Univerzita Pavla Jozefa Šafárika v Košiciach

Prírodovedecká fakulta

URČOVANIE AUTORSTVA NEZNÁMEHO SLOVENSKÉHO TEXTU

ANALÝZA A NÁVRH RIEŠENIA

Študijný odbor: **Informatika**
Školiace pracovisko: **Ústav informatiky**
Vedúci záverečnej práce: **prof. RNDr. Stanislav Krajči, PhD.**

Košice 2021 **Bc. Radovan Fuska**

1 Úvod

Problematike určovania autorstva sa ľudia venujú už dlho. Uplatňuje sa hlavne vo forenznej lingvistike, ale môže sa používať aj na zisťovanie plagiátorstva. Určovaním autorstva sa ľudia zaoberajú už od čias stredoveku, kedy učenci hľadali spôsoby, ako priradiť texty starodávnym autorom [1]. V súčasnosti existuje mnoho spôsobov na analýzu textu za účelom určovania autorstva, ale väčšina z nich je vyvinutá pre svetové jazyky, ako napríklad angličtina. Hoci existujú metódy aj pre slovanské jazyky, pre slovenský jazyk ich poznáme veľmi málo. Väčšinu existujúcich metód vhodných pre slovenčinu tvoria všeobecné metódy, ktoré nie sú závislé na jazyku. Cieľom tejto práce je analýza a vyhodnotenie efektívnosti existujúcich metód, návrh novej metódy a jej porovnanie s už existujúcimi metódami.

1.1 Špecifikácia problému

Pri určovaní autorstva sa snažíme textu priradiť jeho autora. Využívame predpoklad, že texty od jedného autora obsahujú nejakú vnútornú podobnosť a zároveň sa líšia od textov od iných autorov. V lingvistike sa tento fenomén nazýva idiolekt – konkrétna podoba jazyka jednej osoby. Teda na určovanie autorstva sa budeme snažiť takéto znaky nájsť, kvantifikovať a pomocou nich nájsť nejakú metriku na odlišovanie textov od rôznych autorov. Z toho vyplýva, že na určovanie autorstva je potrebná vzorka textov od hľadaného autora, aby sa na nej mohli znaky idiolektu autora odmerať a vytvoriť akýsi osobný model podoby jazyka („odtlačok“) pre daného autora.

Pri určovaní autorstva budeme rozlišovať tri varianty tohto problému:

- V prvom variante máme k dispozícii zoznam možných kandidátov a nám stačí len vybrať, ktorý z tohto zoznamu napísal daný text.
- V druhom variante tiež máme zoznam autorov, ale pripúšťame aj variantu, že ani jeden z týchto kandidátov nie je autorom textu, teda že autor textu sa nenachádza v zozname kandidátov.
- V treťom variante poznáme kolekciu existujúcich textov od kandidátskeho autora a máme iba určiť, či je, alebo nie je autorom daného textu.

Tieto varianty si vyžadujú rozdielne algoritmy na ich výpočet, konkrétne prvý variant si vyžaduje algoritmus pre klasifikáciu do viacerých tried, druhý variant algoritmus pre

klasifikáciu do viacerých tried, ale s možnosťou nulovej klasifikácie, zatiaľ čo tretí variant vyžaduje binárnu klasifikáciu. Existujú známe algoritmy na prvý a tretí variant, avšak algoritmy pre druhý variant nie sú až také rozšírené.

Najvšestrannejšie sú algoritmy na binárnu klasifikáciu, pretože sa dajú použiť na všetky tri varianty. Na tretí variant sa dajú použiť priamo. Na prvý a druhý variant sa dajú použiť tak, že sa vytvorí samostatný model pre každého autora a neznámy text sa ohodnotí pomocou modelov každého autora. Ak nastane zhoda u viac ako jedného autora alebo ak v prvom variante nenastane zhoda u žiadneho autora, ako výsledok sa použije ohodnotenie s najväčšou pravdepodobnosťou, resp. zhodou medzi autorom a neznámym textom. Toto sa však dá použiť iba pri algoritmoch, ktorých výstupom nie je len binárna hodnota, ale reálne číslo, ktoré sa na konci používa na určenie klasifikácie.

Pre druhý variant sa dá použiť aj kombinácia algoritmu na klasifikáciu do viacerých tried a algoritmus na binárnu klasifikáciu. Algoritmom na klasifikáciu do viacerých tried by sa vybral najlepší kandidát na autora. Algoritmom na binárnu klasifikáciu by sa rozhodlo, či neznámy text je dostatočne podobný, resp. zhodný so všetkými autormi, resp. s hociktorým z množiny kandidátov. Nevýhodou tohto prístupu je to, že na takéto natrénovanie binárneho klasifikátora sú potrebné aj texty od autorov, ktorí nie sú súčasťou zoznamu kandidátov. Toto by sa však dalo vyriešiť obmenou tohto prístupu, keď namiesto jedného binárneho klasifikátora by sa použil samostatný klasifikátor pre každého autora, ktorý by len potvrdzoval alebo zamietal klasifikáciu od prvého klasifikátora, resp. rozhodoval by medzi výsledkom „konkrétny kandidát“ a „žiaden z kandidátov“.

2 História

2.1 Prvé snahy

Pri prvých snahách o určenie autorstva textu dominoval prístup, keď sa vedci snažili nájsť akýsi autorský znak – jeden znak (väčšinou číslo), ktorého hodnota by vedela jednoznačne určiť autora textu. Výsledkom tohto snaženia však nebolo nájdenie takéhoto znaku, ale skôr nájdenie invariantných vlastností textu (napríklad Zipfovo pravidlo [1]). Medzi takýchto vedcov patrí napríklad Mendenhall, ktorý sa snažil použiť závislosť medzi dĺžkou slov a ich frekvenciou, a Yule, ktorý sa snažil použiť dĺžku viet

ako jediný znak na určenie autorstva textu [1].

2.2 Zovšeobecnený prístup

Po tom, ako sa jednorozmerný prístup na ohodnotenie textu ukázal ako nie veľmi efektívny, sa začalo uplatňovať viacrozmerné ohodnotenie textu. Pri týchto metódach sa využíval nielen jednoduchý skalár, ale aj vektor hodnôt vyskladaných z rôznych znakov a viacrozmerných hodnôt jedného znaku. Príkladmi autorov takýchto prístupov sú Mosteller a Wallace, ktorí využili frekvencie funkcionálnych slov s metódou naivný bayesovský klasifikátor [1].

Príkladom takýchto snáh je aj známa práca od J. F. Burrowsa [6]:

V tomto článku autor analyzuje frekvencie tridsiatich najčastejšie sa vyskytujúcich slovných typov v anglickom texte.

