Klasifikacija pravnih dokumenata na osnovu više oznaka

Mihajlo Perendija Petar Bašić

Uvod

Klasifikacija bilo kakvih dokumenata je neophodan i zahtevan posao koji mora biti odrađen u svakom poslovnom procesu, a naročito u domenu prava. U svrhe olakšavanja pretrage, pronalaska, upoređivanja i analize zgodno je posedovati dokumente klasifikovane na osnovu oznaka koje su im dodeljene. Domen prava kao takav zahteva postojanje velikog broja ovakvih oznaka, a proces njihovog dodeljivanja pojedinačnom dokumentu podrazumeva čitanje celokupnog dokumenta od strane domenski stručnog lica.

Napredak tehnologija iz domena obrade prirodnog jezika, u trenutnom stepenu razvoja, omogućava razvijanje softvera koji bi služili kao ispomoć u procesu označavanja dokumenta davanjem predloga za konkretne oznake. Uloga čoveka bi se svela na kontrolu i odabir najboljih oznaka čime bi se celokupan posao velikim delom automatizovao i ubrzao.

Oblast kompjuterske obrade pravnih dokumenata je još uvek mlada i kao takva ima mnoštvo problema koje je potrebno prevazići. Jedan od njih je i nedostatak adekvatnih skupova podataka usled velikog početnog napora koji je potrebno uložiti od strane stručnjaka. Napredak ipak postoji. U ovom radu će biti korišćen skup podataka koji je rezultat rada grčkih naučnika sa Atinskog univerziteta, a koji se sastoji od približno 24 hiljada pravnih akata iz zakonodavstva EU, kojima su pridružene oznake. Oznake su preuzete iz EUROVOC-a (Evropski rečnih pravnih pojmova), a njihov broj je oko 3600.

Glavni ciljevi ovog rada su istraživanje, vršenje eksperimenata i identifikacija najboljih metoda za predstavu pravnih dokumenata i za određivanje njima pripadajućih oznaka. Prilikom vršenja eksperimenata uzeta je u obzir i specifična struktura pravnih dokumenata koji su često izdeljeni na nekoliko sekcija.

Naredna poglavlja bave se korišćenim skupom podataka i metodama reprezentacije i klasifikacije pravnih tekstova koje su isprobane. Naposletku su prikazani i upoređeni rezultati, opisani problemi sa kojima su se autori susreli i donet zaključak na osnovu celokupnog istraživanja.

Skup podataka

Korišćeni skup podataka potiče iz EURULEX57k i predstavlja njegov podskup od približno 24 hiljade označenih dokumenata. Sami dokumenti su zakonodavnog tipa, a tiču se zakona u okviru Evropske unije. Prosečno su dužine 727 reči, a njihova sadržina je izdeljena u 5 sekcija i to: naslov, zaglavlje, recitale (reference), glavni deo i priloge. Ovako izdeljenim dokumentima je pridružen skup oznaka koje se na njih odnose, a zatim su organizovani u JSON datoteke.

U svrhe obučavanja modela mašinskog učenja skup podataka je izdeljen na podatke za treniranje modela (train), za validaciju modela (validation) i podatke za testiranje (test).

Predprocesiranje skupa podataka

Nakon procesa ručnog dodeljivanja oznaka podacima, koje je izvršila Evropska kancelarija za publikacije, dobijen je skup podataka u kome je skoro polovina oznaka dodeljena u manje od 10 dokumenata. Ovakva distribucija oznaka može biti problematična prilikom obučavanja modela zbog njihove nejednake zastupljenosti, te je prvi korak obrade skupa podataka obuhvatio izbacivanje svih oznaka koje se javljaju u malom broju dokumenata. Od početnog broja od oko 3600 oznaka, ovim postupkom došlo se do skupa od 1289 najzastupljenijih oznaka, dok je izbacivanjem dokumenata koji su bili označeni samo izbačenim oznakama, eliminisano njih oko trideset.

