# Automatizirana provjera sukladnosti politika privatnosti na webu s Općom uredbom o zaštiti podataka korištenjem strojnog učenja

**Šeša, Rene**

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 471

# AUTOMATIZIRANA PROVJERA SUKLADNOSTI POLITIKA PRIVATNOSTI NA WEBU S OPĆOM UREDBOM O ZAŠTITI PODATAKA KORIŠTENJEM STROJNOG UČENJA

Rene Šeša

Zagreb, lipanj 2024.

SVEUČILIŠTE U ZAGREBU
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

DIPLOMSKI RAD br. 471

# AUTOMATIZIRANA PROVJERA SUKLADNOSTI POLITIKA PRIVATNOSTI NA WEBU S OPĆOM UREDBOM O ZAŠTITI PODATAKA KORIŠTENJEM STROJNOG UČENJA

Rene Šeša

Zagreb, lipanj 2024.

**SVEUČILIŠTE U ZAGREBU**
**FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA**

Zagreb, 4. ožujka 2024.

# DIPLOMSKI ZADATAK br. 471

| | |
|---|---|
| Pristupnik: | **Rene Šeša (0036506503)** |
| Studij: | Računarstvo |
| Profil: | Znanost o podacima |
| Mentor: | doc. dr. sc. Marko Horvat |

Zadatak: **Automatizirana provjera sukladnosti politika privatnosti na webu s Općom uredbom o zaštiti podataka korištenjem strojnog učenja**

Opis zadatka:

Zaštita osobnih podataka na webu važan je čimbenik prilikom korištenja web stranica. Legalni načini prikupljanja, skladištenja i odrade podataka definirani su Općom uredbom o zaštiti podatka (GDPR). Svaki pravni subjekt koji prikuplja podatke na webu dužan je napisati politiku privatnosti koja objašnjava koji podaci se prikupljaju i na koji način se obrađuju. Ručno pregledavanje sukladnosti politike privatnosti s Općom uredbom o zaštiti podataka troši značajnu količinu vremena i materijalnih resursa te postoji potreba za razvojem programske potpore koja bi ubrzala taj postupak. Modeli strojnog učenja za analizu teksta omogućuju pretraživanje, usporedbu i kategorizaciju teksta te su pogodni za analizu pravnih dokumenata. Cilj diplomskog rada je razviti aplikaciju za dohvaćanje sadržaja politika privatnosti iz web stranica i automatiziranu provjeru sukladnosti politike privatnosti s Općom uredbom o zaštiti podataka korištenjem modela strojnog učenja za analizu teksta. Trenirati model strojnog učenja na prikupljenom skupu dobro napisanih politika privatnosti pohranjenih u relacijskoj bazi podataka. Objektivno utvrditi kvalitetu razvijenog modela strojnog učenja za analizu teksta i brzinu analize politika privatnosti koja može dovesti do značajne uštede vremena i materijalnih resursa te omogućiti kvalitetniju zaštitu prava korisnika prilikom korištenja web stranica. Vizualizirati rezultate evaluacije. Prikazati arhitekturu aplikacije i bitne isječke izvornog programskog koda uz potrebna dodatna objašnjenja i dokumentaciju. U diplomskom radu prikazati komponente izrađene programske podrške te definirati sva programska sredstva i potrebne postupke. Citirati korištenu literaturu i navesti dobivenu pomoć.

Rok za predaju rada: 28. lipnja 2024.

## Sadržaj

Uvod ..... 1

1. Opća uredba o zaštiti podataka ..... 2

1.1. Sadržaj Opće uredbe o zaštiti podataka ..... 4

1.1.1. Opće odredbe ..... 4

1.1.2. Načela ..... 6

1.1.3. Prava ispitanika ..... 8

1.1.4. Voditelj obrade i izvršitelj obrade ..... 9

1.1.5. Prijenosi osobnih podataka trećim zemljama ili međunarodnim organizacijama, Neovisna nadzorna tijela, Suradnja i konzistentnost i Pravna sredstva, odgovornost i sankcije ..... 11

1.1.6. Odredbe u vezi s posebnim situacijama obrade, Delegirani akti i provedbeni akti i Završne odredbe ..... 12

1.2. Kazne za kršenje Opće uredbe o zaštiti podataka ..... 13

1.3. Odabrane točke Opće uredbe o zaštiti podataka ..... 14

2. Sustav za dohvaćanje internetskih stranica ..... 15

2.1. PoliPy ..... 16

3. Vektorska reprezentacija teksta ..... 17

3.1. OpenAI text-embedding-3-small model ..... 20

3.2. SPLADE++ model ..... 22

4. Mjere za usporedbu sličnosti teksta ..... 24

5. Klasifikacijski modeli strojnog učenja ..... 27

5.1. Klasifikacija korištenjem modela linearne regresije ..... 28

5.2. Klasifikacija korištenjem modela logističke regresije ..... 31

6. Rezultati ..... 33

6.1. Model text-embedding-3-small ..... 36

6.2. Model SPLADE++ ..... 38

6.3. Usporedba rezultata ..... 39

Zaključak ..... 40

Literatura ..... 41

Sažetak ..... 42

Summary ..... 43

Skraćenice ..... 44

Privitak ..... 45

# Uvod

Procjenjuje se da je u 2023. godini generirano preko 100 zetabajta podataka[1]. Tolika količina podataka predstavlja značajno povećanje u odnosu na ranije godine. Procjenjuje se da će se količina podataka koje ljudi generiraju nastaviti povećavati u budućnosti. Dio tih podataka je izrazito osjetljiv poput medicinskih, biometrijskih i sigurnosnih podataka te postoji potreba za posebnim oprezom prilikom rukovanja s njima. Da bi doskočile ovom gorućem problemu većina svjetskih zemalja je izradila zakonski okvir koji definira načine prikupljanja, obrade i skladištenja podataka.

Na razini Europske Unije, 25. svibnja 2018. godine implementirana je Opća uredba o zaštiti podataka (GDPR)[2]. Cilj uredbe je zaštita i regulacija podataka i osobne privatnosti a između ostaloga definira i propise koji se odnose na razmjenu podataka sa zemljama trećeg svijeta. Unutar uredbe definirano je pravilno prikupljanje, skladištenje i obrada podataka te kazne za kršenje navedenih regulativa. Među ostalim uredba obuhvaća i politike privatnosti web stranica koje su središnja tema ovog diplomskog rada.

Ručno pregledavanje sukladnosti politika privatnosti s Općom uredbom o zaštiti podataka iziskuje znatnu količinu vremena te je vidljiva potreba za automatizacijom te aktivnosti. U ovome diplomskom radu to se pokušava riješiti korištenjem metoda strojnog učenja. Izrađena je programska potpora koja prikuplja politike privatnosti s weba te analizira njihovu sukladnost s Općom uredbom o zaštiti podataka u nekoliko odabranih točaka uredbe. Odabrane točke i način provjeravanje sukladnosti je objašnjen u poglavlju 1 Opća uredba o zaštiti podataka.

Cilj ovoga diplomskog rada je razviti programsku potporu koja provjerava je li politika privatnosti sukladna s Općom uredbom o zaštiti podataka. Za odredbe Opće uredbe o zaštiti podataka koje su odabrane kao pogodne za automatiziranu provjeru, izvršit će se provjera sukladnosti politike privatnosti s Uredbom te joj se dodijeliti odgovarajuća oznaka: sukladna/nije sukladna po određenoj odredbi Opće uredbe o zaštiti podataka. U završnom dijelu ovoga diplomskog rada prikazani su rezultati analize te uspješnost klasifikacije politika privatnosti za svaku od odabranih odredbi. Za kraj donesen je zaključak o primjenjivosti razvijene programske potpore u procesu automatizacije provjere sukladnosti politika privatnosti na webu te su iznesena daljnja razmatranja.

# 1. Opća uredba o zaštiti podataka

Opća Uredba o zaštiti podataka (engl. *The General Data Protection Regulation – GDPR*) je zakonodavni okvir kojem je glavni cilj sveobuhvatna zaštita i regulacija obrade podataka na razini cijele Europske Unije[3]. Prilikom obrade podataka građana Europske Unije dio pravila navedenih u Uredbi vrijede i izvan granica Europske Unije te im je cilj zaštiti prava i dostojanstva njenih građana svugdje u svijetu. Opća uredba o zaštiti podataka implementirana je na razini Europske Unije 25. svibnja 2018. godine na razini svih zemalja članica te se odnosi na tvrtke, pojedince i organizacije koje se nalaze na području Europske Unije i izvan nje. Općoj Uredbi o zaštiti podataka prethodi Direktiva o zaštiti podataka. Direktiva o zaštiti podataka je zakon koji je stupio na snagu u Europskoj Uniji 1995. godine. Cilj Direktive o zaštiti podataka bio je definirati zakonodavni okvir unutar kojeg se obrađuju podaci, ali nagli razvoj IT sektora doveo je do raznih oblika obrade podataka koji nisi bili regulirani od strane direktive. Kako bi doskočila tom problemu, Europske Unija odlučila je izraditi novu uredbu. Izrada Opće Uredbe o zaštiti podataka je završila 14. travnja 2016.

Zaštita prava pojedinaca je jedna od ključnih zadaća Opće uredbe o zaštiti podataka. Centralna bit Opće uredbe o zaštiti podataka je odnos voditelja obrade i subjekta obrade podataka. Termin subjekt obrade podataka označava osobu čiji se podaci obrađuju. To primjerice mogu biti posjetitelji  određene web stranice. GDPR jasno definira da svaki pojedinac čiji podaci se prikupljaju ili obrađuju ima pravo biti informiran o tome te u svakome trenutku ima pravo pristupa podacima koji se prikupljaju. Prilikom prikupljanja podataka strana koja prikuplja podatke naziva se voditelj obrade podataka. To je osoba koja odlučuje kako i zašto će se određeni podaci prikupljaju i obrađuju. U velikim korporacijama za to je najčešće izabran jedan od zaposlenika, u manjim obrtima je najčešće sam vlasnik firme, a u slučaju privatnih web stranica vlasnik domene. Ponekad postoji više istovremenih voditelja obrade, a u određenim situacijama funkciju voditelja obrade zamjenjuje izvršitelj obrade. Sve obveze voditelja obrade odnosno izvršitelja obrade podrobnije su razrađena u Poglavlju iv., dok su prava ispitanika čiji se podaci obrađuju poput prava na brisanje, ispravak i ograničenje obrade podataka detaljno su razrađena i objašnjena u Poglavlju iii. Sva prava ispitanika, odnosno obveze voditelja obrade ili izvršitelja obrade moraju pratiti generalna načela Opće uredbe o zaštiti podataka koja su definirana u Poglavlju ii.

Prije samog prikupljanja podataka kontrolor podataka je dužan tražiti privolu od strane subjekta obrade podataka. Privola mora biti dana svojevoljno te mora biti jednoznačna i specifična. Zahtjev za pristankom mora biti napisan jednostavnim jezikom te se mora moći lako razlikovati od ostalih upita. Subjekt obrade podataka mora moći povući prethodno dan pristanak u bilo kojem trenutku, a kontrolor podataka mora poštovati njegovu odluku. Posebna pozornost posvećena je osobama koje nisu napunile 16 godina te ne mogu same dati privolu već im je za to potrebno dopuštenje roditelja ili skrbnika. Neovisno je li subjekt obrade podataka punoljetna ili maloljetna osoba, voditelj obrade je dužan čuvati dokumentaciju o pristanku od strane subjekta obrade podataka. Svi uvjeti te izuzeća od ovoga pravila detaljnije su opisana u Poglavlju ii.

Opća uredba o zaštiti podataka počiva na sedam glavnih principa: zakonitost, poštenje i transparentnost, točnost, minimizacija podataka, ograničenje pohrane, integritet i povjerljivost, odgovornost te ograničenje namjene. Razumijevanje glavnih principa predstavlja ključ za odgovorno postupanje s podacima. Da bi se to postiglo, potrebno je uložiti vrijeme i resurse u kontinuirano educiranje svih subjekata koji sudjeluju u procesu obrade podataka. Načini na koje je to moguće ostvariti detaljnije su opisani u poglavlju 1.1.

Zaštita podataka i regulacija obrade glavne su ciljevi Opće uredbe o zaštiti podataka, međutim postoje situacije kada ta razina restrikcije iz objektivnih razloga nije moguća jer onemogućava normalno funkcioniranje određenih tijela i organizacija, ili proizvodi neželjene nuspojave. U takvim situacijama govorimo o izuzeću od GDPR-a. Primjerice postoje situacije kada je za normalno funkcioniranje policije, vojske, pravosuđa i obavještajnih službi potrebno koristiti zakonito presretanje podataka kako bi se osigurala sigurnost građana te bi slijepo postupanje u skladu s Općom uredbom o zaštiti podataka onemogućavalo kvalitetno djelovanje navedenih službi. Neke od preostalih situacija kada dolazi do izuzeća predstavljaju podaci o umrlim osobama koje reguliraju nacionalni propisi, određeni podaci koji se prikupljaju u statističke ili znanstvene svrhe i odnosi između zaposlenika i poslodavaca koji su regulirani dodatnim zakonima. Za situacije u kojima dolazi do povrede GDPR-a, a koje se ne smatraju izuzećima propisane su kazne navedene u poglavlju 1.2. Dijelovi Opće uredbe za obradu podataka za koje je procijenjeno da su pogodni za automatiziranu provjeru sukladnost izdvojeni su u poglavlju 1.3.

## 1.1. Sadržaj Opće uredbe o zaštiti podataka

Opća uredba o zaštiti podataka sastoji se od 99 članaka podijeljenih u xi. poglavlja. U sljedećim poglavljima dan je kratak sažetak svakog poglavlja. Za potrebe ovog diplomskog rada najbitnija poglavlja su Poglavlje i., Poglavlje ii., Poglavlje iii. i Poglavlje iv. jer se u njima definiraju opće odredbe Uredbe, kao i generalna načela prema kojima Uredba funkcionira. Za kasniju analizu posebno su bitna prava ispitanika obrađena u Poglavlju iii. jer predstavljaju plodno tlo za automatizaciju procesa provjere sukladnosti politika privatnosti s GDPR-om kao i Poglavlje iv. koje potanko definira prava i obaveze voditelja obrade.

