Intrinzična detekcija plagijata

Viher, Helena

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
SMJER INFORMATIKA (istraživački)
Ak. god. 2019./2020.

Helena Viher

Intrinzična detekcija plagijata

Diplomski rad

Mentor: doc. dr. sc. Vedran Juričić

Zagreb, listopad 2020.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Sadržaj

Uvod

Iako su napredak informacijskih tehnologija i olakšan pristup informacijama imali mnogo pozitivnih učinaka na ljudsko društvo, istovremeno je došlo do javljanja određenih negativnih posljedica. Jedna od njih svakako je plagiranje, koje je postojalo i ranije, no zbog brzog i jednostavnog pristupa besplatnim izvorima informacija postaje sve učestalija pojava. Plagiranje u posljednjih nekoliko desetljeća postaje gotovo svakodnevna pojava na svim razinama školovanja, a dovodi do lošijeg usvajanja gradiva te nepravednog nagrađivanja učenika i studenata za radove koje nisu samostalno izradili. Plagiranje predstavlja veliki problem i u akademskim krugovima u kojima osim problematike vezane uz preuzimanje zasluga za tuđi rad, ono može i dovesti do širenja netočnih informacija, ukoliko se u radu iz kojeg se plagira javljaju neispravni podaci. No, zbog velike količine radova koju je potrebno pregledavati, detaljno provjeravanje svakog pojedinačnog rada postaje teško izvedivo. Iz ovog razloga javlja se potreba za razvijanjem sustava za automatsku detekciju plagijata koji su u mogućnosti obraditi velike količine podataka u kratkom vremenskom razdoblju te profesore i ostale ispitivače upozoriti na potencijalno plagiranje unutar predanog teksta. Cilj ovog rada je dati pregled različitih načina definiranja plagiranja, oblika u kojima se plagiranje može javiti te predstaviti sustave za automatsku detekciju plagijata s naglaskom na pristupe i metode u intrinzičnoj detekciji plagijata. U radu će se također razmotriti prednosti i nedostaci intrinzičnih sustava za detekciju plagijata, prikazat će se upotreba metoda i pristupa koji se koriste u intrinzičnoj detekciji plagijata u drugim područjima te načini za rješavanje i prevenciju plagiranja.

1. Što je plagiranje?

Plagiranje je moguće definirati na više načina, no najjednostavnije rečeno, ono podrazumijeva preuzimanje tuđeg pisanog rada i pripisivanje tog djela samome sebi (Encyclopædia Britannica). Osim plagiranja čitavog rada, moguće je plagirati samo jedan njegov dio ili pak samo ideju kojom se rad bavi. Također, osim iz samo jednog, moguće je plagirati i iz više radova odjednom. Plagiranje pisanih radova obuhvaća sljedeće radnje (Plagiarism.org):

- predaja tuđeg rada kao vlastitog
- kopiranje riječi ili ideja druge osobe bez citiranja
- izostavljanje navodnika kod direktnog citiranja
- davanje lažnih ili krivih informacija o izvoru iz kojeg se citira
- zamjena riječi unutar rečenice sinonimima pri čemu se zadržava izvorna struktura rečenice bez citiranja izvora iz kojeg potječe
- preuzimanje velike količine sadržaja iz drugog rada što rezultira tekstom koji gotovo da ne sadrži originalne ideje i mišljenja, bez obzira na to jesu li izvori pravilno citirani ili ne.

Pisani radovi nisu jedini izvor iz kojeg se može plagirati jer plagiranje likovnih radova, programskog koda, glazbe, videozapisa i sl. nije rijetkost. Povrh toga, važno je napomenuti da plagiranje ne mora biti namjerno te u nekim slučajevima do njega dolazi zbog neznanja autora i neupućenosti u pravilan način citiranja (Weber-Wulff, 2016). Autori, na primjer, mogu slučajno zaboraviti dodati citat ili pak mogu citirati krivi izvor (Maurer, Kappe, & Zaka, 2006) Uz to, kvalitativni poremećaji pamćenja, poznati još i kao skriveno sjećanje ili kriptomnezija, mogu "dovesti do plagiranja, jer je osoba uvjerena da su njene ideje izvorne" (Hrvatska enciklopedija, 2020), a zapravo se radi o nesvjesno potisnutim sadržajima.

Plagiranje je danas izuzetno raširen problem, a neka od područja u kojima se javlja uključuju akademske krugove, znanstvena istraživanja, novinarstvo, patente i književna djela (Oberreuter & Velásquez, 2013). Plagiranje se također učestalo javlja tijekom školovanja, a ostvarivanje odličnog uspjeha na temelju radova i zadaća koje učenici i studenti nisu samostalno izradili ima velik utjecaj na ekstrinzičnu motivaciju pri učenju i usvajanju novih vještina (Foltýnek, Gipp & Meuschke, 2019). Ono, između ostalog, predstavlja i problem za profesore koji dobivaju krive povratne informacije o usvojenosti gradiva što nadalje dovodi do nezasluženih beneficija za plagijatora (Foltýnek et al., 2019). Prema istraživanju

Internacionalnog centra za akademski integritet (engl. *International Center for Academic Integrity*), provedenom u razdoblju od 2002. do 2015. godine, u kojem je sudjelovalo preko 70 000 učenika iz 24 različite srednje škole u Sjedinjenim Američkim državama, 95% ispitanika priznalo je da je počinilo neki oblik varanja tijekom školovanja. Pri tome njih 64% priznaje varanje na testovima, a 58% ih priznaje da je u svojim radovima plagiralo (ICAI). Još jedno istraživanje o raširenosti plagiranja provedeno u Sjedinjenim Američkim Državama, na preko 71 000 studenata preddiplomskih i preko 11 200 studenata diplomskih studija, u razdoblju od 2002. do 2005. godine, pokazalo je da se plagiranje kopiranjem i lijepljenjem sadržaja, parafraziranjem te falsificiranjem bibliografije javlja veoma često, odnosno jedna četvrtina do polovica studenata preddiplomskog studija te jedna četvrtina studenata diplomskog studija priznaje da je plagirala na ovaj način (McCabe, 2005). U Hrvatskoj su pak istraživanja pokazala da čak dvije trećine studenata farmacije i medicinske biokemije smatra da plagiranje nije teški prekršaj etičkog kodeksa u akademskim krugovima (Pupovac, Bilić-Zulle, Mavrinac, & Petrovečki, 2010). Ovakav stav prema plagiranju predstavlja veliki problem za obrazovne ustanove u Hrvatskoj. Zbog činjenice da bi znanstveni sustav trebao počivati na "iskrenosti, povjerenju i na stvaralačkoj sumnji u ispravnost vlastitih i tuđih nalaza" (Rumboldt, 2014, str. 233), plagiranje u akademskoj zajednici također predstavlja veliki problem. U zapadnim se zemljama na plagiranje u akademskim krugovima gleda kao na nečasno akademsko ponašanje te zakonski prijestup. Plagiranje u akademskim krugovima uključuje izmišljanje i falsificiranje istraživačkih rezultata, počasno autorstvo u kojem se kao ključni autor dodaje osoba koja za rad nije doprinijela ili je pridonijela vrlo malo, unajmljivanje gostiju-pisaca te pisaca-duhova, krađu tuđih ideja tijekom procesa recenziranja i mentorstva te ostale zlouporabe istraživačkog, nastavničkog i znanstvenog položaja (Cerjan-Letica & Letica, 2008).

1.1. Povijest plagiranja

Riječ plagiranje potječe od latinskog pojma *plagiarius* čije je značenje "otimatelj", a ovaj pojam izveden je od latinskog naziva za mrežu kojom se hvata divljač (lat. *plaga*) (Merriam-Webster). Pojam *plagiarius* u kontekstu "otimanja riječi" prvi koristi rimski pjesnik Martial koji u svojim epigramima često optužuje druge pjesnike za kopiranje i prisvajanje njegovih djela (Mira Seo, 2009). Plagiranje je, naravno, postojalo i prije Martialovog vremena, no pogled na njega uvelike se razlikovao od onoga koji se javlja u suvremenom društvu.

Lynch (2006) u svom članku *The Perfectly Acceptable Practice of Literary Theft: Plagiarism, Copyright, and the Eighteenth Century* navodi kako se kroz prošlost riječi i ideje nisu smatrale privatnim vlasništvom te se "posuđivanje" od majstora umjetnosti čak i poticalo. U nastavku navodi da su priče i likovi u klasičnoj umjetnosti uglavnom bili javno vlasništvo, odnosno potjecali su iz narodnih mitova i legendi koji sami po sebi nemaju poznatog autora. Istovremeno se umjetnike koji su izmišljali nove i originalne ideje često smatralo arogantnima. Ograničen pristup književnim i umjetničkim djelima, slaba pismenost stanovništva te odsutnost tehnologija i alata za brzo kopiranje djela u velikim količinama sve do pojave tiskarskog stroja dodatno su utjecali na ovakav pogled te se na kopiranje djela i "posuđivanje" ideja gledalo kao na način kroz koji se te ideje mogu proširiti i popularizirati (Bailey, 2019).

Smatra se da se moderni pogled na plagiranje prvi puta javio 1601. godine kada je engleski pjesnik i dramaturg Ben Jonson riječ plagiranje (engl. *plagiarism*) upotrijebio u kontekstu "književne krađe" (Isaacs, 2011, str. 1), a kao riječ u engleskom jeziku definirao ju je u svom rječniku Samuel Johnson 1755. godine (Lynch, 2006). Randall (2001) u svojoj knjizi *Pragmatic Plagiarism: Authorship, Profit, and Power* pokret prosvjetiteljstva, uspon kapitalizma te pojačani individualizam koji dovode do pojave zakona o autorskim pravima navodi kao velike utjecaje na suvremeno shvaćanje plagiranja. No, kao što je već ranije spomenuto, najveći utjecaj na rasprostranjenost plagiranja danas imao je tehnološki razvoj ljudskog društva. Pri tome je naročito velik utjecaj imao razvoj informacijskih tehnologija koji je omogućio brz i jednostavan pristup velikoj količini informacija (Williams, Nathanson, & Paulhus, 2010). Pojava *copy-paste* funkcije također je uvelike olakšala plagiranje (Majstorović, 2016). No, razvoj tehnologije istovremeno je omogućio automatsku detekciju plagijata (Williams et al., 2010). Prvi sustavi za automatsku detekciju plagijata javili su se 1970-ih godina, a temeljili su se na obradi prirodnog jezika (Abid, Usman, & Ashraf, 2017).

1.2. Podjela plagiranja

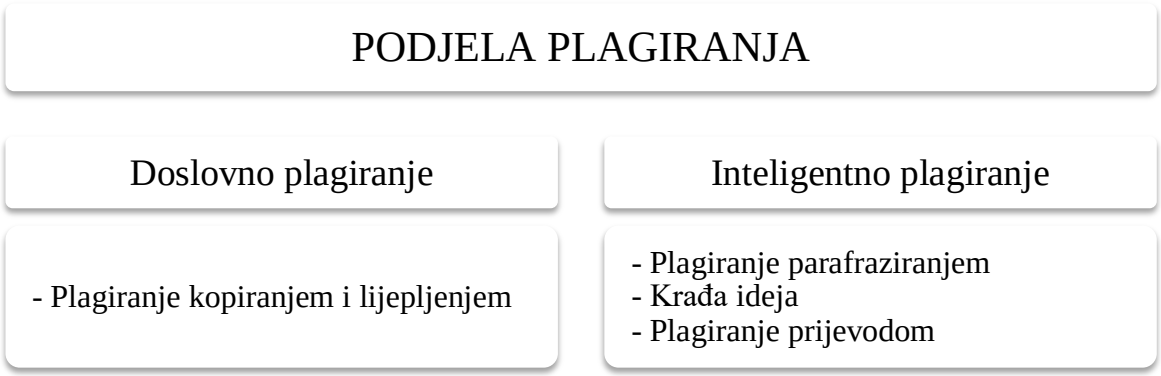
Poznavanje oblika u kojima se plagiranje može javiti te klasifikacija istih čini izuzetno važan aspekt u detekciji plagijata. Osim što predstavlja pomoć u prepoznavanju plagiranja unutar teksta, ono omogućava i izgradnju posebno prilagođenih sustava koji su u mogućnosti automatski detektirati određene oblike plagiranja. Raširenost plagiranja odražava

se i u velikom broju različitih oblika u kojima se plagiranje može javiti, a u pisanim radovima najčešće nailazimo na:

- **plagiranje tehnikom kopiranja i lijepljenja** (engl. *copy-paste plagiarism*) koje spada u jedan od najčešćih oblika plagiranja, a podrazumijeva izravno preuzimanje teksta iz izvora bez ikakvih dodatnih izmjena te bez pravilnog citiranja ili navođenja rada i autora u popisu literature (Alzahrani, Salim, & Abraham, 2012)
- **plagiranje parafraziranjem** (engl. *paraphrasing*) koje se odnosi na oblik plagiranja u kojem se tekst izmjenjuje u dovoljnoj mjeri da ne bude prepoznat kao plagiran, no istovremeno se zadržava izvorna ideja te neke od riječi i fraza (Oberreuter i Velásquez, 2013)
- **plagiranje prijevodom** (engl. *translated plagiarism*) koje obuhvaća radnje u kojima plagijator ručno ili uz pomoć računala prevodi tekst iz jednog jezika u drugi s ciljem prikrivanja plagiranja (Gipp, Meuschke, & Beel, 2011)
- **plagiranje strukture** (engl. *plagiarism of structure*) koje se odnosi na preuzimanje strukture tuđeg rada, tj. njegovog slijeda poglavlja, odlomaka, riječi u rečenici, koraka u rasuđivanju i sl. (myDU)
- **samoplagiranje** (engl. *self-plagiarism*) koje podrazumijeva oblik plagiranja u kojem autor svoje prethodno predane ili objavljene radove predstavlja kao nove, bez citiranja svojih prijašnjih radova (Clough, 2000)
- **zaboravljene ili zastarjele poveznice na izvore** (engl. *forgotten or expired links to references*) koje podrazumijevaju citate koji se nalaze unutar navodnika te imaju oznake za fusnotu, no informacije o samom izvoru ili nedostaju ili su poveznice koje vode do njega zastarjele (Maurer et al., 2006).
- **dezinformacije u fusnotama** (engl. *misinformation of references*) koje se odnose na namjerno postavljanje krivih fusnota na citat te postavljanje poveznica koje usmjeravaju na rad koji ne postoji (Maurer et al., 2006)
- **plagiranje ideja** (engl. *idea plagiarism*) koje se odnosi na oblik plagiranja u kojem plagijator iz drugog rada preuzima ideje i argumente, ali ih prikazuje svojim riječima i strukturalno organizira na drugačiji način, a pri tome ne citira rad iz kojih ih je preuzeo (Adam & Suharjito, 2014)
- **umjetničko plagiranje** (engl. *artistic plagiarism*) koje podrazumijeva predstavljanje tuđih ideja u nekom drugom mediju, npr. slika, zvučni ili video zapis plagira se tako da se preoblikuje u tekstualni oblik bez navođenja izvora (Maurer et al., 2006)

- **tehničko prikrivanje** (engl. *technical disguise*) koje se odnosi na tehnike kojima se pokušavaju eksploatirati nedostaci i slabosti u sustavima za detekciju u svrhu onemogućavanja automatske detekcije plagijata, a mogu uključivati zamjenu znakova grafički identičnim simbolima iz drugog pisma, dodavanje nasumičnih znakova u bijelom fontu u tekst, itd. (Heather, 2010).

Vrste plagiranja moguće je klasificirati na više načina s obzirom na njihova svojstva poput složenosti, razine jezika koju obuhvaćaju te lakoće detekcije. Alzahrani et. al (2012) kao dva glavna oblika plagiranja s obzirom na način na koji je plagiranje provedeno navode **doslovno plagiranje** (engl. *literal plagiarism*) te **inteligentno plagiranje** *(*engl. *intelligent plagiarism*). Pri tome, doslovno plagiranje podrazumijeva oblike plagiranja u kojima plagijator nije posvetio previše vremena skrivanju čina plagiranja, a obuhvaća radnje u kojima se fragmenti teksta preuzimaju izravno iz izvora tehnikom kopiranja i lijepljenja. Kada govorimo o inteligentnom plagiranju Alzahrani et. al (2012), prikazanom na Slici 1., navode da se zapravo radi o činu u kojem plagijator na dosjetljiv način pokušava prikriti plagiranje, a ova vrsta plagiranja može uključivati parafraziranje izvornog teksta na leksičkoj ili sintaktičkoj razini, krađu ideja te plagiranje prijevodom.