Originalni dokumenti sadrže mnoštvo karaktera i/ili skupa karaktera koji nisu od velikog značaja za njihovu semantiku, kao na primer zagrade, crtice, apostrofi, i navodnici, odnosno svi karakteri koji nisu alfanumerici. Takvi karakteri nisu pogodni ni za računarsku reprezentaciju teksta, stoga je izvršeno njihovo uklanjanje. Dodatno, neki od dokumenata su sadržali karaktere enkodirane na pogrešan način (non-UTF8) te su i oni uklonjeni.

Priroda jezika je takva da se vrlo slične reči predstavljaju drugačije u zavisnosti od konteksta u kome se koriste, kao na primer nastavci reči za množinu ili pripadnost. Ukoliko bi svaki oblik reči bio uzet u obzir prilikom mašinske obrade, a ukoliko se razmatraju metode koje se zasnivaju na statističkoj obradi dokumenata, bila bi potrebna znatno veća količina resursa i vremena, dok bi dobijeni rezultati bili neznatno bolji, a potencijalno i lošiji. Iz ovih razloga je za potrebe nekih od reprezentacija nad tekstom izvršena tzv. lematizacija ili stemovanje svake reči. Oba od navedenih pristupa u osnovi svode neku reč na njen korenski oblik.

Podaci su nadalje, za potrebe pojedinih pristupa, obrađeni uklanjanjem tzv. „stop“ reči (stop words), odnosno reči koje ne nose značenje same po sebi, a javljaju se u velikom broju u svakom dokumentu. Ove reči su najčešće prilozi, predlozi, članovi i veznici.

Naposletku, isprobana je i metoda kojom se iz teksta izdvajaju samo imenice i glagoli, kao reči koje nose najveću količinu informacija.

U zavisnosti od izbora metode za vektorizaciju (predstavu teksta u numeričkom obliku) i metode mašinskog učenja, izvršene su i isprobane razne kombinacije prethodno navedenih metoda predprocesiranja.

Metode vektorizacije

Tekstualni oblik predstave dokumenta nije pogodan za mašinsku obradu, te ga je potrebno izraziti numerički. Numerička predstava nekog teksta je najčešće vektor koji sadrži njemu specifične vrednosti.

TF-IDF

TF-IDF je statistička mera koja procenjuje koliko je neka reč relevantna za neki dokumenti u kolekciji dokumenata. Objedinjuje dve metrike:

TF metriku koja određuje koliko se često neki pojam pojavljuje u dokumentu (term frequency) i

IDF metriku koja pokazuje specifičnost pojma za dokument predstavljenu kao inverznu funkciju broja dokumenata u kojima se pojam pojavljuje. Primer na skupu podataka koji se ovde razmatra može biti reč (skraćenica) „EU“, koja se pojavljuje vrlo često u svim dokumentima iz korpusa. Ukoliko se ta reč u nekom dokumentu pojavi više puta nego što je prosek, računarski model, ukoliko koristi isključivo predstavu reči zasnovanu na TF metrici, može doneti zaključak da se taj dokument po nečemu izdvaja, što nije slučaj. IDF mera u takvim slučajevima smanjuje uticaj koji takvi pojmovi mogu imati na konačno donošenje odluka.

Drugim rečima, TF-IDF dodeljuje pojmu *i* težinu u dokumentu *j* koja je:

- Najveća kada se pojam *i* pojavljuje puno puta u malom broju dokumenata (dajući tim dokumentima moć razlikovanja od ostalih)
- Manja kada se pojam *i* pojavljuje manje puta u dokumentima ili se pojavljuje u velikom broju dokumenata (što daje manju relevantnost pojma)
- Najmanja kada se pojam *i* pojavljuje u gotovo svim dokumentima

Da bismo dokument predstavili u vektorskom obliku, potrebno je odrediti TF-IDF meru za svaku reč u tom dokumentu, a u odnosu na celokupni korpus. Vektor kojim se predstavljaju dokumenti je dugačak onoliko koliko ima jedinstvenih reči u celom skupu podataka. Svaka dimenzija vektora predstavlja jednu od reči. Vektorska reprezentacija jednog dokumenta je jedan takav vektor koji poseduje TF-IDF vrednosti na pozicijama koje odgovaraju rečima koje se u dokumentu i pojavljuju.