### 1.1.1. Opće odredbe

Uredba započinje s Poglavljem i. Opće odredbe u kojemu se definiraju glavni ciljevi odredbe poput zaštite prava pojedinaca prilikom obrade osobnih podataka i slobode kretanja osobnih podataka unutar Europske Unije. Za glavno područje primjene navedeno je:

> „Ova se Uredba primjenjuje na obradu osobnih podataka koja se u cijelosti obavlja automatizirano te na neautomatiziranu obradu osobnih podataka koji čine dio sustava pohrane ili su namijenjeni biti dio sustava pohrane.“ [4]

U Poglavlju i. također je jednoznačno navedeno da Uredba odnosi na sve subjekte unutar Europske Unije neovisno o tome vrši li se obrada podataka unutar same Europske Unije te na sve osobne podatke ispitanika unutar Europske Unije, ako se koriste za nuđenje roba ili usluga ili, ako se koriste za praćenje njihovog ponašanja unutar Europske Unije. Poglavlje i. završava s 28 definicija korištenih termina, a za potrebe ovoga diplomskog rada izdvojeno je sljedećih 9:

1) **„osobni podaci”** znači svi podaci koji se odnose na pojedinca čiji je identitet utvrđen ili se može utvrditi („ispitanik”); pojedinac čiji se identitet može utvrditi jest osoba koja se može identificirati izravno ili neizravno, osobito uz pomoć identifikatora kao što su ime, identifikacijski broj, podaci o lokaciji, mrežni identifikator ili uz pomoć jednog ili više čimbenika svojstvenih za fizički, fiziološki, genetski, mentalni, ekonomski, kulturni ili socijalni identitet tog pojedinca; „ [4]

2) **„obrada”** znači svaki postupak ili skup postupaka koji se obavljaju na osobnim podacima ili na skupovima osobnih podataka, bilo automatiziranim bilo neautomatiziranim sredstvima kao što su prikupljanje, bilježenje, organizacija, strukturiranje, pohrana, prilagodba ili izmjena, pronalaženje, obavljanje uvida, uporaba, otkrivanje prijenosom, širenjem ili stavljanjem na raspolaganje na drugi način, usklađivanje ili kombiniranje, ograničavanje, brisanje ili uništavanje;“ [4]

3) **„ograničavanje obrade”** znači označivanje pohranjenih osobnih podataka s ciljem ograničavanja njihove obrade u budućnosti; “ [4]

4) **„voditelj obrade”** znači fizička ili pravna osoba, tijelo javne vlasti, agencija ili drugo tijelo koje samo ili zajedno s drugima određuje svrhe i sredstva obrade osobnih podataka; kada su svrhe i sredstva takve obrade utvrđeni pravom Unije ili pravom države članice, voditelj obrade ili posebni kriteriji za njegovo imenovanje mogu se predvidjeti pravom Unije ili pravom države članice; „ [4]

5) **„izvršitelj obrade”** znači fizička ili pravna osoba, tijelo javne vlasti, agencija ili drugo tijelo koje obrađuje osobne podatke u ime voditelja obrade; „ [4]

6) **„treća strana”** znači fizička ili pravna osoba, tijelo javne vlasti, agencija ili drugo tijelo koje nije ispitanik, voditelj obrade, izvršitelj obrade ni osobe koje su ovlaštene za obradu osobnih podataka pod izravnom nadležnošću voditelja obrade ili izvršitelja obrade; [4]

7) **„privola”** ispitanika znači svako dobrovoljno, posebno, informirano i nedvosmisleno izražavanje želja ispitanika kojim on izjavom ili jasnom potvrdnom radnjom daje pristanak za obradu osobnih podataka koji se na njega odnose; [4]

8) **„povreda osobnih podataka”** znači kršenje sigurnosti koje dovodi do slučajnog ili nezakonitog uništenja, gubitka, izmjene, neovlaštenog otkrivanja ili pristupa osobnim podacima koji su preneseni, pohranjeni ili na drugi način obrađivani; [4]

9) **„nadzorno tijelo”** znači neovisno tijelo javne vlasti koje je osnovala država članica u skladu s člankom 51.; [4]

### 1.1.2. Načela

U ovom poglavlju definiraju se generalna načela obrade podataka. Osobni podaci Moraju biti zakonito i transparentno obrađivani te ih se smije obrađivati samo u svrhu koja je bila ranije naznačena. Mora se voditi računa o točnosti i ažurnosti podataka kao i o načelu smanjenja količine podataka koje kaže da podaci moraju biti ograničeni samo na ono što je nužno i ništa više od toga. Načelo ograničenja pohrane ograničava oblik podataka na način da se pomoću njih može identificirati ispitanika samo određen vremenski period. Svi podaci moraju biti obrađivani na siguran način koji osigurava da neće doći do uništenja, gubitka ili oštećenja podataka uslijed organizacijskih ili tehničkim mjera te kada god je to moguće moraju biti enkriptirani i pseudoanonimizirani.

Prilikom prikupljanja podataka na webu, u većini slučajeva, obrada podataka postaje zakonita tek onoga trenutka kada:

1. „ispitanik je dao privolu za obradu svojih osobnih podataka u jednu ili više posebnih svrha;“ [4]

U slučaju specifičnih situacija kao što je sklapanje ugovora ili izvršavanja zadaća javnog interesa privola nije potrebna. Pritom je svakoj državi članici ostavljena mogućnost postavljanja dodatnih uvjeta zakonitosti obrade. Članak 7 navodi da privola mora biti jasno razlučiva od ostalih pitanja te napisana jednostavnim jezikom. Pritom mora vrijediti da:

> „Kada se obrada temelji na privoli, voditelj obrade mora moći dokazati da je ispitanik dao privolu za obradu svojih osobnih podataka.“ [4]

Privola mora biti dana dragovoljno te ispitanik u svakom trenutku mora moći povući privolu. Pritom vrijedi da je obrada podataka koja je nastupila prije povlačenja privole i dalje zakonita neovisno o povlačenju privole.

Posebna pravila vrijede za obradu podataka djece i maloljetnika. Privola koja je dana dobrovoljno ne mora biti zakonita. To vrijedi u situaciji kada ju daje osoba mlađa od 16 godina. Da bi obrada podataka osobe mlađe od 16 godina bila zakonita potrebno je dobiti odobrenje od roditelja ili staratelja maloljetnika. Državama članicama dana je sloboda na dodatno spustiti dobnu granicu, ali ona ne smije biti manja od 13 godina, a u Republici Hrvatskoj ona iznosi 16 godina. Od voditelja obrade se očekuje da uloži pravovaljane napore u provjeru je li privola od osoba mlađih od 16 godina uistinu odobrena od strane roditelja.

U poglavlju ii. unutar Članka 9. istaknuta su posebna pravila prilikom obrade posebnih kategorija osobnih podataka te se izričito navodi:

> „Zabranjuje se obrada osobnih podataka koji otkrivaju rasno ili etničko podrijetlo, politička mišljenja, vjerska ili filozofska uvjerenja ili članstvo u sindikatu te obrada genetskih podataka, biometrijskih podataka u svrhu jedinstvene identifikacije pojedinca, podataka koji se odnose na zdravlje ili podataka o spolnom životu ili seksualnoj orijentaciji pojedinca.“ [4]

Da bi se ozakonila obrada takve vrste podataka ispitanik mora dati izričitu privolu za obradu istih. Izuzete su situacije kada je obrada od životnog interesa ispitanika (primjerice posebne situacije obrade genetskih podataka u medicinske svrhe), a ispitanik nije u mogućnosti dati privolu za sebe ili drugu osobu. Također su izuzete situacije kada je obrada nužna za ostvarenje pravnih zahtjeva, za provođenje strategija preventivne medicine, javnog interesa u području zdravlja i sl. Dodatno, članak 10. definira obradu osobnih podataka koji se odnose na kaznene osude i kažnjiva djela te navodi:

> „Obrada osobnih podataka koji se odnose na kaznene osude i kažnjiva djela ili povezane mjere sigurnosti na temelju članka 6. stavka 1. provodi se samo pod nadzorom službenog tijela ili kada je obrada odobrena pravom Unije ili pravom države članice kojim se propisuju odgovarajuće zaštitne mjere za prava i slobode ispitanikâ. Svaki sveobuhvatni registar kaznenih osuda vodi se samo pod nadzorom službenog tijela vlasti.“ [4]

Obrada u određenim situacijama ne zahtijeva identifikaciju. U takvim situacijama voditelj odrade nije dužan identificirati ispitanika. Ako je voditelj obrade u stanju dokazati da nije u mogućnosti identificirati ispitanika, dužan je o tome obavijestit ispitanika, ako postoji takva mogućnost. Ako je obveza identificiranja ispitanika čiji su se podaci obradili prestala važiti, voditelj obrade nije dužan zadržavati ili obrađivati dodatne informacije u vrhu identificiranja ispitanika, ako je jedina svrha tih radnji poštivanje Opće uredbe o zaštiti podataka.

### 1.1.3. Prava ispitanika

Poglavlje iii. posebno je zanimljivo za ovaj rad. U ovome poglavlju su definira sva prava ispitanika čiji se podaci obrađuju. Članak 13 definira koje sve informacije je voditelj obrade dužan pružiti ispitaniku u trenutku prikupljanja osobnih podataka da bi obrada bila zakonita:

> „(a) identitet i kontaktne podatke voditelja obrade i, ako je primjenjivo, predstavnika voditelja obrade;
>
> (b) kontaktne podatke službenika za zaštitu podataka, ako je primjenjivo;
>
> (c) svrhe obrade radi kojih se upotrebljavaju osobni podaci kao i pravnu osnovu za obradu;
>
> (d) ako se obrada temelji na članku 6. stavku 1. točki (f), legitimne interese voditelja obrade ili treće strane;
>
> (e) primatelje ili kategorije primatelja osobnih podataka, ako ih ima; i
>
> (f) ako je primjenjivo, činjenicu da voditelja obrade namjerava osobne podatke prenijeti trećoj zemlji ili međunarodnoj organizaciji te postojanje ili nepostojanje odluke Komisije o primjerenosti, ili u slučaju prijenosâ iz članaka 46. ili 47. ili članka 49. stavka 1. drugog podstavka upućivanje na prikladne ili odgovarajuće zaštitne mjere i načine pribavljanja njihove kopije ili mjesta na kojem su stavljene na raspolaganje.“ [4]

Voditelj obrade također je dužan obznaniti ispitaniku vremenski rok u kojem će osobni podaci biti pohranjeni te ukazati mu da u svakome trenutku ima pravo zatražiti pristup osobnim podacima, odnosno ispravak ili brisanje prikupljenih podataka. Voditelj obrade mora ispitaniku obznaniti da ima pravo podnijeti prigovor odgovarajućem nadzornom tijelu te postojanje bilo automatiziranog donošenja odluka, ako se isto koristi. Sva navedena prava vrijede i u situacijama kada podaci nisu dobiveni direktno od strane ispitanika.

Ispitanik također ima pravo na prigovor koje podrazumijeva mogućnost podnošenja prigovora u posebnim situacijama kada postoje uvjerljivi razlozi u svrhu ostvarivanja prava i slobode ispitanika ili u svrhu obrane pravnih zahtjeva. Ako se podaci obrađuju u svrhu marketinga, ispitanik uvijek ima pravo priložiti prigovor. Države članice mogu ograničiti opseg obaveza voditelja obrade ako one predstavljaju opasnost po nacionalnu i javnu sigurnost te ako zabrana osigurava zaštitu demokratskog društva.

### 1.1.4. Voditelj obrade i izvršitelj obrade

U poglavlju iv. preciziraju se prava i obveze voditelja odrade i izvršitelja obrade. Poglavlje definira standarde tehničke i integrirane zaštite podataka koji podrazumijevaju da voditelj vodi računa o modernim tehnikama zaštite podataka te provodi odgovarajuće organizacijske mjere. U posebnim situacijama kada postoje dva ili više voditelja obrade koji se dogovore oko svrhe i načina obrade podataka oni se smatraju zajedničkim voditeljima obrade. U tom slučaju se dogovorom može odrediti kontaktna točka za ispitanike obrade, ali ispitanik mora moći ostvariti sva svoja prava kod svih voditelja koji sudjeluju u obradi podataka.

Kada voditelj odrade ne provodi obradu podataka direktno u svoje ime, već za to koristi jednog ili više izvršitelja odrade za izvršitelja/e obrade vrijedi da moraju poštovati jednaku razinu tehničke i integrirane zaštite podataka. Da bi obrada podataka od strane izvršitelja obrade bili zakonita, voditelj obrade i izvršitelj obrade moraju obradu definirati ugovorom ili drugim pravnim aktom. Izvršitelj obrade ne smije obradu podataka prepustiti trećoj strani odnosno:

> „Izvršitelj obrade ne smije angažirati drugog izvršitelja obrade bez prethodnog posebnog ili općeg pisanog odobrenja voditelja obrade. U slučaju općeg pisanog odobrenja, izvršitelj obrade obavješćuje voditelja obrade o svim planiranim izmjenama u vezi s dodavanjem ili zamjenom drugih izvršitelja obrade kako bi time voditelju obrade omogućio da uloži prigovor na takve izmjene.“ [4]

U svrhe posebnih aktivnosti obrade, izvršitelju obrade se iznimno dopušta angažiranje drugog izvršitelja obrade, ali odgovornost za sve propuste ostaje na originalnom izvršitelju obrade imenovanom od strane voditelja obrade. Svaki voditelj obrade ili izvršitelj voditelja obrade dužan je voditi detaljnu evidenciju o kategorijama podataka koje obrađuje, svrsi obrade, opisima i kategorijama ispitanika obrade te kategorijama primatelja kojima su podaci otkriveni. Prilikom svake obrade mora biti evidentirano ime voditelja obrade te njegovi kontakt podaci.

Kada se procijeni da je postoji visoka razina rizika da bi obrada podataka mogla prouzročiti povrede prava i slobode pojedinca, voditelj obrade dužan je provesti procjenu učinaka predviđenih postupaka zaštite osobnih podataka. Svi postupni procjene rizika moraju biti javno objavljeni od strane nadzornog tijela. Voditelj obrade i izvršitelj obrade u

svakome trenutku dužni su surađivati s nadzornim tijelima ako ista pošalju zahtjev. Pritom je posebno važno da:

> „U slučaju povrede osobnih podataka voditelj obrade bez nepotrebnog odgađanja i, ako je izvedivo, najkasnije 72 sata nakon saznanja o toj povredi, izvješćuje nadzorno tijelo nadležno u skladu s člankom 55. o povredi osobnih podataka, osim ako nije vjerojatno da će povreda osobnih podataka prouzročiti rizik za prava i slobode pojedinaca. Ako izvješćivanje nije učinjeno unutar 72 sata, mora biti popraćeno razlozima za kašnjenje.“ [4]

Osim nadzornog tijela, voditelj obrade je u slučaju povrede osobnih podataka dužan obavijestiti i pojedinca čiji se podaci obrađuju također bez vremenske odgode. U slučaju kada se procijeni da ako voditelj obrade podataka ne primijeni određenu mjeru ublažavanja rizika da je moguće da će nastupiti situacija visokog rizika obrade podataka, voditelj obrade dužan je savjetovati se s nadzornim tijelom kako bi se osigurala maksimalna zaštita osobnih podatka. Nadzorno tijelo dužno je savjetovati voditelja obrade pisanim putem unutar 8 tjedana od zaprimanja zahtjeva.

U slučajevima kada obradu osobnih podataka provodi tijelo javne vlasti ili javno tijelo, obrada podataka zahtjeva redovito praćenje ispitanika ili se prilikom obrade u velikoj mjeri postupa s podacima koji se tiču kaznenih djela, voditelj obrade i izvršitelj obrade dužni su imenovati službenika za zaštitu podataka. Ispitanici čiji se osobni podaci obrađuju, u svakome trenutku, imaju pravo kontaktirati službenika za zaštitu podataka u svrhu ostvarivanja svojih prava. Službenik za zaštitu podataka mora biti pravovremeno upućen u sva pitanja osobnih podataka od voditelja obrade i izvršitelja obrade te su mu dužni osigurati sva potrebna sredstva za izvršavanje njegove dužnosti.

Opća uredba o zaštiti podataka propisuje i najmanje zadaće koje službenik za zaštitu podataka mora obavljati te precizira da je službenik za zaštitu podatak dužan informirati i savjetovati voditelja obrade ili izvršitelja obrade te njegove zaposlenike o njihovim obvezama koje pokriva Opća uredba o zaštiti podataka. Kada god je to potrebno, službenik za zaštitu osobnih podataka dužan je pružiti savjet te surađivati s nadzornim tijelom. Za kraj se navodi da je službenik za zaštitu podataka dužan voditi računa o postojećem riziku prilikom obrade podataka te uzeti u obzir kontekst, opseg i prirodu obrade ispitanikovih osobnih podataka.

