikavimui pritaikyta ir eksperimentiškai patikrinta tikslumo metrika.

2.7 pav. pavaizduota klasifikacijos ir ataskaitos ruošimo sekos diagrama. Ji rodo, kaip prieinama prie autoriaus nustatymo sprendimo, laikant jį sąveikomis laiko juostoje. Kai tik galima, svarbūs požymiai išskiriami, antspaudai uždedami ir klasifikacija, taikant iš anksto apmokytus klasifikatorius, bei ataskaita pateikiama kartu. Kompiuterio ištekliai būtų panaudojami efektyviau, o rezultatai gauti per trumpesnį laiką.



2.7 pav. Klasifikavimo ir ataskaitos generavimo sekų diagrama

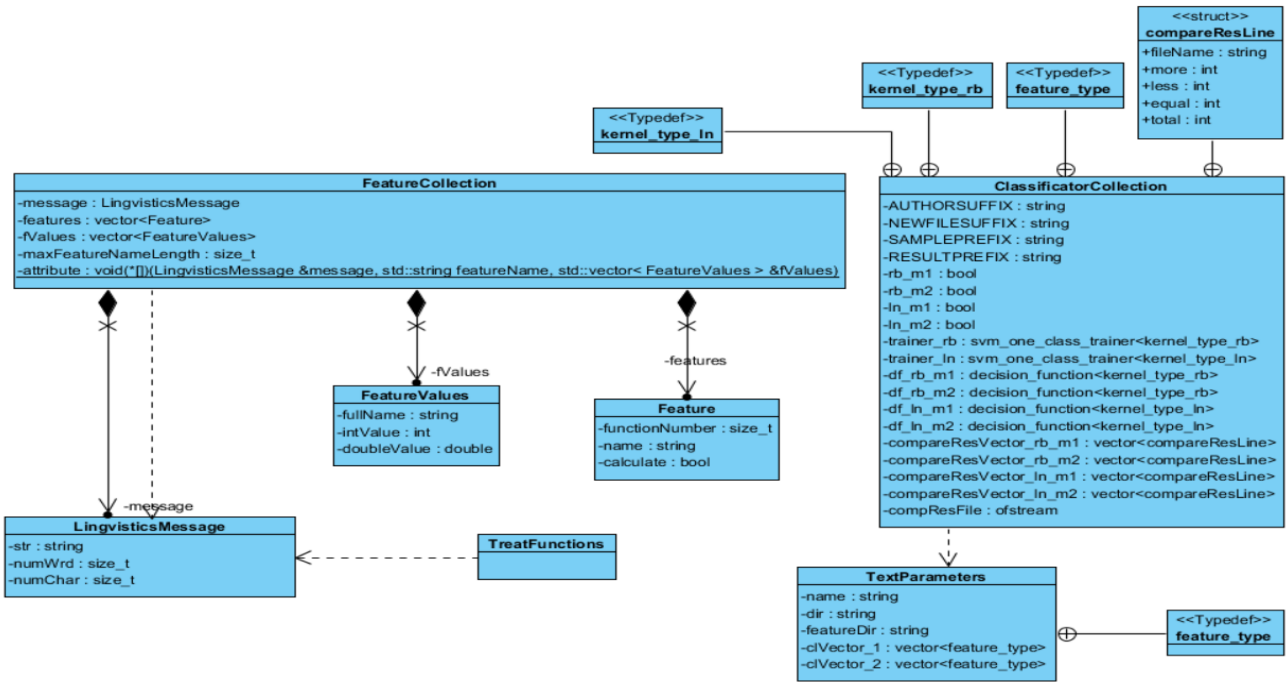
2.8 pav. atvaizduota klasifikacijos ir ataskaitos rengimo sąveikos diagrama. Ji rodo, kaip prieinama prie autoriaus nustatymo sprendimo, kuris laikomas sudėtiniu veiksmų, kurias reikia atlikti, planu.



2.8 pav. Klasifikavimo ir ataskaitos generavimo sąveikos diagrama

2.3.5. Sistemos struktūra

2.9 pav. pavaizduota sistemos klasių diagrama



2.9 pav. Sistemos struktūra (klasių diagrama)

2.4. Standartizuotos tikslumo įvertinimo metrikos

Autorystės nustatymo sistemos dažnai yra įvertinamos naudojant įvairias, standartizuotas tikslumo įvertinimo metrikas.

Paprasčiausia tikslumo metrika yra apskaičiuojama skiriant visą dėmesį pirmajam įtariamam autoriui ir apskaičiuojant vidutinę tikimybę tiesiogiai atspėti šį pirmąjį autorių kaip tikrąjį (Winner-Takes-All, WTA).

$$WTA = \frac{1}{|T^k|}\sum_{T^k} H(rank(a_{true}) = 1)$$

Kur $rank(a_{true})$yra tikrojo autoriaus pozicija, H – Hevisaido funkcija, o T^k – tikrinamų tekstų rinkinys.

Kita, laisvesnė tikslumo įvertinimo metrika yra apskaičiuojama nustatant ar tikrasis autorius yra įtariamų autorių sąraše, nepriklausomai nuo autoriaus pozicijos sąraše. Ši metrika vadinama sąrašo tikslumu (angl. List Precision, LP) ir yra aprašoma žemiau nurodyta išraiška.

$$LP(L) = \frac{1}{|T^k|}\sum_{T^k} H(rank(a_{true}) \leq L)$$

Kur L yra įtariamų autorių sąrašo ilgis.

Kitos naudojamos metrikos remiasi autorių eilės įvertinimu, kur skiriama svarba kuo aukštesnei tikrojo autoriaus pozicijai įtariamų autorių eilėje. Šios metrikos gali būti apibrėžiamos žemiau nurodyta išraiška.

$$\pi \leftarrow \sum_{k=1}^{K} gain(k) \cdot discount(k)$$

Kur pelno (gain) funkcija yra dydis, susietas su tikrojo autoriaus pozicija k autorių eilėje, o nuolaidos (discount) funkcija – dydis, nepriklausantis nuo autoriaus ir susietas su pozicija k.

Keletas skirtingų pelno-nuolaidos funkcijų metrikų, kurios remiasi atvirkštinio vidurkio rangu, normalizuotu DCG, ir rangu paremtu tikslumu, trumpai aprašomos žemiau.

Atvirkštinio vidurkio rangas yra statistinis matas, skirtas įvertinti bet kokį procesą, kuris pateikia sąrašą galimų atsakų į užklausų pavyzdį, surikiuotą pagal teisingumo tikimybę. Žemiau

pateikiama šio mato išraiška, interpretuojama kaip tikrojo autoriaus atvirkštinio rango, pagal pavyzdinių tekstų rinkinį, vidurkis.

$$MRR = \frac{1}{|T^K|}\sum_{T^K}\frac{1}{rank(a_{true})}$$

DCG metrika leidžia apskaičiuoti teisingai atpažintų autorių skaičių ir naudoja logaritminę funkciją progresyviam žemiau sąraše esančių autorių svarbos sumažinimui.

$$DCG = \frac{1}{|T^K|}\sum_{T^K}\frac{1}{\log_2(rank(a_{true})+1)}$$

Rangu paremtas tikslumas yra metrika, kuri priskiria efektyvumo balą rangui, skaičiuojant autoriaus pozicijų reikšmių geometriškai svorinę sumą su monotoniškai mažėjančiais svoriais geometriniame pasiskirstyme, apibrėžtame išlaikymo parametru p, $0 \leq p < 1$, kur mažesnės p reikšmės suteikia didesnę įtaką autoriams, esantiems aukščiau range, o didesnės p reikšmės išplečia svorį žemiau sąraše esantiems autoriams, bet abiem atvejais su visais autoriais range, įeinančiais į galutinį balą.

$$RPB = \frac{1}{|T^K|}(1-p)\sum_{T^K}p^{rank\,(a_{true})-1}$$

2.5. Išvados

1. Sudarytas elektroninio diskurso autorystės nustatymo metodas, pritaikytas lietuvių kalbai, kuriame naudojami ir bendrieji lingvistiniai požymiai ir tik lietuvių kalbai būdingi požymiai, su galimybe pasirinktinai naudoti norimus jų rinkinius.
2. Sudarytame metode naudojami keli dimensiškumo sumažinimo algoritmai bei mašinos mokymosi klasifikatoriai, kurių rezultatai yra apibendrinami ir įvertinami galutinio sprendimo priėmimo modulyje.
3. Keletos dimensiškumo sumažinimo algoritmų bei mašinos mokymosi klasifikatorių architektūra taip pat leidžia įvertinti ir naudoti efektyviausius jų derinius.
4. Sudaryto metodo architektūra yra pritaikyta lygiagrečiam veikimui, ir jos realizacija gali efektyviai dirbti su keletos branduolių procesoriais.

3. ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMO METODO REALIZACIJA

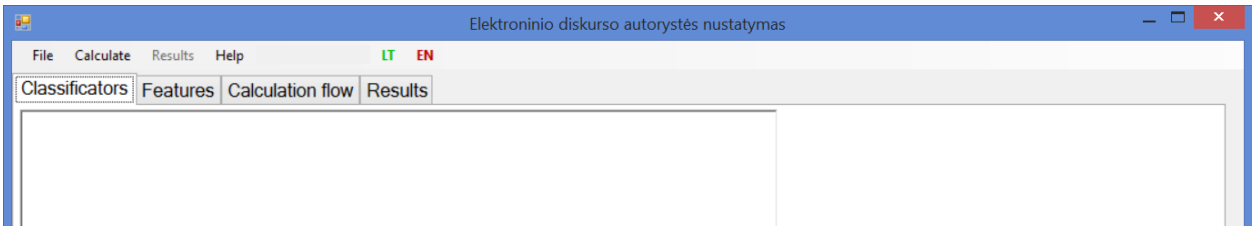
3.1. Trumpas metodo realizacijos aprašymas

Dėl didelės metodo apimties, realizacijoje panaudotas vienos klasės atraminių vektorių (SVM) mašinos mokymosi klasifikatorius. Šis klasifikatorius gali veikti keturiais klasifikavimo režimais, naudojančiais tiesinę arba radialinio pagrindo branduolio funkciją. Šiuos režimus galima nustatyti naudojantis grafine vartotojo sąsaja arba nustatymų failais. Sistemoje naudojami tiek bendrieji, tiek lietuvių kalbai būdingi požymiai, kuriuos taip pat galima pasirinkti.

3.2. Grafinės vartotojo sąsajos bandymams sudarymas

Programos grafinės sąsajos langas paleidžiamas du kartus pele paspaudus programos šakniniame kataloge esantį failą **AuthorshipAnalysis.exe.**

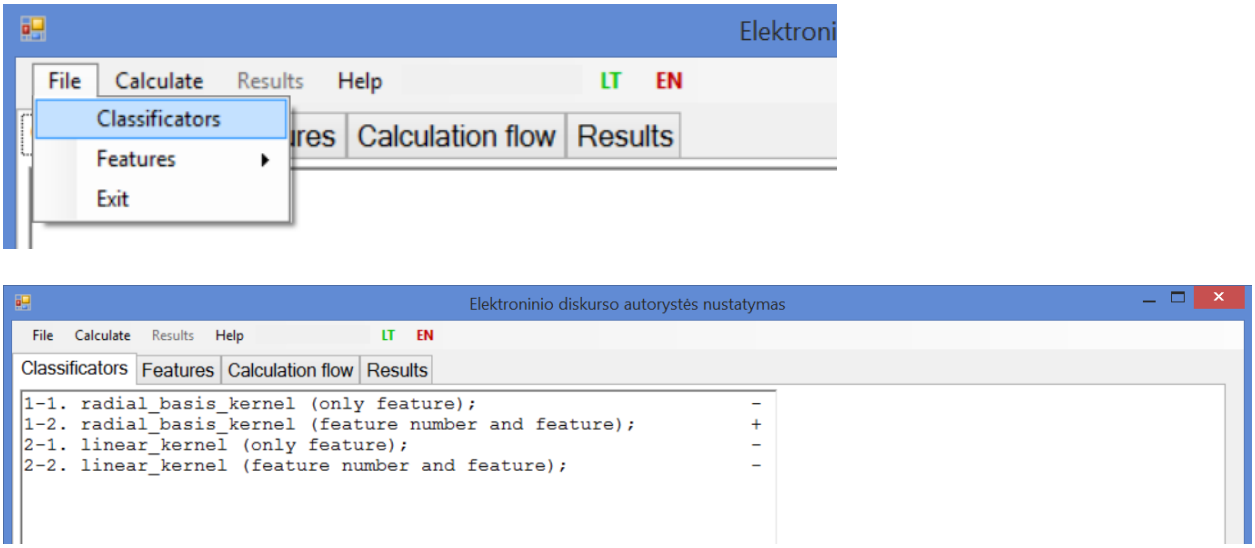
Programos grafinės sąsajos langas pateiktas 3.1 paveiksle.



3.1 pav. Pagrindinis grafinės sąsajos langas

3.2.1. Klasifikavimo režimų nustatymas

Norint nurodyti kurie klasifikavimo režimai bus naudojami programos vykdymo metu, reikia atverti klasifikavimo režimų nustatymų failą. Tą galima padaryti programos meniu juostoje pasirinkus *File →Classificators* (pav. 3.2).

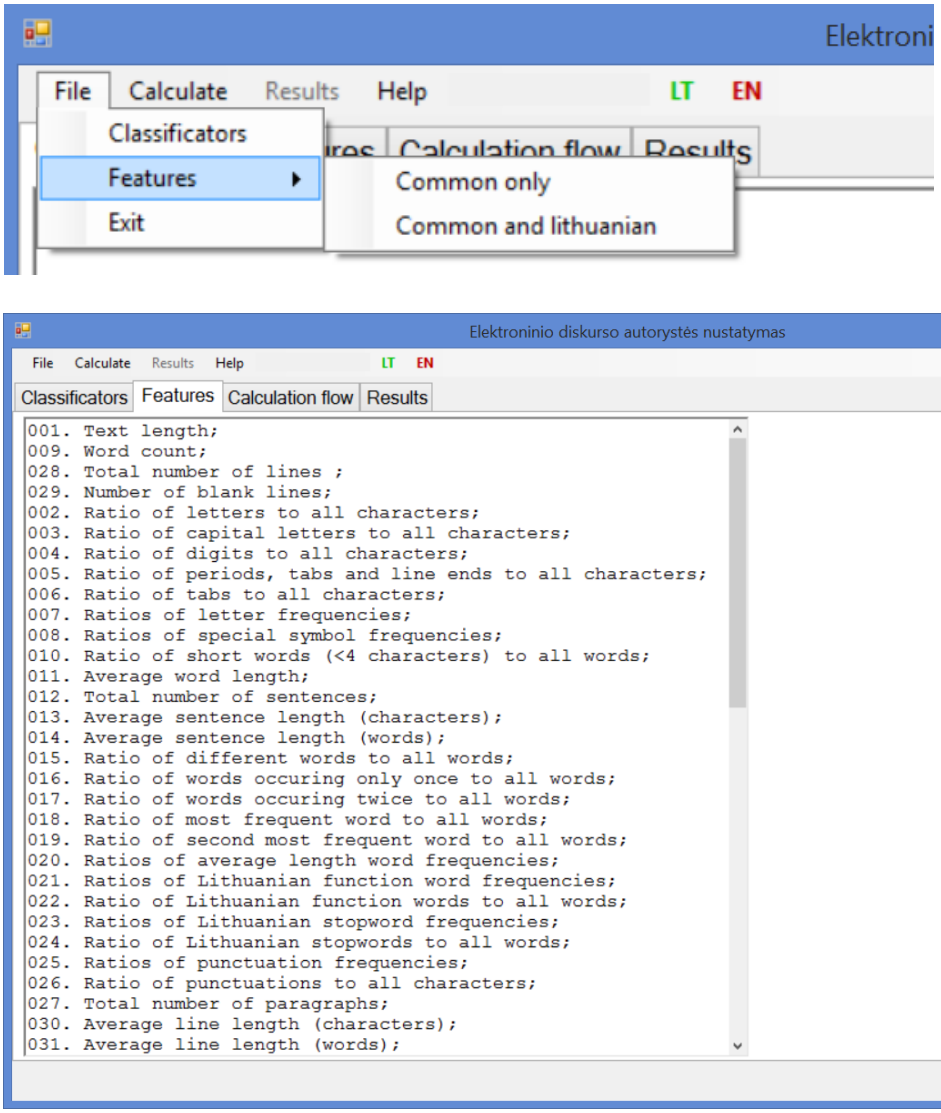


3.2 pav. Klasifikavimo režimų pasirinkimas

Atvėrus klasifikavimo režimų nustatymo failą, galima nurodyti (pažymėti +), kurie klasifikavimo režimai bus naudojami programos vykdymo metu.