Word2vec

Ovaj pristup, na osnovu celog korpusa reči, generiše vektorski prostor od tipično nekoliko stotina dimenzija. Word2vec model, korišćenjem neuronske mreže, rekonstruiše lingvistički kontekst reči, odnosno uzima u obzir u kom kontekstu se neka reč pojavljuje u tekstu. Svakoj reči u korpusu se zatim dodeljuje odgovarajući vektor iz takvog prostora. Sličnost reči je tako predstavljena udaljenošću njihovih vektora u prostoru. Krajnja vektorska reprezentacija jednog dokumenta se dobija agregacijom vektora svake reči koja se u njemu pojavljuje. Dodatno, prilikom agregacije, vektoru svake reči može biti dodat ili oduzet značaj. Ovaj značaj može biti predstavljen kao jednostavna frekvencija pojavljivanja reči u dokumentu ili pak njena TF-IDF vrednost.

Vektori koji predstavljaju pojmove u vektorskom prostoru zavise od podataka nad kojima je vršeno treniranje odnosno generisanje tog prostora. Isti pojam može imati slične, ali pak različite reprezentacije u vektorskom prostoru koji je generisan na generalnom korpusu

dokumenata u odnosu na onaj treniran nad domenski specifičnim skupom podataka. Iz ovog razloga se u ovom radu fokus stavlja na specifičan Word2vec model, treniran nad korpusom pravnih dokumenata, a koji se naziva Law2vec.

BERT

Pretrenirani modeli zasnovani na transformerima, kao što je BERT i njegove varijante, su se pokazali kao vrlo uspešni u mnogim NLP zadacima nad generičkim skupovima podataka (skup tekstova sa Vikipedije itd). Specifičnost BERT modela (Bidirectional Encoder Representations for Transformers) je način generisanja interne reprezentacije pojmova neke rečenice iz teksta. Arhitektura modela je takva da se kao ulaz očekuje čitava rečenica, gde se zatim nad svakoj od reči vrše transformacije koje usko zavise od ostalih reči u rečenici. Ovo generalno znači da model kreira reprezentaciju pojma na osnovu konteksta u kome se on pojavljuje. Dobar primer su sinonimi. Reč „kosa“ može imati potpuno različita značenja u različitim kontekstima, stoga se očekuje da i njena računarska reprezentacija bude drugačija u različitim situacijama. BERT model ovakve probleme uspešno prevazilazi, ali se naravno pojavljuju problemi kod skupova podataka koji su vrlo specifični za neki domen. Ukoliko se razmatra skup pravnih dokumenata može se zdravorazumski uvideti razlika u kontekstima u kojima se određene reči i/ili rečenice pojavljuju. Reprezentacija pravnih pojmova, ili pak običnih reči koje se pojavljuju u kontekstu pravnog dokumenta bi trebalo da bude malo drugačija. Sa druge strane neki pojmovi iz domena prava se ne javljaju u svakodnevnom govoru, te modeli obučavani nad opštim, svakodnevnim tekstovima se vrlo verovatno sa njima nisu ni „susreli“. Ove male razlike u reprezentacijama mogu rezultovati velikim greškama prilikom konačne primene modela. Zbog navedenih razloga sada imamo pristup mnoštvu domenski specifičnih BERT modela, a u ovom radu se fokus stavlja na model pretreniran nad skupom pravnih dokumenata LegalBERT.