Slika 1. Podjela plagiranja prema Alzahrani et. al (2012)

Foltýnek et al. (2019) vrste plagiranja dijele na nešto drugačiji način te ih klasificiraju s obzirom na razinu prikrivenosti. U svom radu navode pet razina plagiranja u kojima u obzir uzimaju razinu jezika koju plagiranje obuhvaća:

1. **plagiranje s očuvanjem znakova** (engl. *characters-preserving plagiarism*) koje podrazumijeva doslovno plagiranje, odnosno plagiranje tehnikom kopiranja i lijepljenja
2. **plagiranje s očuvanjem sintakse** (engl. *syntax-preserving plagiarism*) koje uključuje tehničko prikrivanje te zamjenu riječi u rečenicama iz izvornog teksta sinonimina
3. **plagiranje s očuvanjem semantike** (engl. *semantics-preserving plagiarism*) u koje spada plagiranje parafraziranjem te prijevodom
4. **plagiranje s očuvanjem ideja** (engl. *idea-preserving plagiarism*) koje podrazumijeva plagiranje strukture rada te krađu koncepata i ideja
5. **skriveno autorstvo** (engl. *ghostwriting*) u kojem je autor plaćen da napiše neki tekst, no autorstvo se potom pripisuje nekoj drugoj osobi.

Pri tome oblike plagiranja koji spadaju u niže razine autori navode kao lakše za detektiranje uz pomoć sustava za automatsku detekciju plagijata, no napominju da oni na višim razinama često predstavljaju veliki problem za takve sustave. Tako na primjer skriveno autorstvo u kojem plagijator plaća drugu osobu da mu napiše originalni rad, spada u jedan od oblika plagiranja čija je detekcija izuzetno teška (Foltýnek et al., 2019). Plagiranje ideja, strukture te plagiranje u višejezičnom okruženju također spadaju u oblike plagiranja koje je izuzetno teško detektirati (Mozgovoy, Kakkonen, & Cosma, 2010).

2. Detekcija plagijata

U pisanim djelima plagiranje je moguće detektirati na temelju određenih karakteristika koje se javljaju u plagiranim dijelovima teksta. Znakovi koji mogu upućivati na plagiranje su sljedeći (Clough, 2003):

- uporaba naprednog ili tehničkog vokabulara koji je iznad onog koji se očekuje od autora djela
- veliko poboljšanje u stilu pisanja u usporedbi s prijašnjim radovima autora
- nedosljednosti unutar samog teksta koje mogu uključivati promjene u vokabularu, stilu i kvaliteti pisanja
- nesuvisli dio teksta čiji je tok nedosljedan u usporedbi s ostatkom rada, a može ukazivati na odlomak koji je kopiran iz postojećeg elektroničkog izvora
- velika sličnosti između dva ili više predanih radova, a uključuje i sličnosti u stilu pisanja
- pravopisne i gramatičke pogreške koje se istovremeno javljaju u više različitih tekstova
- tzv. "viseći" citati (engl. *dangling references*) koji se javljaju unutar teksta, ali ne i u popisu literature
- nedosljednosti u stilu citiranja koje ukazuju na kopiranje iz drugog, najčešće elektroničkog izvora.

Plagirane odlomke koji su izravno preuzeti iz izvornog teksta, tzv. *copy-paste* tehnikom, također je moguće uočiti na temelju karakteristika kao što su (Montgomerie, 2019):

- iznenadne promjene u veličini fonta, tipu ili boji slova
- promjene u izgledu navodnika i ostalih interpunkcijskih znakova
- promjene u marginama dokumenta
- promjene u razmaku između redaka
- prisutnost poveznica unutar teksta koje nisu uklonjene pri kopiranju i lijepljenju odlomka, a često vode prema izvoru teksta iz kojeg se plagira
- neobični razmaci unutar i između redaka u dokumentu koji dovode do "slamanja" teksta unutar dokumenta
- iznenadne promjene u postavkama jezika između odlomaka.

Unatoč tome što je plagiranje u većini radova relativno lako detektirati ručno na temelju promjena u stilu, kvalitete pisanja i sl. čak i bez uporabe referentnih materijala (Meyer zu

Eissen, Stein & Kulig, 2007) zbog velikog broja školskih, akademskih i drugih radova koje je potrebno pregledati javlja se potreba za razvijanjem automatskih sustava za detekciju plagijata. Kao što je već ranije spomenuto, dva glavna pristupa u automatskoj detekciji su ekstrinzična i intrinzična detekcija plagijata koji na različite načine pristupaju detekciji plagiranja unutar teksta (Potthast, Stein, Eiselt, Barron-Cedeno & Rosso, 2009).

Ekstrinzični pristup rad za koji se sumnja da je plagiran uspoređuje s referentnom zbirkom tekstova, a uspješnost ovog pristupa izravno je povezana s veličinom i potpunosti zbirke koja se koristi za usporedbu sa sumnjivim dokumentom (Foltýnek et al., 2019). Zbog tehnoloških ograničenja vezanih uz brzinu računalne obrade podataka sustavi za ekstrinzičnu detekciju plagijata rad koji se provjerava uglavnom ne uspoređuju s čitavom bazom, već se prvo odabire set dokumenata za koje se smatra da su najvjerojatniji kandidati plagiranja, a taj se set tek potom uspoređuje s radom za koji se sumnja da je plagiran (Chitra & Rajkumar, 2015). Prvi korak u kojem se odabire set dokumenata poznat je pod nazivom izlučivanje kandidata (engl. *candidate retrieval*) ili pak kao izlučivanje izvora (engl. *source retrieval*), dok se drugi korak, tj. usporedba sumnjivog dokumenta s prethodno odabranim setom, naziva detaljnom analizom (engl. *detailed analysis*), a ovaj korak poznat je još i kao poravnavanje tekstova (engl. *text alignment*) (Potthast, Gollub, Hagen, Tippmann, Kiesel, Rosso, Stamatatos, & Stein, 2013).

Osim što kvaliteta rezultata detekcije uvelike ovisi o dobroj i potpunoj bazi referentnih radova, ekstrinzična detekcija plagijata često nailazi na probleme kada se radi o složenijim oblicima plagiranja kao što su plagiranje parafraziranjem, prijevodom, krađa ideja i sl. (Mozgovoy et al., 2010). Ovi nedostaci potaknuli su na razvijanje intrinzičnih sustava koji plagiranje detektiraju na temelju promjena u stilu pisanja unutar samog dokumenta za koji se sumnja da je plagiran (Zechner, Muhr, Kern & Granitzer, 2009).

3. Intrinzična detekcija plagijata

Koncept intrinzične detekcije plagijata prvi puta predstavili su Meyer zu Eissen i Stein 2006. godine, a definirali su ga kao detekciju plagijata unutar samog rada na temelju stilističkih svojstava koja su jedinstvena za svakog pojedinačnog autora (Meyer zu Eissen & Stein, 2006). Oni, također, kao specifičnost ovog pristupa navode činjenicu da ne postoji potreba za uporabom referentne baze podataka. Naime, pri korištenju ovakve baze u ekstrinzičnom pristupu postoji mogućnost da baza ne sadrži izvorni tekst iz kojeg je rad plagiran što dovodi do nemogućnosti detekcije plagijata.

Glavna razlika između ekstrinzične i intrinzične detekcije je u tome što ekstrinzična detekcija plagiranje otkriva na temelju sličnosti između sumnjivog i drugih dokumenata, dok intrinzična detekcija radi na način sličan način kao i ručna detekcija plagijata, tj. ona pronalazi razlike u stilu pisanja između rečenica, odlomaka i poglavlja te ostale nedosljednosti i nepravilnosti unutar dokumenta za koji se sumnja da je plagiran (Foltýnek et al., 2019). Intrinzični pristup se umjesto ekstrinzičnog koristi u slučajevima kada ne postoji reprezentativni referentni korpus te kada bi računarska obrada izuzetno velikog korpusa oduzela previše vremena, a ponekad se koristi i kao predkorak u ekstrinzičnoj detekciji plagijata kako bi se osigurala što veća ušteda vremena i resursa (Stamatatos, 2009). Uspješnost ovog pristupa u detekciji plagijata ovisi o dvije pretpostavke:

1. stil svakog autora je jedinstven i prepoznatljiv te se u dovoljnoj mjeri može razlikovati od stilova drugih autora (Foltýnek et al., 2019)
2. mora postojati jedan "glavni" autor koji je napisao barem 70% rada (Kuznetsov, Motrenko, Kuznetsova, & Strijov, 2016).

Kuznetsov et al. (2016) navode da ovaj tradicionalni pristup intrinzičnoj detekciji nailazi na probleme kada ne postoji glavni autor djela, već je čitavo djelo sačinjeno od velike količine tekstualnih fragmenata koje su napisali različiti autori, a problemi se također javljaju kada je broj autora čiji se fragmenti nalaze unutar teksta nepoznat.

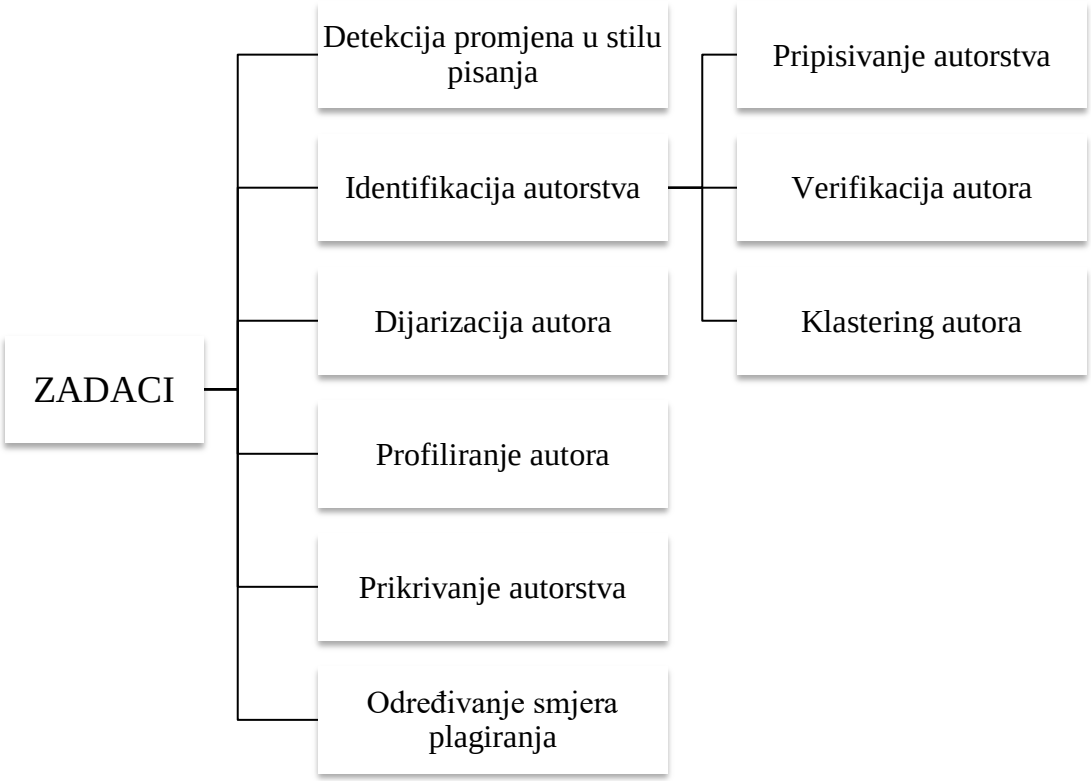
3.1. Zadaci intrinzične detekcije plagijata

Intrinzična detekcija plagijata, kao što je prikazano na Slici 2., bavi se nizom različitih zadataka i problema, no zajednička karakteristika svakog od njih je uska povezanost s odgovaranjem na pitanje autorstva. Identifikacija autorstva (engl. *authorship identification*) te

detekcija promjena u stilu pisanja (engl. *style change detection*) (Stamatatos, Rangel, Tschuggnall, Stein, Kestemont, Rosso, & Potthast, 2018) dva su osnovna zadatka intrinzičnog pristupa na kojima počivaju ostali zadaci i podzadaci intrinzične detekcije.

Detekcija promjena u stilu pisanja čini temelj ovog pristupa koji nadalje omogućava detekciju plagiranja unutar teksta te određivanje autorstva tekstualnih fragmenata koji se obrađuju. Identifikacija autorstva, poznata još kao i klasifikacija autora (engl. *author classification*) sastoji se od nekoliko podzadatka: **pripisivanja autorstva** (engl. *authorship attribution*) (Stein, Rosso, Stamatatos, Koppel, & Agirre, 2009), **verifikacije** (engl. *author verification*) te **klasteringa autora** (engl. *author clustering*) (Foltýnek et al., 2019).

Uz navedeno, intrinzična detekcija bavi se zadacima poput **dijarizacije autora** (engl. *author diarization*), **profiliranja autora** (engl. *author profiling*), **prikrivanja autorstva** (engl. *author obfuscation*) te **određivanja smjera plagiranja** (engl. *plagiarism direction identification*). Važno je napomenuti da se ovisno o autoru znanstvenog članka pojmovi autora i autorstva koriste naizmjenično, no imaju isto značenje, npr. pojmovi identifikacija autora i identifikacija autorstva označavaju isti zadatak.



Slika 2. Zadaci intrinzične detekcije plagijata

3.1.1. Detekcija promjena u stilu pisanja

Detekcija promjena u stilu pisanja, poznata još i kao detekcija proboja stila pisanja (engl. *style breach detection*), prvi je i osnovni zadatak intrinzične detekcija plagijata, a podrazumijeva detektiranje fragmenata teksta koji se stilski razlikuju jedni od drugih te se bavi definiranjem granica između njih (Karas, Spiewak & Sobecki, 2017). Promjene u stilu pisanja uglavnom se ispituju na razini odlomka, no nije rijetkost da se plagiranje unutar dokumenta javi na razini rečenice, tj. da svaka od rečenica unutar odlomka ima zasebnog autora (Hourrane & Benlahmer, 2019). Detektirane stilske promjene u rečenicama ili odlomcima mogu ukazivati na prisutnost plagiranja unutar dokumenta što ovaj zadatak čini izuzetno važnim za postupak intrinzične detekcije plagijata. Zadatak detekcije promjena u stilu pisanja u pravilu se odvija u tri koraka (Safin & Kuznetsova, 2017):

1. **segmentacija teksta** na temelju sheme za segmentaciju koja se može odnositi na paragrafe, rečenice koje se preklapaju, znakove ili pak *n*-grame unutar teksta
2. **izračun stilometrijskih mjera** za svaki od segmenata definiranih u prethodnog koraku
3. **pronalaženje vrijednosti koje najbolje definiraju autorov stil** kako bi se omogućila detekcija segmenata teksta koji mu ne pripadaju.

3.1.2. Identifikacija autorstva

Zadatak identifikacije autorstva, kao što je već spomenuto, uglavnom se dijeli na nekoliko podzadataka, a podrazumijeva podjelu na pitanja vezana uz pripisivanje autorstva te verifikaciju autora (Stein, Koppel, & Stamatatos, 2007). Glavni izvor informacija kod postupka identificiranja autorstva je, dakako, autorov stil pisanja (Polydouri, Vathi, Siolas, & Stafylopatis, 2018), no kako bi njegova detekcija bila moguća potrebno je prikupiti kvalitetne referentne zbirke svakog od autora koji se ispituje (Foltýnek et al., 2019).

Pitanje klasteringa autora uz pripisivanje i verificiranje autorstva također spada u jedan od podzadataka kojima se identifikacija autora bavi. Ovaj korak provodi se nakon detekcije promjena u stilu pisanja**,** a podrazumijeva grupiranje dokumenata ili fragmenata teksta prema autoru (Karas et al., 2017). U ovom se koraku dokumenti, odnosno odlomci teksta uparuju, a parovi se potom međusobno uspoređuju s ciljem izračuna mjere sličnosti

njihovih stilističkih svojstava kako bi se fragmenti (ili pak čitavi dokumenti) mogli pripisati pojedinačnim autorima (Foltýnek et al., 2019).