3.2.2. Požymių nustatymas

Norint nurodyti kokie požymiai bus naudojami programos skaičiavimų metu, reikia atverti požymių nustatymų failą. Tą galima padaryti programos meniu juostoje pasirinkus *File →Features* (pav. 3.3).



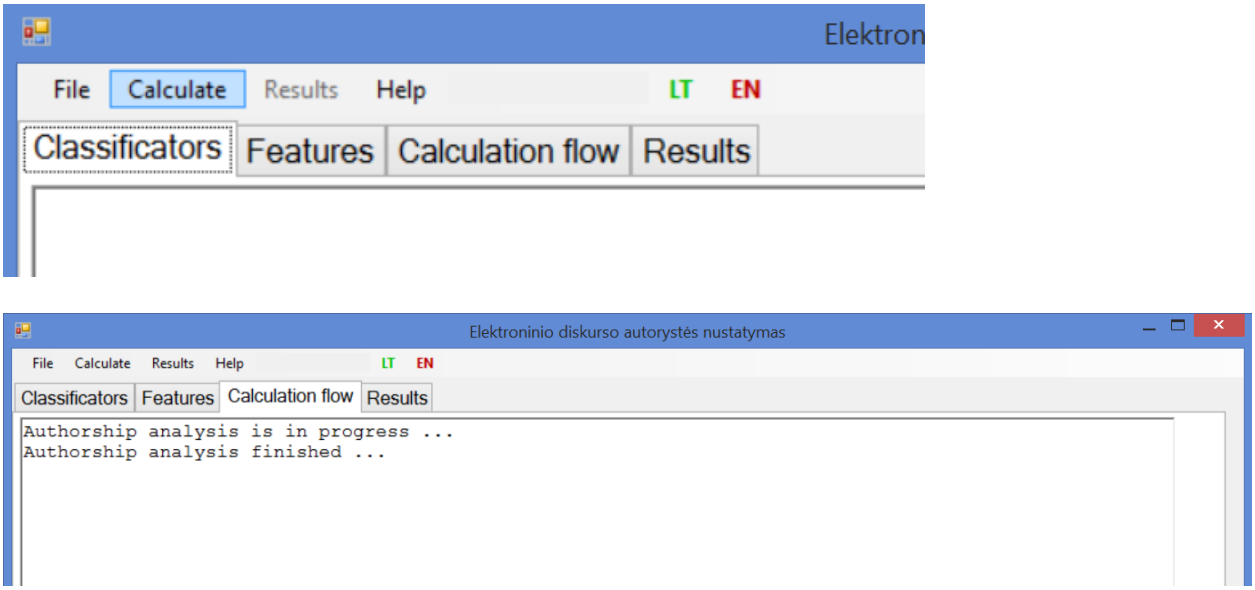
3.3 pav. Lingvistinių požymių pasirinkimas

Atvėrus požymių nustatymo failą, galima nurodyti (pažymėti +), kurie požymiai bus naudojami programos vykdymo metu.

3.2.3. Programos skaičiavimų vykdymas

Prieš pradedant programos skaičiavimų vykdymą, reikia programos kataloge „AnalysisText“ patalpinti tiriamo teksto failą, o kataloge „TextDB“ – tekstų bazės failus. Visi šie failai turi būti su plėtiniu *.info

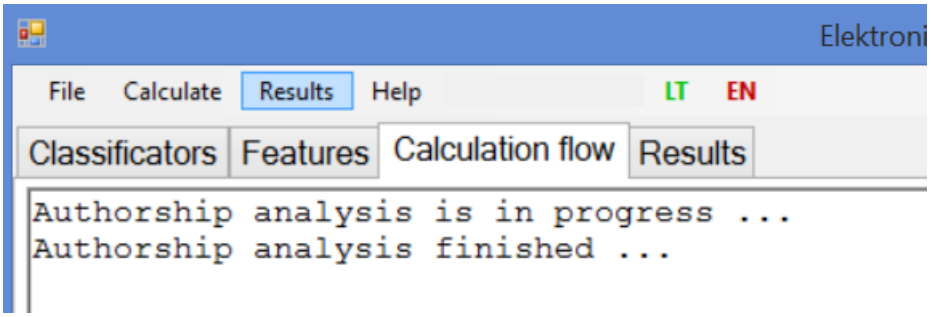
Atlikus nustatymus bei patalpinus tiriamo teksto bei tekstų bazės failus reikiamuose programos kataloguose, galima pradėti programos skaičiavimų vykdymą. Tam programos meniu juostoje paspauskite „Calculate“ (pav. 3.4).

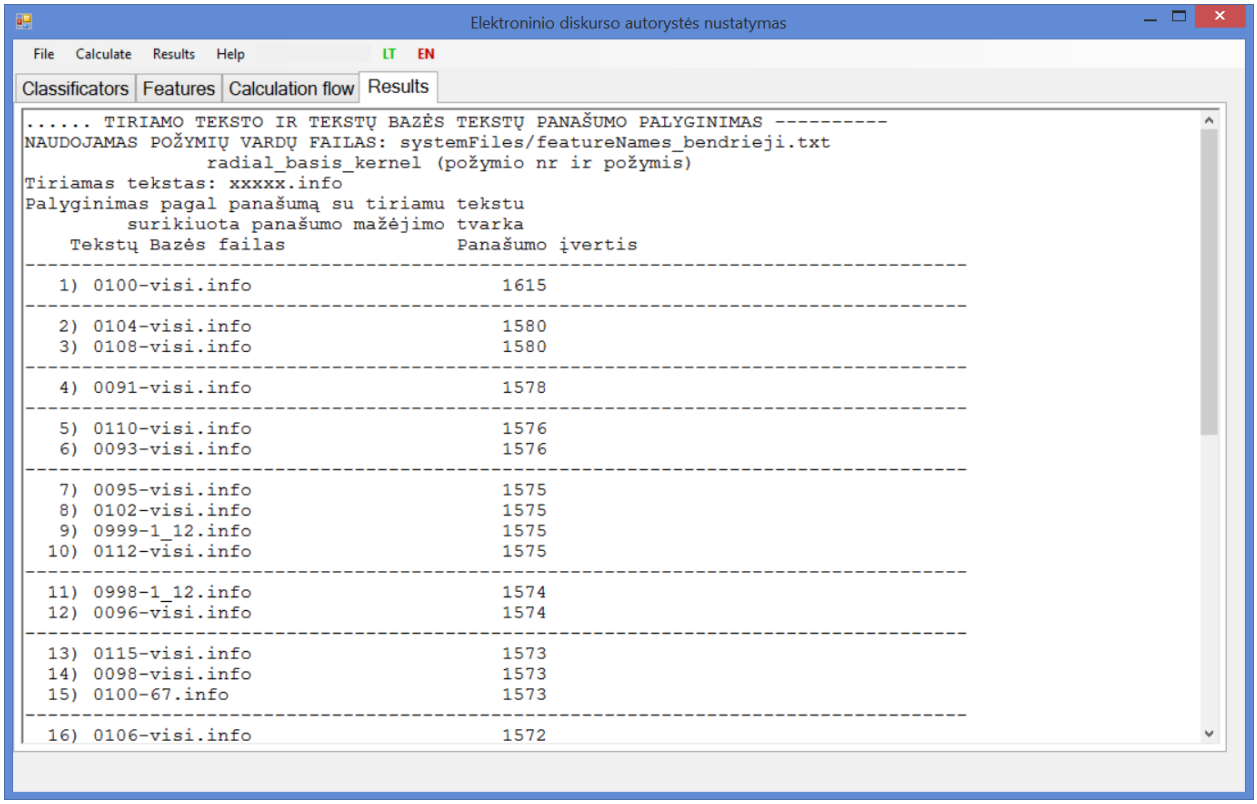


3.4 pav. Skaičiavimų inicijavimas

3.2.4. Rezultatų peržiūra

Programai baigus vykdyti skaičiavimus, galima peržiūrėti gautus rezultatus. Tam programos grafinės sąsajos lange, meniu juostoje paspauskite „Results“ (pav. 3.5).