Metode klasifikacije – pridruživanja oznaka

Za savladavanje problema kojim se bavi ovaj rad, pridruživanje višestrukih oznaka pravnim dokumentima, mogu se iskoristiti razne kombinacije metoda veštačke inteligencije. Generalni pristup se pak svodi na tri okvirne faze: predprocesiranje dokumenta, kreiranje njegove računarske reprezentacije, odnosno vektorizacija i naposletku sama klasifikacija dokumenta, odnosno pridruživanje njemu najverovatnijih oznaka.

Imajući u vidu, pre svega, ograničenja u vidu računarskih resursa koji su bili na raspolaganju prilikom istraživanja, a o čemu će biti više reči kasnije, ovaj rad obuhvata primenu i eksperimentisanje sa ograničenim brojem i kombinacijom metoda vektorizacije i modela mašinskog učenja.

Metode predprocesiranja teksta i za generisanje vektorske reprezentacije koje su iskorišćene su već prethodno navedene u poglavljima „predprocesiranje skupa podataka“ i „metode vektorizacije“.

Posao dodeljivanja višestrukih oznaka dokumentu, odnosno njegova klasifikacija može biti generalno odrađen korišćenjem dva pristupa.

Prvi pristup podrazumeva transformaciju problema dodeljivanja višestrukih oznaka u problem dodeljivanja jedne klase/oznake. Ovim pristupom bi za svaku oznaku bio kreiran algoritam, odnosno nezavisni model koji donosi sud da li ta oznaka pripada specifičnom dokumentu. Agregacijom rezultata svih modela za svaku oznaku bi bio generisan konačni skup oznaka koje mogu biti dodeljene dokumentu. Kompleksnost primene ovog pristupa raste sa porastom broja oznaka koje mogu biti dodeljene dokumentima, te u slučaju skupa podataka koji se razmatra u ovom radu ne predstavlja praktično rešenje.

Drugim pristupom bi rešenje problema zahtevalo od algoritma generisanje verovatnoće za svaku oznaku, odnosno koliko je koja oznaka verovatna za konkretni dokument.

Algoritmi, odnosno modeli koji bi mogli biti korišćeni obuhvataju: SDGClassifier, Logističku regresiju, SVM, KNN, XGBoost, LSTM, „klasičnu“ neuronsku mrežu i razne varijacije RNN i CNN kao što su BIGRU, HAN, CNN-LWAN i druge.

Iz praktičnih razloga i na osnovu zaključaka iz relevantne literature ovaj rad koristi manji podskup navedenih metoda. Algoritmi koji su istraženi i isprobani su: potpuno povezana neuronska mreža sa jednim skrivenim slojem, KNN, XGBoost, LSTM i (prošireni) LegalBERT.

Potpuno povezana neuronska mreža

FCNN (fully connected neural network) je tip veštačke neuronske mreže čija arhitektura podrazumeva da su svi čvorovi, odnosno neuroni, jednog sloja potpuno povezani sa čvorovima narednog sloja. Za zadatke u ovom radu ulaz u prvi sloj neuronske mreže je neka od reprezentacija dokumenta dok se na izlazu očekuju verovatnoće za svaku pojedinačnu oznaku. Vršena su podešavanja parametara modela u cilju otkrivanja najadekvatnije arhitekture.

KNN

KNN (K nearest neighbours) je jedan od starijih i dobro poznatih algoritama nadgledanog mašinskog učenja korišćen za zadatke klasifikacije i regresije. U osnovi, algoritam pretpostavlja da se slični podaci, predstavljeni u n-dimenzionalnom prostoru, nalaze blizu jedan drugog. Predviđanje pripadnosti nekog podatka se onda svodi na određivanje grupe podataka kojima je on najbliži.