Zadatak pripisivanja autorstva bavi se pitanjem utvrđivanja autorstva nad anonimnim dokumentima pomoću referentnih setova tekstova više različitih autora, tj. na temelju setova dokumenata pojedinačnih autora potrebno je zaključiti koji je od autora napisao koji anonimni tekst (Vartapetiance & Gillam, 2014). S obzirom na to nalazi li se set pravog autora u referentnoj zbirci setova možemo govoriti o slučaju sa zatvorenim (engl. *closed-set*) ili otvorenim setom (engl. *open-set*) (Stamatatos et al., 2018). Naime, kada se radi o zatvorenom setu, pravi autor sigurno je jedan od autora čiji se setovi nalaze u referentnoj zbirci koja se koristi za ispitivanje, no u otvorenom setu postoji mogućnost da radovi pravog autora nisu prisutni unutar zbirke.

Verifikacija autora također se bavi pitanjem utvrđivanja autorstva nad dokumentima čiji autor nije poznat, no za razliku od pripisivanja autorstva, ovaj zadatak koristi referentni set dokumenata samo jednog autora na temelju kojeg je potrebno zaključiti je li autor tog seta ujedno i autor anonimnog dokumenta koji se ispituje (Vartapetiance & Gillam, 2014). Verifikacija je klasifikacijski problemu s jednom klasom (engl. *one-class classification problem*) zbog činjenice da dokumenti koji se ispituju ili pripadaju ili ne pripadaju jednom određenom autoru (Koppel & Schler, 2004). Detekcijom odskočnika (engl. *outliera*), odnosno značajki stila koje se ne poklapaju s onima u referentnom setu, ovaj pristup ima mogućnost donošenja zaključka o tome radi li se doista o tekstu koji je napisao jedan autor ili se pak radi o plagiranom tekstu koji pripada nekom potpuno drugom autoru (Foltýnek et al., 2019). Koppel i Schler (2004) verifikaciju autora smatraju vrlo složenim kategorizacijskim problemom. Naime, za razliku od zadatka pripisivanja autorstva u kojem se tekstovi poznatih autora jednostavno uspoređuju jedni s drugima, kod verifikacije autorstva u svrhu usporedbe pomoću sustava za intrinzičnu detekciju plagijata potrebno je odrediti stil pisanja koji autoru pripada te onaj koji mu ne pripada. Određivanje autorovog "ne-stila" koji u zadovoljavajućoj mjeri odstupa od autorovog stvarnog stila izuzetno je teško, što uvelike otežava postupak usporedbe, a samim time i postupak verificiranja autora. Kao dodatni problem Koppel i Schler (2004) navode mogućnost svjesne ili podsvjesne promjene u stilu pisanja s obzirom na žanr kojem tekst pripada, a uz to naglašavaju da nije rijetkost da autorov stil kroz određeno vremensko razdoblje na temelju utjecaja iz okoline spontano evoluira. Intrinzična detekcija plagijata i verifikacija autorstva usko su povezani te se smatra da zapravo čine "dvije strane istog novčića" (Stein et al., 2011, str. 78).

3.1.3. Dijarizacija autora

Pojam dijarizacije autora izvorno potječe iz područja dijarizacije govornika (engl. *speaker diarization*) koje se bavi identifikacijom različitih govornika u zvučnom zapisu na temelju razlika u zvučnim frekvencijama glasa koje su jedinstvene za svaku pojedinačnu osobu (Elamine, Mechti & Belguith, 2017). Zadatak dijarizacije autora u kontekstu intrinzične detekcije plagijata sličan je onome kojim se bavi dijarizacija govornika, odnosno postupkom dijarizacije pokušavaju se detektirati različiti autori unutar jednog dokumenta. Za razliku od tradicionalnog pristupa intrinzičnoj detekciji plagijata u kojoj je jedan od preduvjeta za uspješnost detekcije postojanje glavnog autora koji je samostalno napisao većinu rada, dijarizacija autora bavi se slučajevima kada ne postoji glavni autor te nije poznato koliki udio dokumenta čine tekstualni fragmenti različitih autora (Kuznetsov et al., 2016). Kuznetsov et al. (2016) s obzirom na autorstvo dijarizaciju dijele na tri podzadatka :

1. **tradicionalna intrinzična detekcija plagijata** u kojoj je jedan autor napisao više od 70% rada
2. **dijarizacija s poznatim brojem autora** gdje je poznat broj autora unutar teksta
3. **dijarizacija s nepoznatim brojem autora** gdje nije poznato koliko različitih autora postoji unutar teksta.

3.1.4. Profiliranje autora

Profiliranje autora je zadatak koji se bavi određivanjem profila autora, tj. autorovog spola, dobi, materinjeg jezika pa čak i osobnosti na temelju stilističkih karakteristika njegovih tekstova (Rosso, 2015). U ovu svrhu moguće je koristiti niz karakteristika kao što su bogatstvo vokabulara, redoslijed riječi u rečenici, gramatička točnost, prisutnost žargona, a (Rosso, 2015) kao stilističke karakteristike također navodi učestalost interpunkcijskih znakova, velikih slova te citiranja koje se zajedno s oznakama vrsta riječi (engl. *Part-of-Speech tags*) te detekcijom sadržajnih svojstava (engl. *content-based features*) pomoću latentne semantičke analize, tf–idf mjere (engl. *term frequency–inverse document frequency*), i sl. koriste za detekciju profila autora. Profiliranje autora može biti od koristi sustavima za detekciju plagijata jer bi se na temelju profila jednog ili više autora koji su prisutni unutar teksta moglo zaključiti je li autor koji predaje svoj rad taj rad plagirao od drugog učenika ili studenta, s interneta ili pak iz nekog drugog izvora, što može skratiti vrijeme potrebno za

pronalaženje izvora plagiranja te dokazivanje da je rad doista plagiran. No, nažalost, sustavi koji se bave profiliranjem autora još uvijek nisu dovoljno dobro razvijeni za praktičnu uporabu te ostvaruju točnost samo do 80% (Potthast, Rangel, Tschuggnall, Stamatatos, Rosso, & Stein, 2017).

3.1.5. Prikrivanje autorstva

Za razliku od prijašnjih zadataka čija je cilj bio detekcija autorstva, prikrivanje autorstva bavi se upravo suprotnim problemom (Bevendorff, Potthast, Hagen, & Stein, 2019). Njegova svrha je, kao što i sam naziv govori, prikrivanje identiteta pravog autora. Potreba za razvijanjem ovakvih sustava javila se kao odgovor na veoma precizne moderne sustave za detekciju autorstva koji uvelike ograničavaju privatnost autora, novinara i aktivista, koji svoje tekstove žele objaviti anonimno (Mahmood, Safiq, & Srinivasan, 2020). Ovakvi sustavi najprije detektiraju svojstva unutar teksta na temelju kojih je moguće identificirati autora, a tekst se potom izmjenjuje metodama adicije i redukcije pri čemu se koriste heuristički modeli za parafraziranje (Bevendorff et al., 2019). Trenutno postojeće metode za prikrivanje autorstva uglavnom nailaze na problem zadržavanja konzistentnosti između rečenica, tj. moguće je detektirati razliku između rečenica koje su stilski izmijenjene uz pomoć sustava za prikrivanje od rečenica koje su ostale neizmijenjene (Mahmood et al., 2020). Poznavanje ove činjenice te izrada sustava za detekciju koji je upoznat s postojanjem metoda za prikrivanje autorstva čine važan aspekt koji se mora u obzir pri ispitivanju sumnjivog dokumenta.

3.1.6. Određivanje smjera plagiranja

Jedan od problema koji se javlja kod detekcije plagijata je **određivanje smjera plagiranja**. Naime, kada postoje dva dokumenta koja sadrže dva identična fragmenta teksta nije uvijek očito koji je dokument original, a koji plagirao iz njega (Bensalem, Rosso, & Chikhi, 2019). Ova pojava može stvarati značajne probleme pri detekciji plagijata bez obzira na pristup koji se koristi. Tako, na primjer, u ekstrinzičnoj detekciji plagijata, ako se plagirani dokument slučajno našao u referentnoj bazi originalnih radova, može doći do toga da sustav pogriješi u označavanju plagiranog rada. Grozea i Popescu (2010) oznake datuma na radovima navode kao jedan od načina za detekciju smjera plagiranja, no pri tome napominju da one nisu uvijek dostupne te da ih je istovremeno veoma lako krivotvoriti. Prema tome čak i ako je datum predaje ili objave rada vidljiv, ne znači da se u njega može imati povjerenja. Iz

ovog razloga potrebno je pronaći novi izvor za izlučivanje informacija o tome koji je rad izvornik, a koji plagiran. U ovu svrhu Grozea i Popescu (2010) koriste računalno detektiran stil pisanja. Oni smatraju da, ako je dostupna dobro izračunata stilometrijska mjera, ona se može koristiti za detekciju smjera plagiranja. Naime, ako stil ostatka teksta u dokumentu u dovoljnoj mjeri odstupa od stila tekstualnog fragmenta koji se ispituje, moguće je zaključiti da je autor dokumenta taj fragment plagirao.

3.2. Intrinzična detekcija u drugim područjima

Metode i pristupi koji se koriste u intrinzičnoj detekciji, osim za detektiranje plagijata, mogu se koristiti u čitavom nizu područja koja se bave pitanjima autorstva poput profiliranja autora na temelju njegovih objava u marketinške svrhe (Rosso, Rangel, Potthast, Stamatatos, Tschuggnall, & Stein, 2016), identificiranja plagiranja i "recikliranja" teksta, odnosno ponovne uporabe dijelova teksta iz starih članaka, radova i sl. (Chong. 2013) te za dokazivanje autorstva u znanstvenim i književnim djelima (Oberreuter, L'Huillier, Rios, & Velasquez, 2011). Intrinzična detekcija plagijata u ovakvim situacijama može biti od izuzetne koristi jer se radovi koji su pripisani jednom autoru mogu stilometrijski usporediti jedni s drugima, a na temelju dobivenih rezultata moguće je dobiti odgovor na to radi li se doista o radovima samo tog jednog autora. U kontekstu književnih djela intrinzična detekcija, nadalje, može se koristiti se za određivanje žanrova literarnih djela (Elahi & Muneer, 2018). Intrinzična detekcija uporabu pronalazi i u forenzičkoj lingvistici, tj. u području zakona i kriminalističkih istraga. U digitalnoj forenzici koristi se za detektiranje i dokazivanje kibernetičkog kriminala (engl. *cyber crime*), autore zlonamjernog programskog koda moguće je otkriti uz pomoć stilometrijskih mjera (Claburn, 2018). Potencijal također postoji u područjima koja se bavi analizom patenata, s obzirom da je detekcija plagiranja ideja jedan od zadatka intrinzičnog pristupa. Vartapetiance i Gillam (2014) navode da je stilometriju moguće koristiti za prepoznavanje obmana (engl. *deception detection*), a ovo područje obuhvaća detekciju laži i prevara u medijima te na internetu. U svom radu *Deception detection: dependable or defective* navode neka od područja kojima se detekcija obmana bavi, a to su prepoznavanje lažnih ocjena proizvoda i usluga na mrežnim stranicama, predatora na društvenim mrežama, lažnih profila, detekcije neželjene pošte te ostalih obmana i prevara u digitalnom okruženju.

4. Stilometrija

Stilometrija, najjednostavnije rečeno, podrazumijeva kvantifikaciju stila (Meyer zu Eissen et al., 2006). Kao što je već ranije spomenuto, svaki individualni autor kroz život razvija svoj jedinstveni stil pisanja koji se odlikuje karakteristikama koje ga razlikuju od stilova svih drugih autora. Neke od tih značajki mogu biti: duljina rečenica, raspored riječi u rečenici, uporaba interpunkcijskih znakova te individualni vokabular koji uključuje veznike, čestice, usklike, strane riječi i sl. (Meyer zu Eissen et al., 2006). Stilometrijsku mjeru moguće je izračunati na temelju navedenih karakteristika. S obzirom na činjenicu da se intrinzična detekcija zasniva na prepoznavanju promjena u stilu pisanja unutar rečenica i odlomaka u tekstu, stilometrija čini izuzetno važan i nezaobilazan aspekt čitavog procesa.

Zbog velike količine stilističkih svojstava koje je moguće izlučiti iz teksta javlja se potreba za njihovim klasificiranjem. Stilometrijska svojstva moguće je podijeliti na više načina, a svaka od kategorija uglavnom se bavi jednom razinom teksta, od znakova do rečenica, pa sve do semantike čitavog djela. Meyer zu Eissen et al. (2006) smatraju da se stilometrijska svojstva temelje na semiotičkim, tj. znakovnim značajkama teksta te navode sljedeću podjelu:

1. **statistika teksta:** bavi se tekstom na razini znakova, a uključuje svojstva poput broja interpunkcijskih znakova i duljinu riječi unutar dokumenta
2. **sintaktička svojstva:** odnose se na obradu teksta na razini rečenica, a uključuju duljinu rečenica i uporabu funkcijskih riječi (veznici, prijedlozi, članovi, itd.)
3. ***Part-of-Speech* svojstva:** koja se koriste za kvantifikaciju vrsta riječi koje se koriste u tekstu, a može se odnositi na broj pridjeva ili zamjenica
4. **setovi zatvorenih klasa riječi (engl. *closed-class word sets*):** koji se koriste za izračun broja riječi koje se smatraju specifičnima zbog njihove složenosti, a također se može raditi o stranim ili stop riječima (engl. *stop-words*)
5. **strukturalna svojstva:** koja se koriste za prikaz organizacije teksta, a odnose se na duljinu odlomaka i poglavlja.

Stamatatos (2009) stilometrijska svojstva teksta klasificira na nešto drugačiji način te u svom radu razlikuje:

1. **leksička svojstva (engl. *lexical features*):** koja se odnose na frekvenciju riječi, *n*-grame riječi, bogatstvo rječnika i sl.

2. **znakovna svojstva (engl. *character features*):** koja uključuju vrste znakova, znakovne *n*-grame itd.
3. **sintaktička svojstva (engl. *syntactic features*):** koja se odnose na *Part-of-Speech* frekvencije, vrste fraza itd.
4. **semantička svojstva (engl. *semantic features*):** koja se odnose na sinonime, semantičke zavisnosti unutar teksta i sl.
5. **svojstva specifična za određenu primjenu (engl. *application-specific features*):** koja podrazumijevaju strukturna svojstva te svojstva specifična za određeni sadržaj i jezik.

Ova svojstva u stilometrijskom postupku prvo se izlučuju na globalnoj razini, tj. razini čitavog dokumenta, a potom i na lokalnoj razini koja se može odnositi na rečenice ili pak odlomke unutar teksta (Seaward & Matwin, 2009). Meyer zu Eissen et al. (2006) navode da je na temelju stilometrijskih svojstava moguće izvesti formule za kvantifikaciju stila. Pri tome ističu da se formule za kvantifikaciju mogu podijeliti s obzirom na to radi li se o formulama koje se odnose na samog autora i njegov tekst ili pak čitatelja tog teksta. Gotovo sve formule usmjerene su na kvantifikaciju razine školovanja koja je potrebna za pisanje i čitanje određenih tekstova. Meyer zu Eissen et al. (2006) formule specifične za autora teksta klasificiraju na one koje se koriste za izračunavanje bogatstva rječnika, a uključuju *Honore's R, Yules's K* i *Sichel's S* formulu te na formule za izračunavanje složenosti i razumljivosti teksta, koje pak uključuju *Flesch Reading Ease, Miyazaki Readability Index* te *Passive Sentences Readability Score* formule. Kao formule usmjerene na čitateljevu sposobnost razumijevanja teksta navode: *Dale-Chall* formulu, *Powers-Sumner-Kearl Grade, Bormuth Grade Level, Flesch-Kincaid Grade Level, Gunning Fog Index, Coleman-Liau Grade Level, Fry Readability Formula* te *The Army's Automated Readability Index (ARI)*. Ove formule veoma su korisne u intrinzičnoj detekciji plagijata s obzirom na to da promjene u broju pasivnih rečenica, leksičkoj raznolikosti, čitljivosti teksta, duljini rečenica u riječima te riječi u slogovima mogu biti indikator plagiranja.

4.1. Formule za kvantifikaciju stila koje se odnose na autora

Honore's R (Honore, 1979) i ***Yule's K*** (Yule, 1944) formule, kao što je ranije istaknuto, koriste se za određivanje mjere bogatstva rječnika, a rade na način da organiziraju riječi u rječnike prema bogatstvu i složenosti te potom računaju frekvenciju tih riječi unutar

teksta koji se ispituje (Alsallal, Iqbal, Amin, & James, 2013). Nedostatak ove dvije formule je to što točnost njihovog rezultata uvelike ovisi o duljini dokumenta ili odlomka koji se ispituje (Meyer zu Eissen et al., 2006).