3.5 pav. Rezultatų peržiūra

3.2.5. Rezultatų paaiškinimas

Rezultatų pavyzdys:

```
...... Pavyzdžio ir autorių tekstų df reikšmių palyginimas ----------
NAUDOJAMAS POŽYMIŲ VARDŲ FAILAS: systemFiles/featureNames_bendrieji_ir_lietuviski.txt
                radial_basis_kernel (požymio nr ir požymis)
Pavyzdžio autorius: 0100
Pavyzdžio autoriaus eilės numeris tarp tiriamų autorių: 3
Palyginimas su pavyzdžio decision funkcijos reikšmėmis
                        Surikiuoti rezultatai
    Autoriaus failas                  > sk.    < sk.     = sk.   Bendras sk.
-------------------------------------------------------------------------------
   1) 0999-1_12                        11       752      2135      2898
   2) 1000-5                            5       752      2141      2898
   3) 0100-67                           4       752      2142      2898
-------------------------------------------------------------------------------
   4) 0998-1_12                         9       753      2136      2898
-------------------------------------------------------------------------------
   5) 0997-1_12                         5       754      2139      2898
-------------------------------------------------------------------------------
   6) 0200-1_4                          3       756      2139      2898
   7) 0201-36_45                        3       756      2139      2898
   8) 0995-1_12                         2       756      2140      2898
-------------------------------------------------------------------------------
   9) 0996-1_12                        12       759      2127      2898
```

****

Eilutėje „NAUDOJAMAS POŽYMIŲ VARDŲ FAILAS“, nurodyta pagal kokį požymių nustatymų failą buvo vykdyti skaičiavimai.

Žemiau esančioje eilutėje nurodyta koks klasifikavimo režimas buvo naudojamas (pavyzdžio atveju tai radial_basis_kernel(požymio nr ir požymis)).

Eilutėje „Pavyzdžio autorius:“, nurodytas tiriamo teksto/autoriaus identifikatorius.

Žemiau pateikiamas surikiuotas tekstų bazės failų sąrašas pagal panašumą į tiriamą tekstą (1 įrašas panašiausias). Rikiavimas atliekamas pagal panašumo įvertį „< sk.“ didėjimo tvarka, kadangi kuo šis įvertis yra mažesnis, tuo tekstas yra panašesnis į tiriamą tekstą.

3.3. Išvados

1. Dėl didelės metodo apimties, realizuota sistema su vienu mašinos mokymosi klasifikatoriumi.
2. Metodo realizacijoje naudojamas vienos klasės atraminių vektorių (SVM) mašinos mokymosi klasifikatorius su dviejomis branduolio funkcijomis ir keturiais galimais klasifikavimo režimais.
3. Sistemos realizacija gali automatiškai atlikti kelis autorystės nustatymo tikrinimus pagal nurodytus nustatymų failus. Taip pat galima naudoti iš anksto išsaugotus nustatymų failus greitam autorystės nustatymo tikrinimo paleidimui.
4. Realizuota grafinė vartotojo sąsaja lietuvių ir anglų kalbomis.

4. ELEKTRONINIO DISKURSO AUTORYSTĖS NUSTATYMO METODO BANDYMAI

4.1. Duomenų rinkinys

Duomenų rinkinys buvo sudarytas iš interneto komentarų, surinktų iš lietuviško naujienų portalo Delfi (http://www.delfi.lt) ir apimantis 8 mėnesių laikotarpį nuo 2015 m. sausio mėn. iki 2015 m. rugpjūčio mėn. Šie komentarai buvo parašyti anoniminių vartotojų, išreiškiant jų nuomones apie straipsnius. Visi teksto fragmentai, turintys nelietuviškos abėcėlės raidžių (išskyrus skyrybos ženklus ir skaitmenis), taip pat atsakymai į komentarus bei meta informacija buvo pašalinti, paliekant tik gryną tekstą, be to, trumpesni nei 30 simbolių (neskaitant tuščios erdvės simbolių) ilgio tekstai nebuvo įtraukti. Šio duomenų rinkiniotekstus parašė iš viso 1000 skirtingų autorių. Jo charakteristikos apibendrinamos lentelėje.

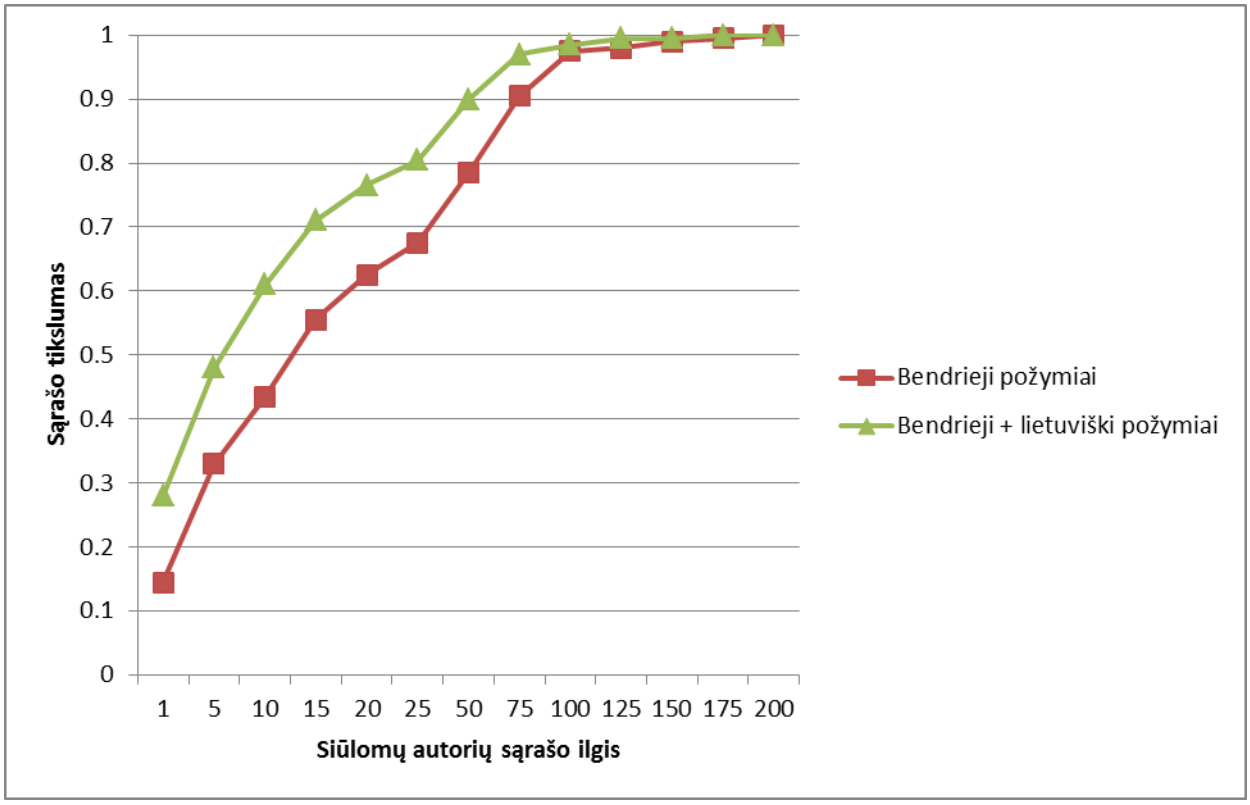
4.1 lentelė. Duomenų rinkinio charakteristikos

Charakteristika	Reikšmė
Trumpiausio teksto ilgis (simboliais)	3543
Trumpiausio teksto ilgis (žodžiais)	504
Ilgiausio teksto ilgis (simboliais)	119169
Ilgiausio teksto ilgis (žodžiais)	15874
Vidutinis teksto ilgis (simboliais)	15866
Vidutinis teksto ilgis (žodžiais)	2267

Klasifikavimui buvo naudojamas vienos klasės atraminių vektorių mašinos (SVM) klasifikatorius iš DLIB C++ bibliotekos. Šis klasifikatorius naudojo RBF branduolį su numatytomis parametrų reikšmėmis.