### 1.1.5. Prijenosi osobnih podataka trećim zemljama ili međunarodnim organizacijama, Neovisna nadzorna tijela, Suradnja i konzistentnost i Pravna sredstva, odgovornost i sankcije

Sljedeća četiri poglavlja Opće uredbe o zaštiti podataka odnose se na teme prijenosa osobnih podataka trećim zemljama ili međunarodnim organizacijama, radu neovisnih nadzornih tijela te međusobnoj suradnji i konzistentnosti istih te pravnih sredstava i sankcija za prekršitelje. Poglavlje v. definira načela prijenosa zemljama trećeg svijeta te definira da je prijenos podataka zakonit samo ako se događa iz primjerenih razloga, a obrada podataka i dalje mora biti sa svim načelima Uredbe. Ovo poglavlje je bitno jer osigurava zaštitu podataka građana Europske Unije i izvan granica Europske Unije te sprječava zaobilaženje Opće uredbe o zaštiti podataka jednostavnim prebacivanjem obrade van njezinih granica.

Svaka država članica dužna je imati barem jedno neovisno nadzorno tijelo koje prati primjene Opće uredbe o zaštiti podataka prateći naputke u Poglavlju vi. U Hrvatskoj tu dužnost izvršava AZOP(Agencija za zaštitu osobnih podataka). Agencija za zaštitu podataka započela je sa svojim radom 2004. godine, a njezin zadatak je promicanje javne svijesti o pravilima i zaštitnim mjerama osobnih podataka kao i savjetovanje državnih institucija u vezi sa zaštitom prava ispitanika čiji se podaci obrađuju. Agencija je dužna rješavati pritužbe dobivene od strane tijela, organizacija ili ispitanika obrade te izvještavati o napretku i ishodu istrage, a u slučaju kršenja Opće uredbe o zaštiti podataka ima pravo privremeno ili trajno ograničiti ili zabraniti obradu podataka, izdati službenu opomenu voditelju obrade ili izvršitelju obrade, narediti prekid protoka podataka prema zemljama trećeg svijeta te, u slučajevima teškog kršenja Uredbe, izreći novčanu kaznu prekršiteljima.

Na razini Hrvatske djeluje AZOP, a na razini Europske Unije te Norveške, Lihtenštajna i Islanda djeluje EDPO (engl. *European Data Protection Bord*) - Europski odbor za zaštitu podataka sa sjedištem u Bruxellesu. Zadaća odbora je pružati opće smjernice prilikom tumačenja Uredbe, donošenje zaključaka o dosljednosti u predmetima koji se odnose na dvije ili više zemalja članica te savjetovanje Europske komisije o pitanjima zaštite osobnih podataka. [5] Europski odbor za zaštitu podataka donosi i smjernice te savjetuje nacionalna nadzorna tijela po pitanju načela međusobne suradnje i konzistentnosti istaknutima u Poglavlju vii. Opće uredbe o zaštiti podataka. U situacijama kršenja Uredbe svaki ispitanik ima pravo podnijeti pritužbu odgovarajućem nadzornom tijelu te može

ovlastiti neprofitno tijelo ili organizaciju da podnese pritužbu u njegovo ime. Smjernice za kažnjavanje prekršitelja istaknute su u poglavlju viii. Slučajevi kršenja Opće uredbe o zaštiti podataka i iznosi kazni izrečenih prekršiteljima navedeni su u poglavlju 1.2. ovoga rada.

### 1.1.6. Odredbe u vezi s posebnim situacijama obrade, Delegirani akti i provedbeni akti i Završne odredbe

Posljednja tri poglavlja Opće Uredbe o zaštiti podataka tiču se odredba u vezi s posebnim situacijama obrade te delegiranim i provedbenim aktima. Termin posebne situacije označava situacije kada dolazi do konflikta javnog dobra i prava na zaštitu osobnih podataka. Primjer takve situacije su podaci koji se iznose kako bi se očuvala sloboda izražavanja i informiranja te članak 85 unutar poglavlja ix. navodi:

> „Države članice zakonom usklađuju pravo na zaštitu osobnih podataka u skladu s ovom Uredbom s pravom na slobodu izražavanja i informiranja, što uključuje obradu u novinarske svrhe i svrhe akademskog, umjetničkog ili književnog izražavanja.“ [4]

Posebna pravila vrijede i u kontekstu zaposlenja i vjerskih zajednica. U kontekstu zaposlenja obrada podataka je zakonita i bez privole, ali pritom mora vrijediti da poštuje legitimne interese i temeljna prava ispitanika te brine o zaštiti ljudskog dostojanstva ispitanika. U kontekstu vjerskih zajednica Uredba dopušta primjenu protokola i pravila o zaštiti podataka korištenih od strane crkve i drugih vjerskih zajednica dokle god su ona usklađena s istom.

Završna dva poglavlja definiraju izvršavanje delegiranja ovlasti te odnos direktive s drugim direktivama i sporazumima na razini Europske Unije. Članak 92. navodi da se:

> „Ovlast za donošenje delegiranih akata dodjeljuje se Komisiji podložno uvjetima utvrđenima u ovom članku..“ [4]

te da joj se ovlasti dodjeljuju na neodređeno vrijeme počevši od 24. svibnja 2016. godine. Tijela koja u svakome trenutku može opozvati ovlasti su Europski parlament ili Vijeće te je Komisija dužna istovremeno priopćiti Europskom parlamentu i Vijeću svaki doneseni delegirani akt. Delegirani akt stupa na snagu samo ako Vijeće i Europski parlament, unutar tri mjeseca, ne podnesu nikakav prigovor ili izvijeste Komisiju da do prigovora neće doći.

Prije Opće uredbe o zaštiti podataka primjenjivala se Direktiva 95/46/EZ o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom protoku takvih podataka. Stupanjem Opće uredbe o zaštiti podatka ona prestaje vrijediti što je navedeno u članku 94:

„Direktiva 95/46/EZ stavlja se izvan snage s učinkom od 25. svibnja 2018.“ [4]

Uredba završava člankom 99 koji precizira da se Opća uredba o zaštiti podataka počinje primjenjivati 25. svibnja 2018. godine.

## 1.2. Kazne za kršenje Opće uredbe o zaštiti podataka

Postoje dva glavna razreda kazni za kršenje GDPR-a. Kazne su podijeljene s obzirom na godišnja primanja kompanija. Svrha podjele kazni u dva zasebna razreda istovremeno sprečavanje budućeg kršenja Uredbe, a opet pripaziti na težinu kazne kako ne bi došlo do gušenja gospodarstva Europske Unije. Također uzima se u obzir način i opseg kršenja Opće uredbe o zaštiti podataka te sukladno situaciji primjenjuje različiti iznos kazne. Kazne su izražene na dva načina: kao fiksni iznos ili kao postotak godišnjeg prihoda organizacije.

Prvi razred kazni predstavljaju kazne koje su do 10 mil. eura ili 2% svih godišnjih prihoda organizacije, što god od navedenog je veće. Ovaj razred kazni uglavnom je namijenjen manjim kompanijama. Tipični razlog za izricanje ovog razreda kazni su uzastopni manji prekršaji. Primjerice, organizacija može biti kažnjena do 10 mil. eura ili 2% svih godišnjih prihoda u situacijama kada je utvrđeno da unutar organizacije nije imenovan službenik za zaštitu podataka ili prilikom uzastopnih propusta u održavanju točne evidencije.

Drugi razred kazni predstavljaju kazne koje su do 20 milijuna eura ili 4% svih godišnjih prihoda organizacije, što god od navedenog je veće. Ovaj razred kazni uglavnom je namijenjen multinacionalnim kompanijama. Kazne iz ovog razreda izriču se u situacijama ozbiljnih kršenja Opće uredbe o zaštiti podataka. Primjerice, kada se dokaže svjesna zloporaba podataka ili u situacijama kada dolazi do proboja podataka od treće strane.

Do 2024. godine dvije kompanije kojima su izrečene najveće kazne su Meta i Amazon. U svibnju 2023. godine, donesena je odluka od strane Irske komisije za zaštitu podataka (DPC) o kažnjavanju kompanije Meta. Razlog kažnjavanja je bila nedostatna razina zaštite podataka europskih korisnika prilikom transfera podataka u Sjedinjena Američke Države. Postignut je novi rekord u iznosu kazne te je Meta kažnjena iznosom od 1.2 mlrd. eura. U svibnju 2021. godine Luksemburška nacionalna komisija za zaštitu podataka (CNDP) kaznila je Amazon s iznosom od 746 mil. eura. Nakon što je više od 10000 osoba podnijelo prijavu, provedena je istraga te je utvrđeno da Amazonov sustav za ciljano oglašavanje na neregularan način obrađuje podatke svojih kupaca. Kazna izrečena Amazonu je druga najveća kazna po iznosu koja je nekoj organizaciji izrečena radi kršenja GDPR-a.

## 1.3. Odabrane točke Opće uredbe o zaštiti podataka

U poglavlju 1.1 dan je sažeti prikaz Opće uredbe o zaštiti podataka. Uredba je opsežan dokument koji definira pravila koja moraju poštovati voditelj obrade ili izvršitelj obrade, prava ispitanika obrade te zadaće nadzornih tijela, ali nisu sve navedena pravila i smjernice prikladne za automatiziranu obradu. Za automatiziranu provjeru koja se provodi u ovome diplomskom radu izabrano je 5 pravila definiranih unutar Uredbe za koja je procijenjeno da su pogodna za analizu. Pregled odabranih pravila dan je u tablici 1.1:

Tablica 1.1 Odabrana pravila GDPR-a pogodna za automatiziranu analizu

| Pravilo | Primjer kršenja pravila | Predloženi način provjere sukladnosti politike privatnosti s navedenim pravilom |
|---|---|---|
| Politika privatnosti mora sadržavati kontakt voditelja obrade | Unutar sadržaja politike privatnosti nije naveden kontakt voditelja obrade | Pretraga za kontakt podacima voditelja obrade |
| Ispitaniku mora biti obaviješten o pravu na povlačenje privole | Ispitaniku nije obznanjeno pravo na povlačenje privole | Pretraga za odlomkom koji obznanjuje ispitaniku pravo na povlačenje privole |
| Voditelj obrade dužan je navesti svrhu obrade podataka | Svrha obrade podataka nije navedena | Pretraga za odlomkom koji navodi svrhu obrade |
| Voditelj obrade dužan je navesti kategoriju osobnih podataka koji se obrađuju | Kategorija osobnih podataka koji se obrađuju nije navedena | Pretraga za odlomkom koji navodi kategoriju osobnih podataka koji se obrađuju |
| Ispitanik mora biti obaviješten o pravu na brisanje i ispravak prikupljenih podataka. | Ispitaniku nije obznanjeno pravo na brisanje i ispravak prikupljenih podataka. | Pretraga za odlomkom koji obznanjuje ispitaniku pravo na brisanje i ispravak prikupljenih podataka |

Opis korištenih metoda prilikom analize dan je u poglavlju 5, a dobiveni rezultati i komentari analize nalaze se u poglavlju 6.

# 2. Sustav za dohvaćanje internetskih stranica

*Web scraping* je sistematizirano dohvaćanje sadržaja (u medijskom ili tekstualnom obliku) s web stranica. Cilj *web scrapinga* je na automatizirani način dohvatiti podatke s ciljanog skupa web stranica. To se obično postiže upotrebom programske potpore u obliku programskih alata ili skripti koje pristupaju web stranicama a nazivaju se *web scraperi*. Korištena programska potpora identificira relevantne podatke, izdvaja ih, strukturira u pogodnom formatu te ih pohranjuje za daljnju upotrebu. Najčešći formati u kojima se pohranjuju prikupljeni podaci su JSON i CSV, ali ovisno o primjeni može se odlučiti i za druge oblike pohrane. Na taj način omogućuje se značajna ušteda vremena, jer je cijeli proces automatiziran, te se olakšava analiza podataka.

*Web scraping* se može koristiti u čitavom nizu područja, a trenutno najčešće pronalazi svoju primjenu u području obrade prirodnog jezika (engl. *Natural Language Processing – NLP*). Druga popularna primjena *web scrapinga* je prilikom istraživanja tržišta. Kompanije se odlučuju na automatizirano prikupljanje podataka o svojim konkurentima kako bi u stvarnom vremenu mogle lakše donositi kvalitetne odluke. *Web scraping* svoju primjenu također pronalazi i u području monitoringa. Primjerice, može se koristiti za praćenje sadržaja na društvenim mrežama što kompanijama i poduzećima olakšava analizu korisnika koji koriste njihove usluge ili ih spominju u svojim objavama i komentarima.

Jedan od alata koji se koristi za generalnu primjenu *web scrapinga* je biblioteka Beautiful Soup. Beautiful Soup je Python biblioteka koja je dizajnirana na način da je intutitivna i jednostavna za korištenje. Autor biblioteke je Leonard Richardson, a namijenio ju je za dohvaćanje podatka, rudarenje podataka i *web scraping*. Pogodna je za primjenu prilikom korištenja strojnog učenja jer osim što dohvaća podatke istovremeno ih i čisti i formatira te na taj način ubrzava proces treniranja modela strojnog učenja.

Za potrebe ovog diplomskog rada korištena je slična Python biblioteka Polipy. PoliPy je Python biblioteka specijalizirana za *web scraping* politika privatnosti s web stranica. Kako je to sastavni dio ovog rada dana joj je prednost nad bibliotekama za generalnu primjenu *web scrapinga*. Funkcionalnosti biblioteke opisane su u sljedećem poglavlju.

## 2.1. PoliPy

PoliPy je biblioteka razvijena u programskom jeziku Python od strane istraživača sa sveučilišta Berkeley kao projekt odjela Berkeley Lab for Usable and Experimental Security (BLUES) [6]. Biblioteka pruža sučelje unutar komandne linije kao i API koji se može koristiti za analizu, web scraping i parsiranje politika privatnosti od interesa. Polipy biblioteka može se jednostavno instalirati korištenjem naredbe `pip install polipy`.