Sichel's S formula (Sichel, 1986), za razliku od njih, u tekstu uspoređuje TTR (engl. *Type-Token Ratio*) (Williamson, 2014), koji se koristi za izračunavanje varijacije vokabulara u tekstu, sa svim tokenima, tj. riječima unutar teksta ili odlomka koji se ispituje, pri čemu formula za izračun TTR glasi:

TTR = (broj tipova/broj tokena) * 100

Pri tome TTR označava već ranije spomenut *Type-Token Ratio*, broj tipova označava riječi koje su u tekstu pojavljuju samo jedanput, a broj tokena označava broj svih riječi u tekstu. Ova formula na taj način zapravo izračunava leksičku raznolikost teksta (Williamson, 2014). Sam je pristup matematički veoma složen te ostvaruje puno bolje rezultate na tekstu s malim brojem tokena, a najbolje radi na tekstovima koji sadrže do nekoliko tisuća tokena (McKee, Malvern, & Richards, 2000).

Flesch Reading Ease razvio je R. Flesch 1948. godine (Readability Formulas). Cilj njegovog pristupa bila je izrada jednostavnije formule za izračunavanje razumljivosti teksta koja za izračun mjere čitljivosti koristi samo dvije varijable: 1) prosječnu dužinu rečenica s obzirom na broj riječi te 2) prosječan broj slogova u riječi (van de Rakt, 2019). Sama formula glasi (Readability Formulas):

RE = 206.835 – (1.015 x ASL) – (84.6 x ASW)

Pri čemu RE označava lakoću čitljivosti, ASL prosječan broj riječi u rečenici, a ASW prosječan broj slogova po riječi. Na temelju ovih varijabli tekst se ocjenjuje ocjenom od 0 do 100 pri čemu ocjena 100 označava lako čitljive tekstove, a 0 označava izuzetno teško čitljive tekstove (van de Rakt, 2019). *Flesch Reading Ease* prvenstveno se koristi za ocjenjivanje čitljivosti tekstova namijenjenih djeci školske dobi (Readability Formulas).

Miyazaki's Readability Index 1999. godine razvio je J. Greenfield, a za izračun indeksa čitljivosti promatra se broj znakova po riječi te broj riječi u rečenicama (Greenfield, 1999). Kao i prethodna formula ovaj pristup također koristi sustav ocjenjivanja od 1 do 100 (Greenfield, 2003). Mjera koja se dobiva pomoću *Miyazaki's Readability Indexa* prilagođena je čitateljima kojima je engleski nije materinji jezik (Crossley, Greenfield, & McNamara, 2008).

Passive Sentences Readability Score razinu čitljivosti izračunava usporedbom broja pasivnih i aktivnih rečenica unutar teksta pri čemu se pasivne rečenice u engleskom jeziku definiraju kao rečenice koje sadrže pasivni oblik glagola "biti" koji je popraćen glagolom i prošlim glagolskim oblikom (RFP-Templates). Rezultat se potom prikazuje kao postotak pasivnih rečenica u tekstu pri čemu niži postotak označava lakše čitljiv tekst (RFP-Templates).

4.2. Formule za kvantifikaciju stila koje se odnose na čitatelja

Dale-Chall Formula osmislili su E. Dale i J. Chall 1948. godine, a radi se o jednoj od formula koje se koriste za izračun čitateljeve sposobnosti razumijevanja teksta. Ova formula temelji se na *Flesch Reading Ease* formuli, no prilagođena je učenicima od četvrtog razreda osnovne škole nadalje te odraslim osobama (Readability Formulas). Dale i Chall (1948) formulu definiraju na sljedeći način:

DC = 0.1579 * (složene riječi/riječi * 100) + 0.0496 * (riječi/rečenice)

Dale-Chall formula svojevremeno bila je jedinstvena jer, za razliku od ostalih pristupa, umjesto korištenja duljine riječi za određivanje čitljivosti i složenosti teksta koristi popis riječi koje su poznate većini učenika u četvrtom razredu osnovne škole te se često javljaju u tekstovima namijenjenima njihovoj dobi (Dale & Chall, 1948). Ovaj je popis izvorno sadržavao 763 riječi, a kasnije je proširen na 3000 riječi (Chall & Dale, 1995).

Power-Sumner-Kearl Readability Formula je formula temeljena na *Gunning Fog Indexu*, a predstavili su je 1958. godine R. Powers, W. A. Sumner i B. E. Kearl. Radi se o još jednoj formuli usmjerenoj na izračunavanje čitljivosti tekstova namijenjenih djeci i to u dobi od sedam do deset godina (Powers, Dunmer, & Kearl, 1958), a svojstva koja uzima u obzir su broj riječi, prosječna duljina rečenica te broj slogova u tekstu (Khan, 2020). *Power-Sumner-Kearl Readability Formula* sastoji se od dva dijela, a prvi glasi (Readability Formulas):

GL = 0.0778(ASL) + 0.0455(NS) – 2.2029

Pri čemu GL označava razinu školovanja u Sjedinjenim Američkim Državama. ASL prosječnu duljinu rečenica u riječima, a NS broj svih slogova u segmentu teksta koji se ispituje podijeljen s brojem svih riječi u segmentu. Drugi dio formule glasi (Readability Formulas):

RA = 0.0778(ASL) + 0.0455(NS) + 2.7971

U ovoj formuli RA označava dob kojoj je prilagođena težina čitanja teksta, ASL, kao i u prethodnoj formuli označava prosječnu duljinu rečenice, dok NS ponovno označava prosječan broj suglasnika. Važno je napomenuti da ova formula ostvaruje relativno loše rezultate u literaturi koja je namijenjena za djecu iznad 10 godine (Khan, 2020).

Bormuth Grade Level osmislio je je J. R. Bormuth 1966. godine, a njegov pristup za razliku od prethodno navedene formule kao svojstvo ne promatra broj slogova i riječi u rečenici već broj znakova u tekstu (Bormuth, 1966). On u svom radu tvrdi da promjene u broju varijabli unutar teksta te vokabular i dužina rečenica mogu imati utjecaj na razumijevanje teksta. Njegova formula usmjerena je na dokumente akademske prirode kao što su npr. školski udžbenici, a sama formula glasi (RFP-Templates):

BGL = 0.886593 - (AWL x 0.03640) + (AFW x 0.161911) - (ASL x 0.21401) -

- (ASL x 0.000577) - (ASL x 0.000005)

U formuli AWL označava prosječnu duljinu riječi u znakovima, AFW broj jednostavnih riječi koje se javljaju u tekstu prema Dale-Chall listi od 3000 jednostavnih riječi podijeljen s brojem svih riječi u tekstu, a ASL označava prosječan broj riječi u rečenici.

Flesch-Kincaid Readability Grade Level prvi je razvio R. Flesch 1940-ih, a za potrebe Američke vojske. Njegovu su formulu doradili J. Peter Kincaid i njegov tim 1975. godine na temelju rezultata testova provedenih na osoblju u vojnoj službi (Kincaid, Fishburne, Rogers, & Chisson, 1975). Ova formula prvenstveno se bavi tekstovima namijenjenim djeci u sedmom i osmom razredu osnovne škole (RFP-Templates). Temelj ovog pristupa je broj slogova i broj riječi u rečenici, a njegov rezultat je ocjena od 1 do 18 koja korelira razini edukacije u Sjedinjenim Američkim Državama koja je potrebna za razumijevanje teksta koji se ispituje (Kelly, 2017). Formula ovog pristupa glasi (RFP-Templates):

FKRS = (0.39 x ASL) + (11.8 x ASW) - 15.59

Pri čemu ASL označava prosječnu duljinu rečenice u riječima, a ASW prosječan broj slogova u riječima. Ova formula koristi se kao standardizirani test Ministarstva obrane američke vlade (RFP-Templates).

Gunning Fog Readability poznatu još i pod nazivom *Fog Index* (FOG) osmislio je R. Gunning 1952. godine, a radi se o prvoj formuli koja nije bila prvenstveno usmjerena na pitanja vezana uz školstvo (Postscripts). Gunningov (1952) rad bio je usmjeren na novinske

članke za koje je smatrao da su prepuni "magle" (engl. *fog*), tj. da je njihov tekst nepotrebno složen i nerazumljiv. Iz ovog razloga on predstavlja formulu koja je bila mnogo jednostavnija od formula koje su joj prethodile te za izračun mjere čitljivosti koristi samo dva svojstva teksta, a to su prosječan broj riječi u rečenici te postotak "teških" odnosno složenih riječi. Sama formula glasi (Readability Formulas):

$$GL = 0.4\ (ASL + PHW)$$

Pri čemu ASL označava prosječnu duljinu rečenice prema broju riječi, a PHW predstavlja postotak složenih riječi u tekstu. Ocjene koje se dobivaju ovom formulom slično kao i druge formule odnose se na razinu edukacije koja je potrebna za razumijevanje teksta, a ocjena može iznositi od 5 do 17 ("Robert Gunning's Fog Readability Formula", 2004).

Coleman-Liau Readability Score osmislili su M. Coleman i T. L. Liau 1975. godine kao kritiku na dotadašnje pristupe koji su čitljivost teksta određivali na temelju broja slogova (Kelly, 2017). Oni umjesto broja slogova za određivanje čitljivosti teksta koriste broj znakova u riječima te rečenicama unutar teksta, a sama formula za izračun čitljivosti (CLGL) glasi (RFP-Templates):

$$CLGL = (5.89 \times (AWL / ASL)) - (((3 \times ANS) / 1000) \times ASL) - 15.8$$

Pri tome; AWL označava prosječnu dužinu riječi u znakovima, ASL označava prosječnu dužinu rečenica prema broju riječi, a ANS označava prosječan broj rečenica.

Fry Readability Formula, poznata još i kao *The Fry Graph* nastala je 1960-ih, a njen je autor E. Fry (Khan, 2020). Fry (1977) navodi da je za izračun čitljivosti pomoću ove formule najprije potrebno, iz tri nasumična segmenta teksta od 100 riječi, izlučiti broj slogova u svakoj od riječi te broj rečenica, a potom odrediti njihovu prosječnu vrijednost. Dobivene vrijednosti se zatim upisuju na graf na kojem se iz presjecišta prosječnog broja slogova te prosječnog broja rečenica može razlučiti razina čitljivosti teksta koji se ispituje. Ova formula najčešće se koristi za određivanje konsenzusa čitljivosti tekstova regulatornih namjena, kao što su tekstovi vezani uz zdravstvo, npr. upute o lijeku, kako bi se osiguralo da ih šira populacija može razumjeti s lakoćom (Readability Formulas). Ova formula općenito je popularan izbor za izračunavanje čitljivosti zbog svoje jednostavnosti te točnosti njenih rezultata (Khan, 2020).

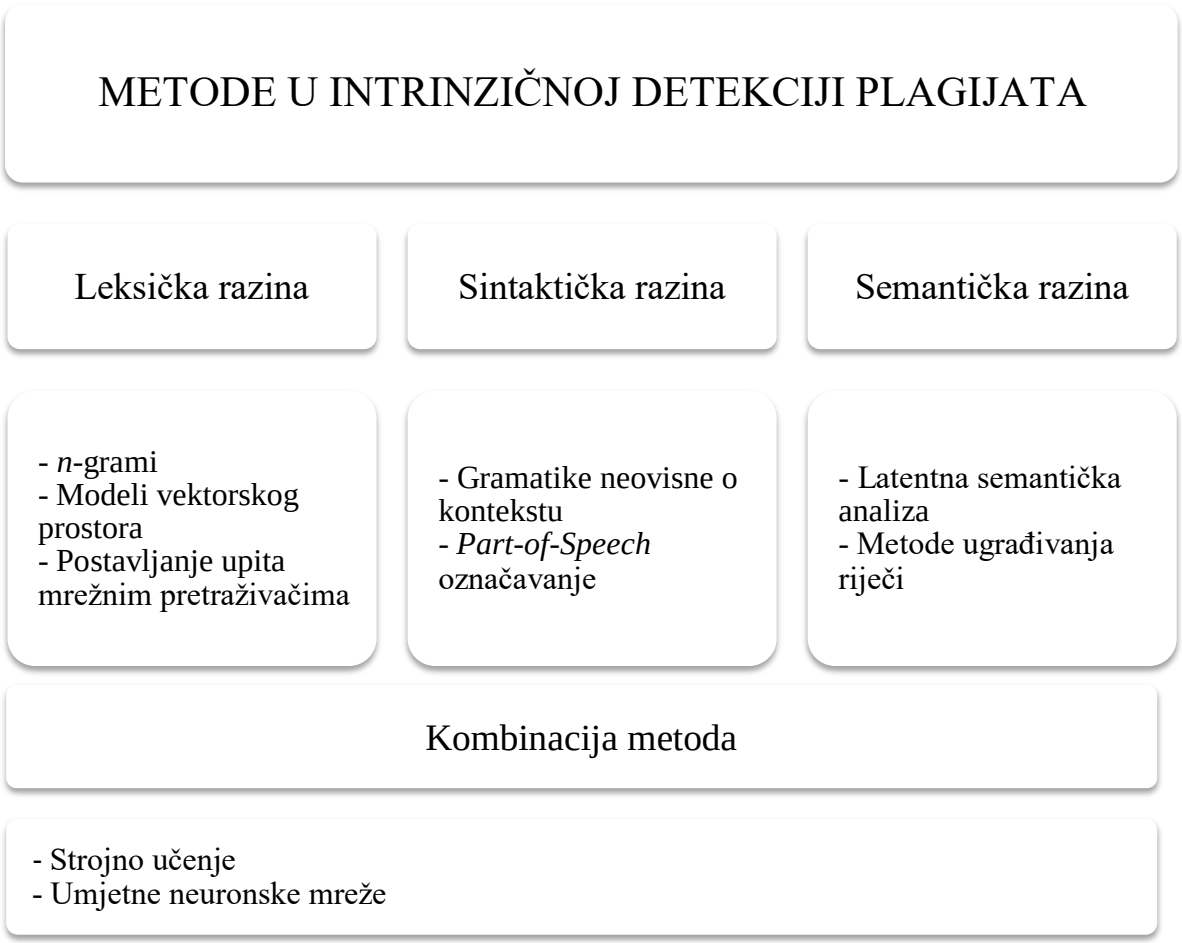
Automated Readability Index (ARI) izradili su E. A. Smith i R. J. Senter 1967. godine za potrebe Američke vojske (Cantos-Gómez & Almela, 2019). Za izračun indeksa čitljivosti korištena je modificirana električna pisaća mašina kojom se brojala prosječna duljina riječi i

rečenica (Kincaid & Delionbach, 1973). Njen rezultat korelira s razinama školovanja te dobi čitatelja koja je potrebna za razumijevanje teksta, a može iznositi od 1 do 12 (Readability Formulas). Formula za izračun ovog indeksa glasi (Cantos-Gómez & Almela, 2019):

ARI = 4.71 * (znakovi/riječi) + 0.5 * (riječi/rečenice) - 21.43

5. Suvremene metode u intrinzičnoj detekciji plagijata

Stilometrijska analiza u svojim počecima uglavnom se izvodila ručno, no u novije vrijeme razvoj tehnologije i informacijskih znanosti omogućili su izračunavanje stilometrijske mjere uz pomoć računala. Osim što je analiza na taj način postala brža i točnija, postalo je moguće i analizirati puno veće količine informacije. Informacijske tehnologije te područja poput obrade prirodnog jezika, strojnog učenja te umjetnih neuronskih mreža danas su od osobite važnosti za metode koje se koriste u intrinzičnoj detekciji plagijata. Suvremene metode i pristupi prikazani su na Slici 3. i uglavnom se klasificiraju prema razini jezika na kojoj analiziraju tekstove za koje se sumnja na plagiranje, a radi se o metodama temeljenim na **leksičkim**, **sintaktičkim** i **semantičkim** svojstvima. Sustavi za intrinzičnu detekciju plagijata često koriste kombinaciju ovih metode u svrhu postizanja boljih rezultata pri detekciji (Foltýnek et al., 2019).