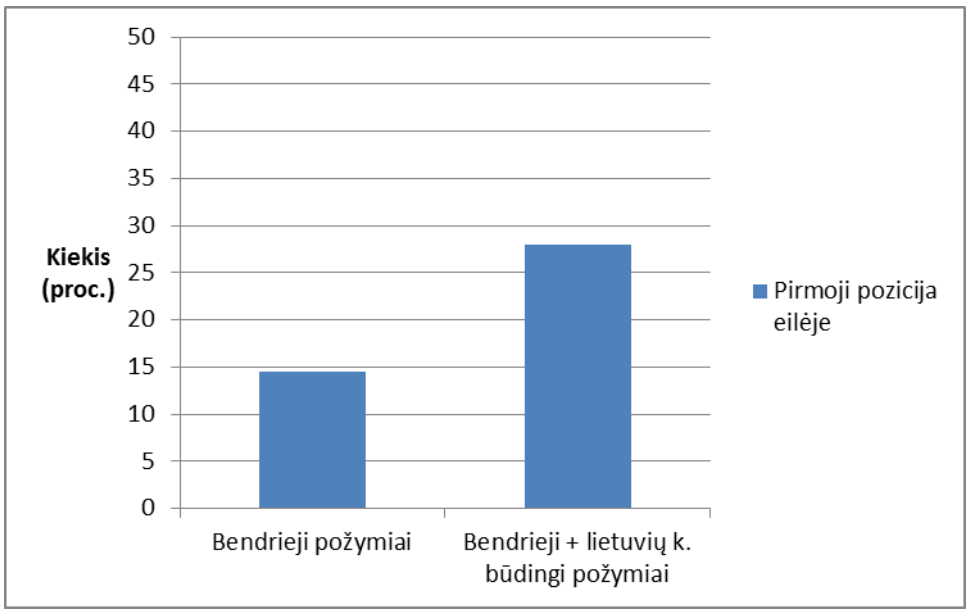
Šių eksperimentų metu, 80 proc. tekstų buvo naudojama klasifikatoriaus apmokymui, o 20 proc. – tikrinimui. Vienos klasės klasifikavimo rezultatai buvo naudojami galimų autorių eilės sudarymui. Pirmiausia, autorių eilė buvo sudaryta, atliekant klasifikavimą tik su bendraisiais požymiais. Autorių eilė buvo sudaroma remiantis požymių skaičiumi, kuris tam tikrą autorių labiau priskiria prie tiriamo nežinomo teksto. Po to buvo pridėti lietuvių kalbai būdingi požymiai ir tikrinama ar tikrojo autoriaus pozicija pakilo galimų autorių eilėje. Paveiksle vaizduojami eksperimentų (sąrašo tikslumo atžvilgiu) rezultatai naudojant tik bendruosius požymius bei bendruosius ir lietuvių kalbai būdingus požymius kartu. Rezultatai rodo ženkliai padidėjusį tikslumą, kuomet pridedami lietuvių kalbai būdingi požymiai.

Žemiau esančiame paveiksle pavaizduotas sąrašo tikslumo metrikos įvertinimas. Jame matomas tikrųjų autorių, kurie buvo suklasifikuoti nurodytuose pozicijų diapazonuose, kiekis. Grafike matyti, kad lietuviški požymiai ženkliai padidina tikslesnėse pozicijose suklasifikuotų tikrųjų autorių skaičių.



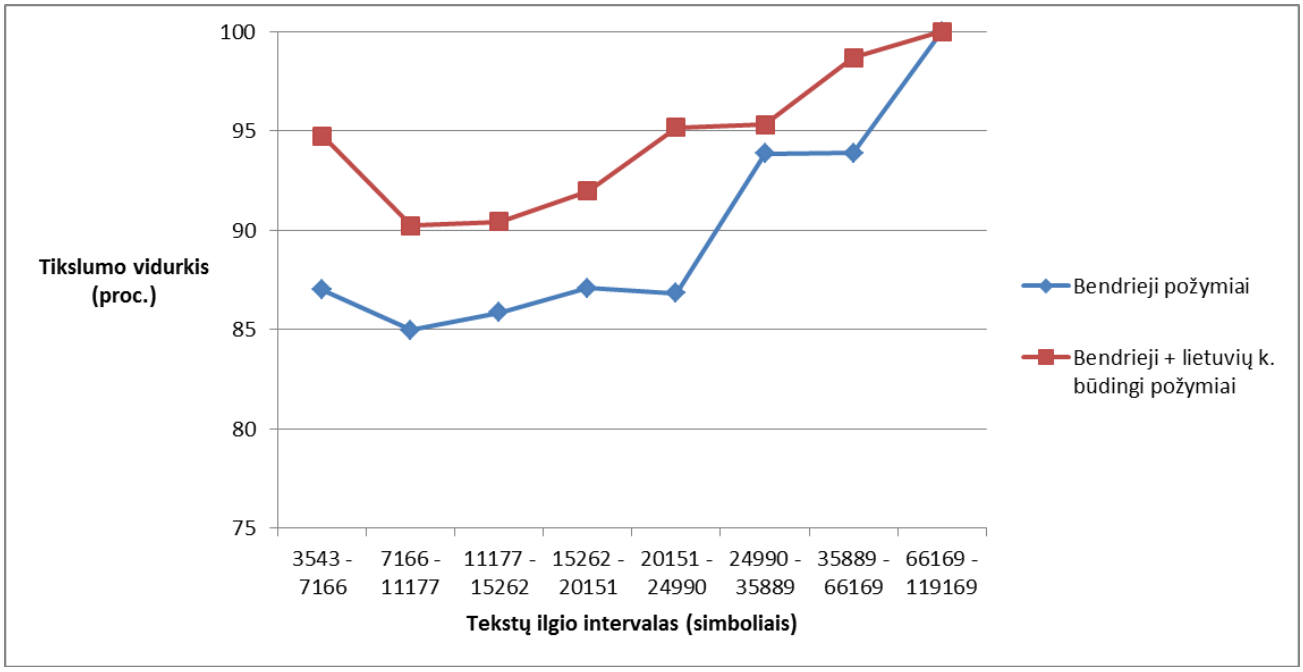
4.1 pav. Eksperimentų sąrašo tikslumo atžvilgiu rezultatai

Žemiau esančiame paveiksle pavaizduotas tikrųjų autorių, suklasifikuotų pirmose pozicijose, kiekis procentais. Naudojant tik bendruosius lingvistinius požymius, beveik 15 proc. tikrųjų autorių buvo atpažinti pirmose pozicijose. Pridėjus papildomus, tik lietuvių kalbai būdingus požymius, šis skaičius išaugo beveik dvigubai.