Nakon instalacije biblioteka se može koristiti na dva načina. Prvi način je preko komandne linije. Primjerice ako želimo dohvatiti politiku privatnosti s web stranice https://docs.github.com/en/github/site-policy/github-privacy-statement to se može učiniti sa sljedećih nekoliko linija koda:

```
$ cat policies.txt

https://docs.github.com/en/github/site-policy/github-privacy-
statement(link is external)

$ polipy policies.txt -s
```

Kôd 2.1 – Dohvaćanje politike privatnosti korištenjem naredbenog retka

Drugi način je pokretanjem slijedeće Python skripte:

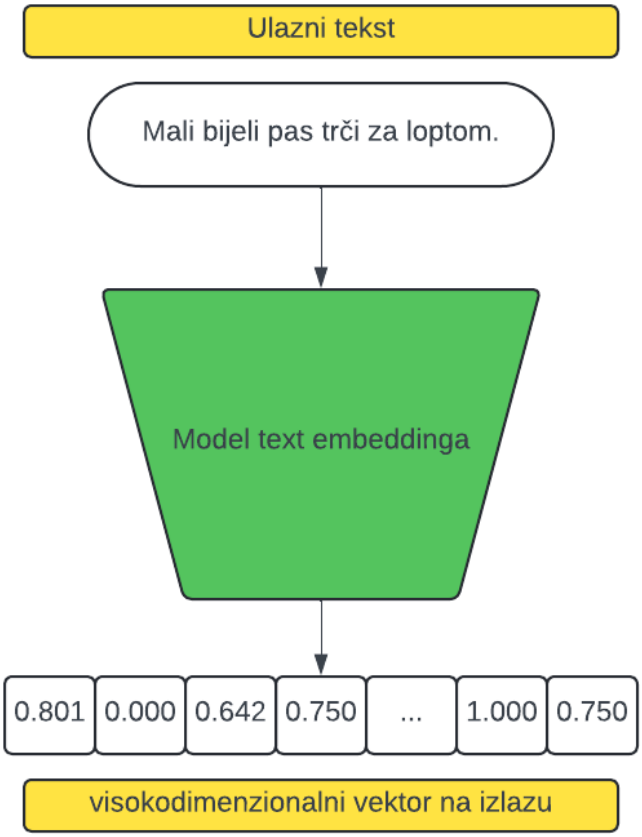
```
import polipy

url = 'https://docs.github.com/en/github/site-policy/github-privacy-
statement(link is external)'

result = polipy.get_policy(url, screenshot=True)

result.save(output_dir='.')
```

Kôd 2.2 – Dohvaćanje politike privatnosti pokretanjem Python skripte

U oba slučaja rezultat je jednak. U odredišnom direktoriju generiraju se četiri datoteke: .html datoteka koja sadržava izvorni kod web stranice na kojoj se nalazi politika privatnosti, .meta datoteka koja sadrži meta podatke, .png datoteka koja sadrži snimak zaslona web stranice u trenutku kada je došlo do prikupljanja podataka i .json datoteku u kojoj je pohranjen sam sadržaj politike privatnosti u tekstualnom obliku. Za daljnju analizu najbitnija je .json datoteka jer se unutar nje nalazi tekst koji će se očistiti i razdvojiti po paragrafima. Nakon toga će se svaki paragraf pretvoriti u vektor korištenjem tehnika vektorske reprezentacije riječi koje su objašnjene u poglavlju 3. Jednom kada podaci poprime vektorski oblik spremni su za uspoređivanje i daljnju analizu korištenjem metoda strojnog učenja

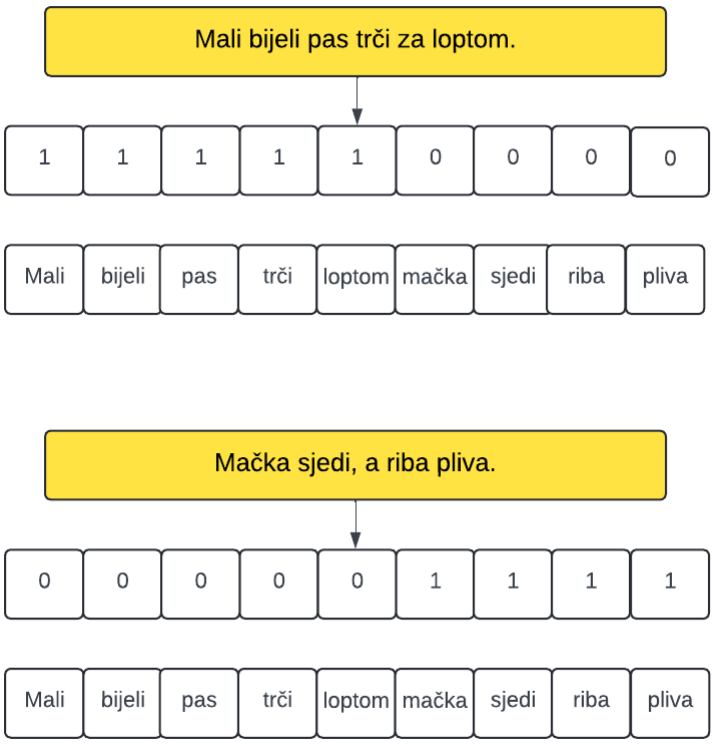
# 3. Vektorska reprezentacija teksta

Vektorska reprezentacija(engl. e*mbedding)* je proces stvaranja prikaza određenog objekta pomoću vektora. Objekt od interesa može na primjer biti slika, tekst, zvuk ili video. Vektorska reprezentacija riječi (engl. *text embedding*) je vektorski prikaz određene riječi, rečenice ili pak cijeloga paragrafa. Ideja je ulazni tekst pretvoriti u izlazni vektor fiksne dimenzije. Model koji se koristi za vektorsku reprezentaciju riječi najčešće koristi neuronske mreže kako bi generirao izlazni vektor. Izlazni vektor je niz decimalnih brojeva pomoću kojih se pokušava opisati ulazni tekst, a tipičan vektor ima preko 1000 dimenzija. Glavni cilj vektorske reprezentacije riječi je pretvoriti semantičko i sintaktičko značenje teksta u numerički oblik pogodan za uspoređivanje i analizu korištenjem strojnog učenja. Shematski prikaz vektorske reprezentacije riječi dan je niže na ovoj stranici na slici 3.1.




Sl. 3.1 Shematski prikaz procesa vektorske reprezentacije riječi: ulazni tekst „Mali bijeli pas trči za loptom.“ se korištenjem modela za vektorsku reprezentaciju riječi pretvara u izlazni vektor koji je sačinjen od niza decimalnih brojeva

Postoje razne tehnike vektorske reprezentacije riječi. Jedna od njih je reprezentacija teksta kao vreće riječi(engl. *bag-of-words*). Cilj tehnike je razlomiti tekst na riječi te prebrojati koliko često se koja riječ pojavljuje u tekstu. Prije prebrojavanja tekst se čisti od zaustavnih riječi (eng. *stopwords*) koje imaju gotovo isključivo gramatičku funkciju, ili se često javljaju u određenom jeziku te vrlo malo doprinose semantičkom značenju. U hrvatskom jeziku to su primjerice riječi poput „a“, „je“, „i“, „ni“, „to“ i „ta“. U engleskom jeziku to su riječi poput „the“, „a“, „an“, „be“, „as“, „to“ i „just“. Ne postoji samo jedan rječnik zaustavnih riječi te ovisno o primjeni koriste se različiti rječnici. Primjer rječnika koji se koriste za rad s tekstovima na engleskom jeziku su iSpell i Snowball, a u ovome radu se koristi rječnik koji je sastavni dio biblioteke Nltk. Također, iz teksta se izbacuju dijakritički znakovi te se ignorira razlika između velikih i malih slova. Primjerice za paragraf „Mali bijeli pas trči za loptom. Mačka sjedi, a riba pliva“ koristio bi se rječnik od 9 riječi [„mali“, „bijeli“, „pas“, „trči“, „loptom“, „mačka“, „sjedi“, „riba“, „pliva“] jer korišteni rječnik zaustavnih riječi izbacuje sve prijedloge i veznike (riječi „za“ i „a“). Rečenica „Mali bijeli pas trči za loptom.“ se u slučaju tog rječnika može prikazati kao vektor [1 1 1 1 1 0 0 0 0], a rečenica „Mačka sjedi, a riba pliva“ kao vektor [0 0 0 0 0 1 1 1 1]. Oba vektora prikazani su na slici 3.2. te je ispod svakog vektora prikazan korišteni rječnik.




Sl. 3.2 Vektorski prikaz rečenice dobiven modelom vreća riječi

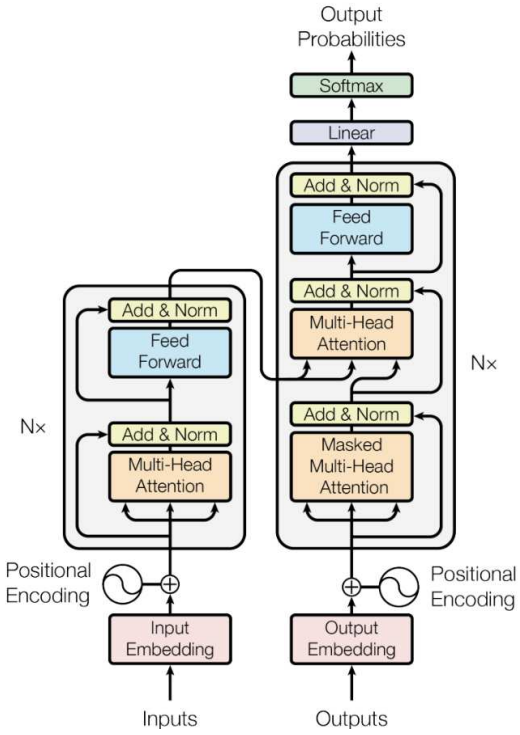
Model vreća riječi krasi jednostavnost implementacije i intuitivnost korištenja, ali generalnu upotrebu ne daje pretjerano dobre rezultate. U navedenom primjeru paragraf se podijelio u dvije rečenice: „Mali bijeli pas trči za loptom.“ i „Mačka sjedi, a riba pliva“. Te dvije rečenice imaju različito značenje, ali su utoliko slične što govore o životinjama koje obavljaju neke aktivnosti. Ako usporedimo vektore koje je proizveo model: [1 1 1 1 1 0 0 0 0] i [0 0 0 0 0 1 1 1 1], možemo primijetiti da smo dobili dva apsolutno komplementarna vektora. Takav odnos između dva vektora indicira značajno razlikovanje u sadržaju, a sadržaj početne dvije rečenice nije u takvom odnosu. Ovaj model je pogodan za analizu velikih dokumenata i pretraživanje duplikata. Mana modela je da u potpunosti ignorira poredak i odnose između riječi te ne prepoznaje sinonime kao riječi različitog oblika, ali istoga značenja. Da bi se doskočilo ovim problemima i postigla veća kvaliteta vektorske reprezentacije riječi razvijeni su napredniji modeli.

Moderni modeli koji se koriste za vektorsku reprezentaciju riječi mogu koristiti guste vektore (engl. *dense vectors*) ili rijetke vektore (engl. *sparse vectors*). Gusti vektori tipično se sastoje od nekoliko tisuća decimalnih brojeva među kojima je rijetko koja veličina postavljena na nulu, dok se rijetki vektori tipično sastoje od nekoliko desetaka tisuća decimalnih brojeva te je većina njihovih vrijednosti postavljena na nulu. Prednosti rijetkih vektora su efikasnije korištenje memorije, skalabilnost te manja osjetljivost na šum (budući da koriste manji broj značajki u odnosu na guste vektore), a glavna mana im je što uglavnom postižu manju vremensku efikasnost u odnosu na guste vektore. Prednosti gustih vektora su kompatibilnost s većinom biblioteka koje se koriste za strojno učenje te vremenska efikasnost, a mane su smanjena skalabilnost te veći memorijski zahtjevi u odnosu na rijetke vektore. Glavni kriterij odabira uglavnom se svodi na pitanje kompatibilnosti s korištenim algoritmima te odnos vremenske i memorijske efikasnosti. U ovome radu se eksperimentira i s jednom i s drugom vrstom vektora. U trenutku pisanja ovoga rada dva najnovija modela za vektorsku reprezentaciju riječi razvijena od strane kompanije OpenAI su „text-embedding-3-small“ model čija standardna veličina vektora (bez redukcije dimenzionalnosti) iznosi 1536 dimenzija i text-embedding-3-large model s veličinom vektora od 3072 dimenzija. Oba modela koriste guste vektore kojima je velika većina dimenzija različita od nule. U ovome radu korišten je model „text-embedding-3-small“, koji je detaljnije opisan u poglavlju 3.1., te model SPLADE++ koji koristi rijetke vektore veličine 30522 dimenzija a opisan je u poglavlju 3.2.

## 3.1. OpenAI text-embedding-3-small model

Model text-embedding-3-small je pušten u komercijalnu upotrebu 25. siječnja 2024. od strane američke tvrtke OpenAI. Koristi guste vektore veličine 1536 dimenzija[7]. Model je razvijen sa namjerom da se koristi za različite svrhe poput unaprjeđenja algoritama pretraživanja, poboljšanja kvalitete sustava preporuka, otkrivanje izdvojenica u podacima i klasifikacije tekstova što ga potencijalno čini dobrim izborom za potrebe ovoga rada. Testovi preciznosti pokazali su da text-embedding-3-small model postiže 62.3% prilikom testiranja kvalitete vektorske reprezentacije riječi na engleskom jeziku uz korištenje MTEB(engl. *Massive Text Embedding Benchmark*) skupa podataka, te 44.0% prilikom testiranja kvalitete embeddinga teksta koji se sastoji od više različitih jezika uz korištenje MIRACL(engl. *Multilingual Information Retrieval Across a Continuum of Languages*) skupa podataka. Iako podržava vektorsku reprezentaciju riječi nekolicine jezika, hrvatski jezik nije među njima te je za daljnji rad s politikama privatnosti na hrvatskom jeziku potrebno obaviti prijevod na engleski jezik. Za to je korištena Python biblioteka Translators [8]. Modelu se pristupa preko API-ja.

Model koristi neuronsku mrežu koja je bazirana na transformer arhitekturi. Transformer arhitektura prvi puta se pojavljuje u radu izdanom 2017. godine od strane kompanije Google pod naslovom „*Attention is all you need*“ [9], a prikazana je na slici 3.3:




Sl. 3.3 Transformer arhitektura, predstavljena u Google-ovom radu „*Attention is all you need*“ [9]

Prednosti u odnosu na tradicionalne modele vektorske reprezentacije riječi su što ova arhitektura uspijeva uhvatiti kontekst u kojemu se riječi i rečenice nalaze prilikom vektorizacije teksta, paralelnu obradu podataka te brže treniranje modela. Transformer arhitektura je podijeljena u dvije glavne cjeline prikazane na slici 3.3: enkoder prikazan lijevo te dekoder prikazan desno.

Svrha enkodera je jednom dobiveni skup tokena (engl. *input embedding*) iz originalnog ulaznog teksta(inputs) pretvoriti u vektor koji uzima u obzir i kontekst u kojemu se riječi nalaze. Dakle na ulazu dobiva tekst, a na izlazu vraća numerički vektor. Sastavljen je od 6 slojeva enkodera koji izvršavaju različite funkcije. Obrada ulaznog teksta izvodi se u četiri koraka. U prvome koraku tekst prolazi kroz prvi enkoder te se svaka riječ umeće u zasebni vektor. U drugome koraku se izvršava pozicijsko kodiranje (engl. *positional encoding*) te se vektori dobiveni u prvom koraku obogaćuju sa informacijom gdje se koja riječ nalazi unutar teksta. Treći korak se sastoji od nekoliko dijelova. Vektor sada mora proći kroz takozvani „*Multi-Headed Self-Attention*“ mehanizam koji omogućuje modelu ulazne riječi staviti u odnos. Vektor zatim prolazi kroz proces normalizacije te se izlaz iz ovoga sloja (*Add & Norm*) zbraja s ulazom u prvi enkoder. Idući blok „*Feed-Forward*“ zadužen je za dodatno podešavanje težina vektora pomoću neuronske mreže. Nakon što je vektor prošao kroz sve faze trećeg koraka, vektor se ponovno normalizira te se za kraj u četvrtom koraku prosljeđuje kao izlaz iz enkodera.