Slika 3. Metode u intrinzičnoj detekciji plagijata

5.1. Metode temeljene na leksičkoj analizi

Metode koje plagiranje detektiraju na temelju leksičkih svojstava teksta dokumente analiziraju na razini riječi i znakova, a u intrinzičnoj detekciji plagijata gotovo isključivo rade na razini znakova (Foltýnek et al., 2019). U nekim od metoda koje se koriste, prije provođenja analize nad leksičkim svojstvima sumnjivog dokumenta, najprije je potrebno provesti predobradu teksta koja može uključivati: tokenizaciju, korjenovanje riječi, odstranjivanje interpunkcijskih znakova te preoblikovanje čitavog teksta u mala slova (Chong & Specia, 2011). Tehnike koje se koriste u metodama temeljenim na leksičkim svojstvima teksta mogu se podijeliti na (Foltýnek et al., 2019): 1) **metode temeljene na *n*-gramima**, 2) **modele vektorskog prostora** (engl. *vector space models*) te 3) **metode koje uključuju postavljanje upita mrežnim pretraživačima**.

Koncept ***n*-grama** potječe iz područja obrade prirodnog jezika, a podrazumijeva niz koji se sastoji od *n* broja riječi ili znakova (Kumar, 2017). Bitne odrednice metoda temeljenih na *n*-gramima su njihova neovisnost o jeziku i domeni teksta, uz to na tekstu prije analize nije potrebno provesti predobradu te su robusne na šumove koji se u plagiranom tekstu mogu pojaviti u obliku parafraziranja i zamjene riječi sinonimima (Stamatatos, 2009). Algoritmi za detekciju plagijata koji se temelje na znakovnim *n*-gramima spadaju pod najuspješnije metode u intrinzičnoj detekciji plagijata, a koriste se u zadacima pripisivanja i verifikacije autorstva (Kuta & Kitowski, 2014). Ove metode najbolje rezultate ostvaruju u slučajevima kada plagijator tekst preuzima izravno iz izvora, a relativno su uspješne i u detekciji skrivenog autorstva te plagiranja u kojem se zadržava izvorna sintaksna struktura (Foltýnek et al., 2019). Metode temeljene na *n*-gramima u intrinzičnoj detekciji plagijata prvi su koristili Graham, Hirst i Marthi (2005) za prepoznavanje stilističke nedosljednosti između odlomaka unutar teksta. U svrhu detekcije koristili su bigrame znakova, a za izračunavanje različitosti između odlomaka koristili su kosinusnu udaljenost. U svom radu zaključili su da metode temeljene na *n*-gramima najbolje rezultate ostvaruju u dugačkim tekstovima dok u tekstovima manjeg opsega ostvaruju znatno lošije rezultate. Kao odgovor na ovaj problem Kuta i Kitowski (2014) predlažu nove parametre za optimizaciju postupka. Na temelju svog istraživanja zaključili su da je za ostvarivanje boljih rezultata potrebno koristiti *n*-grame u kojima *n* iznosi 4 ili 5, za razliku od metoda u drugim radovima koje uglavnom koriste bigrame i trigrame. Također, smatraju da je potrebno povećati broj znakova pri definiranju veličine odsječaka, a uključivanje *n*-gramskih svojstava promjenjive duljine nadalje

poboljšava rezultate dobivene ovom metodom, osobito ako se radi o metodama koje koriste odsječke teksta manjeg opsega.

Modeli vektorskog prostora spadaju u još jednu od metoda koje se temelje na leksičkim svojstvima teksta. U intrinzičnoj detekciji plagijata koriste se za detekciju proboja stila, klastering autora te verifikaciju autorstva (Foltýnek et al., 2019). Model vektorskog prostora tekst prikazuje u *n*-dimenzionalnom prostoru, a broj dimenzija ovisi o zadanom broju svojstava koja se ispituju unutar teksta (Towards Data Science). Radi se o algebarskom modelu koji se sastoji od dva koraka (Data Science Central):

1. prikazivanje teksta kao vektora riječi
2. transformacija vektora u brojčani format koji je računalno čitljiv te se na njega mogu primijeniti tehnike za rudarenje podataka.

U prvom koraku tekst se najprije mora rastaviti na riječi na kojima se vrši predobrada, a uključuje uklanjanje stop riječi, interpunkcijskih znakova, posebnih znakova, itd. Riječi se potom prikazuju kao vektori. Vektori se nakon toga upisuju u matricu pojmova u dokumentu (engl. *term document matrix*), a pojmovima je pri tome potrebno odrediti težinu (engl. *weight*) uz pomoć tf-idf mjere kako bi se odredila njihova relevantnost unutar dokumenta. Sličnost između dva vektora u ovom postupku računa se uz pomoć kosinusne sličnosti. Neki od autora koji metode vektorskog prostora u svrhu intrinzične detekcije plagijata u svojim radovima koriste su Castro, Adame, Pelaez i Muñoz (2015), Karaś, Śpiewak i Sobecki (2017), Kocher (2016) te Oberreuter i Velásquez (2013). Modeli vektorskog prostora, osim u leksičkim metodama, koriste se i u metodama koje se temelje na semantičkoj analizi teksta (Foltýnek et al., 2019).

Iako se **postavljanje upita mrežnim pretraživačima** često koristi kao jedan od koraka u ekstrinzičnoj detekciji plagijata za izlučivanje kandidata, odnosno dokumenata koji su potencijalni izvori plagiranja (Kanjirangat & Gupta, 2016), ova tehnika uporabu pronalazi i u intrinzičnoj detekciji plagijata. U intrinzičnoj detekciji mrežni pretraživači mogu se koristiti u metodama za detektiranje "općeg varalice" (engl. *General Impostors Method*) (Foltýnek et al., 2019). Foltýnek et al. (2019) navode da je u ovoj tehnici najprije potrebno iz sumnjivog dokumenta izlučiti ključne riječi uz pomoć kojih se pretražuju mrežne stranice u potrazi za radovima slične tematike. U pronađenim radovima se potom kvantificira stil pisanja koji se zatim uspoređuje s stilom pisanja u dokumentu koji se ispituje na plagiranje. Ova metoda omogućava razlikovanje stilometrijskih mjera koje su karakteristične za

određenu temu od onih koje su karakteristične za autora čije se djelo ispituje što olakšava detekciju promjena u stilu pisanja unutar sumnjivog dokumenta (Khonji & Iraqi, 2014). Neki od autora koji se bave ovim metodama su Seidman (2013), Moreau, Jayapal, Lynch i Vogel (2015) te Koppel i Winter (2014).

5.2. Metode temeljene na sintaktičkoj analizi

Metode temeljene na sintaktičkoj analizi podrazumijevaju algoritme koji rečenice rastavljaju na tokene te analiziraju strukturu svake riječi unutar rečenice s obzirom na njen položaj u samoj rečenici (Adam & Suharjito, 2014). Sintaktička svojstva koja se obično uzimaju u obzir pri analizi uključuju duljinu rečenice (Yule, 1944), funkcijske riječi (Burrows, 2002) te raspodjelu *Part-of-Speech* (*PoS*) oznaka i određenih klasa riječi (Stein and Meyer zu Eissen, 2007). U intrinzičnoj detekciji plagijata sintaktička analiza koristi se za izravnu usporedbu dokumenata, izračun frekvencije *Part-of-Speech* oznaka, frekvencije *Part-of-Speech* oznaka za *n*-grame te prikaz integriranih sintaktičkih grafova (Foltýnek et al., 2019). Metode koje se koriste uključuju **gramatike neovisne o kontekstu** (engl. *Context-Free Grammar, CFG*) te ***Part-of-Speech* označivanje.**

Gramatike neovisne o kontekstu spadaju u područje obrade prirodnog jezika, a definiraju se kao popis pravila koji obuhvaćaju sve gramatički točno složene rečenice nekog jezika pri čemu svako pravilo ima lijevu stranu koja određuje sintaktičku kategoriju te desnu stranu koja definira alternativne komponente (Bowdoin). Ove gramatike rade na način da raščlanjuju (engl. *parsing*) rečenice unutar teksta na tokene uz pomoć kojih se potom analizira struktura svake riječi s obzirom na njen položaj unutar rečenice (Adam & Suharjito, 2014). Rezultat koji se dobiva ovim procesom prikazuje se u obliku parsemskog stabla (engl. *parse tree*) (Hrvatski mrežni rječnik). Parsemska stabla nadalje omogućavaju vizualni prikaz raščlanjenih rečenica. Mjera sličnosti između riječi u ovim metodama računa se uz pomoć sintaktičkih stabala zavisnosti (engl. *syntactic dependency trees*) (Adam & Suharjito, 2014). Ovaj pristup u svom radu koriste Tschuggnall i Specht (2013), a njihov sustav plagiranje prepoznaje na temelju nedosljednosti u sintaktičkoj strukturi rečenica koje se prepoznaju uz pomoć Gaussove armature normalne raspodjele (engl. *Gaussian normal distribution fitting*). Pri tome primjećuju da je ova metoda najprikladnija za kratke dokumente koji sadrže do 300 rečenica, a najbolje rezultate postiže u dokumentima s 50 ili manje rečenica.

***Part-of-Speech* oznake** također se koriste za određivanje sintaktičke strukture rečenica u intrinzičnoj detekciji plagijata (Gómez-Adorno, Sidorov, Pinto, & Markov, 2015) Ovaj pristup primjenjuje iste metode kao i gramatike nezavisne o kontekstu, tj. rečenice se rastavljaju na tokene, a tokenima potom pridodaju *PoS* oznake koje označavaju vrste riječi (Adam & Suharjito, 2014). U intrinzičnoj detekciji *PoS* oznake često se koriste kao jedno od stilometrijskih svojstava (Foltýnek et al., 2019). Mjera sličnosti između riječi se računa uz pomoć sljedeće formule (Alzahrani et. al, 2012):

Sličnost = broj uparenih riječi s identičnim POS oznakama / broj uparenih riječi

Nedostatak ovih metoda je nemogućnost detektiranja plagiranja parafraziranjem zbog promjena u strukturi same rečenice koje se događaju kada plagijator mijenja redoslijed riječi u rečenici ili pak zamjenjuje neke od riječi sinonimima ili drugim sličnim riječima (Adam & Suharjito, 2014). Iz razloga što metode temeljene na *PoS* označavanju ne ostvaruju dobre rezultate u složenijim oblicima plagiranja te ih je iz ovog razloga sintaktičke metode potrebno kombinirati sa metodama koje se temelje na semantičkoj analizom (Adam & Suharjito, 2014). Metode temeljene na *PoS* oznakama u svojim radovima koriste Maitra, Ghosh i Das (2016) te Afroz, Islam, Stolerman, Greenstadt i McCoy (2014).

5.3. Metode temeljene na semantičkoj analizi

Nije rijetkost da osobe koje plagiraju pri preuzimanju tuđeg teksta izmjenjuju riječi, fraze i čitave rečenice s ciljem prikrivanja plagiranja (Yousf, Ahmad, & Nasurllah, 2013). S obzirom da prethodno navedene leksičke i sintaktičke metode ne postižu zadovoljavajuće rezultate u detekciji ovakvih složenih oblika plagiranja javlja se potreba za razvijanjem metoda koje se temelje na analizi značenja, tj. semantike teksta (Adam & Suharjito, 2014). U metodama koje se temelje na semantičkoj analizi čitav postupak zasniva se na pretpostavci da semantička sličnost dva odlomka ovisi o postojanju semantički sličnih jedinica koje se unutar tih odlomaka javljaju u identičnom ili gotovo identičnom kontekstu (Foltýnek et al., 2019). Osim na razini odlomaka, semantičku analizu moguće je provesti i na razini rečenica te čitavih dokumenata. Metode koje se koriste za semantičku analizu u intrinzičnoj detekciji plagijata uključuju **latentnu semantičku analizu** (engl. *Latent Semantic Analysis, LSA*) te **metode ugrađivanja riječi** (engl. *Word Embedding*).

Latentna semantička analiza koristi se za izračunavanje sličnosti između riječi usporedbom kosinusne vrijednosti koja je dobivena usporedbom vektora tih riječi, a što je

njena vrijednost manja, to su riječi sličnije (Adam & Suharjito, 2014). Adam i Suharjito (2014) navode da algoritmi za semantičku analizu teksta dokument najprije moraju analizirati, a potom se na temelju provedene analize izrađuje se *n*-dimenzionalna matrica u kojoj svaki redak predstavlja jednu riječ, a stupci mogu predstavljati dokument, odlomak, rečenicu, itd. Nakon ovog koraka slijedi izrada vektora temeljenog na lingvističkoj mjeri dobivenoj iz slične riječi koja se uspoređuje s riječi koja se trenutno ispituje. Matricu izrađenu u prvom koraku zatim je potrebno rastaviti uz pomoć dekompozicije singularnog vektora (engl. *Singular Vector Decomposition, SDV*) ili sličnom tehnikom za redukciju dimenzionalnosti matrice koje odstranjuju manje relevantne riječi te tako smanjuju broj redaka u matrici, a pri tome zadržavaju distribuciju sličnosti između stupaca (Foltýnek et al., 2019) što nadalje omogućava izračun semantičke sličnosti između riječi koje se smatra najreprezentativnijima za dokument koji se ispituje.

Metode ugrađivanja riječi bave se "ugrađivanjem" značenja koje je izraženo u tekstualnom kontekstu dokumenta (Nugaliyadde, Wong, Sohel, & Xie, 2019). Umjesto provođenja analize nad samim riječima koje se ispituju, analizu isključivo provode nad riječima koje ih okružuju (Foltýnek et al., 2019). Pri tome se biraju riječi koje se nalaze u blizini pojma koji se ispituje zbog pretpostavke da su one relevantnije za kontekst od riječi koje su udaljenije. Naime, dva pojma koja se unutar teksta često javljaju u blizini jedan drugome također će se javljati bliže i u vektorskom prostoru koji se koristi za određivanje sličnosti (Foltýnek & Glendinning, 2015). Ovaj pristup u intrinzičnoj detekciji plagijata koristi se za identificiranje parafraziranja, detekciju promjena u stilu pisanja te klastering autora (Foltýnek et al., 2019).

5.4. Kombiniranje metoda za detekciju

S obzirom na to da svaka od prethodno navedenih metoda ima svoje prednosti i nedostatke te da autorov stil pisanja sadrži obilježja stilometrijskih svojstava koja je moguće opisati s leksičke, sintaktičke i semantičke razine, u sustavima za intrinzičnu detekciju plagijata gotovo se uvijek koristi kombinacija više različitih metoda (Foltýnek et al., 2019). **Strojno učenje** te **umjetne neuronske mreže** spadaju u neke od ovakvih sustava.

Strojno učenje u intrinzičnoj detekciji plagijata često se koristi kao pomagalo pri odabiru svojstava koja najbolje opisuju autorov stil pisanja (Stamatatos, Daelemans, Verhoeven, Juola, López-López, Potthast,& Stein, 2015). Pristupi temeljeni na strojnom

učenju ostvaruju izuzetno dobre rezultate (Foltýnek et al., 2019) te predstavljaju obećavajuće područje u intrinzičnoj detekciji plagijata. Sustave za strojno učenje možemo podijeliti na: učenje s nadzorom (engl. *supervised machine learning*), učenje bez nadzora (engl. *unsupervised machine learning*) te ojačano učenje (engl. *reinforcement learning*).

Učenje s nadzorom podrazumijeva metode koje sustavu za strojno učenje omogućavaju pristup podacima iz uzorka koji se obrađuje te podatke o ishodu do kojeg se želi doći (PioneerLabs). Pri tome, podaci koji se unose u sustav moraju biti označeni (engl. *labeled*) kako bi algoritam kojeg sustav koristi prilikom učenja mogao doći do zaključaka o odnosima između svojstava koja se ispituju te njihovih oznaka, a ti dobiveni odnosi između njih nazivaju se modelima (GoogleDevelopers). Cilj ovog pristupa je generirati modele koji će moći donositi točne zaključke o novim podacima s kojima sustav nije bio upoznat u procesu učenja (Towards Data Science). Pristupe u učenju s nadzorom dodatno se može podijeliti na klasifikaciju i regresiju, a neki od algoritama koji se koriste uključuju (PioneerLabs): linearnu regresiju (engl. *linear regression*), naivne Bayesove metode (engl. *Naïve Bayes*), stabla odlučivanja (engl. *decision trees*) te metode potpornih vektora (engl. *support vector machines*). Metode potpornih vektora najpoznatiji su pristup temeljen na strojnom učenju u intrinzičnoj detekciji plagijata, a za verifikaciju autora koriste se u procesu poznatom pod nazivom razotkrivanje (engl. *unmasking*) (Foltýnek et al., 2019). Ova tehnika temelji se na klasifikatoru koji se trenira raspoznati razlike u stilu pisanja između sumnjivog dokumenta te seta dokumenata čiji je autor poznat. Klasifikator plagiranje unutar sumnjivog dokumenta detektira odstranjivanjem najznačajnih stilističkih svojstava nakon čega ponovno provodi klasifikaciju. Ukoliko, nakon toga, njegova preciznost značajno padne, sustav zaključuje da je autor sumnjivog dokumenta jednak autoru referentnog seta dokumenata. U suprotnom slučaju zaključuje da imaju različite autore (Foltýnek et al., 2019). Neki od autora koji koriste ovu metodu za intrinzičnu detekciju plagijata su Feng i Hirst (2013), Hürlimann, Weck, Berg, Šuster i Nissim (2015) te Koppel i Winter (2014).