Autor poukazuje na odlišnosť jednotlivých častí prózy a na jej unikátnosť v tomto ohľade. Autor argumentuje, naproti tvrdeniam ostatných učencov, že individuálny štýl autora textu sa v texte prejavuje nie len v malých rozdieloch v slovách, ale dá sa pozorovať vo väčšej miere pomocou štatistickej analýzy.

Na prisudzovanie autorstva textu autor používa metódu, ktorá je odlišná od stylometrie a nerozlišuje medzi obsahovými slovami a funkcionálnymi slovami – slovami, ktoré plnia obsahovú funkciu, a slovami, ktoré plnia gramatickú funkciu. Autor priznáva možnosť, že zvyky autora textu sa skôr odrazia v slovách, nad ktorými autor textu nemá vedomú kontrolu, avšak argumentuje, že na účely literárneho kriticizmu asistovaného počítačom by sa obsahové slová nemali vynechať.

Autor používa tridsať najčastejšie sa vyskytujúcich slovných typov z textov zo šiestich noviel od Jane Austen na vytvorenie frekvenčného profilu autora textu, ktorý potom porovnáva s profilmi textov noviel od iných autorov ako Henry James, E. M. Forster a Georgette Heyer. Tento profil používa aj na porovnanie naratívu jednotlivých textov od Jane Austen navzájom. Týchto tridsať najčastejšie sa vyskytujúcich slovných typov spolu tvorí asi dve pätiny všetkých slov použitých v textoch. Autor predpokladá, že tieto dve pätiny slov musia mať výpovednú hodnotu o celom texte.

Autor pomocou Pearsonovho korelačného koeficientu vypočítal korelácie medzi jednotlivými naratívmi v textoch noviel od Jane Austen a ukázalo sa, že jednotlivé naratíva

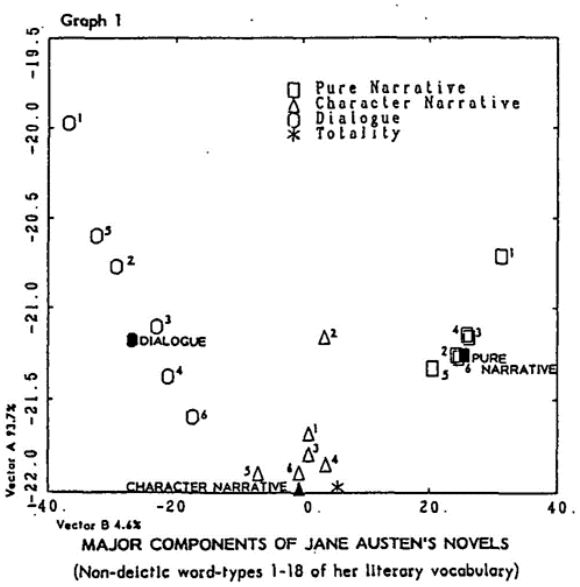
najbližšie korelujú s naratívami rovnakého typu, a teda sú nezávislé na obsahu textu či gramatických vlastností jednotlivých slov.

	PN	pn1	pn2	pn3	pn4	pn5	pn6	CN	cn1	cn2	cn3	cn4	cn5	cn6	D	d1	d2	d3	d4	d5
NA pn1	990																			
SS pn2	994	983																		
PP pn3	998	986	995																	
MP pn4	998	983	989	993																
E pn5	993	975	979	988	995															
P pn6	998	989	989	994	997	993														
CN	960	933	961	956	954	963	962													
NA cn1	950	941	953	945	939	943	951	981												
SS cn2	926	918	946	925	911	906	925	964	963											
PP cn3	954	924	961	952	949	954	953	993	968	960										
MP cn4	967	936	964	964	963	973	968	995	964	944	991									
E cn5	940	912	941	934	934	946	943	996	977	961	983	986								
P cn6	954	931	958	949	948	955	956	996	985	968	991	987	990							
D	860	818	861	854	855	878	862	960	941	908	947	944	975	955						
NA d1	784	755	780	773	778	810	788	898	899	846	877	873	921	897	975					
SS d2	834	781	842	829	832	853	835	946	916	892	940	932	960	940	988	952				
PP d3	863	818	877	865	853	869	863	960	937	925	956	947	969	953	985	938	978			
MP d4	883	853	879	874	879	899	887	967	960	914	952	950	980	964	993	971	974	968		
E d5	822	776	819	814	819	851	827	932	908	871	913	916	952	925	992	976	973	967	979	
P d6	902	866	900	895	898	917	905	979	964	928	967	967	987	975	991	954	971	977	991	980

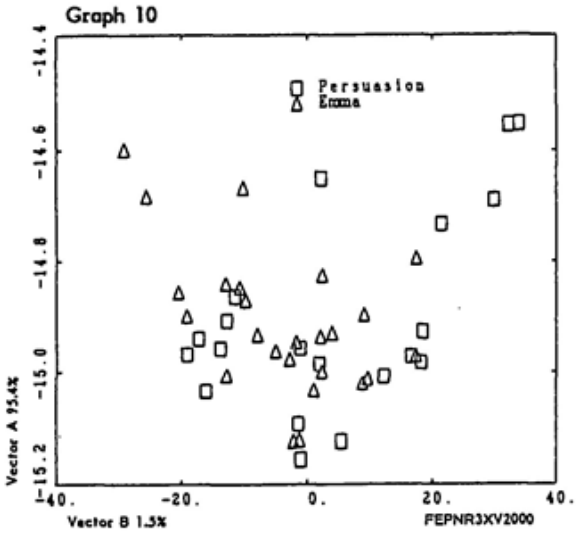
NB. For more compact representation, the coefficients are expressed as whole numbers and not in the usual form of decimal fractions like +0.963

Tabuľka 1: Korelácie jednotlivých naratív v textoch noviel od Jane Austen.

Autor na týchto koreláciách potom použil dekompozíciu na extrahovanie vlastných vektorov a vlastných čísel z matice korelácií. Ukázalo sa, že najväčší prínos majú prvé dva vektory. Podľa nich autor potom zakreslil do 2D grafu jednotlivé naratíva. Takto sa ukázalo, že naratíva tvoria malé skupinky podľa ich typu. Najrozľahlejšia – najväčšie rozprestretá skupinka je skupinka dialógov postáv. Autor tvrdí, že toto je spôsobené odlišným charakterom jednotlivých postáv novely. Z grafu 1 je taktiež vidno, že najužšiu skupinku tvorí čistý naratív autora textu. Toto naznačuje, že Jane Austen počas jej literárnej kariéry zostala konzistentná vo svojom naratíve.