Izlaz iz enkodera predstavlja ulaz u dekoder. Ako promotrimo građu dekodera na slici 3.3, možemo primijetiti da se on sastoji od sličnih blokova kao i enkoder. Razlika su po jedan dodatni „*Multi-Headed Self-Attention*“ blok i „*Add & Norm*“ blok. Dakle prva dva koraka kod dekodera su isti kao kod enkodera. Nakon toga dolazi do zbrajanja izlaza iz dekodera i enkodera. U ovome koraku dolazi do dodatnog obogaćivanja vektora što omogućuje dekoderu identificirati ključne dijelove koji su nastali kao izlaz iz enkodera. Vektor zatim prolazi kroz 4 bloka koja su istovjetna onima u enkoderu te imaju istu ulogu. Predzadnji blok u enkoderu je linearni klasifikator (*Linear Classifier*) koji oblikuje izlaznu veličinu vektora. U modelu text-embedding-3-small to je 1536 dimenzija kolika je i veličina izlaznog rječnika linearnog klasifikatora. Za kraj se koristi aktivacijska funkcija softmax:

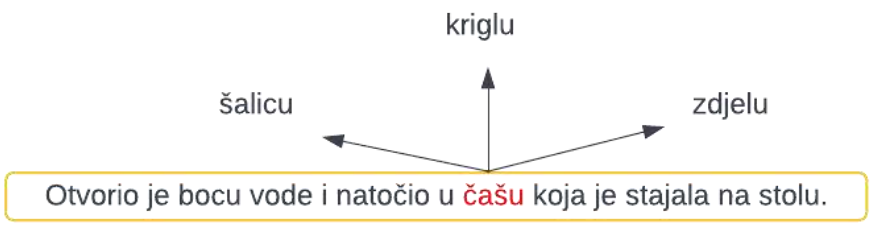
$$Softmax(xi) = \frac{e^{xi}}{\sum_j e^{xj}} \qquad (1)$$

Njezina uloga je normalizacija izlaznih veličina vektora na raspon od 0 do 1, gdje 1 predstavlja veličine od maksimalnog značaja, a 0 veličine koje su zanemarive.

## 3.2. SPLADE++ model

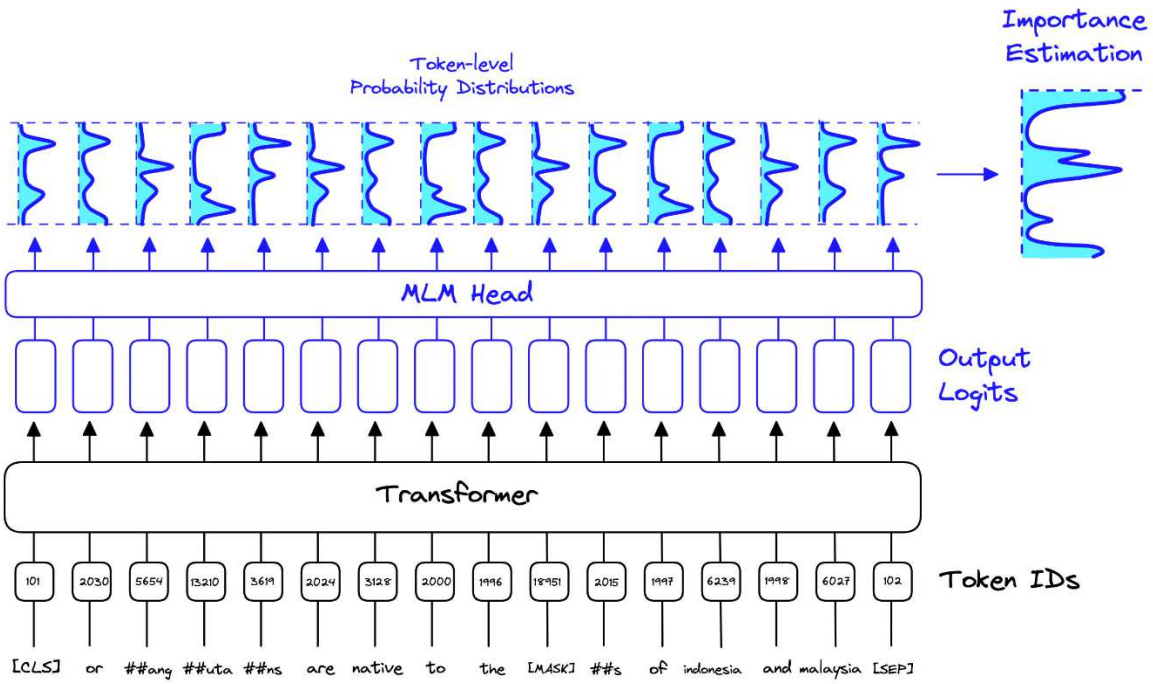
Za razliku od modela text-embedding-3-small opisanog u prethodnom poglavlju, model SPLADE++ koristi rijetke vektore. SPLADE++ je napredna verzija modela SPLADE (*Sparse Lexical and Expansion Model*). Model je primarno razvijen za efikasno pretraživanje informacija te je kao takav pogodan za automatiziranu analizu sukladnosti politika privatnosti. Kao i model text-embedding-3-small, SPLADE++ koristi transformer arhitekturu. SPLADE++ je originalno razvijenu svrhu analize i pretraživanja teksta. Kako je jedna od glavnih zadaća ovoga rada klasifikacija i dohvaćanje paragrafa koji dokazuju da je politika privatnosti sukladna s određenim pravilom iz Opće uredbe o zaštiti podataka, SPLADE++ se nameće kao potencijalni model kandidat koji bi tu zadaću mogao uspješno izvršavati.

Jedna od posebnosti ovoga modela je to što koristi BERT-ov (*Bidirectional Encoder Representations from Transformers*) MLM (*Masked Language Model*) sloj kojemu je glavna zadaća poboljšati leksičko indeksiranje tako što integrira skup proširenih semantičkih značajki u rijetki vektorski prostor. To se postiže procesom maskiranja gdje se jedna riječ maskira i pokušava zamijeniti s drugim riječima u svrhu boljeg razumijevanja konteksta u kojemu se riječ nalazi. Primjerice u rečenici „Otvorio je bocu vode i natočio u čašu koja je staja na stolu“ MLM sloj će maskirati riječ „čašu“ i pokušati predvidjeti, na osnovu ostalih riječi u rečenici, koja riječ iz BERT-ovog vokabulara, koji je veličine 30522 riječi, najbolje zamjenjuje maskirani token. Riječ koja odgovara najvećem broju u izlaznom vektoru bit će izabrana kao predikcija tokena MLM sloja. Prikaz procesa maskiranja određene riječi u rečenici dan je na slici 3.4:




Sl. 3.4 Primjer maskiranja riječi u MLM sloju: u rečenici „Otvorio je bocu vode i natočio u čašu koja je stajala na stolu“ model maskira riječ „čašu“ i pokušava ju zamijeniti s drugim riječima sličnog značenja poput „šalicu“, „kriglu“ i „zdjelu“ kako bi bolje shvatio kontekst u kojem se riječ nalazi te na temelju toga dao predikciju riječi, iz rječnika, koja najbolje zamjenjuje maskiranu riječ

Arhitektura modela SPADE++ prikazana je na slici 3.5. Kao i kod modela text-embedding-3-small ulazni tekst prvo prolazi kroz proces tokenizacije. Jednom dobiveni tokeni se reprezentiraju vektorom te se prosljeđuju centralnom transformeru (blok *Transformer*). Zadaća transformera je ulazni vektor dodatno obogatiti kontekstualnim informacijama u kojima se riječi nalaze te na izlazu dati obogaćeni vektor. U svrhu dodatnog obogaćivanja sadržaja predikcijskog vektora za što kvalitetnije SPLADE++ koristi i dodatni MLM sloj (*MLM Head blok*) koji na izlazu daje 30522 vrijednosti koje predstavljaju distribucije vjerojatnosti procjene važnosti svake pojedine riječi iz BERT-ovog rječnika. Većina pozicija izlaznog vektora bit će postavljena na vrijednost 0, dok će ih samo nekolicina biti različita od 0. Prednost ovakvog rijetkog vektora u odnosu na gusti vektor je mogućnost pohrane u obliku rječnika čime se postiže ušteda memorije jer se pohranjuju samo vrijednosti koje su različite od 0. Druga prednost nad gustim vektorima je značajno brže računanje sličnosti s drugim vektorima. Primjerice za računanje kosinusne sličnosti potrebno pomnožiti svega nekoliko pozicija da bi se moglo usporediti dva vektora jer su sve ostale pozicije postavljene na 0 te se mogu ignorirati. Dobiveni rezultati analize politika privatnosti korištenjem ovog pristupa i pristupa koji je korišten u modelu text-embedding-3-small dana je u poglavlju 6.




Sl. 3.5 Arhitektura modela SPLADE++[10]

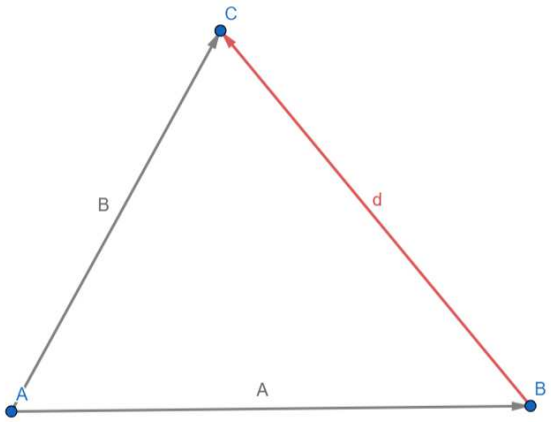
# 4. Mjere za usporedbu sličnosti teksta

Nakon provedene vektorske reprezentacije riječi raspolažemo s vektorima koji su pogodni za operacije uspoređivanja. Postavlja se pitanje koji način usporedbe vektora je optimalan kako bi semantička sličnost originalnih ulaznih tekstova ostala maksimalno sačuvana. Ovisno o primjeni i korištenim modelima za vektorsku reprezentaciju riječi postoje različite mjere koje se primjenjuju kako bi se utvrdila sličnost odnosno razlike između dvaju ili više vektora. U nastavku ovoga poglavlja opisane su tri takve mjere: euklidska udaljenost, skalarni produkt i kosinusna sličnost.

Euklidska udaljenost d u općenitom smislu predstavlja udaljenost dviju točaka u n-dimenzijskom prostoru. Euklidska udaljenost između 2 vektora $A=(a_1,a_2,\ldots,a_n)$ i $B=(b_1,b_2,\ldots,b_n)$ prikazana je na slici 4.1 te je definirana izrazom (2):

$$d(A,B) = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2} \qquad (2)$$

te je omeđena na intervalu [0, +∞>. Udaljenost 0 predstavlja dva potpuno istovjetna vektora, dok veliki iznosi udaljenost (u odnosu na veličinu vektora A i B) označavaju značajno različite vektore. Korištenje euklidske udaljenost je pogodno kada se uspoređuju tekstovi koji koriste slične rječnike. Problem korištenja euklidske udaljenosti je što je ona jako osjetljiva na veličinu ulaznih tekstova. Poseban problem nastaje kada je jedan tekst značajno duži od drugoga te je tada iznos euklidske udaljenosti velik odnosno ukazuje na veliku različitost između tekstova, iako oni mogu biti izrazito sličnoga značenja.




Sl. 4.1 Euklidska udaljenost d vektora A i B jest duljina vektora d za koji vrijedi da je A+d=B

Dužina vektora A=($a_1$,$a_2$,…,$a_n$) označava se sa |A| i definirana je relacijom (3):

$$|A| = \sqrt{(a_1)^2 + (a_2)^2 + \cdots + (a_n)^2} \quad (3)$$

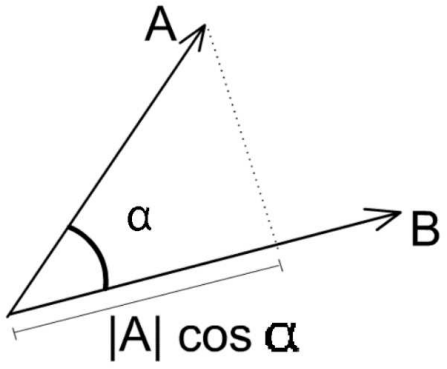
Skalarni produkt vektora A=($a_1$,$a_2$,…,$a_n$) i B=($b_1$,$b_2$,…,$b_n$) prikazani na slici 4.2 koji međusobno zatvaraju kut α definiran je kao umnožak duljine vektora A i duljine ortogonalne projekcije vektora B na vektor A. Skalarni produkt vektora A i B može se izračuna na dva načina. Prvi način je koristeći jednadžbu (4):

$$A * B = |A| * |B| * \cos \alpha \quad (4)$$

Drugi način je koristeći jednadžbu (5):

$$A * B = \sum a_i * b_i = a_1 * b_1 + b_2 * b_2 + \cdots + a_n * b_b \quad (5)$$

Izlaz iz funkcije skalarnog produkta je definiran na intervalu <-∞, +∞>. Kada su vektori međusobno okomiti vrijednost skalarnog produkta iznosi nula. Veliki iznosi skalarnog produkta predstavljaju veliku sličnost, dok negativni iznosi izražavaju suprotnost među njima. Glavna mana skalarnog produkta slična je glavnoj mani euklidske udaljenosti: jako je osjetljiv na veličinu vektora te je skalarni produkt dobar izbor kada se uspoređuju vektori sličnih veličina i korištenih rječnika. Drugi problem je što je semantičku suprotnost među tekstovima teško modelirati na ovaj način te ju se češće doživljava kao suprotnost umjesto kao sličnost.



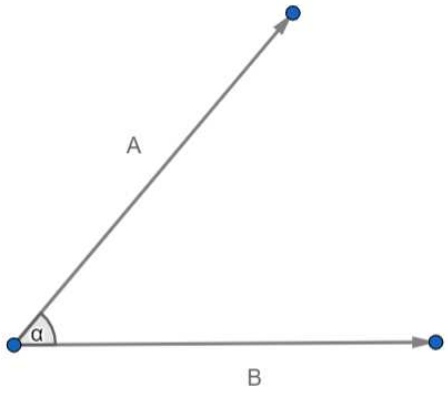

Sl. 4.2 Skalarni produkt vektora A i B je umnožak duljine vektora B i duljine ortogonalne projekcije vektora A na vektor B

Kosinusna sličnost jedna je od najčešće korištenih mjera prilikom usporedbe sličnosti dvaju ili više vektora nastalih vektorskom reprezentacijom tekstualnih podataka. Kosinusna sličnost definirana je na intervalu [-1, 1] te pritom vrijedi da kosinusna sličnost 1 označava savršeno podudaranje 2 potpuno jednaka ili barem proporcionalna vektora, dok kosinusna sličnost -1 označava 2 savršeno suprotna ili suprotno proporcionalna vektora. Ova mjera u potpunosti ignorira duljinu vektora koji se uspoređuju i promatra samo kut među njima. Ovo svojstvo je posebice dobrodošlo prilikom uspoređivanja velikih odlomaka teksta s kojima se radi analiza u ovome diplomskom radu.

Za vektore A i B prikazane na slici 4.3 i odgovarajuće duljine vektora |A| i |B| kosinusna sličnost može se izračunati korištenjem relacije (6):

$$\cos\alpha = \frac{A * B}{|A| * |B|} \qquad (6)$$

Drugo svojstvo koje je dobrodošlo jest vremenska složenost $O(n^2)$ koja omogućuje brzo generiranje sličnost među vektorima što je izrazito bitno kada se uspoređuju veliki vektori (primjerice model SPLADE++ koristi vektore veličine 30522 dimenzija).