Za konkretan problem iz ovog rada korišćen je derivat originalnog KNN algoritma pod nazivom ML-KNN, gde se prefiks ML odnosi na višestruke oznake (multi-label). Ovaj izmenjeni algoritam pronalazi najbliže podatke (njih K) podatku za koji se trenutno određuju oznake. Od pronađenih sličnih podataka se zatim preuzimaju njihove poznate oznake. Za novi podatak se, zatim, iz ovog skupa oznaka, MAP (maximum a posteriori) principom određuje skup njemu odgovarajućih oznaka. MAP princip je u osnovi zasnovan na Bajesovoj teoremi za procenu distribucije i parametara koji najbolje objašnjavaju date podatke.

Gradient boosting algoritam

Gradient boosting je tehnika mašinskog učenja za rešavanje problema klasifikacije i regresije, koja generiše model za predviđanje u formi ansambla ili niza „slabih“ prediktivnih modela (koji su često stabla odlučivanja). Slab model se ovde odnosi na onaj model koji je u stanju da reši samo vrlo jednostavan problem. Takav model može biti dobar za donošenje dela odluke u nekom procesu, ali ukoliko se posmatra celokupni problem, model ne daje zadovoljavajuće rezultate. Ideja gradient boosting-a je korišćenje mnoštva ovakvih slabih modela, od kojih je svaki dobar za rešavanje različitog dela problema u odnosu na druge. Pretpostavka je, onda, da bi kombinacijom rezultata svih slabih modela u ansamblu bila dobijena konačna, tačna odluka.

Za potrebe ovog rada korišćena je implementacija gradient boosting algoritma iz biblioteke XGBoost.

LSTM

Prilikom obrade prirodnog jezika pogodno je, u jednom vremenskom koraku, posedovati informacije iz drugih vremenskih koraka. Jedan od često korišćenih modela je RNN koji za mnoge probleme daje dobre rezultate, ali ima manu koja ograničava njegovu upotrebu za poslove NLP-a, a naziva se nestajući gradijent. Gradijenti, odnosno vrednosti kojima se ažuriraju težine unutar neuronske mreže, imaju tendenciju da se smanjuju tokom vremena. Ova pojava rezultuje vrlo sporim učenjem, odnosno prestankom učenja slojeva mreže čiji je gradijent vrlo mali nakon nekog vremena. Od ovog problema prvo „stradaju“ prvi slojevi mreže, što znači da u dužim sekvencama mreža „zaboravlja“ ranija saznanja i za nju se kaže da ima „kratkotrajno pamćenje“. „Zaboravljanje“ koje se javlja, često predstavlja problem prilikom obrade teksta jer se može izgubiti kontekst koji je nekad vrlo važan u prirodnom jeziku.

LSTM (long short term memory) model je pokušaj otklanjanja prethodnog problema, izmenom „ćelija“, tj. čvorova mreže na takav način da čvor može biti naučen da pamti određene informacije dok druge zaboravlja. Ove informacije su predstavljene kao stanje čvora koje se propagira kroz celokupnu mrežu produžavajući njeno pamćenje (memoriju). Ovakvo ponašanje ćelija LSTM modela se postiže usložnjavanjem njihove interne arhitekture tako da se sastoje od kombinacije niza kapija i slojeva sa naučenim težinama, čime se reguliše protok informacija kroz lanac sekvenci modela. Detalji načina interakcije ovih slojeva izlaze iz okvira ovog rada, te neće biti detaljnije razmatrani. Važna karakteristika bitna za domen obrade prirodnog jezika, govoreći na višem nivou apstrakcije, je mogućnost modela za pamćenjem i propagiranjem konteksta.

Eksperimenti i rezultati

Rešavanje problema klasifikacije pravnih dokumenata na osnovu višestrukih oznaka isprobano je kombinacijama metoda predprocesiranja, vektorizacije i klasifikacije, koje su navedene u prethodnim poglavljima. U ovom poglavlju daje se pregled svake od istraženih kombinacija, dobijenih rezultata i sažetak zapažanja do kojih su autori došli prilikom istraživanja.