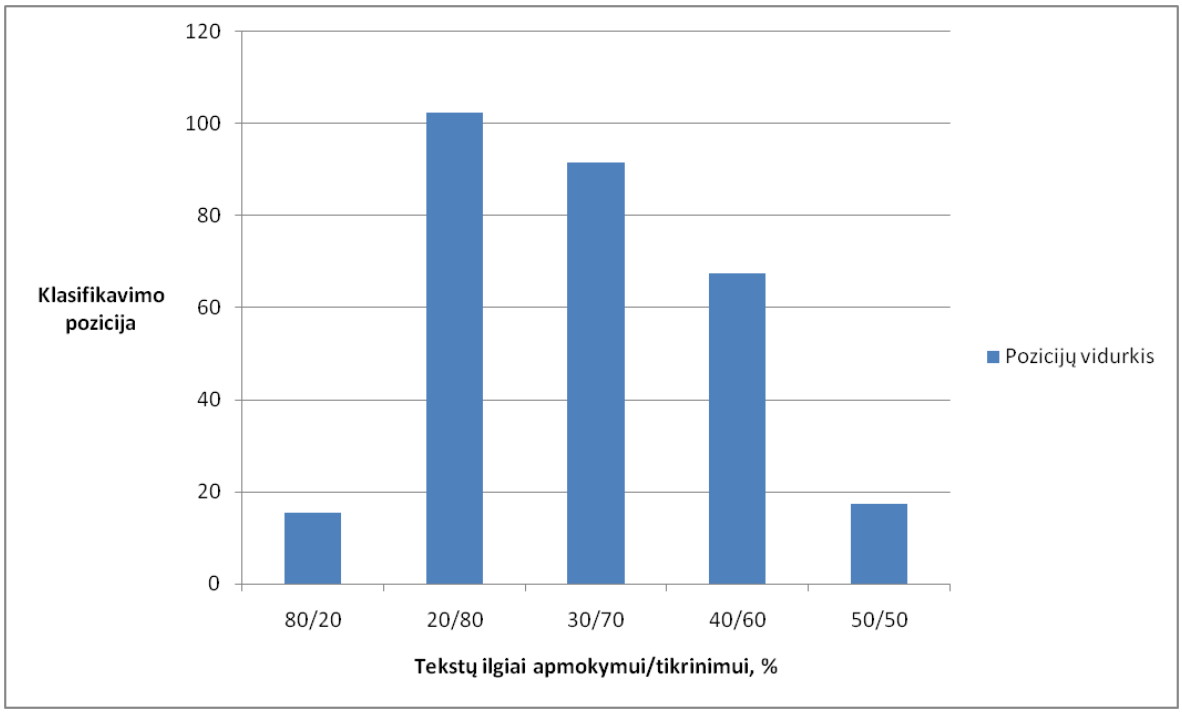
4.2 pav. Tikrojo autoriaus atpažinimo pirmoje autorių eilės pozicijoje tikslumas

Žemiau esančiame paveiksle matoma klasifikavimo tikslumo priklausomybė nuo tekstų ilgių. Kaip matyti, ilgi tekstai duoda geresnius autorystės atpažinimo rezultatus, papildomi, lietuvių kalbai būdingi požymiai ženkliai pagerina klasifikavimo tikslumą. Naudojant tekstus, kurių ilgis viršija 66000 simbolių, gaunamas itin didelis tikslumas.



4.3 pav. Tikslumo priklausomybės nuo tekstų ilgio rezultatai

Taip pat buvo atlikti eksperimentai, nustatant tekstų ilgių padalinimo įtaką klasifikavimo pozicijų vidurkiui. Žemiau esančiame paveiksle pavaizduoti šių eksperimentų rezultatai. Geriausi rezultatai buvo pasiekti naudojant 80% tekstų ilgio sistemos apmokymo procesui ir likusius 20% autorystės tikrinimo procesui (grafike žemesnė pozicijų vidurkio reikšmė reiškia geresnį rezultatą, t.y. aukštesnę poziciją autorių eilėje). Tai dažniausiai literatūroje rekomenduojamas padalinimo santykis teksto autorystės identifikavimo sistemoms. Šis santykis ir buvo naudojamas atliekant kitus eksperimentus šiame darbe. Taip pat neblogi rezultatai buvo gauti naudojant 50/50% padalinimo santykį, kitais tekstų ilgių padalinimo atvejais buvo gauti kur kas prastesni rezultatai.



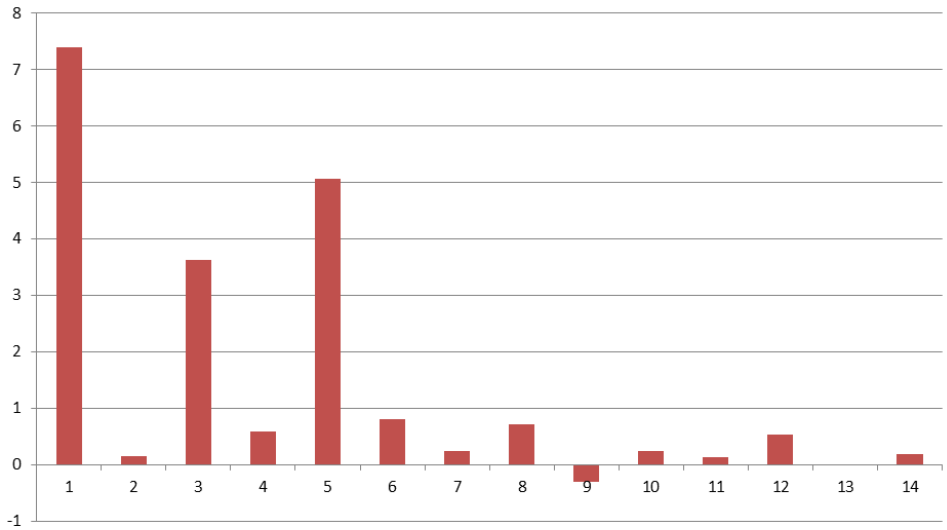
4.4 pav. Tekstų ilgių padalinimo įtaka klasifikavimo pozicijų vidurkiui

Tyrimų rezultatų įvertinimas naudojant standartizuotas tikslumo metrikas, pateiktas lentelėje. Šios metrikos plačiau aprašytos antrajame skyriuje. Kaip matyti, lietuvių kalbai būdingi požymiai visais atvejais padidino autorystės nustatymo tikslumą.

4.2 lentelė. Standartizuotos tikslumo įvertinimo metrikos

Metrika	Tik bendrieji požymiai	Bendrieji + lietuvių k. būdingi požymiai	Tikslumo padidėjimas
WTA	0.145	0.280	0.135
MRR	0.237	0.391	0.154
DCG	0.385	0.513	0.131
RPB (p=0.8)	0.288	0.455	0.167

Taip pat buvo atlikti eksperimentai, naudojant skirtingas lietuvių kalbai būdingų požymių rinkinio poaibes ir apskaičiuojant vidutinį tikrojo autoriaus pozicijos eilėje pokytį (pagerėjimą). Šių eksperimentų rezultatai vaizduojami paveiksle, pagal kurį matoma, jog aukščiausias pagerėjimas pasiekiamas naudojant funkcinius žodžius (1-as stulpelis), stabdančius žodžius (3-ias stulpelis) ir žodžių galūnes (5-as stulpelis).



4.5 pav. Lietuvių kalbai būdingų požymių įvertinimas

1 – funkcinių žodžių santykiniai dažniai; 2 – funkcinių žodžių ir visų žodžių santykis; 3 – stabdančių žodžių santykiniai dažniai; 4 – stabdančių žodžių ir visų žodžių santykis; 5 – žodžių su nurodytomis galūnėmis santykiniai dažniai; 6 – lietuvių kalbai nebūdingų simbolių bigramų santykiniai dažniai; 7 – lietuvių kalbai nebūdingų bigramų ir visų simbolių bigramų santykis; 8 – priešdėlio „ne“ santykinis dažnis; 9 – lietuviškų raidžių santykiniai dažniai; 10 – lietuviškų raidžių ir visų raidžių santykis; 11 – sutrumpinimų santykiniai dažniai; 12 – sutrumpinimų ir visų žodžių santykis; 13 – palyginimų santykiniai dažniai; 14 – palyginimų ir visų žodžių santykis.

4.2. Išvados

1. Atlikti eksperimentai parodė, kad lietuvių kalbai būdingi lingvistiniai požymiai ženkliai pagerina autorystės nustatymo tikslumą.
2. Eksperimentų metu nustatyta, kad didžiausią įtaką tikslumo pagerėjimui duoda lietuviškų funkcinių ir stabdančių žodžių bei žodžių su nurodytomis galūnėmis santykiniai dažniai.
3. Bandymų rezultatai parodė, kad lietuviškų raidžių santykiniai dažniai nežymiai sumažina autorystės nustatymo tikslumą. Tyrimų metu nustatyta, kad taip yra dėl to, kad tas pats autorius skirtinguose tekstuose gali rašyti ir su lietuviškomis raidėmis, ir be jų, o tai įtakoja autoriaus lingvistinių savybių aprašą.