Sl. 4.3 Kosinusna sličnost dva vektora je omjer skalarnog produkta vektora A i B i umnoška njihovih duljina |A| i |B|

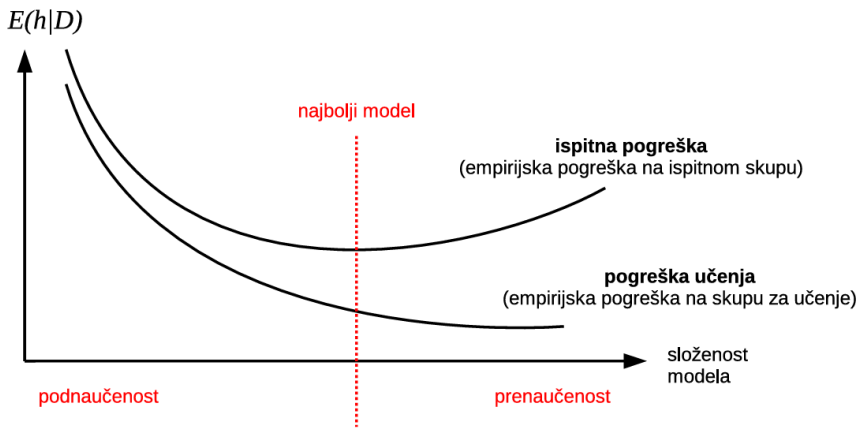
Upravo je radi tih razloga, prilikom interne analize sličnosti među tekstovima, u sklopu ovoga rada korištena upravo kosinusna sličnost odabrana za daljnju analizu sukladnosti politika privatnosti s Općom uredbom o zaštiti podatka. Odabrana je također i radi prirode korištenih modela za vektorsku reprezentaciju riječi te je najsmisleniji izbor od svih navedenih mjera sličnosti vektora. Razlog tome je način na koji modeli za vektorsku reprezentaciju riječi, opisani u poglavlju 3.1 i 3.2, oblikuju izlazni vektor na temelju teksta koji im je doveden na ulazu.

# 5. Klasifikacijski modeli strojnog učenja

Klasifikacija je vrsta nadziranog strojnog učenja. Nadzirano strojno učenje (engl. *supervised learning*) je vrsta strojnog učenja u kojemu se algoritmu predaje označeni skup podatka koji se koristi za treniranje modela. Ulaz u model se tipično označava s vektorom $X=(x_1,x_2,\ldots,x_n)$, a izlaz iz modela sa vektorom $Y=(y_1,y_2,\ldots,y_n)$. Cilj modela je pomoću određenog algoritma pronaći funkciju h (hipotezu) koja najbolje opisuje vezu između ulaznog i izlaznog skupa podataka. Skup podataka se dijeli na podatke za treniranje modela i skup za testiranje modela. Svrha te podjele je naučiti model klasificirati podatke kojima zna kategoriju kojoj pripadaju kako bi bio u stanju klasificirati podatke čija kategorija nije poznata. Evaluacija modela provodi se analizom preciznosti modela na skupu podataka za trening i testiranje analizom funkcije pogreške E. Model radi utoliko bolje što je empirijska pogreška manja. Empirijska pogreška klasifikacijskog modela tipično je zadana izrazom (7):

$$E(h|D) = \frac{1}{N}\sum_{i=1}^{N}|h(xi) - yi| \qquad (7)$$

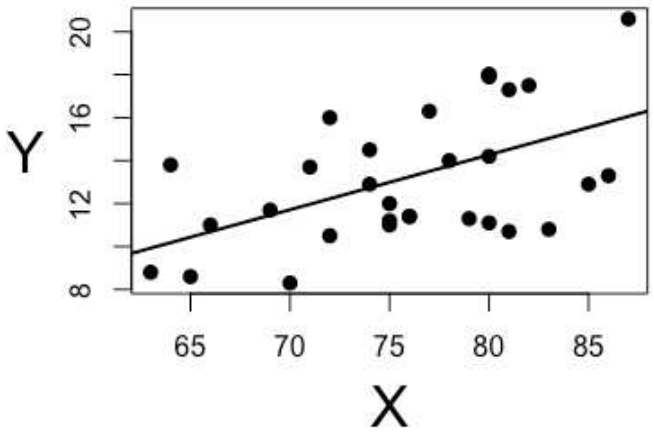
Te je ograničena na intervalu od 0 do 1, gdje 0 predstavlja preciznost od 0%, a 1 preciznost od 100%. Preciznost modela izmjerena na skupu za treniranje uglavnom se razlikuje od preciznosti dobivene na skupu za testiranje. Tipično se na skupu za treniranje postiže veća preciznost te je potrebno izabrati optimalan omjer preciznosti na skupu za treniranje i skupu za testiranje kako ne bi došlo niti do podnaučenosti niti do prenaučenosti modela. Prikaz omjera preciznosti na skupu za treniranje i testiranje dan je na slici 5.1:




Sl. 5.1 Proces odabira modela strojnog učenja na temelju iznosa funkcije ispitne pogreške i pogreške učenja [11]

## 5.1. Klasifikacija korištenjem modela linearne regresije

Linearna regresija je algoritam strojnog učenja koji se uglavnom koristi za predviđanje budućih vrijednosti određene numeričke vrijednosti. Može se koristiti i za klasifikaciju kao što je slučaj u ovome radu. Grafički prikaz regresijskog pravca, jednostavne linearne regresije, koji najbolje opisuje vezu između ulaznog skupa primjera X i njihovih pripadnih oznaka Y dan je na slici 5.2:




Sl. 5.2 Jednostavna linearna regresija [12]: na slici je prikazan regresijski pravac koji najbolje opisuje vezu ulaznog skupa primjera X i njihovih pripadnih oznaka Y

Ideja linearne regresije je pronaći funkciju h koja najbolje opisuje vezu između ulaznog skupa podataka X i izlaznih oznaka Y. Za tu zadaću koristiti se funkcijom gubitka koja se naziva kvadratni gubitak te je onda funkcija empirijske pogreške zadana izrazom (8):

$$E(h|D) = \frac{1}{2}\sum_{i=1}^{N}(h(xi) - yi)^2 \qquad (8)$$

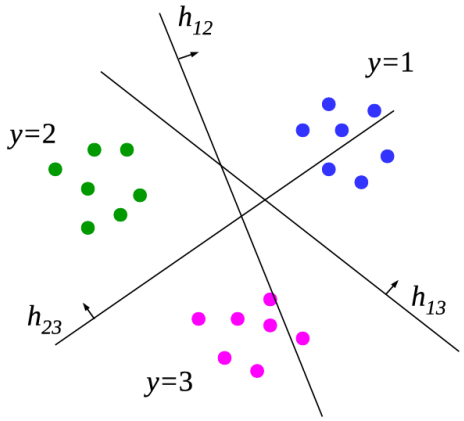
Kvadratni gubitak ima nekoliko pogodnosti, a glavna pogodnost je da omogućava pronalaženje tražene hipoteze h u zatvorenoj formi. Svojstva kvadratnog gubitka su da jako kažnjava podatke koji su značajno udaljeni od predviđeni vrijednosti, a slabo kažnjava podatke koji su blizu predviđenih vrijednosti. To za zadaće klasifikacije obično stvara značajne probleme, ali budući da se u ovome radu koriste modeli čiji izlazni vektori koriste

vrijednosti unutar intervala [0, 1] pretpostavka je da u ovoj analizi neće predstavljati prepreku jer sva odstupanja koja su manja od 1 kvadriranjem postaju još manja.

Prilikom klasifikacije podataka moguće je više pristupa. OVO(engl. *One Versus One*) pristup klasificira podatke tako da im dodjeljuje oznaku jednu od dvije moguće kategorije. Za odvajanje podatka potrebno je za svaki par izlaznih kategorija imati zasebnu granicu. U slučaju kada je ukupan broj kategorija označen s K, ukupan broj granica G ( različitih hipoteza h) zadan je izrazom (9):

$$G = \frac{K * (K - 1)}{2} \tag{9}$$

Svakom podatku se dodjeljuje oznaka 0 ako ne pripada određenoj kategoriji te 1 ako joj pripada. OVO pristup prikazan je na slici 5.3



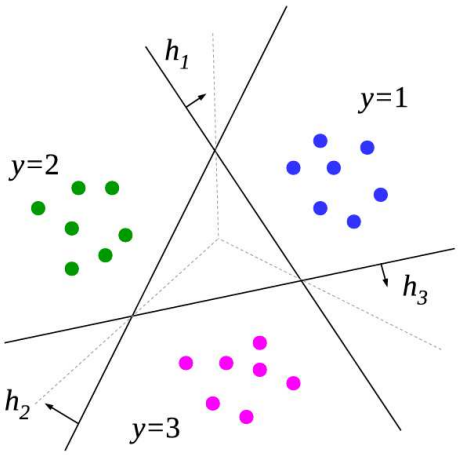

Sl. 5.3 Višeklasni model u shemi OVO (engl. *One Versus One*) [11]: grozd y=1 predstavlja podatke koji imaju oznaku klase 1, grozd y=2 podatke koji imaju oznaku klase 2, a grozd y=3 podatke koji imaju oznaku klase 3. Funkcija $h_{12}$ predstavlja granicu između klasa 1 i 2, $h_{13}$ granicu između klasa 1 i 3, a $h_{23}$ granicu između klasa 2 i 3

Za potrebe klasifikacije u ovome radu iznos granice koji razdvaja dvije klase 0 i 1 može se postaviti na fiksni iznos od 0.5. Svi ulazni podaci imaju oznaku 0 ili 1 te jednom kada model linearne regresije odredi oznaku promatranoga podatka dovoljno je usporediti dobivenu oznaku s udaljenosti među promatranim klasama. U slučaju da je model dodijelio oznaku koja je veća od 0.5 to znači da je podatak bliži klasi 1 nego klasi 0 te će mu se dodijeliti oznaka 1, a ako je dodijeljena oznaka manja od 0.5 znači da je podataka bliži klasi 0 te će mu se dodijeliti odgovarajuća oznaka.

Drugi pristup klasifikacija podataka je OVR (engl. *One Versus Rest*). Za razliku od OVO klasifikatora OVR za svaki podataka pokušava odrediti je li dio određene klase ili bilo koje druga svih ostalih mogućih klasa koje su u modelu koriste. Prilikom klasifikacije podataka uglavnom mu je potreban manji broj granica nego OVO pristupu. Broj granica G koje koristi klasifikator OVO zadan je izrazom (10):

$$G = K \tag{10}$$

Odnosno broj potrebnih granica za razdvajanje klasa G potpuno je jednak promatranom broju klasa K. Pristup OVR prikazan je na slici 5.3:




Sl. 5.4 Višeklasni model u shemi OVR(engl. *One Versus Rest*) [11]: grozd y=1 predstavlja podatke koji imaju oznaku klase 1, grozd y=2 podatke koji imaju oznaku klase 2, a grozd y=3 podatke koji imaju oznaku klase 3. Funkcija $h_1$ predstavlja granicu između klase 1 i svih ostalih klasa, $h_2$ granicu između klase 2 i svih ostalih klasa, a $h_3$ granicu između klase 3 i svih ostalih klasa

Iako koristi manje granica te omogućuje brže treniranje modela, OVR pristup osjetljiv je na nebalansiranost među klasama ulaznih podataka. Upravo zato u ovome radu korišten je pristup OVO radi svoje robusnosti na nebalansiranost među klasama ulaznim podacima koja je neporecivo prisutna prilikom klasifikacije paragrafa politika privatnosti (većina paragrafa ne pripada niti jednom od promatranih kategorija) te jednostavnosti implementacije.

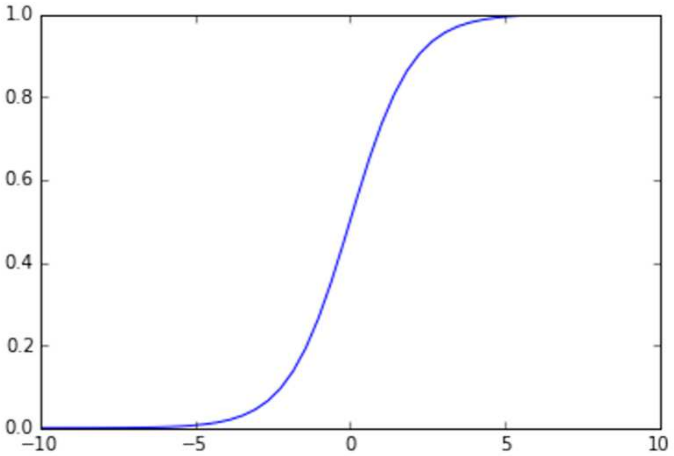
## 5.2. Klasifikacija korištenjem modela logističke regresije

Logistička regresija jedna je od temeljnih metoda strojnog učenja koja je namijenjena problemima klasifikacije. Krase je jednostavnost implementacije, visoka razina preciznosti te robusnost na oblik ulaznih podataka. Za razliku od linearne regresije logistička regresija originalno nije zamišljena kao metoda klasifikacije podataka u više promatranih kategorija već je taj broj uglavnom ograničen na klasifikaciju između dvije klase. U tom slučaju govorimo o binarnoj logističkoj regresiji te je upravo ona jedan od dva modela strojnog učenja koji se koriste u ovome radu.

Posebnost logističke regresije jest da za aktivacijsku funkciju koja na izlazu daje oznaku promatranom podatku koristi sigmoidalnu funkciju koja je prikazana na slici 5.4, a zadana je izrazom (11):

$$f(x) = \frac{1}{1 + e^{-x}} \tag{11}$$

Ako se u promatranom skupu podataka nalaze dvije klase klasa 0 i klasa 1, prilikom određivanja klasa određenog ulaznog podatka sigmoidalna funkcija će na izlazu dati broj koji je ograničen na intervalu [0, 1] te taj broj predstavlja procijenjenu šansu od strane modela da podatak pripada klasi 1. Primjerice, izlaz sigmoidalne funkcije od 0.67 predstavlja 67% šanse da podatak pripada klasi 1 te 33% šanse da pripada klasi 0.




Sl. 5.5 sigmoidalna funkcija [11]

Druga razlika između linearne regresije i logistička regresija jest ta da logistička regresija nema rješenje u zatvorenoj formi. Za potrebe pronalaženja optimalne hipoteze koja razdvaja klase koristi se gradijentnim spustom koji je definiran izrazom (12):

$$b = a - \gamma * \nabla f(a) \qquad (12)$$

Valja napomenuti da prilikom korištenja gradijentnog spusta treba biti na oprezu jer ovisno o stopi koraka gradijentni spust može i ne mora konvergirati ka najboljem rješenju.

Za funkciju gubitka u ovome radu koristi se unakrsna entropija definirana izrazom (13):

$$L\big(y, h(x)\big) = -y * \ln\big(h(x)\big) - (1 - y) * \ln(1 - (h(x)) \qquad (13)$$

Gubitak unakrsne entropije mjeri koliko dobro predviđene vjerojatnosti dobivene na izlazu iz modela odgovaraju stvarnim oznakama ulaznog skupa podataka x. Efikasnost gubitaka unakrsne entropije leži u činjenici da izrazito kažnjava netočno klasificirane podatke za koje je model predvidio veliku pouzdanost klasifikacije.