TF-IDF & neuronska mreža

Dodeljivanje oznaka pravnim dokumentima u ovom eksperimentu započinje pripremom podataka za njihovu vektorizaciju. Pripremljeni podaci se zatim vektorizuju korišćenjem TF-IDF tehnike i tako predstavljeni prosleđuju na ulaz neuronskoj mreži koja sprovodi njihovu klasifikaciju.

Prvi korak obuhvata pripremu podataka. Najpre, uzimajući u obzir strukturu pravnih dokumenata nad kojima se vrše eksperimenti, izvršena je selekcija delova dokumenta koji se uzimaju u obzir. Prethodno je navedeno da se svaki dokument sastoji iz 5 glavnih delova (naslov, zaglavlje, reference, glavni deo i prilozi). Eksperiment je ponavljan više puta, gde je prilikom svakog ponavljanja u dalju obradu uziman podskup od ovih 5 delova dokumenata.

Od ranije spomenutih metoda ovde je nad svakim podatkom (dokumentom) primenjeno: izbacivanje svih ne-alfanumerika, izbacivanje „stop“ reči, izbacivanje brojeva, stemovanje ili lematizacija reči, a od preostalih reči su izdvojene samo imenice i glagoli. Ovako je dobijen dosta pročišćen niz reči, ali je zadržana generalna semantika i značenje dokumenta. Prilikom ove obrade se iz dokumenta izdvajaju i njemu dodeljene oznake koje se transformišu u „one-hot-encoded“ oblik koji će biti korišćen u procesu treniranja neuronske mreže.

Sledeći korak obuhvata generisanje numeričke predstave svakog dokumenta u obliku vektora. Nad predprocesuiranim dokumentom, koji je sada predstavljen kao skup reči, se primenjuje algoritam dobijanja vektora uz pomoć TF-IDF mere koji je opisan u poglavlju TF-IDF metoda vektorizacije.

Međurezultat u tom trenutku se može predstaviti kao niz vektora od kojih svaki predstavlja jedan dokument iz korpusa.

Krajnji korak obuhvata obučavanje neuronske mreže nad pripremljenim vektorima dokumenata i njima odgovarajućim „one hot encoded“ vektorima oznaka.

Eksperiment je ponavljan upotrebom različitih parametara i arhitektura neuronske mreže, koji su davali i različite rezultate. Isprobane su neuronske mreže sa:

- 1 i 2 skrivena sloja
- 2048/1024/512 čvorova skrivenog sloja
- Relu/selu/tanh/exponencijalna aktivacionom funkcijom skrivenog sloja
- Binary Crossentropy i Weighted binary crossentropy loss funkcijom
- Sigmoidalnom aktivacionom funkcijom izlaznog sloja

Od svih isprobanih najbolje se pokazala neuronska mreža sa jednim skrivenim slojem sa 2048 čvorova, selu aktivacionom funkcijom i sa weighted binary crossentropy loss funkcijom koja nudi podešavanje odnosa preciznosti i odziva (*precision* & *recall*).

Eksperimenti su vršeni sa kombinacijama prethodnih mogućnosti za pripremu podataka, vektorizaciju i izbor neuronske mreže. Sve kombinacije su dale zadovoljavajuće rezultate koji se međusobno vrlo malo razlikuju. Kombinacija koja je dala najbolje rezultate u pogledu preciznosti, odziva i F mere je ona koja:

- U obzir uzima 3 dela dokumenta i to naslov, zaglavlje i reference,
- Koristi sve navedene metode predprocesiranja, uz odabir lematizacije i
- Neuronske mreže koja je prethodno navedena kao najbolja.

Iako ovaj pristup daje najbolje rezultate, treba napomenuti da su se eksperimenti u kojima su korišćeni samo naslov i zaglavlje dokumenta kao i stemovanje umesto lematizacije, pokazali kao vrlo uspešni. Njihovi rezultati ne odudaraju mnogo u pogledu vrednosti za preciznost, odziv i F meru, dok se vreme potrebno za obradu podataka i dobijeni vektori znatno smanjuju.