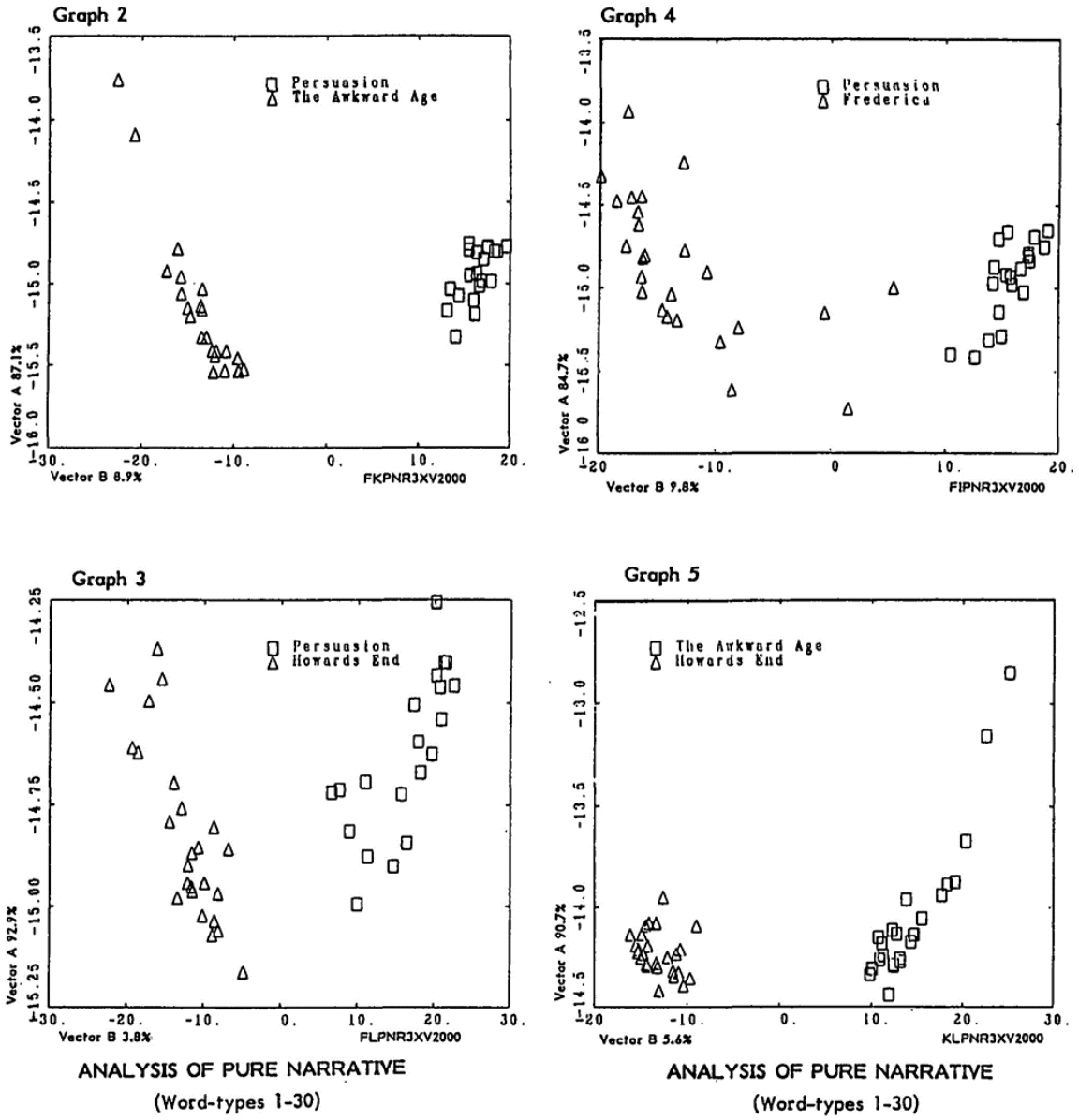


Graf 1: Porovnanie rôznych naratív v dielach od Jane Austen.



Graf 2: Porovnanie naratív dvoch diel od rovnakého autora.

Autor sa potom zameria iba na čistú naratívu autorov noviel. Taktiež autor rozdelí diela na dvetisíc-slovné časti na účely porovnania jednotlivých častí ako sú úvod a záver diel. Štatistiky týchto častí sa potom v poradí pospájajú do šesťtisíc slovných prekrývajúcich sa blokov. Toto má účel odstránenia štatistických výkyvov z jednotlivých individuálnych častí. Takto vytvorené štatistiky sa potom navzájom rovnakým spôsobom porovnajú.

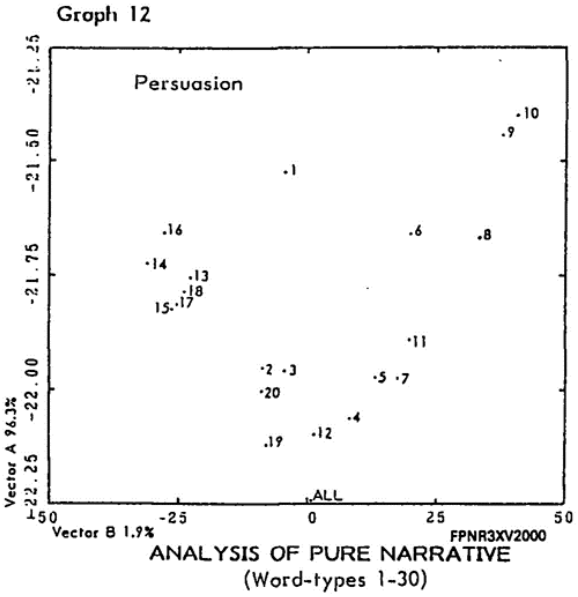


Graf 3–6: Porovnanie čistých naratív diel od rôznych autorov.

Z grafov 3–6 je vidieť, že jednotliví autori majú svoje vlastné naratíva a tie sú navzájom odlišné.

Autor taktiež porovnáva aj čistú naratívu dvoch rôznych diel od jedného autora. Na grafe 2 vidieť že tieto naratíva si sú podobné a niektoré časti sa prekrývajú.

V závere autor porovnáva aj jednotlivé časti novely *Persuasion* (*Anna Elliotová*) samostatne. Z nich je možné vidieť vývoj štýlu autora textu a jeho zmeny v priebehu diela.



Graf 7: Porovnanie jednotlivých častí novely *Persuasion*.

2.3 Moderné prístupy

V súčasnej dobe sa už zväčša upustilo od ručného hľadania znakov a vzorcov na kvantifikáciu textu a prešlo sa na automatické metódy – metódy využívajúce strojové učenie [1].

3 Súčasný stav

V súčasnosti existuje viacero odskúšaných metód na určovanie autorstva. Veľká časť týchto metód je špecifická pre anglický jazyk. Toto je pravdepodobne zapríčinené tým, že anglický jazyk je vo svete rozšírený, pre jeho analýzu existuje veľké množstvo dát a má relatívne ľahkú štruktúru, vhodnú na počítačové spracovanie bez potreby veľkého predspracovania.