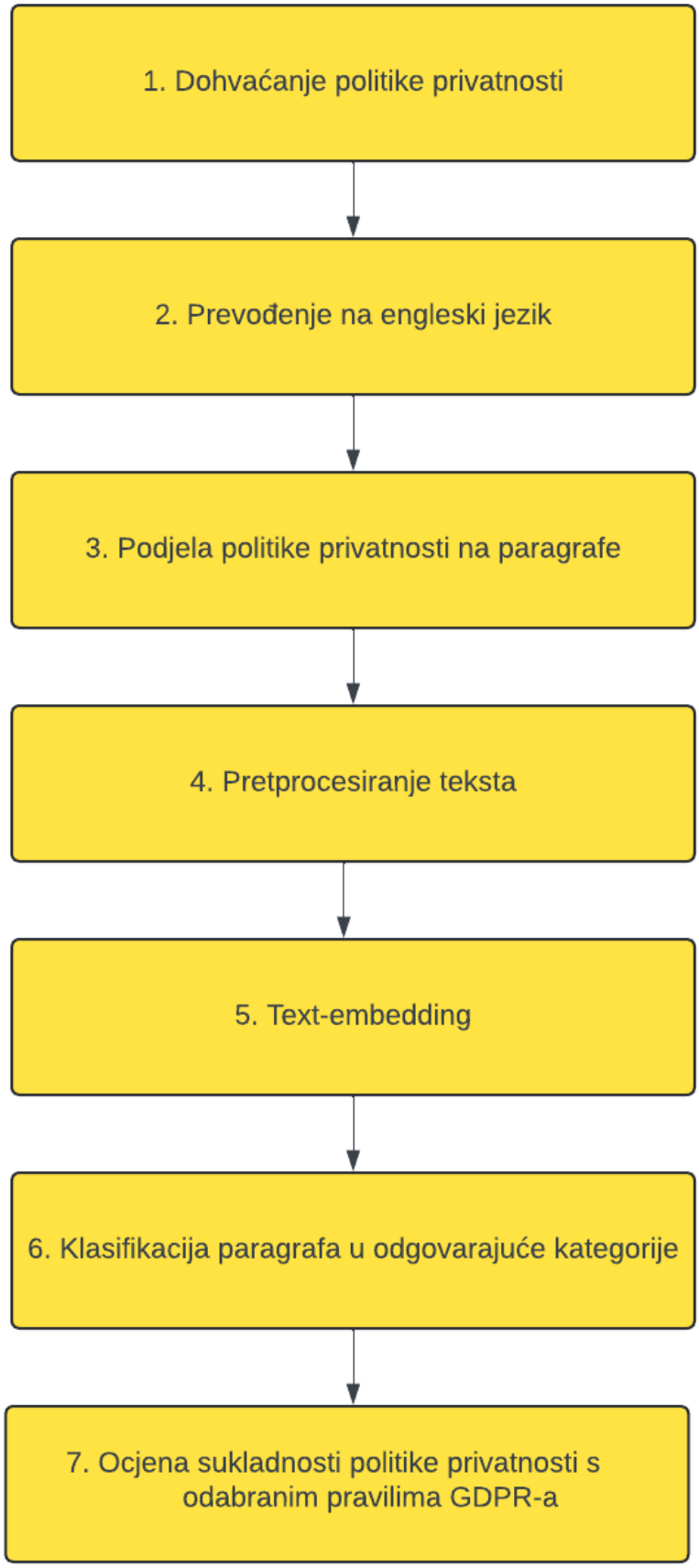
Za kraj ovog poglavlja valja još napomenuti da će za potrebe ovoga rada granica među klasama će u slučaju korištenja logističke regresije biti postavljena na istu vrijednost kao i kod korištenja linearne regresije opisane u prethodnom poglavlju te će iznositi 0.5. Oznake će se dodjeljivati promatranom podatku na način da se usporedi vrijednost na izlazu iz sigmoidalne funkcije za promatrani podatak s navedenom granicom. Prilikom korištenja modela logističke regresije korištena je implementacija LogisticRegression() koja je sastavni dio Python biblioteke sklearn, a uspješnost primjene modela logističke regresije na problem klasifikacije paragrafa politika privatnosti i dobiveni rezultati nalaze se u poglavlju 6.

# 6. Rezultati

Na slici 6.1 nalazi se dijagramski prikaz procesa analize sukladnosti politike privatnosti s GDPR-om. Proces analize podijeljen je u 7 koraka:

1. U prvom koraku odvija se dohvaćanje politike privatnosti s weba korištenjem Python biblioteke Polipy
2. Budući da korišteni modeli za vektorsku reprezentaciju riječi su namijenjeni za tekstove na engleskom jeziku, politika privatnosti se prevodi s hrvatskog jezika na engleski jezik korištenjem Python biblioteke Translators
3. U trećem koraku se politika privatnosti na engleskom jeziku dijeli na paragrafe koji će se u nastavku analize zasebno koristiti
4. Četvrti korak predstavlja pretprocesiranje teksta. Paragrafi se čiste od dijakritičkih znakova i zaustavnih riječi te se cijeli tekst prebacuje u oblik u kojem se koriste isključivo mala slova. Za ovaj proces korištenje su Python biblioteke: Re, Nltk i String
5. Sada je vrijeme za vektorsku reprezentaciju riječi korištenjem ili modela text-embedding-3-small, ili modela SPLADE++
6. U šestom koraku se vrši klasifikacija analiziranog, odnosno pokušava se ustanoviti odnosi li se analizirani paragraf na jedno od 5 odabranih pravila GDPR-a:
    a. Politika privatnosti mora sadržavati kontakt voditelja obrade
    b. Ispitaniku mora biti obaviješten o pravu na povlačenje privole
    c. Voditelj obrade dužan je navesti svrhu obrade podataka
    d. Voditelj obrade dužan je navesti kategoriju osobnih podataka koji se obrađuju
    e. Ispitanik mora biti obaviješten o pravu na brisanje i ispravak prikupljenih podataka
7. Nakon što se obrade svi paragrafi daje se ocjena sukladnosti politike privatnosti za svih pet kategorija analiziranih u šestom koraku

Za treniranje modela strojnog učenja korištena je baza podataka od 50 politika privatnosti. Politike privatnosti su pohranjene u relacijskoj bazi podataka MySQL. Iz prikupljenih politika su izvađeni paragrafi koji se odnose na 5 analiziranih pravila GDPR-a. Ostali paragrafi korišteni su kao podaci koji ne pripadaju niti jednoj kategoriji te ih se također koristi prilikom treniranja modela. Prije korištenja podataka, podaci su obrađeni na istovjetan način kao i prilikom automatizirane analize određene politike privatnosti. Modeli su pohranjeni korištenjem Python biblioteke Pickle te ih prilikom provjere politike privatnosti nije potrebno nanovo trenirati već ih je dovoljno samo učitati njihove težine.




Sl. 6.1 Dijagramski prikaz procesa analize sukladnosti politike privatnosti s GDPR-om

Prilikom treniranja modela koji provjerava je li određeni paragraf označava da je politika privatnosti sukladna s Uredbom ili ne koristio se skup od 125 primjera za učenje za svaku kategoriju te 30 testnih paragrafa. Modeli su bazirani na dva modela za vektorsku reprezentaciju riječi: text-embedding-3-small i SPLADE++. Od metoda strojnoga učenja korištene su linearna regresija i logistička regresije. Svi modeli su imali 100% preciznost klasifikacije paragrafa na skupu za učenje, a rezultati na skupu za testiranje dani su u tablicama 6.1 i 6.2:

Tablica 6.1 Preciznost klasifikacije testnih paragrafa korištenjem modela text-embedding-3-small i linearne ili logističke regresije

| Broj točno klasificiranih politika privatnosti | Kontakt voditelja obrade | Pravo na povlačenje privole | Svrha obrade podataka | Kategorija osobnih podataka | Pravo na ispravak i brisanje | ukupno |
|---|---|---|---|---|---|---|
| Linearna regresija | 30/30 (100,00%) | 28/30 (93,33%) | 27/30 (90,00%) | 25/30 (83,33%) | 28/30 (93,33%) | 138/150 (92,00%) |
| Logistička regresija | 29/30 (96,67%) | 29/30 (96,67%) | 28/30 (93,33%) | 28/30 (93,33%) | 29/30 (96,67%) | 143/150 (95,33%) |

Iz priloženoga vidimo da oba modela daju zadovoljavajuće rezultate te nema naznaka da su modeli prenaučeni na skupu za učenje. Prilikom korištenja modela text-embedding-3-small malo bolji rezultati klasifikacije paragrafa su postignuti korištenjem metode logističke regresije. Preciznost korištenjem te metode iznosi visokih 95.33% uspješno klasificiranih paragrafa po svim kategorijama, dok se korištenjem linearne regresije postižu malčice slabiji rezultati klasifikacije. Prilikom korištenja linearne regresije postiže se 92.00% po svim promatranim kategorijama.

U tablici 6.2. primjećujemo da kada se za text-embedding koristi model SPLADE++ postižu se bolji rezultati klasifikacije testnih paragrafa prilikom korištenja i linearne regresije i logističke regresije nego kada se koristi model text-embedding-3-small. Prilikom korištenja linearne regresije postiže se preciznost od 96.00%, a prilikom korištenja metoda logističke regresije zavidnih 97.33%. Iako se prilikom treniranja modela postigla preciznost od 100%

u oba slučaja, budući da model s visokom preciznošću klasificira testne paragrafe po svim promatranim kategorijama, nema znakova prenaučenosti modela. Iz rezultata klasifikacije paragrafa zaključuje se da su modeli u stanju s visokom preciznošću klasificirati paragrafe te da model SPLADE++ pokazuje neznatno bolje rezultate od modela text-embedding-3-small. U poglavljima 6.1 i 6.2 dan je prikaz rezultata provjere 21 politike privatnosti te postignute preciznosti provjere sukladnosti politika s GDPR-om.

Tablica 6.2 Preciznost klasifikacije testnih paragrafa korištenjem modela SPLADE++ i linearne ili logističke regresije

| Broj točno klasificiranih paragrafa | Kontakt voditelja obrade | Pravo na povlačenje privole | Svrha obrade podataka | Kategorija osobnih podataka | Pravo na ispravak i brisanje | ukupno |
|---|---|---|---|---|---|---|
| Linearna regresija | 30/30 (100,00%) | 28/30 (93,33%) | 27/30 (90,00%) | 30/30 (100,00%) | 29/30 (96,67%) | 144/150 (96,00%) |
| Logistička regresija | 30/30 (100,00%) | 28/30 (93,33%) | 29/30 (96,67%) | 30/30 (100,00%) | 29/30 (96,67%) | 146/150 (97,33%) |

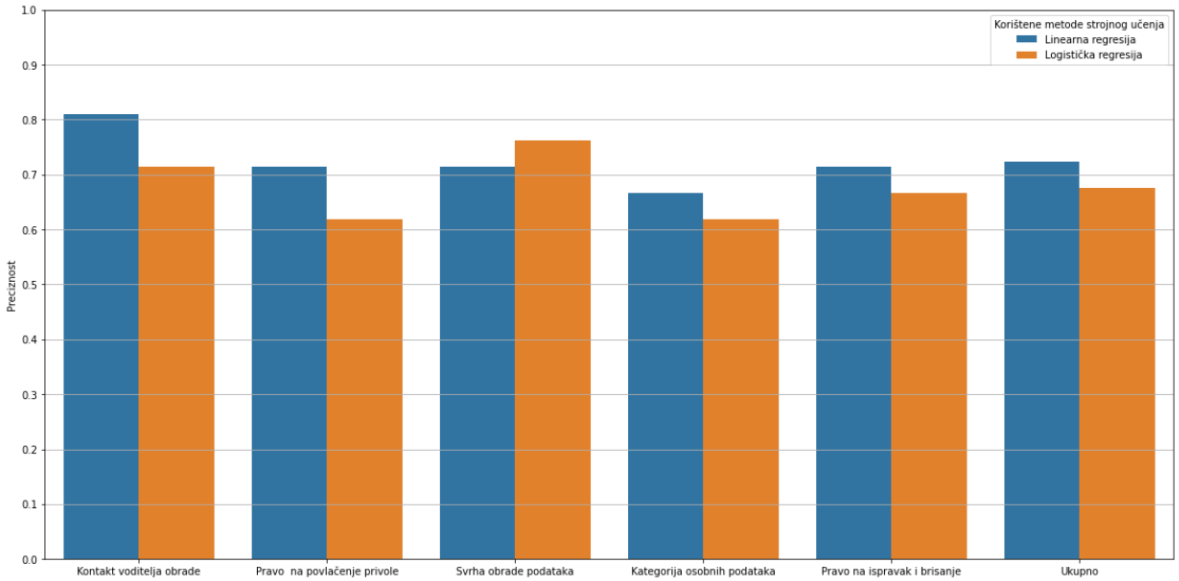
## 6.1. Model text-embedding-3-small

U tablici 6.1 nalaze se rezultati automatizirane provjere 21 politike privatnosti po navedenim kategorijama korištenjem modela text-embedding-3-small. Da bi politika privatnosti bila označena sukladnom s određenom kategorijom, mora postojati barem jedan paragraf unutar politike koji je prilikom klasifikacije označen da pripadnik te kategorije. Na razini provjere cijele politike vidimo nešto skromnije rezultate nego prilikom klasifikacije pojedinih paragrafa. Prilikom korištenja modela text-embedding-3-small najveća je preciznost od 72.38% postignuta korištenjem linearne regresije prilikom provjere svih promatranih kategorija provjere. Pritom je model bio posebno uspješan u provjeri sadržava li politika privatnosti kontakt voditelja te je u tome bio uspješan u 80.95% slučajeva, a najmanje uspješan prilikom provjere jesu li navedene kategorije osobnih podataka koji se prikupljaju te ostvario preciznost od 66.67%.

Tablica 6.3 Preciznost dobivena korištenjem modela text-embedding-3-small i linearne ili logističke regresije

| Broj točno klasificiranih politika privatnosti | Kontakt voditelja obrade | Pravo na povlačenje privole | Svrha obrade podataka | Kategorija osobnih podataka | Pravo na ispravak i brisanje | ukupno |
|---|---|---|---|---|---|---|
| Linearna regresija | 17/21 (80,95%) | 15/21 (71,43%) | 15/21 (71,43%) | 14/21 (66,67%) | 15/21 (71,43%) | 76/105 (72,38%) |
| Logistička regresija | 15/21 (71,43%) | 13/21 (61,90%) | 16/21 (76,19%) | 13/21 (61,90%) | 14/21 (66,67%) | 71/105 (67,62%) |

Grafički prikaz rezultata koji se nalaze u tablici 6.3, korištenjem vertikalnih stupaca, dan je na slici 6.1:




Sl. 6.1 Grafički prikaz rezultata, dobivenih korištenjem modela text-embedding-3-small i linearne ili logističke regresije, korištenjem vertikalnih stupaca

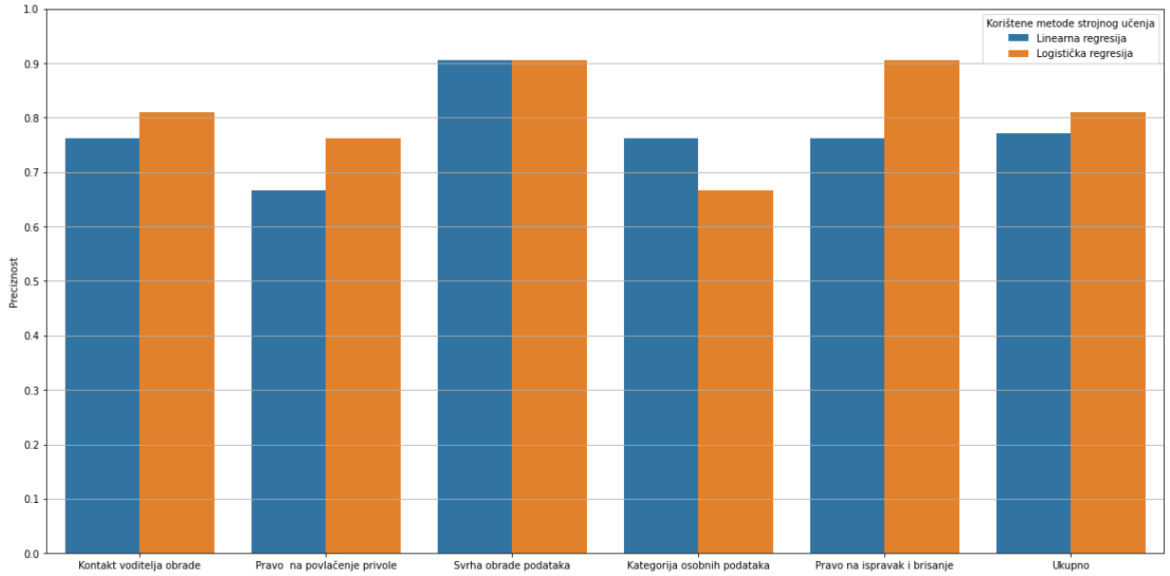
## 6.2. Model SPLADE++

Model SPLADE++ postigao je značajno veću preciznost provjere nego korištenjem modela text-embedding-3-small. Model je ostvario preciznost od 80.95% prilikom korištenja logističke regresije te 77.14% prilikom korištenja linearne regresije tijekom klasifikacije svih promatranih kategorija provjere. Posebice je uspješan prilikom provjere je li navedena svrha prikupljenih podataka prilikom čega ostvaruje preciznost od 90.48%, a najmanje je uspješan prilikom provjere je li navedena kategorija osobnih podataka koji se obrađuju te tada ostvaruje preciznost od 66.67%. Iz navedenog zaključujemo da model SPLADE++ najlošije rezultate ostvaruje upravo prilikom izvršavanje tog zadatka ali je u odnosu na model text-embedding-3-small i dalje uspješniji u 4.76% slučajeva izraženo apsolutno. Model ostvaruje zadovoljavajuće rezultate te je pogodan za zadaću automatizirane provjere politika privatnosti.