Ono što je generalni zaključak ovog eksperimenta je dosta dobra uspešnost primenjenih metoda u rešavanju zadatog problema, nad skupom podataka koji se razmatra, iako su sve korišćene metode jednostavne i lake za razumevanje, te ne predstavljaju najsavremenija rešenja u svetu NLP-a.

Treba navesti i iznenađujuće zapažanje u vezi sa informativnošću delova pravnog dokumenta. Naime, vidimo da je najbolji rezultat dobijen korišćenjem samo naslova, zaglavlja i referenci iz dokumenata, što znači da glavni deo dokumenta, koji bi zdravorazumski trebalo da nosi i najveće značenje, zapravo ima mali uticaj na konačno rešenje, a čak je i lošije za rezultat. Ovakvo ponašanje se može objasniti time da korišćenjem glavnog dela dokumenta generisani vektori postaju znatno veći (vektor ukoliko se koriste svi delovi dokumenta – 90000 jedinstvenih reči u korpusu dokumenata, dok najbolji rezultat radi sa 28000 jedinstvenih reči).

Law2vec & neuronska mreža

Eksperimenti rađeni uz pomoć Word2vec modela za vektorsku reprezentaciju dokumenta se po strukturi ne razlikuju mnogo od prethodnog pristupa. Generalni pristup takođe obuhvata isprobavanje raznih kombinacija za pripremu podataka i parametara neuronske mreže, dok su vektori dokumenata generisani uz pomoć postojećeg Law2Vec modela.

Prilikom pripreme podataka primenjene su sve ranije navedene metode predprocesiranja osim ekstrakcije imenica i glagola, a izostavljeni su i procesi stemovanja i lematizacije. Ovakav pristup se poklapa sa metodama predprocesiranja koje su primenjene prilikom obučavanja Law2vec modela, a koje spominju njegovi autori u pratećoj dokumentaciji.

Za generisanje vektorske reprezentacije dokumenta isprobana su dva pristupa koja se oslanjaju na strukturu dokumenta.

Prvi pristup podrazumeva objedinjavanje svih delova dokumenta u jedan, određivanje vektora za svaku reč uz poštovanje njenog značaja za dokument (u vidu njene frekventnosti u dokumentu ili njene TF-IDF vredosti). Krajnji vektor koji reprezentuje dokument je tada predstavljen kao agregacija vektora svih reči iz tog dokumenta.

Drugim isprobanim pristupom se prvo generišu vektori za odgovarajuće delove dokumenta (primena prvog pristupa na svaki deo posebno), a zatim se dobijeni vektori konkateniraju u jedan veći. Pretpostavka je da ovako dobijeni vektor nosi više značenja u sebi, poštujući strukturu pravnog dokumenta. Sa druge strane, dobijena reprezentacija dokumenta je višestruko veće dimenzionalnosti što zahteva mnogo veću upotrebu resursa i vremena.

Vektori se, na kraju, šalju kao ulaz za obučavanje neuronske mreže, zajedno sa „one hot encoded“ oznakama za svaki dokument.

Rezultati eksperimenata pokazali su da je razlika u korišćenju različitih navedenih pristupa vrlo mala, ali se ipak ističe onaj u kome je:

- Korišćena TF-IDF vrednost za ponderisanje vrednosti vektora, umesto frekventnosti reči,
- Primenjen drugi pristup za generisanje vektora dokumenta i
- Korišćena neuronska mreža sličnih karakteristika kao u prethodnom eksperimentu, gde je razlika jedino u broju čvorova skrivenog sloja koji je ovde 1024.