Učenje bez nadzora, za razliku od učenja s nadzorom, sustavu za strojno učenje ne osigurava pristup označenim setovima podataka za treniranje algoritma (GoogleDevelopers). Umjesto toga, sustav sam na temelju neoznačenih podataka donosi zaključke o njihovim međuodnosima te prepoznaje uzorke koji često nisu očiti ljudskom ispitivaču (PioneerLabs). Učenje bez nadzora koristi se za detekciju anomalija te klastering (PioneerLabs) pa se tako u intrinzičnoj detekciji može koristiti za detekciju promjena u stilu pisanja i identifikaciju autora. Algoritmi koji se koriste u ovom obliku strojnog učenja uključuju (PioneerLabs):

hijerarhijski klastering (engl. *hierarchical clustering*), skriveni Markovljev model (engl. *Hidden Markov model*), samoorganizirajuće mape (engl. *self-organising maps*) te miješane gausovske modele (engl. *Gaussian mixture models*) . Jedan od sustava za leksičku detekciju plagijata koji koristi strojno učenje bez nadzora u svom radu predstavio je Khan (2017). Ovaj sustav koristi klasifikacijski pristup bez nadzora za detekciju i označavanje odsječaka teksta u kojima se javljaju značajne promjene u stilu pisanja. Sustav pri tome koristi kombinaciju leksičkih, sintaktičkih te svojstava specifičnih za određeni sadržaj. S obzirom da sustav ima sposobnost detekcije promjena u stilu na razini svake pojedinačne rečenice, ovaj pristup izuzetno je koristan u situacijama kada broj autora unutar dokumenta nije poznat te kada ispitivač nije siguran gdje se nalaze granice između odsječaka teksta koji imaju različitog autora. Pristup strojnom učenju bez nadzora u svom radu također predstavljaju Ranatuna, Atukorale i Hewagamage (2011). Oni za detekciju plagijata koriste Kohonenove samoorganizirajuće mape (engl. *Kohonen Self Organising Maps*) te analiziraju stilistička svojstva dokumenta, autorovo bogatstvo rječnika te sintaktička i *part-of-speech* svojstva dobivena parsiranjem teksta.

Ojačano učenje strojnom učenju pristupa na način da agenta, tj. algoritam, trenira interakcijom te davanjem pozitivne ili negativne povratne informacije (PioneerLabs). Ukoliko agent za podatke koje dobiva kroz ulaznu jedinicu daje ispravan odgovor, sustav ga "nagrađuje" skalarnom vrijednosti ili nekom drugom funkcijom za nagrađivanje (Medium). Agentov algoritam je pri tome izrađen na način da pokušava maksimizirati količinu nagrade koju dobiva (Sutton & Barto, 1998). Sustav se tako uči rješavanju zadatka metodom pokušaja i pogrešaka (engl. *trial-and-error*) u stvarnom okruženju, a ne na prethodno pripremljenim i označenim bazama podataka (Medium).

Umjetne neuronske mreže jedan su od pristupa u strojnom učenju i umjetnoj inteligenciji (engl. *Artificial Intelligence*) u kojem računalo uči izvršavati određene zadatke na temelju analize podataka koji se nalaze u setu za treniranje, a ti podaci su uglavnom prethodno ručno označeni (Hardesty, 2017). Dok se biološki mozak sastoji se od kombinacije neurona i sinapsi koje ih povezuju, umjetna neuronska mreža sastoji se od čvorova i veza s pripadajućim težinskim vrijednostima (engl. *weight*). Umjetna neuronska mreža može sadržavati od nekoliko tisuća do nekoliko milijuna gusto povezanih čvorova za obradu podataka (Hardesty, 2017). Pojedinačni čvorovi sami po sebi izuzetno su jednostavni, no njihovim kombiniranjem te stvaranjem više-slojevitog sustava omogućava se rješavanje izuzetno složenih zadataka. Proces učenja kod neuronskih mreža Ujević Andrijić (2019, str.

219) opisuje kao “iterativni postupak podešavanja (optimiziranja) vrijednosti težinskih faktora na osnovu pogreške između proračunate vrijednosti modelom i stvarne vrijednosti mjerene veličine”. Učenje se pri tome odvija prema nekom od prethodno definiranih pravila učenja matematičke logike (DataFlair) kao što su npr. korelacijsko pravilo učenja (engl. *correlation learning rule*), delta pravilo učenja (engl. *delta learning rule*), pravilo učenja perceptrona (engl. *perceptron learning rule)*, itd. Umjetne neuronske mreže mogu se koristiti za klasifikaciju, prepoznavanje uzoraka, obradu prirodnog jezika, ekstrakciju svojstava i sl., a mogućnost odrađivanja ovih zadataka čini umjetne neuronske mreže izuzetno zanimljivim područjem za intrinzičnu detekciju plagijata (Towards AI). U novije vrijeme također se javljaju evoluirajuće umjetne neuronske mreže (engl. *evolutionary neural networks*) s mogućnošću samoorganiziranja, samoprilagođavanja te samostalnog učenja (Ding, Li, Su, & Yu, 2013). Pristup intrinzičnoj detekciji plagijata temeljen na ovakvim neuronskim mrežama u svom radu predstavlja Curran (2009). Njegov pristup temelji se na obradi stilometrijskih svojstava u tekstu te kombinira neuronske mreže s računicom za evoluciju (engl. *evolutionary computation*), tj. genetskim algoritmom (engl. *genetic algorithm)* koji sustavu omogućava da, prema potrebi, istovremeno izmjenjuje vlastitu arhitekturu te parametre težine u svrhu pronalaženja optimalne mreže za detekciju plagijata. Pristupe temeljene na umjetnim neuronskim mrežama u svojim radovima također koriste Hourrane i Benlahmer (2019) u kombinaciji sa sintaktičkim stablima u svrhu bogatog ugrađivanja stila (engl. *rich style embedding*) te Al-Sallal, Iqbal, Palade, Amin i Chang (2019) koji u svom sustavu višeslojne perceptronske neuronske mreže kombiniraju s tehnikama latentne semantičke analize, stilometrije te vrećama riječi (engl. *bag of words, BOW*) u svrhu poboljšanja postupka kvantifikacije stila.

6. Prednosti i nedostaci intrinzične detekcije plagijata

Najveća prednost intrinzične detekcije svakako je već ranije spomenuta činjenica da detekcija plagijata uz pomoć ovog pristupa ne ovisi o referentnoj zbirci radova. Zahvaljujući tome, intrinzična detekcija može se koristiti u situacijama kada referentna zbirka nije dostupna te kada se sumnja na složenije oblike plagiranja poput plagiranja prijevodom, parafraziranjem te skriveno autorstvo koje nije moguće detektirati uz pomoć ekstrinzičnih sustava za detekciju plagijata. Uz to, stilometrija, a samim time i intrinzična detekcija plagijata te zadaci kojima se ona bavi, imaju široku uporabu u područjima vezanim uz detekciju plagiranja kao što su određivanje i dokazivanje autorstva u književnim djelima, detektiranje literarnih žanrova, detekcija obmana, profiliranje autora na temelju njegovog stila u marketinške svrhe i sl.

Unatoč tome što intrinzična detekcija predstavlja veoma obećavajuće područje u automatskoj detekciji plagijata, ovaj pristup još je uvijek nedovoljno razvijen da bi se u njega u njega moglo imati potpuno povjerenje da će detektirati potencijalno plagiranje unutar teksta. Jednu od najvećih zapreka u ovom pristupu predstavlja problem vezan uz duljinu teksta koji se ispituje. Naime, za izračunavanje stilometrijske mjere na globalnoj razini, tj. na razini čitavog dokumenta, potrebna je dovoljno velika količina teksta. Kada govorimo o stilometriji zapravo govorimo o statističkom pristupu detekciji plagijata koji ostvaruje puno kvalitetnije rezultate sa što većom količinom podataka (Suchomel & Brandejs, 2015). Razlog nedovoljne količine teksta je činjenica što se na plagiranje uglavnom provjerava samo jedan rad autora, a nije rijetkost da autorovi ostali radovi ispitivačima i sustavima za detekciju nisu dostupni. Ova situacija osobito je česta ako se radi o radovima učenika i studenata za koje nije uobičajeno da se trajno pohranjuju nakon predaje i ocjenjivanja. Ukoliko se radovi ipak pohranjuju javlja se niz pitanja vezanih uz autorska prava prema kojima je za pohranu svakog pojedinačnog rada potrebno imati pisanu dozvolu autora bez obzira na svrhu pohranjivanja (Weber-Wulff, 2016). Također, ako postoji zbirka referentnih radova određenog autora, javlja se dodatni problem vezan uz prirodnu promjenu autorovog stila pisanja kroz njegov život pod utjecajem vanjskih i unutarnjih faktora koji mogu uključivati razinu školovanja, utjecaje na stil koji proizlaze čitanja i proučavanja tuđih radova, samostalno ili potpomognuto proučavanje vlastitih pogrešaka u pisanju u svrhu njegovog poboljšanja, i sl. Promjene u stilu pisanja također se mogu javiti zbog izmjena koje su se dogodile pri reviziji članka (Alican, 2012). Ovakve promjene mogu imati značajan utjecaj na postupak kvantifikacije stila autora

pri čemu se jednog autora tekstovi nastali u različitim vremenskim razdobljima mogu greškom označiti kao radovi različitih autora.

Još jedan problem javlja se iz razloga što intrinzična detekcija plagijata, osim na globalnoj, često mora djelovati i na lokalnoj razini, pri čemu obrađuje male segmente teksta za koje se sumnja da su plagirani te da pripadaju različitim autorima. Sustavima za detekciju to ponovno predstavlja problem zbog statističke prirode stilometrijskog pristupa. Naime radi se o situacijama u kojima je potrebno kvantificirati stilove više različitih autora na temelju često oskudnih i ograničenih primjera njihovog stila koji se u plagiranom dokumentu potencijalno javljaju u samo jednom odlomku ili čak u samo jednoj rečenici. Uz to, u većini slučajeva ne postoje informacije o tome koji su dijelovi teksta plagirani te koliki postotak dokumenta čine što dodatno otežava detekciju (Suchomel & Brandejs, 2015).

Svi navedeni nedostaci mogu dovesti do problema u samim sustavima za detekciju plagijata od kojih su neki već u komercijalnoj upotrebi. Osim što sustavi za detekciju plagijata mogu propustiti detektirati plagiranje, jednako tako postoji mogućnost da originalni dio teksta kojeg autor nije plagirao sustav označi kao plagiran. Problemi se također mogu javiti kada jezik koji je osnova sustava, najčešće engleski, nije autorov materinji jezik. U ovakvim situacijama događalo se da sustav na temelju promjena u leksičkoj raznolikosti teksta te uporabi stop riječi čak i pravilno citirane dijelove teksta označava kao plagirane (Guida, 2019). Ovakve pogreške i propusti mogu imati izravne posljedice za sve osobe koje su uključene u proces detekcije plagijata. Zbog toga, izuzetno je važno naglasiti da "sustav sam po sebi ne može odrediti je li tekst plagiran ili ne te on jedino može ukazati na potencijalno plagiranje unutar dokumenta" (Weber-Wulff, 2016, str. 8). Dosadašnja istraživanja ukazuju na to da, iako sustavi za automatsku detekciju plagijata mogu biti od izuzetne koristi za profesore i druge ispitivače, prijeko je potrebno da se radovi dodatno ručno pregledaju te da se tek tada presudi o tome radi li se doista o plagiranom djelu ili ne (Weber-Wulff, 2016).

7. Prevencija plagiranja

Kada govorimo o prevenciji plagiranja edukacija o tome što je plagiranje te edukacija o etičkom ponašanju u akademskim krugovima svakako bi trebala predstavljati prvi i nezaobilazni korak. Neznanje autora te nepoznavanje oblika plagiranja i pravilnih načina citiranja mogu dovesti do toga da autor nije ni svjestan da je počinio plagiranje. Iz ovog razloga važno je da učitelji i profesori jasno definiraju što se smatra plagiranjem te da svoje učenike i studente upoznaju s službenim stilovima citiranja. Također je izuzetno važno da učitelji i profesori pri tome budu dosljedni u svojim stavovima i sankcijama, jer ukoliko učenici i studenti vide da akademska nečestitost ostaje nesankcionirana veća je vjerojatnost da će se i kod njih javiti takvo ponašanje (Kukolja Tardi, Tardi, & Đogaš, 2012).

Uz edukaciju se sankcioniranje plagiranja općenito smatra dobrim oblikom prevencije, no sankcije pri tome svakako trebaju ovisiti o težini prekršaja. Težina bi trebala ovisiti o postotku plagiranog teksta te obliku plagiranja koji se javlja unutar rada, odnosno sankcije za predaju tuđeg završnog rada kao vlastitog ne bi smjele biti jednake onima kada autor slučajno izostavi citiranje. Unatoč tome što na hrvatskim sveučilištima postoje sustavni pristupi prevenciji i rješavanju plagiranja smatra se da "još uvijek ne postoji strateški i sustavan pristup ovom problemu" (Birkić, Celjak, Cundeković, & Rako, 2017, str. 6). Uz to Rumboldt (2014) smatra da bi akademska zajednica trebala postati kritičnija prema samoj sebi pri čemu se pažnja najprije treba usmjeriti na recenziranje članaka, uređivanje znanstvenih i stručnih časopisa te drugih publikacija poput zbornika i kvalifikacijskih radova.

Prevenciji plagiranja kroz edukaciju može se pristupiti još od školske dobi, a jedan uspješan primjer je CARNetov priručnik *Citiranje u digitalnom okruženju* (2018) koji se bavi zakonskim i etičkim aspektima korištenja informacija i zaštitom autorskih prava pri čemu objašnjava osnovne pojmove, pravne odredbe iz nacionalnog zakonodavstva te međunarodne smjernice. Kako bi se izbjeglo nenamjerno plagiranje zbog nepoznavanja pravila citiranja, u priručniku su opisani najčešće korišteni sustavi i stilovi citiranja te su objašnjena osnovna pravila i definicije citiranja. Posebno poglavlje posvećeno je vrstama alata za upravljanje bibliografskim bilješkama i stvaranju vlastitih baza znanja. Ovaj priručnik je izdan u sklopu projekta "e-Škole: Uspostava sustava razvoja digitalno zrelih škola" za korištenje u osnovnom i srednjem obrazovanju. Podizanje razine svijesti o štetnosti plagiranja sigurno će imati pozitivne učinke na smanjenje učestalosti namjernog i nenamjernog plagiranja.

Uspješnost ove i sličnih inicijativa trebalo bi potvrditi kroz provođenje adekvatnih istraživanja. Uz to, potrebno je provesti i istraživanja o motivaciji plagijatora. Na primjer jedan od razloga za samoplagiranje je umjetno uvećavanje broja citiranja autorovih radova (Collberg & Kobourov, 2005). Ovaj fenomen javlja se iz razloga što se na broj citiranja gleda kao na indikator uspješnosti istraživanja koji ima velik utjecaj pri zapošljavanju i osiguravanju sredstava za buduće projekte (Foltýnek et al., 2019). Raspoznavanje profila plagijatora također može pomoći u predviđanju i prevenciji plagiranja. McCabe i Trevino (1997) predložili su korištenje opisnih faktora poput dobi, spola, postignuća u školovanju, razine edukacije roditelja, izvannastavnih aktivnosti, ponašanja vršnjaka, strogosti sankcioniranja plagiranja te pružanja potpore za razvoj akademskog integriteta od strane edukacijske ustanove.