Slovenský jazyk, na rozdiel od anglického, obsahuje zložitejšiu gramatiku, najmä ohýbanie slov, ktoré si na úspešné porozumenie vyžaduje náročné predspracovanie, ako napríklad určenie gramatických kategórií či syntaktický rozbor.

4 Znaky textu

Na vytvorenie modelu sa dá použiť viacero rôznych znakov. Znaky sa často rozlišujú podľa ich jazykovednej roviny, do ktorej patria, resp. podľa toho, z akej roviny po-

chádzajú ich vstupné údaje. Existujú však aj znaky, ktoré patria do viacerých rovín, resp. znaky, ktoré len stavajú na informáciách od iných znakov – tzv. metaznaky. Takéto delenie znakov však na samotný výpočet nie je nevyhnutné (a častokrát je len približné) a slúži hlavne len na lepšiu charakterizáciu znaku. V tejto práci sme vychádzali zo znakov pre anglický jazyk, pretože sú najpoužívanejšie a je ich najväčší počet.

4.1 Anglický jazyk

Pre anglický jazyk sú najčastejšie tieto znaky [2]:

- lexikálne znaky
  - slová

    Ide o sledovanie výskytu slov z konkrétneho predom pripraveného zoznamu. Často sa používajú funkcionálne slová.

  - slovné multigramy

    Podobne ako pri slovách, ale ide o krátke postupnosti slov.

  - funkcionálne slová

    Sú to slová, ktoré nenesú obsah textu, ale určujú vzťahy medzi slovami vety. Často sú to slovné druhy, ako napríklad spojky alebo predložky.

  - zámená

    Rozdelenie použitých zámen. Využíva predpoklad, že rôzni ľudia preferujú rôzne zámená [3].

  - modálne slovesá

    Slovesá, ktoré vyjadrujú modalitu – pravdepodobnosť, schoponsť, ... (Sú to napríklad slová *can*, *could*, *may*, ...)

  - slang

    Medzi tieto slová patrí internetový slang, smajlíci či nadávky [3].

- skratky a stiahnuté tvary

  Ide o skratky a stiahnuté tvary anglického jazyka (napríklad *I'm*, *you're*, *etc.*).

- emotikony

  Ide o často používaných smajlíkov [3], [4].

- chyby hláskovania

  Ide o chyby pri hláskovaní, ako napríklad opakujúce sa písmeno, chýbajúce zdvojené písmeno, prehodenie písmen, písmeno navyše, chýbajúce písmeno, zamenené písmeno. Tieto chyby sa dajú hľadať pomocou slovníka na všetkých slovách, ale dá sa zamerať aj na špecifické vybrané chyby [4].

- varianty britskej a americkej angličtiny

  Napríklad alternatívne slová (*vacation*/*holiday*) a varianty v hláskovaní (*color*/*colour*) [4].

- sémantika (mnohoznačnosť, jednoznačnosť)

  Ide o porovnávanie významu slov textu, jednotnosť významu slov v texte. Na tento účel je možné použiť špeciálne vytvorené slovníky, ako je napríklad WordNet [4].

- typy pomenovaných entít

  Ide o relatívnu frekvenciu rôznych typov pomenovaných entít, ako napríklad dátum, miesto, peniaze, číslo, radová číslovka, organizácia, percento, osoba a čas [4].

- znaky na úrovni písmen

  - písmenové multigramy

    Ide o multigamy po sebe idúcich písmen. Pre ich veľký počet zvyčajne ide len o bigramy a trigramy. Aj takýto jednoduchý znak vie čiastočne zachytiť znaky textu ako napríklad: prípony, interpunkcia, čas, stiahnuté tvary, smajlíci, funkcionálne slová vo funkcii predložiek, zámená a spojky, ako aj

alternatívne hláskovania (americká vs. britská angličtina, napr. *-or*/*-our*, *-ise*/*-ize*) [4].

- prípony

  Ide o prípony slov vytvorených pridaním prípony na koreň slova [4].

- interpunkcia

  Ide o frekvenciu interpunkčných znamienok (výkričník, otáznik) a ich násobkov (napríklad !! alebo ???). Taktiež ide aj o ich zlé použitie ako napríklad prítomnosť medzery pred dvojbodkou [4].

- formátovacie znaky

  - dĺžka

    Ide o dĺžku textu, viet a slov, resp. ich priemernú dĺžku, maximálnu dĺžku a ich distribúciu [3], [4].

  - formátovanie textu

    Ide o počet resp. pomer prázdnych riadkov a celkového počtu riadkov [3].

  - pravopis

    Ide o pomer slov s veľkými písmenami k celkovému počtu slov v texte a pomer veľkých písmen k celkovému počtu písmen [3].

  - úvodné/záverečné frázy

    Ide o časté frázy na začiatku a konci správ, najmä e-mailov [3].

- syntaktické znaky

  - vetné členy

    Ide o unigramy, bigramy a trigramy vetných členov [5].

  - syntax

    Ide o závislosti resp. vzťahy medzi jednotlivými vetnými členmi [5] a relatívne frekvencie vetných členov ich samotných a ich pomeru k iným znakom [4].

- iné
  - miery zložitosti

    Ide o miery, ktoré sa snažia vypočítať zložitosť textu napríklad vytvorením modelu jazyka a následného ohodnotenia textu podľa neho [5] alebo pomocou analýzy hĺbky či košatosti syntaktického stromu viet [4].

  - vzdialenosti stredov zhlukov

    Ide o metaznak vytvorený vzdialenosťami vektorov znakov od stredov vytvorených zhlukovacím algoritmom [5].

Niektoré rozoberieme bližšie:

Miery zložitosti

Sú to miery, ktoré sa snažia odhadnúť istú zložitosť textu. Tento prístup sa používal najmä pri prvých snahách o určovanie autorstva pre jeho jednoduchosť.

Patrí tu napríklad priemerná dĺžka slov, resp. rozdelenie dĺžok slov. Tá sa dá merať rôzne v závislosti od zvolenej najmenšej jednotky. Konkrétnym príkladom je počet slabík alebo počet písmen. Používaný je aj priemerný počet slov vo vete [1].

Tieto mierky avšak nie sú veľmi efektívne samostatne. Väčšiu efektivitu dosahujú až v kombinácii s inými znakmi.

Funkcionálne slová

Funkcionálne slová sú slová, ktoré v texte určujú gramatické alebo štrukturálne závislosti jednotlivých slov vety. Sú to väčšinou neplnovýznamové slová. V slovenčine na tento účel slúži hlavne ohýbanie slov. Tieto slová sú nezávislé od obsahu textu. Keďže sú zväčša ovládané gramatikou textu, cielená manipulácia s nimi je nepravdepodobná [1].