4. Atlikti eksperimentai parodė, kad kuo ilgesni tekstai, tuo tikslesni autorystės nustatymo rezultatai.

5. REZULTATAI IR IŠVADOS

1. Kiekvienam autoriui būdingos tam tikros lingvistinės savybės, kurios leidžia jį išskirti iš kitų autorių.
2. Elektroninio diskurso autorystės nustatymo sistemose, autorių savybėms aprašyti yra naudojami įvairių lingvistinių požymių skaitinių reikšmių rinkiniai.
3. Elektroninio diskurso autorystės nustatymo sistemose, autoriams suklasifikuoti pagal tekstų panašumą yra naudojami mašinos mokymosi klasifikatoriai arba panašumo įvertinimo metrikos.
4. Elektroninio diskurso autorystės nustatymo sistemos leidžia ženkliai paspartinti autorystės nustatymo procesą, tačiau galutinį sprendimą priima tyrėjas.
5. Sudarytas lietuvių kalbai pritaikytas elektroninio diskurso autorystės nustatymo metodas, kuriame naudojami bendrieji lingvistiniai požymiai bei lietuvių kalbai būdingi požymiai.
6. Sudarytame metode naudojami keli dimensiškumo sumažinimo algoritmai bei mašinos mokymosi klasifikatoriai, o jų rezultatai apibendrinami galutinio sprendimo priėmimo modulyje.
7. Metodo realizacijoje panaudotas vienos klasės atraminių vektorių (SVM) mašinos mokymosi klasifikatorius su dviejomis branduolio funkcijomis bei keturiais klasifikavimo režimais.
8. Eksperimentai parodė, kad lietuvių kalbai būdingi požymiai ženkliai pagerina autorystės nustatymo (klasifikavimo) tikslumą.
9. Atliekant tyrimus nustatyta, kad tikslumo pagerėjimą labiausiai įtakoja lietuviškų funkcinių ir stabdančių žodžių santykiniai dažniai bei žodžių su lietuviškomis galūnėmis santykiniai dažniai.
10. Eksperimentų rezultatai parodė, kad lietuviškų raidžių santykiniai dažniai nežymiai sumažina klasifikavimo tikslumą, kadangi tas pats autorius gali rašyti tiek su lietuviškomis raidėmis, tiek be jų.
11. Tolimesnis šios srities vystymas galėtų apimti teksto robotų atpažinimą, suklasifikuotų tekstų autorių geografinės padėties nustatymą, apibendrinimą ir atvaizdavimą, bei autorių grupavimą pagal įvairias kategorijas, remiantis jų tekstais.
12. Šis darbas buvo atliktas vykdant projektą „Inovatyvaus kibernetinių nusikaltimų daiktų internete tyrimo metodo sukūrimas ir tyrimas“. Jo metu buvo parašyti straipsniai „Problems of authorship identification of the national language electronic discourse“ [72] ir „Authorship verification for Lithuanian internet comments using national lexical features and one-class classifier“ [73] bei sudalyvauta konferencijoje „Information and Software Technologies 21st International Conference, ICIST 2015, Druskininkai“.

6. LITERATŪRA

[1] A. Sánchez-Moya, O. Cruz-Moya. Whatsapp, Textese, and Moral Panics: Discourse Features and Habits Across Two Generations. Procedia - Social and Behavioral Scienc-es, Vol. 173, 2015, 300-306.

[2] Y.H. Segerstad. Use and adaptation of written language to the conditions of Computer-Mediated Communication. PhD dissertation, Göteborg University, 2002.

[3] C. Thurlow. Generation txt? The sociolinguistics of young people's text-messaging. Discourse Analysis Online 2003, vol. 1, no. 1, 2003.

[4] N. MacLeod, T. Grant. Whose tweet?: authorship analysis of micro-blogs and other short form messages. Electronic Proceedings of the International Association of Forensic Linguists' 10th Biennial Conference, Aston University, Birmingham, UK, July 2011.

[5] A. Voutilainen, 2003. Part-of-speech tagging. In R. Mitkov, editor, The Oxford hand-book of computational linguistics. University Press, Oxford, pp. 219-232

[6] J. Nivre. Logic programming tools for probabilistic part-of-speech tagging. Master's thesis, Växjö University, 2000

[7] D. Bogdanova, A. Lazaridou. Cross-language authorship attribution. The International Conference on Language Resources and Evaluation, 2015-2020, 2014.

[8] M. Potthast, A. Barron-Cedeno, B. Stein, Benno, and P. Rosso. (2011). Cross-language plagiarism detection. Language Resources and Evaluation, 45(1):45-62

[9] M. F. Salvador, P. Gupta, and P. Rosso (2013). Cross-Language plagiarism detection using a multilingual semantic network. In Proceedings of the 35th European confer-ence on Advances in Information Retrieval, ECIR '13, 710-713.

[10] R. Navigli, and S. P. Ponzetto (2010). BabelNet: building a very large multilingual semantic network. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, 216-225.

[11] P. Panicheva, J. Cardiff, and P. Rosso (2010). Personal Sense and Idiolect: Combining Authorship Attribution and Opinion Analysis. In In Proceedings of the 4th Interna-tional Conference on Language Resources and Evaluation, LREC '10

[12] R. Dunn, J. Beaudry, and A. Klavas (1989). Survey of Research on Learning Styles. Educational Leadership, vol. 46, no. 6, pp. 50-58.

[13] R. Bellman, Adaptive Control Processes: A Guided Tour. Princeton University Press, 1961

[14] D. Koller, M. Sahami, Hierarchically classifying documents using very few words. In Proceedings of International Conference on Machine Learning, 1997, pp. 170-178

[15] K. Fuka and R. Hanka. Feature Set Reduction for Document Classification Problems. In: Proc. of IJCAI-01 Workshop: Text Learning: Beyond Supervision, Seattle, 2001

[16] G. Zervas, S.M. Rüger. The curse of dimensionality and document clustering. In Pro-ceedings of the IEEE Searching for Information: AI and IR Approaches, 1999.

[17] L. Pearl, M. Steyvers: Detecting authorship deception: a supervised machine learning approach using author writeprints. LLC 27(2): 183-196 (2012).

[18] M. Brennan, S. Afroz, and R. Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Trans. Inf. Syst. Se-cur. 15, 3, Article 12, 22 pages.

[19] O. deVel, Mining E-mail authorship. ACM International Conference on Knowledge

Discovery and Data Mining (KDD 2000). Workshop on Text Mining.

[20] D. Holmes, Authorship attribution. Computers and the Humanities 28(2) (1994) 87-106

[21] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology, v.60 n.3, p.538-556, 2009.

[22] R. Zheng, J. Li, H. Chen, Z. Huang: A framework for authorship identification of online messages: Writing-style features and classification techniques. JASIST (JASIS) 57(3):378-393 (2006)

[23] J. Graovac. A variant of n-gram based language-independent text categorization. Intel-ligent Data Analysis, Vol. 18, No. 4, 2014

[24] R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, and P. Rosso. 2006. Authorship attribution using word sequences. In Proceedings of the 11th Iberoameri-can conference on Progress in Pattern Recognition, Image Analysis and Applications (CIARP'06). Springer-Verlag, Berlin, Heidelberg, 844-853

[25] E. Stamatatos, On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21(2), 421 (March 2013)

[26] M. Koppel, J. Schler and S. Argamon (2011) Authorship attribution in the Wild. Lan-guage Resources & Evaluation 45, 83-94

[27] T. Reicher, I. Kristo, I. Belsa, and A. Silic. Automatic Authorship Attribution for Texts in Croatian Language Using Combinations of Features. KES (2) 2010: 21-30

[28] S. Argamon, and S. Levitan (2005). Measuring the usefulness of function words for authorship attribution. In Proceedings of the ACH/ALLC Conference, Victoria, BC, Canada, June 2005

[29] C.E. Chaski, Who's at the keyboard? Authorship attribution in digital evidence inves-tigations. International Journal of Digital Evidence 4(1), pp. 1-13 (2005)

[30] O. Hilton, (1993). Scientific Examination of Questioned Documents. Boca Raton, Florida, CRC Press.

[31] G. R. McMenamin (2003). Forensic Linguistics: Advances in Forensic Stylistics. Boca Raton, Florida, CRC Press.

[32] C. Martindale, and D. McKenzie (1995). On the utility of content analysis in author attribution: The Federalist. Computer and the Humanities, 29, 259-270

[33] Y. Palkovskii, A. Belov, I. Muzika: Exploring Fingerprinting as External Plagiarism Detection Method - Lab Report for PAN at CLEF 2010. CLEF (Notebook Pa-pers/LABs/Workshops) 2010

[34] T. Yang, D. Lee. T3: On Mapping Text To Time Series. Proceedings of the 3rd Alberto Mendelzon International Workshop on Foundations of Data Management, Arequipa, Peru, May 12-15, 2009. CEUR Workshop Proceedings 450, CEUR-WS.org 2009, AMW 2009.