Zaključak

Intrinzična detekcija plagijata, kao pristup u kojem se ne koristi referentna baza podataka, posjeduje niz prednosti nad ekstrinzičnom detekcijom, a prvenstveno se radi o mogućnosti detekcije plagijata unutar dokumenta bez potrebe za korištenjem vanjskih resursa. Dva glavna zadatka kojima se ovaj oblik automatske detekcije bavi je identifikacija autorstva te detekcija promjena u stilu pisanja. Pri tome, radi na način da kvantificira stil pisanja unutar dokumenta koji se ispituje uz pomoć stilometrijskih formula koje analiziraju leksička, sintaktička i semantička svojstva teksta. Metode koje se koriste pri intrinzičnoj detekciji temelje se na tehnologijama iz područja obrade prirodnog jezika, strojnog učenja te umjetnih neuronskih mreža. Daljnji razvitak ovih tehnologija svakako će imati pozitivni utjecaj na samu intrinzičnu detekciju plagijata. Intrinzični pristup, uz detekciju plagijata, dodatno omogućava izvršavanje čitavog niza zadataka koji se bave pitanjem autorstva kao što su profiliranje autora u marketinške svrhe, dokazivanjem autorstva u znanstvenim i književnim djelima, određivanje žanrova literarnih djela, prepoznavanje obmana, a uporabu pronalazi i u forenzičkoj lingvistici te digitalnoj forenzici. Unatoč tome što intrinzični sustavi uvelike olakšavaju detekciju plagijata oni još uvijek nailaze na mnogo zapreka i problema u svom radu. Najveći problem nesumnjivo predstavlja duljina teksta s obzirom na to da se u većini slučajeva ispituje samo jedan autorov dokument, a zbog statističke prirode stilometrijskog pristupa manja količina podataka za obradu dovodi do lošijih rezultata. Promjene u stilu pisanja koje su se dogodile tijekom recenzije rada također predstavljaju problem jer mijenjaju autorov izvorni stil pisanja. Izgradnja i održavanje sustava za detekciju plagijata dodatno donose niz financijskih i vremenskih troškova, a uz to su temeljnim korisnicima često zbunjujući i teški za korištenje. Iz ovih razloga važno je naglasiti da sustavi za automatsku detekciju plagijata nisu savršeni te da se na njih treba gledati kao na pomagalo u detekciji plagijata, a konačnu presudu o tome je li određeni rad plagijat ili ne mora donijeti čovjek. Uz to je potrebno da sama akademska zajednica koja se suočava s plagiranjem izradi pravilnike u kojima se točno definira koje se radnje smatraju plagiranjem te sankcije za svaki od prekršaja koji se dogodio, a stvaranje okruženja etičke odgovornosti među studentima i ostalim sudionicima akademske zajednice predstavlja još jedan izuzetno važan korak u prevenciji plagiranja.

Literatura

1. Abid, M., Usman, M., & Ashraf, W. M. (2017). Plagiarism Detection Process using Data Mining Techniques. *International Journal of Recent Contributions from Engineering Science & IT (iJES) 5(4),* str. 68-75.
2. Adam, R., & Suharjito, S. (2014). Plagiarism detection algorithm using natural language processing based on grammar analyzing. *Journal of Theoretical and Applied Information Technology 10(63),* str. 168 - 180.
3. Afroz, S., Islam, A. C., Stolerman, A., Greenstadt, R., & McCoy, D. (2014). Doppelgänger finder: Taking stylometry to the underground. U *Proceedings of the 2014 IEEE Symposium on Security and Privacy* (str. 212–226).
4. Alican, N. F. (2012). *Rethinking Plato: A Cartesian Quest for the Real Plato (Value Inquiry).* New York-Amsterdam: Rodopi.
5. Alsallal, M., Iqbal, R., Amin, S., & James, A. (2013). Intrinsic plagiarism detection using latent semantic indexing and stylometry. U *Proceedings - 2013 6th International Conference on Developments in eSystems Engineering, DeSE 2013.* (str. 145-150). New York, NY: Institute of Electrical and Electronics Engineers Inc.
6. Al-Sallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. Future Generation Computer Systems, 96, str. 700-712.
7. Alzahrani, S. M., Salim, N., & Abraham, A. (2012). Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 42(2),* str. 133-149.
8. Bailey, J. (2019). *5 Historical Moments that Shaped Plagiarism.* Preuzeto 15. svibnja, s https://www.turnitin.com/blog/5-historical-moments-that-shaped-plagiarism
9. Bensalem, I., Rosso, P., & Chikhi, S. (2019). On the use of character n-grams as the only intrinsic evidence of plagiarism. *Language Resources and Evaluation volume, 53,* str. 363–396.
10. Bevendorff, J., Potthast, M., Hagen, M., & Stein, B. (2019). Heuristic Authorship Obfuscation. U *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* str. 1098–1108.

11. Birkić, T., Celjak, D., Cundeković, M., & Rako, S. (2017). *Analiza softvera za otkrivanje plagiranja u znanosti i obrazovanju: inačica 1.1.* Zagreb: Sveučilišni računski centar. Preuzeto 28. rujna 2020, s https://www.srce.unizg.hr/files/srce/docs/CEU/izvjestaj_analiza_softvera_za_otkrivanje_plagiranja_u_znanosti_i_obrazovanju_inacica_1_1.pdf
12. Bormuth, J. R. (1966). Readability: A New Approach. *Reading Research Quarterly, 1 (3),* str. 79-132
13. Bowdoin. (n.d.) *Overview of NLP: Issues and Strategies*. Preuzeto 6. rujna 2020, s http://www.bowdoin.edu/~allen/nlp/nlp1.html
14. Burrows, J. (2002). Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and linguistic computing, 17(3),* str. 267-287.
15. Cantos-Gómez, P. & Almela, Á. (2019). Readability indices for the assessment of textbooks: a feasibility study in the context of EFL. *Vigo International Journal of Applied Linguistics, 16,* str. 31-52.
16. Castro, D., Adame, Y., Pelaez, M., & Muñoz, R. (2015). Authorship verification, combining linguistic features and different similarity functions—Notebook for PAN at CLEF 2015. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop, CEUR Workshop Proceedings 1391.*
17. Cerjan-Letica, G., & Letica, S. (2008). Znanstvena nedoličnost: kako se s njom nositi u Hrvatskoj? *Acta stomatologica Croatica, 42(2),* str. 117-122.
18. Chall, J. S. & Dale, E. (1995). Readability Revisited: The New Dale-Chall Readability Formula. Brookline, MA: Brookline Books.
19. Chitra A., & Rajkumar, A. (2015). Plagiarism Detection Using Machine Learning-Based Paraphrase Recognizer. *Journal of Intelligent Systems, 25(3),* str. 351–359.
20. Chong, M. & Specia, L. (2011). Lexical generalisation for word-level match-ing in plagiarism detection. U *Proceedings of the International Conference Recent Advances in Natural Language Processing, 12-14* (str. 704–709). Hissar: Association for Computational Linguistics.
21. Chong, M. Y. M. (2013). *A study on plagiarism detection and plagiarism direction identification using natural language processing techniques.* Doktorski rad. Preuzeto 5. rujna 2020, s

https://www.semanticscholar.org/paper/A-study-on-plagiarism-detection-and-plagiarism-Chong/f2fee611894aef01c5409d920eea2019ce59940b

22. Claburn, T. (2018). *FYI: AI tools can unmask anonymous coders from their binary executables*. Preuzeto 28. rujna 2020, s https://www.theregister.com/2018/03/16/identifying_anonymous_programmers/
23. Clough, P. (2003). Old and new challenges in automatic plagiarism detection. *National Plagiarism Advisory Service*, str. 391-407.
24. Clough, R. (2000). Plagiarism in natural and programming languages: an overview of current tools and technologies. Preuzeto 20. rujna 2020, s https://ir.shef.ac.uk/cloughie/papers/plagiarism2000.pdf
25. Collberg, C. & Kobourov, S. (2005). Self-plagiarism in computer science. *Communications of the ACM, 48(4),* str. 88–94.
26. Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing Text Readability Using Cognitively Based Indices. *Psycholinguistics for TESOL. 42(3),* str. 475-493.
27. Curran, D. (2009). An Evolutionary Neural Network Approach to Intrinsic Plagiarism Detection. U *AICS 2009: Artificial Intelligence and Cognitive Science* (str. 33-40).
28. Dale E, Chall J. (1948). A Formula for Predicting Readability. *Educational Research Bulletin, 27*, str. 11–20+28.
29. Daniel Karaś, Martyna Śpiewak, and Piotr Sobecki. (2017). OPI-JSA at CLEF 2017: Author clustering and style breach detection—Notebook for PAN at CLEF 2017. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17).*
30. Data Science Central. (2017). *Information retrieval document search using vector space model in R.* Preuzeto 5. rujna 2020, s https://www.datasciencecentral.com/profiles/blogs/information-retrieval-document-search-using-vector-space-model-in
31. DataFlair. (n.d.). *Introduction to Learning Rules in Neural Network.* Preuzeto 22. rujna 2020, s https://data-flair.training/blogs/learning-rules-in-neural-network/
32. Ding, S., Li, H., Su, C., & Yu, J. (2013). Evolutionary artificial neural networks: A review. *Artificial Intelligence Review, 39(3).*

33. Elahi, H., & Muneer, H. (2018). Identifying Different Writing Styles in a Document Intrinsically using Stylometric Analysis. *Zenodo.* Preuzeto 28. rujna 2020, s https://zenodo.org/record/2538334

34. Elamine, M., Mechti, S., & Belguith, L.H. (2017). Intrinsic Detection of Plagiarism based on Writing Style Grouping. *LPKM 2017*.

35. Encyclopædia Britannica. (1998). *Plagiarism.* Preuzeto 12. svibnja 2020, s https://www.britannica.com/topic/plagiarism

36. Feng, W. V. & Hirst, G. (2013). Authorship Verification with Entity Coherence and Other Rich Linguistic Features Notebook for PAN at CLEF 2013. U *CLEF (Working Notes) 2013.*

37. Foltýnek T., Meuschke N., & Gipp B. (2019). Academic Plagiarism Detection: A Systematic Literature Review. *ACM Comput, 52(6),* str. 1-49.

38. Foltýnek, T. & Glendinning, I. (2015). Impact of policies for plagiarism in higher education across europe: Results of the project. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis, 63(1),* str. 207-216.

39. Fry, E. (1977). Fry's Readability Graph: Clarifications, Validity, and Extension to Level 17. *Journal of Reading, 21(3),* str. 242-252.

40. Gipp, B., Meuschke, N., & Beel, J. (2011). Comparative Evaluation of Text- and Citation-based Plagiarism Detection Approaches using GuttenPlag. U *Proceedings of 11th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'11)* (str. 255-258). New York, NY: Association for Computing Machinery.

41. Gómez-Adorno, H., Sidorov, G., Pinto, D., Markov, I. (2015). A graph based authorship identification approach—Notebook for PAN at CLEF 2015. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15).*

42. GoogleDevelopers. (n.d.). *Common ML Problems.* Preuzeto 20. rujna 2020, s https://developers.google.com/machine-learning/problem-framing/cases

43. Graham, N., Hirst, G., & Marthi, B. (2005). Segmenting documents by stylistic character. *Natural Language Engineering, 11(04),* str. 397-415.

44. Greenfield, J. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Philadelphia, PA: Temple University.

45. Greenfield, J. (2003). The Miyazaki EFL Readability Index. *Comparative Culture, 9,* str. 41-49.

46. Grozea, C., & Popescu, M. (2010). Who's the Thief? Automatic Detection of the Direction of Plagiarism. U *International Conference on Intelligent Text Processing and Computational Linguistics. CICLing 2010. Lecture Notes in Computer Science, 6008.* (str. 700-710). Berlin: Springer.

47. Guida, C. T. (2019). Plago: A system for plagiarism detection and intervention in massive courses. Diplomski rad. Preuzeto 29. rujna 2020, s https://smartech.gatech.edu/handle/1853/61787

48. Gunning, R. (1952). *The technique of clear writing.* Toronto : McGraw-Hill.

49. Hardesty, L. (2017). *Explained: Neural networks.* Preuzeto 22. rujna 2020, s https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

50. Heather, J. (2010). Turnitoff: Identifying and fixing a hole in current plagiarism detection software. *Assessment & Evaluation in Higher Education, 35(6),* str. 647–660.

51. Hebrang Grgić, I.,; Ivanjko, T., Melinščak Zlodi, I., Mučnjak, D. (2018). Citiranje u digitalnom okruženju: priručnik. Zagreb: Carnet. Preuzeto 1. listopada 2020, s https://pilot.e-skole.hr/wp-content/uploads/2018/03/Prirucnik_Citiranje-u-digitalnom-okruzenju-1.pdf

52. Honore, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin, 7(2),* str. 172-177.

53. Hourrane, Q. & Benlahmer, E. H. (2019). Rich Style Embedding for Intrinsic Plagiarism Detection. *(IJACSA) International Journal of Advanced Computer Science and Applications, 10(11).*

54. Hrvatska enciklopedija, mrežno izdanje. (2020). *Kriptomnezija*. Preuzeto 1. listopada 2020, s http://www.enciklopedija.hr/Natuknica.aspx?ID=33993

55. Hrvatski mrežni rječnik - Mrežnik. (n.d.). *Pojmovnik.* Preuzeto 7. rujna 2020, s http://ihjj.hr/mreznik/page/pojmovnik/6/

56. Hürlimann, M., Weck, B., Berg, E., Simon, Š., & Nissim, M. (2015). GLAD: Groningen Lightweight Authorship Detection. U *PAN at CLEF 2015.*

57. International Center for Academic Integrity.(n.d.). *Statistics.* Preuzeto 1. listopada 2020, s https://www.academicintegrity.org/statistics/
58. Isaacs, D. (2011). Plagiarism is not OK. *Journal of Paediatrics and Child Health, 47,* 159. Preuzeto 15. svibnja, s Wiley Online Library baze.
59. Kanjirangat, V. & Gupta, D. (2016). A Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *Journal of Engineering Science and Technology Review, 9(4*), str. 150-164.
60. Karas, D., Spiewak, M., & Sobecki, P. (2017). OPI-JSA at CLEF 2017: Author clustering and style breach detection—Notebook for PAN at CLEF 2017. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'17).*
61. Kelly, L. (2017). *The Coleman-Liau Index.* Preuzeto 28. kolovoza 2020, s https://readable.com/blog/the-coleman-liau-index/
62. Kelly, L. (2017). *The Flesch Reading Ease and Flesch-Kincaid Grade Level.* Preuzeto 28. kolovoza 2020, s https://readable.com/blog/the-flesch-reading-ease-and-flesch-kincaid-grade-level/
63. Khan, B. (2020). *Five Methods for Measuring Text Readability.* Preuzeto 25. kolovoza 2020, s https://www.litinfocus.com/5-accurate-methods-for-measuring-text-readability/
64. Khan, J. A. (2017). Style Breach Detection: An Unsupervised Detection Model. *CLEF 2017.*
65. Khonji, M. & Iraqi, Y. (2014). A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)—Notebook for PAN at CLEF 2014. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'14)*.
66. Kincaid, J. P. & Delionbach, L. J. (1973). Validation of the Automated Readability Index: A Follow-Up. *Human Factors: The Journal of Human Factors and Ergonomics Society, 15,* str. 17-20.
67. Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chisson, B. S. (1975). Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. *Institute for Simulation and Training, 56.* Preuzeto 25. kolovoza, s https://stars.library.ucf.edu/istlibrary/56

68. Koppel, M. & Schler, J. (2004). Authorship verification as a one-class classification problem.
69. Koppel, M. & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology 65(1),* str. 178–187.
70. Kukolja Taradi, Sunčana; Taradi, Milan; Đogaš, Zoran. (2012). Croatian medical students see academic dishonesty as an acceptable behaviour: a cross-sectional multicampus study. *Journal of medical ethics, 38(1),* str.376-379.
71. Kumar, P. (2017). *An Introduction to N-grams: What Are They and Why Do We Need Them?* Preuzeto 1. rujna 2020, s https://blog.xrds.acm.org/2017/10/introduction-n-grams-need/
72. Kuta, M. & Kitowski, J. (2014). Optimisation of Character n-gram Profiles Method for Intrinsic Plagiarism Detection. *Artificial Intelligence and Soft Computing Lecture Notes in Computer Science*, 8458, str. 500-511.
73. Kuznetsov, M., Motrenko, A., Kuznetsova, R., & Strijov, V. (2016). *Methods for intrinsic plagiarism detection and author diarization.* CLEF.
74. Lynch, J. (2006). The Perfectly Acceptable Practice of Literary Theft: Plagiarism, Copyright, and the Eighteenth Century. *Write It Now.* Preuzeto 15. svibnja 2020, s https://www.writing-world.com/rights/lynch.shtml
75. Mahmood, A., Safiq, Z., & Srinivasan, P. (2020) *A Girl Has A Name: Detecting Authorship Obfuscation.* Preuzeto 15. rujna 2020, s https://www.researchgate.net/publication/341148446_A_Girl_Has_A_Name_Detecting_Authorship_Obfuscation
76. Maitra, P., Ghosh, S., & Das, D. (2016). Authorship Verification - An Approach based on Random Forest. U *CLEF 2015.*
77. Majstorović, D. (2016). Stavovi studenata korisnika Nacionalne i sveučilišne knjižnice u Zagrebu o plagiranju i javnoj objavi ocjenskih radova. *Vjesnik bibliotekara Hrvatske, 59(3-4),* str. 131-152.
78. Maurer, H., Kappe, F., & Zaka, B. (2006). Plagiarism – A survey. *Journal of Universal Computer Science, 12(8),* str. 1050–1084.
79. McCabe, D. (2005). Cheating among college and university students: A North American perspective. The International Journal for Educational Integrity, 1.