Tento znak meria relatívne rozdelenie týchto slov navzájom. V súčasnosti sa používajú zoznamy obsahujúce stovky slov, ktoré obsahujú slovné druhy ako predložky, spojky, členy, ale taktiež aj plnovýznamové slovné druhy ako sú zámená a modálne slovesá [1].

Syntax, vetné členy

Tento znak využíva syntax, resp. vetné členy. Pri predspracovaní textu sú slovám určené ich slovné druhy, na ktorých sa počíta ich relatívna frekvencia, resp. frekvencie na ich krátkych postupnostiach v kombinácii s konkrétnymi slovnými druhmi [1].

Funkcionálne lexikálne taxonómie

Tento znak vytvára taxonómie postavené na funkcionálnych slovách, ktoré reprezentujú gramatické a sémantické rozdiely medzi triedami funkcionálnych slov v rôznych úrovniach abstrakcie. Vytvára stromy, ktoré v koreni majú slovných druh a každý potomok je označený podtriedou svojho rodiča (napríklad druhy zámen). V listoch sa nachádzajú konkrétne slová. Tieto stromy sú vytvárané zo slov z uzavretej množiny, takže predspracovanie textu na označovanie slov nie je potrebné [1].

Obsahové slová

Na rozdiel od funkcionálnych slov obsahové slová sú slová, ktoré nesú obsah textu. Takéto slová sú teda ovplyvňované preferenciami autora textu (napríklad pri výbere synoným). Jednotlivé preferencie autora sa dajú merať pomocou relatívnych frekvencií synoným. Zvyčajne veľmi zriedkavé slová a slová s rovnomerným rozdelením sa vyraďujú, aby na analýzu zostalo niekoľko tisíc slov. Je možné použiť aj postupnosti takýchto slov. Problémom tohto znaku je, že keďže je závislý na obsahu textu, výsledným rozdielom v analýze môže byť len obsahový rozdiel textov [1].

Písmenové multigramy

Písmenové, resp. znakové n-gramy (zvyčajne pre $n \leq 3$ pre veľký počet kombinácií) počítajú relatívne frekvencie, resp. rozloženie n po sebe idúcich znakov. Vedia zachytiť lexikálne, gramatické a ortografické preferencie autorov bez potreby lingvistického vzdelania na interpretáciu výsledkov. Nie sú závislé na jazyku [1].

Úspešne boli použité v angličtine, holandčine, ruštine, taliančine, gréčtine [1].

4.2 Slovenský jazyk

Na analýzu slovenského textu sa dajú použiť znaky nezávislé na jazyku (zvyčajne sú to znaky pracujúce na úrovni písmen) a znaky špecifické pre slovenčinu, prípadne slovanské jazyky. Mohli by to byť tieto znaky:

Znakové, resp. písmenové

Frekvencie n-gramov písmen. Pri takýchto znakoch si na rozdiel od angličtiny môžeme vybrať, čo budeme považovať za rovnaké písmeno: či písmená líšiace sa len v diakritike budeme považovať za rovnaké alebo odlišné a či budeme brať písmená ako „ch“, „dz“ a „dž“ ako jedno písmeno alebo zoskupenie dvoch písmen. Tiež si môžeme vybrať, či budeme používať aj nepísmenové znaky, ako je interpunkcia, medzera alebo číslice.

Morfologické

Z morfologických znakov by sa dali použiť štatistiky výskytu slov podľa slovných druhov, gramatických kategórií a rôznych delení slov. Pre ohybné slovné druhy sa dajú určiť ich gramatické kategórie. Patria medzi nich aj gramatické kategórie ako rod a pád, ktoré sú na rozdiel od angličtiny špecifické pre slovenčinu. Štatistiky rodu slov by mohli byť obzvlášť zaujímavé, keďže rod, v ktorom autor o sebe píše, je zvyčajne konštantný vo všetkých jeho prejavoch. Pre všetky slovné druhy sa dajú počítať štatistiky rozdelenia do skupín v rámci ich slovného druhu.

Nevýhodou týchto vlastností je potreba predspracovania textu na ich určenie a taktiež prípadná nejednoznačnosť spôsobená viacerými možnými interpretáciami.

Radíme sem aj štatistiky výskytu interpunkcie.

Tieto znaky vedia v istom zmysle merať rozvitosť viet.

Syntaktické

Zo syntaktický znakov by sa dali použiť rôzne typy dĺžok, ako sú dĺžky slov, viet, riadkov, odsekov a iných jednotiek. Taktiež sem vieme zaradiť aj štatistiky druhov viet, vetných členov a ich vzťahov. U vetných členov sa dá vytvoriť strom ich vzájomných závislostí, na ktorom sa dajú vypočítať rôzne metriky, ako hĺbka, vzdialenosť najvzdialenejších listov či košatosť.

Zaradíme sem aj znaky ako počet prázdnych riadkov.

Lexikálne

Štatistiky vybratých konkrétnych slov a slovných spojení, resp. slovných multigramov (vrátane unigramov). Tieto slová môžu byť vopred vybraté alebo sa môžu použiť všetky slová textu.

Taktiež si môžeme vybrať, či budeme zohľadňovať ohýbanie slov – či rôzne tvary slova budeme považovať za jedno slovo, alebo odlišné slová. Neohybné slovné druhy toto neovplyvní, ale pre ohybné slovné druhy v prípade, že sa rozhodneme považovať rôzne tvary slova za jedno slovo, to bude vyžadovať predspracovanie textu.

Chyby

Medzi najosobnejšie znaky môžeme radiť rôzne chyby v texte, obzvlášť ak sú konzistentné pre daného autora. Chyby sa dajú považovať za metaznak o iných znakoch, pretože v podstate ide o prípady, kedy sú porušené pravidlá používania štruktúr textu na rôznych úrovniach, resp. iných znakov. Tieto znaky sa najviac budú vyskytovať v neupravených textoch, ako je emailová komunikácia alebo chatová správa.

Tu vieme rozlišovať medzi chybami na úrovni slov, resp. písmen ako napríklad preklepy a chybami na úrovni viet, ako napríklad neprávne skloňovanie alebo nezhoda v gramatických kategóriách, ako sú rod, číslo, pád u slov tvoriacich vetné členy v určovacom vetnom sklade.