[35] K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972

[36] Y. Qu, G. Ostrouchovz, N. Samatovaz, and A. Geist. Principal component analysis for dimension reduction in massive distributed data sets. In Proceedings of IEEE Interna-tional Conference on Data Mining (ICDM), 2002

[37] M. Koppel, J. Schler, S. Argamon, E. Messeri (2006) Authorship attribution with thou-sands of candidate authors. Proc. of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval Seattle, Washington, pp. 659-660.

[38] M. Koppel, J. Schler, S. Argamon and Y. Winter (2012): The "Fundamental Problem" of Authorship Attribution, English Studies, 93:3, 284-291

[39] A. Huang (2008). Similarity Measures for Text Document Clustering. Proceedings of the

Sixth New Zealand Computer Science Research Student Conference NZCSRSC2008,Christchurch, New Zealand, 49-56

[40] B. Kjell, W.A., Woods, and O. Frieder. Discrimination of authorship using visualiza-tion. Information Processing and Management, 30 (1): (1994), 141-150

[41] C.D. Shaw, J.M. Kukla, I. Soboroff, D.S. Ebert, C.K. Nicholas, A. Zwa, E.L. Miller, and D.A. Roberts, Interactive volumetric information visualization for document cor-pus management. International Journal on Digital Libraries, 2: (1999), 144-156

[42] J. F. Burrows, Delta: a measure of stylistic difference and a guide to likely authorship, Literary and Linguistic Computing 17, pp. 267-287, 2002.

[43] F. Mosteller, and D.L. Wallace (1964). Inference and disputed authorship: The Federal-ist. Reading, MA: Addison-Wesley

[44] L. Antiqueira, T. A. S. Pardo, M. das Gracas Volpe Nunes, O. N. de Oliveira Jr., L. da Fontoura Costa. Some issues on complex networks for author characterization. Proc. of the Int. Joint Conference IBERAMIA/SBIA/SBRN 2006 - 4th Workshop in Infor-mation and Human Language Technology (TIL'2006), 2006. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial 11(36): 51-58

[45] S. Segarra, M. Eisen, A. Ribeiro: Authorship attribution using function words adjacen-cy networks. IEEE International Conference on Acoustics, Speech and Signal Pro-cessing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. IEEE 2013, 5563-5567.

[46] J. Ke and Y. Yao. (2008). Analysing Language Development from a Network, Ap-proach. Journal of Quantitative Linguistics, 15(1):70-99.

[47] J. Leskovec, J. Kleinberg, and C. Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. ACM Trans. Knowl. Discov. Data, 1(1), Article no. 2.

[48] Y. Matsuo, Y. Ohsawa, and M. Ishizuka. (2001). A Document as a Small World. In Proceedings of the Joint JSAI 2001 Workshop on New Frontiers in Artificial Intelli-gence, pp. 444-448, London, UK, UK. Springer Verlag.

[49] F. Chang, E. Lieven, and M. Tomasello. 2008. Automatic evaluation of syntactic learners in typologically-different languages. Cogn. Syst. Res. 9, 3 (June 2008), 198-213

[50] H. Rizvic, S. Martincic-Ipsic, and A. Mestrovic: Network Motifs Analysis of Croatian Literature. CoRR abs/1411.4960 (2014)

[51] H. Wagner, P. Dlotko, M. Mrozek. Computational Topology in Text Mining. Computa-tional Topology in Image Context - 4th International Workshop, CTIC 2012, Bertinoro, Italy, May 28-30, 2012. Proceedings. LNCS vol. 7309, Springer 2012, 68-78.

[52] D. Beeferman; A. Berger; J. Lafferty. A Model of Lexical Attraction and Repulsion. 35th Annual Meeting of the Association for Computational Linguistics.

[53] D. R Amancio Authorship recognition via fluctuation analysis of network topology and word intermittency. J. Stat. Mech. (2015) P03005.

[54] C. Basile, D. Benedetto, E., Caglioti, M. Degli Esposti. 2008. An example of mathe-matical authorship attribution. In: Journal of Mathematical Physics, 49:125211-125230

[55] A.Todirascu, S. Pado, J. Krisch, M. Kisselew, U. Heid. French and German Corpora for Audience-based Text Type Classification, LREC 2012: 1591-1597.

[56] Varela, P., Justino, E., Oliveira, L. S., Verbs and Pronouns for Authorship Attribution, 17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010), 89-92, Rio de Janeiro, Brazil, 2010.

[57] D. Pavelec, L.S. Oliveira, E. Justino, L.V. Batista, Using Conjunctions and Adverbs for Author Verification. Journal of Universal Computer Science (JUCS), 14(18):2967-2981, 2008.

[58] J. Hancke, D. Meurers, and S. Vajjala. 2012. Readability classification for German using lexical, syntactic, and morphological features. In Proceedings of the 24th Inter-national Conference on Computational Linguistics (COLING), pp. 1063-1080, Mum-bai, India

[59] J. Diederich, J. Kindermann, E. Leopold, and G. Paass. (2003) Authorship attribution with support vector machines, Applied Intelligence 19 (1), pp. 109-123.

[60] O.V. Kukushkina, A.A. Polikarpov, and D.V. Khmelev (2001), Using Literal and Grammatical Statistics for Authorship Attribution, Probl. Inf. Transm. 37, 2, 2001, 172-184.

[61] A. Zecevic, M. Utvic. An Authorship Attribution for Serbian. BCI (Local) 2012: 109-112

[62] G. Žalkauskaitė. Idiolect signs in the e-mail. PhD dissertation, Vilnius University, Lithuania (2012)

[63] J. Barragán. Why some hard cases remain unsolved. Legal knowledge based systems. JURIX 93. 1993

[64] T. Grant (2013). TXT 4N6 method, consistency, and distinctiveness in the analysis of SMS text messages. Journal of law and policy, 21 (2), pp. 467-494.

[65] Mohtasseb, H., Ahmed, A.: Two-layered Blogger identification model integrating profile and instance-based methods. Knowledge and Information Systems, 31(1), pp. 1-21 (2012)

[66] Guillén-Nieto V., Vargas- Sierra C., Pardiño- Juan M., Martínez- Barco P. & Armando Suárez -Cueto A. Exploring state-of-the art software for forensic authorship identification. International Journal of English Studies, 8 (1), p.1-28 (2008)

[67] Teahan, W.J., Harper. D.J.: Using compression-based language models for text categoriza-tion. In W.B. Croft, J. Lafferty (Eds.), Language modeling for information retrieval, Springer, pp. 141-165 (2003)

[68] Benedetto, D., Caglioti, E. , Loreto, V.: Language trees and zipping. Phys. Rev. Lett., 88 (4), p. 048702 (2002)

[69] Abbasi A., Chen H.: Applying authorship analysis to extremist-group web forum messages. IEEE Intelligent Systems, 20(5), pp. 67-75 (2005)

[70] Kapociute-Dzikiene, J., Vaassen, F., Daelemans, W., Krupavicius, A.: Improving Topic Classification for Highly Inflective Languages. 24th International Conference on Compu-tational Linguistics, COLING 2012, pp. 1393-1410 (2012)

[71] Napoli, C., Tramontana, E., Lo Sciuto, G., Wozniak, M., Damasevicius, R., Borowik, G.: Authorship Semantical Identification using Holomorphic Chebyshev Projectors. Proc. of 3rd Asia-Pacific Conference on Computer Aided System Engineering (APCASE) (2015)

[72] Venčkauskas, A., Damaševičius, R., Marcinkevičius, R., Karpavičius, A.: Problems of authorship identification of the national language electronic discourse. Information and Software Technologies 21st International Conference, ICIST, 2015, 415-432 (2015)

[73] Venčkauskas, A., Damaševičius, R., Marcinkevičius, R., Kapočiūtė-Dzikienė, J., Karpavičius, A.: Authorship Verification for Lithuanian Internet Comments Using National Lexical Features and One-Class Classifier