Tablica 6.4 Preciznost dobivena korištenjem modela SPLADE++ i linearne ili logističke regresije

| Broj točno klasificiranih politika privatnosti | Kontakt voditelja obrade | Pravo na povlačenje privole | Svrha obrade podataka | Kategorija osobnih podataka | Pravo na ispravak i brisanje | ukupno |
|---|---|---|---|---|---|---|
| Linearna regresija | 16/21 (76,19%) | 14/21 (66,67%) | 19/21 (90,48%) | 16/21 (76,19%) | 16/21 (76,19%) | 81/105 (77,14%) |
| Logistička regresija | 17/21 (80,95%) | 16/21 (76,19%) | 19/21 (90,48%) | 14/21 (66,67%) | 19/21 (90,48%) | 85/105 (80,95%) |

Grafički prikaz rezultata iz tablice 6.4, korištenjem vertikalnih stupaca, dan je na slici 6.2:




Sl. 6.2 Grafički prikaz rezultata, dobivenih korištenjem modela SPLADE++ i linearne ili logističke regresije, korištenjem vertikalnih stupaca

## 6.3. Usporedba rezultata

Uspoređujući rezultate dva promatrana modela možemo primijetiti da je model text-embedding-3-small u dva slučaja precizniji od modela SPLADE++. Prvi slučaj je prilikom korištenja linearne regresije za provjeru je li naveden kontakt voditelja obrade. Model postiže preciznosti od 81% dok model SPLADE++ postiže preciznost od 76.2%. Drugi slučaj je prilikom provjere je li navedena svrha obrade također prilikom korištenja linearne regresije. Model text-embedding-3-small tada postiže preciznost od 71.4%, a model SPLADE++ preciznost od 66.7%. U svim ostalim situacijama model SPLADE++ postiže bolje rezultate od modela text-embedding-3-small. Ako uspoređujemo samo korištene modele za vektorsku reprezentaciju riječi i ignoriramo korištene metode strojnog učenja možemo vidjeti da model SPLADE++ postiže značajno bolje rezultate prilikom provjere svih promatranih kategorija.

Najbolja preciznost, prema dobivenim rezultatima, je postignuta prilikom korištenja modela SPLADE++ u kombinaciji s logističkom regresijom. Prilikom svih provedenih provjera sukladnosti testnih politika ta kombinacija daje najbolje rezultate, osim u situaciji kada se vrši provjera je li na kategorija osobnih podataka koji se prikupljaju. U toj situaciji bolje rezultate ostvaruje kombinacija modela SPLADE++ i linearne regresije te je ona optimalniji izbor za tu zadaću.

# Zaključak

Na temelju dobivenih rezultata prikazanih u poglavlju 6. može se zaključiti da je korištenjem metoda strojnog učenja moguće automatizirati dio procesa provjere politika privatnosti s Općom uredbom o zaštiti podataka. Rezultati prikazani u ovome radu djeluju obećavajuće te ukazuju da su metode i modeli korišteni u ovome radu prikladni za ovu zadaću. Model SPLADE++ koji koristi rijetke vektore u kombinaciji s logističkom regresijom pokazao je najbolje rezultate provjere politika privatnosti s visokom preciznošću od preko 80%.

Navedena preciznost zasigurno se može dodatno poboljšati uporabom većeg skupa podataka. Također treba uzeti u obzir da je na kvalitetu rezultata također utjecala i činjenica da su modeli korišteni za vektorsku reprezentaciju riječi trenirani za rad s tekstovima na engleskom jeziku te su se sve politike privatnosti koje su analizirane u ovome radu prvo prošle proces prijevoda na engleski jezik čime je neosporno narušena kvaliteta rezultata. Trenutni fokus prilikom razvoja jezičnih modela koji se mogu koristiti za vektorsku reprezentaciju riječi je na radu s tekstovima na engleskom jeziku, a jednom kada se razviju kvalitetni modeli koji mogu raditi s hrvatskim jezikom bit će moguće ponoviti ovu analizu u svrhu unaprjeđenja rezultata.

Za kraj ovoga rada želio bih naglasiti da valja ostati na oprezu prilikom automatiziranog provjeravanja sukladnost politika privatnosti s Općom uredbom o zaštiti podataka. Budući da se radi o pravnim dokumentima, za kršenje pravila Opće uredbe o zaštiti podataka je moguće kazneno odgovarati. U tom svjetlu strojno učenje definitivno može ubrzati proces provjere sukladnosti, ali nipošto ne bi trebalo biti potpuna zamjena za čovjeka. Treba težiti izradi alata koji mogu služiti kao pomoć educiranom kadru ljudi. Pritom je prvenstveno važno obratiti pozornost na dobivenu povratnu informaciju od njih te saslušati njihove savjete kako bi sinergija čovjeka i računala bila što kvalitetnija.

# Literatura


[1] Taylor Petroc, Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2020, with forecasts from 2021 to 2025 (2023., studeni). Poveznica: https://www.statista.com/statistics/871513/worldwide-data-created/ ; pristupljeno 25. travnja 2024.

[2] AZOP, Osnovne informacije o usklađivanju s Općom uredbom o zaštiti podataka i Zakonom o provedbi Opće uredbe o zaštiti podataka za voditelje i izvršitelje obrade (poslovne subjekte, organizacije, poduzeća, državna i ostala tijela). Poveznica: https://azop.hr/osnovne-informacije-za-organizacije/ ; pristupljeno 25. travnja 2024.

[3] Regulation (EU) 2016/679 (General Data Protection Regulation) of the European Parliament. Poveznica: https://gdprinfo.eu/hr; Opća uredba o zaštiti podataka. Poveznica: https://www.zakon.hr/z/3112/Op%C4%87a-uredba-o-za%C5%A1titi-podataka---Uredba-%28EU%29-2016-679-; pristupljeno 5. svibnja 2024.

[4] Uredba (EU) 2016/679 Europskog parlamenta i Vijeća od 27. travnja 2016. o zaštiti pojedinaca u vezi s obradom osobnih podataka i o slobodnom kretanju takvih podataka te o stavljanju izvan snage Direktive 95/46/EZ (Opća uredba o zaštiti podataka), Europska Komisija, (27. travnja 2016.). Poveznica: https://eur-lex.europa.eu/legal-content/HR/TXT/?uri=CELEX%3A32016R0679 ;

[5] Europski odbor za zaštitu podataka (EDPB) , (2019, rujan). Poveznica: https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/european-data-protection-board-edpb_hr; pristupljeno 6. lipnja 2024.

[6] Samarin, N., Kothari, S., Siyed, Z., Wijesekera, P., Fischer, J., Hoofnagle, C. and Egelman Polipy: Library For Scraping, Parsing, And Analyzing Privacy Policies, Github, (2021, prosinac). Poveznica: https://github.com/blues-lab/polipy; pristupljeno 4. svibnja 2024.

[7] New embedding models and API updates (25. siječnja 2024.). Poveznica: https://openai.com/index/new-embedding-models-and-api-updates/; pristupljeno 12. lipnja 2024.

[8] Translators, Poveznica: https://pypi.org/project/translators; pristupljeno 10. svibnja 2024.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need. znanstveni rad. Google 2017.

[10] SPLADE for Sparse Vector Search Explained (30. lipnja 2023.). Poveznica: https://www.pinecone.io/learn/splade/; pristupljeno: 2. svibnja 2024.

[11] Jan Šnajder, Bojana Dalbelo Bašić Strojno učenje. Nastavni materijal za potrebe kolegija strojno učenje. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva 2014.

[12] R for ecology, How to do a simple linear regression in R (srpanj 2021) Poveznica: https://www.rforecology.com/post/how-to-do-simple-linear-regression-in-r/ pristupljeno 10. lipnja 2024.

# Sažetak

Automatizirana provjera sukladnosti politika privatnosti na webu s Općom uredbom o zaštiti podataka (GDPR) korištenjem strojnog učenja

U ovome radu istražuju se mogućnosti automatiziranja procesa provjere sukladnost politika privatnosti s Općom uredbom o zaštiti podataka korištenjem metoda strojnog učenja. Odabrano je 5 odredbi GDPR koje su pogodnih za analizu koje svaka politika privatnosti mora imati: kontakt voditelja obrade, pravo ispitanika na povlačenje privole, navedenu svrhu obrade podataka, navedenu kategoriju osobnih podataka koji se obrađuju te obavijest ispitaniku na pravo na brisanje i ispravak prikupljenih podataka. Prilikom prikupljanja politika privatnosti korištena je Python biblioteka Polipy. Za vektorsku reprezentaciju riječi korišteni su modeli OpenAI text-embedding-3-small i SPLADE++. Navedeni modeli su trenirani za engleski jezik pa je za prijevod teksta s hrvatskog jezika na engleski jezik korištena biblioteka Translators. Za klasifikaciju paragrafa politika privatnosti korištene su metode strojnog učenja: logistička regresija i linearna regresija. Na kraju je donesen zaključak na temelju dobivenih rezultati te su navedena daljnja razmatranja.

ključne riječi: GDPR, sukladnost, obrada prirodnog jezika, vektorska reprezentacija riječi, automatizacija

# Summary

Automated compliance check of website privacy policies with the general data protection regulation using machine learning

This paper explores the possibility of automating the process of checking privacy policy compliance with the General Data Protection Regulation by using machine learning methods. 5 parts of the GDPR were selected that are suitable for analysis and that every privacy policy needs to have: contact of the data controller, notification of the subject's right to withdraw the consent, specification of the purpose of data processing, specification of the category of personal data that is processed, and notification to the data subject of the right to delete and correct collected data. The Python library Polipy is used to collect privacy policies. OpenAI text-embedding-3-small and SPLADE++ models are used for text embedding. The mentioned models are trained for English language usage, so the Translators library is used to translate the text from Croatian to English. Machine learning methods are used to classify privacy policy paragraphs: logistic regression and linear regression. Ultimately, a conclusion was made based on the obtained results, and further considerations were given.

keywords: GDPR, compliance, natural language processing, text embedding, automation

# Skraćenice

| | | |
|---|---|---|
| GDPR | *General Data Protection Regulation* | Opća uredba o zaštiti podataka |
| DPC | *Data Protection Commission* | irska komisija za zaštitu podataka |
| CNDP | *National Commission for Data Protection* | Luksemburška nacionalna komisija zaštitu podataka |
| AZOP | *Agencija za zaštitu podataka* | Agencija za zaštitu podataka |
| NLP | *Natural Language Processing* | obrada prirodnog jezika |
| GPL 3.0 | *General Public Licence Version 3* | GNU opća javna licenca |
| MTEB | *Massive Text Embedding Benchmark* | mjera obilne vektorske reprezentacije riječi |
| SPLADE | *Sparse Lexical and Expansion Model* | rijetki leksički i ekspanzivni model |
| MLM | *Masked Language Model* | jezični modeli koji koriste maskiranje |
| API | *Application Programming Interface* | aplikacijsko programsko sučelje |
| BERT | *Bidirectional Encoder Representations from Transformers* | dvosmjerne reprezentacije enkodera iz transformera |

# Privitak

## Instalacija programske podrške

Za pokretanje programske potpore korištene u ovome diplomskom radu potrebno je instalirati sljedeće Python biblioteke:

- translators
- nltk
- string
- re
- pandas
- polipy
- os
- openai
- textwrap
- transformers
- torch
- pickle
- sklearn

Za instalaciju svih navedenih biblioteka dovoljno je iskoristiti sljedeću naredbu:

```
pip install <ime_bibilioteke>
```

To je moguće učiniti unutar naredbenog retka ili direktno unutar Jupyther bilježnice.

Dio biblioteka se koristi samo prilikom treniranja modela te nisu potrebne za automatiziranu provjeru politike privatnosti koje se izvršava unutar datoteke automatizirana_provjera.ipynb

## Upute za korištenje programske podrške

Za potrebe ovoga rada korišteno je nekoliko datoteka koje sadržavaju programski kod. Njihova imena i opis njihove funkcije dan je u nastavku:

- diplomski_rad_preprocesiranje_teksta.ipynb
  - u ovoj datoteci se od ručno prikupljenih paragrafa politika privatnosti stvara ulazni skup pročišćenih tekstualnih podataka na engleskom jeziku koji se koriste za trening modela
- diplomski_rad_kosinusna_sličnost.ipynb
  - ova datoteka služi za interno uspoređivanje dvaju vektora prilikom analize sličnosti dvaju tekstova
- diplomski_rad_splade_embedding.ipynb
  - u ovoj datoteci se treniraju modeli koji koriste vektore dobivene korištenjem SPLADE++ modela. Težine vektora su pohranjene u folderu težine te navedene modele strojnog učenja nije potrebno nanovo trenirati
- diplomski_rad_openai_embedding.ipynb
  - u ovoj datoteci se treniraju modeli koji koriste vektore dobivene korištenjem OpenAI text-embedding-3-small modela. Težine vektora su pohranjene u folderu težine te navedene modele strojnog učenja nije potrebno nanovo trenirati
- diplomski_rad_automatizirana_provjera.ipynb
  - u ovome dijelu programskog koda se vrši automatizirana provjera željene politike privatnosti. Za njezino korištenje nije potrebno pokretati ostale dijelove programskog koda već je dovoljno samo instalirati sve korištene Python biblioteke. Na ulazu u kod korisnik predaje link politike privatnosti koju želi provjeriti, a na izlazu dobiva ocjenu sukladnosti politike u odnosu na 5 kategorija navedenih u poglavlju 1.3