5 Metódy

Určovanie autorstva je v svojej podstate klasifikačný problém. Preto je vhodné na neho použiť klasifikačné metódy. Uvedieme niekoľko populárnych metód [2]:

5.1 Metóda podporných vektorov (Support Vector Machine)

Ide o metódu, ktorá pracuje s vektormi v multidimenzionálnom priestore. Táto metóda sa snaží binárne klasifikovať vektory v priestore tým, že vytvorí hranicu, podľa ktorej sa určí klasifikácia. Jej idea spočíva v tom, že sa vytvoria tzv. podporné vektory, pomocou ktorých sa nájde lineárna hranica (nadrovina), ktorá rozdelí priestor vektorov na polovicu tak, aby na jednej strane boli vektory jednej triedy a na druhej strane vektory druhej triedy. Táto hranica sa určí maximalizovaním vzdialenosti k bodom, ktoré sa nachádzajú najbližšie k tejto hranici.

Keďže metóda rozdeľuje priestor na dve polpriestory nadrovinou, vyžaduje aby trénovacia množina bodov bola lineárne separovateľná. V prípadoch, kedy trénovacia množina nie je lineárne separovateľná, je možné použiť transformácie priestoru a metódu použiť na takto transformovanom priestore.

Keďže táto metóda vytvára binárne klasifikácie je najvhodnejšia na tretí variant problému určovania autorstva.

5.2 Naivný bayesovský klasifikátor

Táto metóda na základe hodnôt vstupného vektora vypočíta pravdepodobnosť klasifikácie daného vektora do každej triedy. Túto pravdepodobnosť vypočíta pomocou apriórnej pravdepodobnosti výskytu každého znaku (prvku vektora) pre danú kategóriu. Pritom výpočet nezohľadňuje možnú závislosť jednotlivých znakov navzájom, a teda predpokladá, že jednotlivé znaky sú nezávislé.

Vzhľadom na to, že táto metóda počíta pravdepodobnosti pre každú kategóriu a teda vytvára klasifikácie pre viacero kategórií, je najvhodnejšia pre prvý variant problému určovania autorstva.

5.3 k najbližších susedov (k nearest neighbour)

Táto metóda klasifikuje vstupné vektory na základe ich vzdialenosti vo vektorovom priestore pomocou trénovacej množiny vektorov. Pre daný vektor metóda vypočíta jeho k najbližších susedov a daný vektor klasifikuje podľa väčšinového hlasovania týchto susedov. Teda daný vektor dostane klasifikáciu, ktorú má väčšina jeho susedov v jeho najbližšom okolí.

Táto metóda je vhodná pre prvý variant problému určovania autorstva.

5.4 Logistická regresia

Táto metóda funguje na rovnakom princípe ako lineárna regresia, ale namiesto lineárnej kombinácie používa logistickú funkciu. Metóda binárne klasifikuje príslušnosť vstupného vektora do danej triedy. Metóda vypočíta lineárnu kombináciu jednotlivých prvkov vstupného vektora, z ktorej pomocou ligistickej funkcie vypočíta pravdepodobnosť príslušnosti vektora do danej triedy.

Keďže metóda počíta pravdepodobnosť klasifikácie vektora do jednej kategórie, s vhodným nastavením hranice pre klasifikáciu príslušnosti je vhodná pre všetky tri varianty problému určovania autorstva.

5.5 Rozhodovacie stromy

Ide o metódu, ktorá používa strom, resp. cestu stromom na klasifikáciu vstupného vektora do jednej z kategórií. Metóda využíva strom, ktorým prechádza od koreňa po vrcholoch, v ktorých sa rozhoduje, ktorým vrcholom bude ďalej pokračovať až k listom, v ktorých sa nachádza výsledná klasifikácia. Vo vrcholoch sa nachádzajú podmienky založené na hodnotách jednotlivých prvkov vstupného vektora. Podľa toho, či daná podmienka je alebo nie je splnená, sa rozhodne, ktorým potomkom sa pokračuje z daného vrcholu. Každý vrchol sa môže vetviť na dvoch alebo viac potomkov.

Počas trénovania sa začína koreňom, ktorý sa postupne rozvetvuje pridávaním, resp. vetvením listov na ďalšie vrcholy o úroveň nižšie v strome. Existuje viacero metrík a algoritmov, pomocou ktorých sa dá určiť ďalšie vetvenie a koniec vetvenia stromu, jedným z nich je napríklad pomocou pomerového informačného zisku rozdelenia podľa atribútu.

Existuje aj variant v ktorom sa používa viacero stromov – les. V tomto variante sa vypočíta klasifikácia pomocou všetkých stromov a výsledná klasifikácia sa určí kombináciou týchto výsledkov, zvyčajne pomocou väčšinového hlasovania.

Keďže táto metóda vie klasifikovať do dvoch, ale aj viacerých kategórií, je vhodná pre prvý a tretí variant problému určovania autorstva.

5.6 Winnowov algoritmus

Ide o metódu, ktorá pracuje na rovnakom princípe ako perceptrón, ale líši sa v jeho učení. Pri predikcii vypočíta lineárnu kombináciu binárnych prvkov vstupného vektora a pomocou naučenej hranie určí binárnu klasifikáciu. Pri učení sa zameriava na aktualizáciu len tých prvkov, ktoré prispeli k nesprávnej klasifikácii (prvkov s hodnotou 0).

Keďže táto metóda poskytuje binárnu klasifikáciu pre konkrétnu triedu, pričom sa zameriava len na znaky, ktoré majú najväčšiu výpovednú hodnotu, je najvhodnejšia pre prvú a tretiu variantu problému určovania autorstva.

Taktiež keďže tento algoritmus sa zameriava len na prvky, ktoré sú pre klasifikáciu významné, jeho naučená klasifikačná funkcia by sa mohla použiť na zistenie užitočnosti použitých znakov.

5.7 Multigramové vyhladzovanie (multigram smoothing)

Ide o metódu, ktorá vytvorí pravdepodobnostný n-gramový model jazyka. Tento model sa vytvára tak, že sa spočítajú všetky n-gramy znakov (resp. iných jednotiek) v zdrojovom texte. Na týchto n-gramoch sa potom vypočítajú podmienené pravdepodobnosti výskytu každého znaku za predpokladu, že bolo spozorovaných $n-1$ predchádzajúcich znakov. Na vytvorenie predikcie, či skúmaný text je modelovaný vytvoreným modelom, sa vytvorí súčin všetkých podmienených pravdepodobností znak po znaku na všetkých n-gramoch textu. Výsledné číslo reprezentuje pravdepodobnosť, že vytvorený model vyprodukuje daný text.

Tento jednoduchý prístup však v praxi často naráža na problém, že nie všetky n-gramy sa vyskytujú v zdrojovom texte. Jedným z prístupov na vyriešenie tohto nedostatku je vyhladzovanie. Pri vyhladzovaní sa namerané podmienené pravdepodobnosti kombinujú s nejakou inou hodnotou tak, aby aj v prípade, že samotná pravdepodobnosť má hodnotu nula, výsledok nebol nulový.