80. McCabe, D. L. & Trevino, L. K. (1997). Individual and Contextual Influences on Academic Dishonesty: A Multicampus Investigation. *Research in Higher Education, 38*, str. 379–396.

81. McKee, G., Malvern, D. D., & Richards, B. J. (2000). Measuring Vocabulary Diversity Using Dedicated Software. *Literary and Linguistic Computing 15(3)*, str. 323-337.

82. Medium. (2019). *Reinforcement Learning algorithms — an intuitive overview*. Preuzeto 21. rujna 2020, s https://medium.com/@SmartLabAI/reinforcement-learning-algorithms-an-intuitive-overview-904e2dff5bbc

83. Merriam-Webster. (n.d.). *Plagiarize*. Preuzeto 12. svibnja 2020, s https://www.merriam-webster.com/dictionary/plagiarize

84. Meyer zu Eissen, S., & Stein, B. (2006). Intrinsic plagiarism detection. U *ECIR'06: Proceedings of the 28th European conference on Advances in Information Retrieval* (str. 565–569). Berlin: Springer-Verlag.

85. Meyer zu Eissen, S., Stein, B., & Kulig, M. (2007) Plagiarism Detection Without Reference Collections. U *Advances in Data Analysis. Proceedings of the 30th Annual Conference of the Gesellschaft für Klassifikation e.V* (str. 359-366). Berlin: Springer.

86. Mira Seo, J. (2009). Plagiarism and Poetic Identity in Martial. *The American Journal of Philology, 130(4)*, str. 567-593. Preuzeto 15. svibnja 2020, s JSTOR baze.

87. Mirco Kocher. 2016. UniNE at CLEF 2016: Author clustering—Notebook for PAN at CLEF 2016. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'16).*

88. Montgomerie, A. (2019). *Ten Signs of "Lifted" Text.* Preuzeto 15. kolovoza 2020, s https://aceseditors.org/news/2019/ten-signs-of-lifted-text

89. Moreau, E., Jayapal, A., Lynch, G., & Vogel, C. (2015). Author verification: Basic stacked generalization applied to predictions from a set of heterogeneous learners—Notebook for PAN at CLEF 2015. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15)*.

90. Mozgovoy, M., Kakkonen, T., & Cosma, G. (2010). Automatic student plagiarism detection: Future perspectives. *Journal of Theoretical and Applied Information Technology, 43(4)*, str. 511–531.

91. myDU, Dominican University. (n.d.). *Plagiarism of Structure.* Preuzeto 24. rujna 2020, s https://jicsweb1.dom.edu/ICS/Resources/Student_Services/Learning_Resources/Writing_Resources/Plagiarism.jnz?portlet=Plagiarism_of_Structure
92. Nugaliyadde, A., Wong, K., Sohel, F., & Xie, H. (2019). Enhancing Semantic Word Representations by Embedding Deep Word Relationships. *ArXiv, abs/1901.07176*.
93. Oberreuter, G. & and Velásquez D. J., (2013). Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style. *Expert Systems with Applications, 40(9),* str. 3756–3763.
94. Oberreuter, G., L'Huillier, G., Rios, S., & Velasquez, J. D. (2011). Approaches for Intrinsic and External Plagiarism Detection - Notebook for PAN at CLEF 2011. U *CLEF 2011*.
95. PioneerLabs. (n.d.) *The Three Types of Machine Learning Algorithms.* Preuzeto 20. rujna 2020, s https://pioneerlabs.io/insights/the-three-types-of-machine-learning-algorithms/
96. Plagiarism.org. (2017). *What is Plagiarism?* Preuzeto 12. svibnja 2020, s https://www.plagiarism.org/article/what-is-plagiarism
97. Polydouri, A., Vathi, E., Siolas, G., & Stafylopatis, A. (2018). An efficient classification approach in imbalanced datasets for intrinsic plagiarism detection. *Evolving Systems, 11*, str. 503-515.
98. Postscripts. (2006). *THE WRITING CLINIC: Clear Writing: How to Achieve and Measure Readability.* Preuzeto 28. kolovoza 2020, s http://notorc.blogspot.com/2006/09/devils-in-details-measuring.html
99. Potthast M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.. (2013). Overview of the 5th International Competition on Plagiarism Detection. U *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop* (CLEF'13).
100. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., & Stein, B. (2017). Overview of PAN'17: Author identification, author profiling, and author obfuscation. U *Proceedings of the 7th International Conference of the CLEF Initiative.*

101. Potthast, M., Stein, B., Eiselt, A., Barron-Cedeno, A., & Rosso, P. (2009) Overview of the 1st International Competition on Plagiarism Detection. U *3rd PAN Workshop : Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing* (str. 9-17).

102. Powers, R. D., Sumner, W. A., & Kearl, B. E. (1958). A recalculation of four adult readability formulas. *Journal of Educational Psychology, 49(2)*, str. 99–105.

103. Pupovac, Vanja; Bilić-Zulle, Lidija; Mavrinac, Martina; Petrovečki, Mladen. (2010). Attitudes toward plagiarism among pharmacy and medical biochemistry students – cross-sectional survey study. *Biochemia medica : časopis hrvatskoga društva medicinskih biokemičara, 20(3)*, str. 307-313.

104. Ranatunga, R. V. S. P. K., Atukorale, A. S, & Hewagamage, K. P. (2011). Intrinsic plagiarism detection with kohonen self organizing maps. U *The International Conference on Advances in ICT for Emerging Regions - ICTer2011*, 125.

105. Randall, M. (2001). *Pragmatic Plagiarism: Authorship, Profit, and Power.* Toronto: University of Toronto Press.

106. Readability Formulas. (n.d.). *The Flesch Reading Ease Readability Formula*. Preuzeto 22. kolovoza, s https://readabilityformulas.com/flesch-reading-ease-readability-formula.php

107. Readability Formulas. (n.d.). *The Fry Graph Readability Formula.* Preuzeto 29. kolovoza 2020, s https://readabilityformulas.com/fry-graph-readability-formula.php

108. Readability Formulas. (n.d.). *The Gunning's Fog Index (or FOG) Readability Formula.* Preuzeto 28. kolovoza 2020, s https://readabilityformulas.com/gunning-fog-readability-formula.php

109. Readability Formulas. (n.d.). *The New Dale-Chall Readability Formula.* Preuzeto 24. kolovoza 2020, s https://readabilityformulas.com/new-dale-chall-readability-formula.php

110. Readability Formulas. (n.d.). *The Powers-Sumner-Kearl Readability Formula.* Preuzeto 25. kolovoza 2020, s https://readabilityformulas.com/powers-sumner-kear-readability-formula.php

111. RFP-Templates. (n.d.). *Bormuth Readability Score*. Preuzeto 25. kolovoza 2020, s http://www.rfp-templates.com/Readability-Scores/Bormuth-Grade-Level

112. RFP-Templates. (n.d.). *Coleman-Liau Readability Score.* Preuzeto 28. kolovoza 2020, s http://www.rfp-templates.com/Readability-Scores/Coleman-Liau-Grade-Level

113. RFP-Templates. (n.d.). *Flesch-Kincaid Readability Score.* Preuzeto 25. kolovoza 2020, s http://www.rfp-templates.com/Readability-Scores/Flesch-Kincaid

114. RFP-Templates. (n.d.). *Passive Sentences Readability Score.* Preuzeto 22. kolovoza 2020, s http://www.rfp-templates.com/readability-scores/passive-sentences

115. Robert Gunning's Fog Readability Formula. (2004, 23. ožujak, br. 8). *Plain Language at Work Newsletter.* Preuzeto 28. kolovoza 2020, s http://www.impact-information.com/impactinfo/newsletter/plwork08.htm

116. Rosso, P. (2015). Author Profiling and Plagiarism Detection. *Russian Summer School in Information Retrieval: Information Retrieval*, str. 229-250.

117. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., & Stein, B. (2016). Overview of PAN'16. U *Experimental IR Meets Multilinguality, Multimodality, and Interaction* (str. 332–350).

118. Rumblodt, Z. (2014). Što je to plagijat u znanosti? *Arhiv za higijenu rada i toksikologiju, 65*, str. 241-244.

119. Safin, K., & Kuznetsova, R. (2017). Style Breach Detection with Neural Sentence Embeddings. *CLEF 2017*.

120. Seaward, L. & Matwin, S. (2009). Intrinsic Plagiarism Detection using Complexity Analysis. U *3rd PAN Workshop : Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing.* (str. 56-61)

121. Seidman, S. (2013). Authorship verification using the impostors method—Notebook for PAN at CLEF 2013. U Forner, P., Müller, H., Paredes, R., Rosso, P. & Stein, B. *Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF13)*.

122. Sichel, H. S. (1986). Word Frequency Distributions and Type-token Characteristics. *The Mathematical Scientist, 11*, str. 45-72.

123. Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for information Science and Technology, 60(3)*, str. 538-556.

124. Stamatatos, E. (2009). Intrinsic Plagiarism Detection Using Character *n*-gram Profiles. U *3rd PAN Workshop : Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing* (str. 46-54).

125. Stamatatos, E., Daelemans, W., Verhoeven, B:, Juola, P., López-López, A., Potthast, M., & Stein, B. (2015). Overview of the Author Identification Task at PAN 2015. *PAN 2015.* Preuzeto 10. rujna 2020, s http://ceur-ws.org/Vol-1391/inv-pap3-CR.pdf

126. Stamatatos, E., Rangel, F., Tschuggnall, M., Stein, B., Kestemont, M., Rosso, P., Potthast, M. (2018). Overview of PAN 2018. Author identification, author profiling, and author obfuscation. *CLEF 2018: Experimental IR Meets Multilinguality, Multimodality, and Interaction, 11018,* str. 267-285.

127. Stein, B. & zu Eissen, S.M. (2007) Intrinsic Plagiarism Analysis with Meta Learning. *PAN 2007,* 276.

128. Stein, B. Lipka, N., Prettenhofer, P. (2011). Intrinsic Plagiarism Analysis. *Language Resources and Evaluation, 45*, str. 63–82.

129. Stein, B., Koppel, M., & Stamatatos, E. (2007). Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07. *ACM SIGIR Forum, 41(2),* str. 68-71.

130. Stein, B., Rosso, P., Stamatatos, E., Koppel, M., & Agirre., E. (2009). Preface. U *3rd PAN Workshop : Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing* (str. 3).

131. Suchomel, Š. & Brandejs, M. (2015). Determining Window Size from Plagiarism Corpus for Stylometric Features. U *CLEF 2015: Experimental IR Meets Multilinguality, Multimodality, and Interaction* (str. 293-299).

132. Sutton, R. & Barto, A. (1998). *Reinforcement Learning: An Introduction.* Cambridge, MA: MIT Press.

133. Towards AI. (2020). *Main Types of Neural Networks and its Applications - Tutorial.* Preuzeto 22. rujna 2020, s https://medium.com/towards-artificial-intelligence/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e

134. Towards Data Science. (2019). *A Brief Introduction to Supervised Learning.* Preuzeto 21. rujna 2020, s https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590

135. Towards Data Science. (2019). *Let's Understand the Vector Space Model in Machine Learning by Modelling Cars.* Preuzeto 5. rujna 2020, s https://towardsdatascience.com/lets-understand-the-vector-space-model-in-machine-learning-by-modelling-cars-b60a8df6684f

136. Tschuggnall, M., & Specht, G. (2013). Detecting Plagiarism in Text Documents through Grammar-Analysis of Authors. *BTW 2013*.

137. Ujević Andrijić, Ž. (2019). Osvježimo znanje: Umjetne neuronske mreže. *Kemija u industriji : Časopis kemičara i kemijskih inženjera Hrvatske, 68(5-6)*, str. 219–220.

138. van de Rakt, M. (2019). *The Flesch reading ease score: why and how to use it.* Preuzeto 22. kolovoza, s https://yoast.com/flesch-reading-ease-score/

139. Vartapetiance, A. & Gillam, L. (2014). Deception detection: dependable or defective? *Deception detection: dependable or defective?, 4,* čl. br. 166.

140. Weber-Wulff, D. (2016). *Plagiarism Detection Software: Promises, Pitfalls, and Practices.* U Bretag, T. (Ur.) Handbook of Academic Integrity (str.625-638). Singapore: Springer.

141. Williams, K. M., Nathanson, C., & Paulhus, D. L. (2010) Identifying and Profiling Scholastic Cheaters: Their Personality, Cognitive Ability, and Motivation. *Journal of Experimental Psychology: Applied, 16(3),* str. 293-307.

142. Williamson, G. (2014). *Type-Token Ratio.* Preuzeto 22. kolovoza 2020, s https://www.sltinfo.com/type-token-ratio/

143. Yousf, S., Ahmad, M., & Nasurllah, S. (2013). A review of plagiarism detection based on Lexical and Semantic Approach. U *2013 International Conference*

on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA).

144. Yule, G. U. (1944). *The statistical study of literary vocabulary.* Cambridge: University Press.

145. Zechner, M., Muhr, M., Kern, R., & Granitzer, M. (2009). External and Intrinsic Plagiarism Detection Using Vector Space Models. U *3rd PAN Workshop : Uncovering Plagiarism, Authorship and Social Software Misuse. 25th Annual Conference of the Spanish Society for Natural Language Processing* (str. 55-63).

Intrinzična detekcija plagijata

Sažetak

Detekcija plagijata podrazumijeva pronalaženje i prepoznavanje plagiranih dijelova teksta unutar pisanih dokumenata kao što su znanstveni radovi, školski eseji i seminarski radovi, no može se provesti i nad programskim kodom, umjetničkim djelima i sl. Pojavom interneta plagiranje je postalo sve prisutnije, a zbog velike količine izvora iz kojih je moguće plagirati, detekcija plagijata bez pomoći računala uvelike je otežana. Kod računalno potpomognute detekcije plagijata moguća su dva pristupa: ekstrinzični i intrinzični. Ekstrinzična detekcija uspoređuje rad s referentnim tekstovima, dok intrinzična detekcija plagirane dijelove teksta prepoznaje na temelju nedosljednosti u tekstu te promjena u stilu pisanja. U radu se daje pregled intrinzične detekcije plagijata, prikazuje njen način rada i pristupi koji se koriste u detekciji promjena u stilu pisanja. Na samom kraju navode se prednosti i nedostaci ovog pristupa te evaluacija uspješnosti u pronalaženju plagiranih dijelova teksta.

Ključne riječi: plagiranje, detekcija plagijata, intrinzična detekcija plagijata, stilometrija

Intrinsic plagiarism detection

Summary

Plagiarism detection involves finding and recognizing plagiarized parts of text within written documents such as scientific papers, school essays, seminar papers, etc., and can also be performed on program code, works of art and other examples of work. With the advent of the Internet, plagiarism has become more prevalent and due to a large number of sources to plagiarize from the detection of plagiarism without the help of a computer is greatly hampered. In computer-assisted plagiarism detection two approaches are possible: extrinsic and intrinsic. Extrinsic detection compares the work with reference texts, while intrinsic detection recognizes plagiarized parts of the text based on inconsistencies in the text and changes in writing styles. This paper provides an overview of intrinsic plagiarism detection, shows how it operates and the approaches used in detecting changes in writing styles. At the very end, the advantages and disadvantages of this approach are stated and the evaluation of its success in finding plagiarized parts of the text.

Key words: plagiarism, plagiarism detection, intrinsic plagiarism detection, stylometry