Jedným z variantov vyhladzovania je Wittenovo-Bellovo vyhladzovanie, v ktorom sa pomocou váženého priemeru kombinujú pravdepodobnosti n-gramov s pravdepodobnosťami $(n-1)$-gramov. Tento algoritmus teda kombinuje podmienené pravdepodobnosti n-gramov z viacerých úrovní.

Častý je aj prípad, kedy sa tieto metódy kombinujú, resp. dochádza k výraznému predspracovaniu znakov pred ich použitím v klasifikačnom algoritme.

6 Miery

Na analýzu a vyhodnotenie správnosti výsledkov z jednotlivých metód sa dá použiť veľa mier. Keďže ide o problém klasifikácie textov, často sa používajú zaužívané miery z oboru získavania informácií (information retrieval) [2].

Tie sú podľa autora A:

- Precision: $\mathrm{P}_A = \frac{|\text{správne priradené}(A)|}{|\text{všetky priradené}(A)|}$
- Recall: $\mathrm{R}_A = \frac{|\text{správne priradené}(A)|}{|\text{dokumenty od autora}(A)|}$
- Harmonický priemer: $\mathrm{F}_1 = \frac{2\mathrm{P}_A\mathrm{R}_A}{\mathrm{P}_A+\mathrm{R}_A}$

Tieto hodnoty potom vypovedajú o kvalite danej metódy na textoch od autora A.

Následne ak chceme vedieť kvalitu danej metódy na množine autorov $\{A_i : i \in \{1, \dots, n\}\}$, vieme podľa miery M vypočítať

- $makropriemer_M(\{A_i\}) = \frac{1}{n} \sum_i M_{A_i}$, kde n je počet autorov
- $mikropriemer_M(\{A_i\}) = \frac{1}{k} \sum_i k_i M_{A_i}$, kde k je počet dokumentov a k_i je počet dokumentov od A_i

Pomocou týchto hodnôt potom vieme navzájom porovnávať jednotlivé metódy.

Je možné však použiť aj klasické metódy založené na chybách, resp. odchýlkach od správneho ohodnotenia (ε). Takéto miery sú napríklad

- Mean Absolute Error – priemer absolútnych chýb

$$\sum \frac{|\varepsilon|}{n}$$

- Root Mean Squared Error – odmocnina priemeru druhých mocnín chýb

$$\sqrt{\sum \frac{\varepsilon^2}{n}}$$

7 Dataset

Dataset možno vytvoriť z rôznych zdrojových textov: zo Slovenského národného korpusu (SNK), z prepisov vystúpení poslancov v Národnej rade alebo z článkov z novín. Najvhodnejšími sa javia články z novín, pretože SNK nie je určený na takéto účely (nedajú sa v ňom prezerať publikácie v ich celom znení, čo znemožňuje jeho využitie na naše účely) a v prepisoch vystúpení poslancov je zachytené hovorené slovo a nie písomný prejav autora, teda napríklad znaky z kategórie Chyby by sa nedali použiť.

Ako zdroj na vytvorenie datasetu sme použili články z novín. Tieto články boli stiahnuté manuálne z webového portálu dňa 9. 12. 2020 pomocou webového rozšírenia. Toto webové rozšírenie sme naprogramovali pre webový prehliadač Mozilla Firefox v programovacom jazyku JavaScript. Webové rozšírenie po načítaní stránky článku v objektovom modeli dokumentu (DOM) vyhľadá jednotlivé odseky článku a ich textový obsah uloží do súboru spolu s menom jeho autora a adresou stránky vo formáte

JSON. Tieto súbory boli potom ďalej spracované pomocou aplikácie napísanej v programovacom jazyku C# bežiacej na platforme .NET Core.

Takýto poloautomatizovaný prístup pomocou webového rozšírenia nám umožnil vybrať si, ktoré články zahrnieme do datasetu bez nadmerného zaťažovania webserveru (keďže stránky sa načítavajú iba tak rýchlo, ako ich používateľ dokáže otvárať) a s istotou, že nenastane chyba v prepise, resp. vo vytváraní lokálnej kópie článku, keďže webové rozšírenie robí exaktné kópie textu znak po znaku (bajt po bajte).

Pri výbere článkov do datasetu sme sa zamerali na autorov s väčším množstvom článkov. Dôvodom je náš predpoklad, že často publikujúci autori majú dobre vypracovaný svoj osobný štýl, a teda ich články sa vyznačujú charakteristickým rukopisom. Väčší objem dát je pre naše účely vhodný aj preto, že sa snažíme využívať aj tie znaky textu, ktoré sa vyskytujú len s malou pravdepodobnosťou (s malou frekvenciou), a teda na ich zachytenie je potrebné väčšie množstvo textov. Väčší objem dát okrem toho vo všeobecnosti zvyčajne zlepšuje výsledky metód strojového učenia.

Výsledkom tohoto zberu bolo vytvorenie datasetu, ktorý obsahuje celkovo 2 154 článkov od 26 autorov.

7.1 Normalizácia

Po vytvorení datasetu prebehla jeho normalizácia.

V prvom kroku boli vyfiltrované články, ktoré pozostávali len z jedného odseku. Tieto články boli v skutočnosti tvorené len popisom alebo komentárom k publikovanému videu alebo fotografii. V tomto kroku bolo vyfiltrovaných 269 článkov a v datasete ostalo 1 885 článkov.

V druhom kroku boli normalizované webové adresy článkov na zjednodušenú formu, a to z dôvodu ich lepšieho spracovania. Webový portál, z ktorého boli články stiahnuté, používa pre jednotlivé publikačné sekcie rôzne domény. Ukázalo sa, že niektoré články sú publikované vo viacerých sekciách a teda sa nachádzajú na viacerých doménach, v dôsledku čoho boli pri zbere stiahnuté viackrát. Riešením bola normalizácia webovej adresy článku na jednotný identifikátor článku, pri ktorej bolo vyfiltrovaných 63 duplicitných článkov.

Do konečnej kolekcie článkov bolo teda umiestnených 1 822 článkov od 20 autorov.

Jednotlivé články potom boli rozdelené na jednotlivé odseky, vety a slová. Pri rozdeľovaní na odseky sa z textov vyfiltrovali odseky, ktoré boli len redaktorskými poznámkami (napríklad o aktualizácii článku). Pri rozdeľovaní na vety a slová boli odstránené špeciálne znaky vrátane číslic a entity, ktoré ich obsahovali (napríklad URI adresy, čísla a ich jednotky).

Štatistiky konečného datasetu

Počet autorov:	20
Počet článkov na autora:	26 až 143
Počet článkov:	1 822
Počet odsekov:	26 854
Počet viet:	61 010
Počet slov:	782 333
Počet písmen:	4 379 687

Priemerný počet odsekov v článku:	14,739
Priemerný počet viet v článku:	33,485
Priemerný počet viet v odseku:	2,272
Priemerný počet slov v článku:	429,381
Priemerný počet slov v odseku:	29,133
Priemerný počet slov vo vete:	12,823
Priemerný počet písmen v článku:	2 403,780
Priemerný počet písmen v odseku:	163,093
Priemerný počet písmen vo vete:	71,786
Priemerný počet písmen v slove:	5,598

8 Tvaroslovník

Na určenie gramatických kategórií slov sme použili Tvaroslovník. Ide o databázu slov zo slovnej zásoby slovenského jazyka a ich tvarov. Táto databáza vie pre zadané slovo, resp. jeho konkrétny tvar určiť jeho gramatické kategórie, ako napríklad rod, číslo, pád. Keďže jeho účelom je určovať vlastnosti jednotlivých tvarov slov, sú v ňom obsiahnuté najmä ohybné slovné druhy.

Databáza je implementovaná v jazyku SQL. Na jej prevádzku sme použili softvér MySQL, ktorý bežal ako lokálna inštancia na rovnakom počítači ako C# aplikácia na

spracovanie datasetu článkov.

9 Výsledky

Pomocou programu napísaného v C# sme spracovali jednotlivé články datasetu a určili na nich viaceré znaky. Podľa spôsobu ich výpočtu sa dajú rozdeliť na tieto kategórie:

- STA
  - štatistiky počtu a dĺžok odsekov, viet, slov, písmen

    Skladajú sa zo sedemdimenzionálneho vektora obsahujúceho tieto komponenty:

    * celkový počet odsekov v článku
    * celkový počet viet v článku
    * celkový počet slov v článku
    * celkový počet písmen v článku
    * priemerný počet viet v odseku
    * priemerný počet slov vo vete
    * priemerný počet písmen v slove
- CW
  - štatistiky počtov unigramov, bigramov a trigramov najčastejších slov z celého datasetu

    Skladajú sa z vektorov počtov 20 najčastejšie sa vyskytujúcich unigramov, bigramov a trigramov slov z celého datasetu v aktuálnom článku.
- CNG
  - štatistiky počtov bigramov a trigramov najčastejších písmen z celého datasetu

Skladajú sa z vektorov počtov 20 najčastejšie sa vyskytujúcich bigramov a trigramov písmen z celého datasetu v aktuálnom článku.

Tieto znaky sme potom v rôznych kombináciách zadali do programu WEKA [7]. V tomto programe sme vyskúšali viaceré metódy strojového učenia:

- J4.8
  - Rozhodovací strom

    Vytvára rozhodovací strom, ktorý, začínajúc v koreni a končiac v listoch, dokáže klasifikovať daný vstup postupným odpovedaním na binárne otázky o jednotlivých atribútoch v jednotlivých vrcholoch.

- NB
  - Naivný Bayes

    Využíva predpoklad, že jednotlivé vstupné atribúty sú nezávislé a kombinuje ich hodnoty pomocou normálneho rozdelenia do výsledného súčinu pravdepodobností.

- NBM
  - Naivný Bayes multinomický

    Variant naivného bayesovského algoritmu, ktorý používa multinomické rozloženie namiesto normálneho.

- SL
  - Logistická regresia

    Jednotlivé atribúty kombinuje podobne ako lineárna kombinácia, avšak vo výsledku používa logistickú funkciu.

- SMO
  - Metóda podporných vektorov

    Natrénuje model podporných vektorov pre každú triedu a pomocou párovej klasifikácie vyberie výslednú klasifikáciu.

Na vyhodnotenie sme použili 10-smernú krížovú validáciu, pri ktorej sa dataset rozdelil na 10 rovnomerných častí. Potom na každej tejto časti prebehla jedna inštancia trénovania a vyhodnocovania. V každej inštancii sa daná časť použila ako testovacia sada a zvyšné časti ako trénovacia sada. Výsledné hodnotenie metódy je potom aritmetický priemer zo všetkých inštancií.

Precision	J4.8	NB	NBM	SL	SMO
STA	0,728	0,356	0,296	0,338	0,353
CW	0,779	0,365	0,424	0,572	0,611
CNG	0,822	0,279	0,382	0,536	0,464
CW+CNG	0,831	0,366	0,539	0,726	0,634
STA+CW+CNG	0,851	0,413	0,591	0,742	0,699

Tabuľka 2: Porovnanie presnosti jednotlivých metód na rôznych kombináciách znakov textu.

F_1	J4.8	NB	NBM	SL	SMO
STA	0,723	0,238	0,249	0,323	0,336
CW	0,771	0,287	0,418	0,551	0,383
CNG	0,816	0,158	0,381	0,515	0,266
CW+CNG	0,829	0,246	0,533	0,713	0,456
STA+CW+CNG	0,847	0,259	0,556	0,732	0,543

Tabuľka 3: Porovnanie F_1 miery jednotlivých metód na rôznych kombináciách znakov textu.

Tabuľky 2 a 3 obsahujú podobné výsledky. Možno v nich vidieť trend, ktorý poukazuje na to, že viac znakov dosahuje lepšie výsledky. Toto by sa dalo vysvetliť tým, že všetky použité metódy patria k metódam strojového učenia – takéto metódy sú zvyčajne schopné efektívne využiť všetky im dostupné užitočné informácie a odignorovať nepotrebné informácie.

Ďalším zistením je skutočnosť, že rozhodovacie stromy dosahujú najlepšie výsledky, zatiaľ čo metóda Naivného Bayesa dosahuje najhoršie výsledky. Možným vysvetlením je to, že jednotlivé znaky nie sú nezávislé, a teda nespĺňajú predpoklad metódy Naivného Bayesa.

Zoznam použitej literatúry

[1] KOPPEL, Moshe; SCHLER, Jonathan; ARGAMON, Shlomo. Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 2009, 60.1: 9-26.

[2] ARGAMON, Shlomo; JUOLA, Patrick. Overview of the international authorship identification competition at PAN-2011. In: *CLEF (Notebook Papers/Labs/Workshop)*. 2011.

[3] KERN, Roman, et al. Vote/veto meta-classifier for authorship identification. In: *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands*. 2011.

[4] TANGUY, Ludovic, et al. A multitude of linguistically-rich features for authorship attribution.

[5] SOLORIO, Thamar, et al. Authorship Identification with Modality Specific Meta Features Notebook for PAN at CLEF 2011. 2011.

[6] BURROWS, John F. Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary & Linguistic Computing*, 1987, 2.2: 61-70.

[7] WITTEN, Ian H., et al. The WEKA workbench. online appendix for „Data Mining: Practical machine learning tools and techniques“. In: Morgan Kaufmann. 2016.