Početna očekivanja od ovog pristupa sa Law2vec modelom, koja su bila dosta velika, su se pokazala previše optimističnim. Zapaženi su lošiji rezultati u odnosu na vektorizaciju dokumenta uz pomoć TF-IDF mere, dok je za treniranje neuronske mreže bio potreban veći broj epoha. Mogu se identifikovati dva potencijalna razloga za ovakve rezultate. Prvi razlog je korišćenje karakterističnih vektora u kombinaciji sa neuronskom mrežom. Vrlo je verovatno da modelu kao što je neuronska mreža više „odgovaraju“ vektori koji na nekim mestima imaju vrednosti a na nekime ne, u odnosu na vektore koji su svuda popunjeni. Drugi potencijalni razlog može biti sama priroda dobijenih vektora, koji dobro predstavljaju internu semantiku dokumenta koja možda nema toliko veliki značaj za konkretne oznake koje mu se dodeljuju.

Ostali eksperimenti

BERT

Na početku istraživanja postojala je pretpostavka da bi klasifikator zasnovan na predstavama dokumenata nastalim pomoću BERT modela treniranom na korpusu pravnih dokumenata dao najbolje rezultate. Čak štaviše, velika uspešnost primene ovog pristupa se pominje u relevantnoj literaturi, ali istraživanja u ovom radu nisu dovela do sličnih rezultata.

Pristup koji koristi LegalBERT model za enkodiranje dokumenata zasniva se na konceptu prenosa znanja (transfer learning), gde se pretrenirani BERT model proširuje klasifikacionim slojem, a zatim trenira nad konkretnim podacima za rešavanje željenog problema. Proširenom modelu se prosleđuju tokenizovani dokumenti, koji nisu prethodno predprocesirani. Priroda BERT modela je da je moguće obraditi najviše 512 tokena, što se pokazao kao veliki problem prilikom rada sa velikim dokumentima, jer je potrebno odbaciti i potpuno zanemariti njihov veći deo.

Rezultati ovakvog pristupa su se pokazali veoma lošim. Uzroci lošeg rada ovog pristupa obuhvataju:

- Nedostatak resursa. Ovo je najznačajniji razlog neuspeha jer razmatrani pristup zahteva veoma veliku količinu računarskih resursa za pokretanje glomaznog proširenog BERT modela. Prvenstveno, tokom eksperimenata, nije bilo moguće pokrenuti model nad više od 200 tokena, što znači da je svaki dokument efektivno skraćen u proseku za 73%. Nadalje, čak i ako se koristi ovako smanjen broj tokena, bilo

bi potrebno više od 9 sati za obučavanje krajnjeg modela, što se pokazalo kao krajnje nepraktično za kontinuirano eksperimentisanje i nameštanje parametara.

- Reprezentacija reči dokumenta potencijalno ne odgovara datom zadatku. Verovatno je da reprezentacija koja se dobija primenom BERT modela nosi puno informacija u vezi sa kontekstom u kojem se neka reč nalazi, što je odlično za poslove predviđanja naredne reči/rečenice, analizu sentimenta i slično, ali ne doprinosi otkrivanju oznake na nivou celog dokumenta.
- Nedostatak podataka u obučavajućem skupu. U relevantnoj literaturi se najčešće spominju brojke od preko 50 hiljada dokumenata, dok je u ovom radu na raspolaganju bilo upola manje.

Law2vec & LSTM

Ovaj pristup je vrlo sličan prethodno opisanom pristupu gde je korišćen Law2vec model sa jednostavnom neuronskom mrežom. Razlika se ogleda u uvođenju LSTM sloja u arhitekturu neuronske mreže. Uvođenje ovog sloja dovodi do drastičnog povećanja kompleksnosti klasifikacionog sloja, te su dobri rezultati, kao i u prethodnom slučaju, izostali, iz vrlo sličnih razloga.

Nedostatak resursa za obučavanje kompleksnog klasifikatora, kao i vrlo veliko vreme obučavanja glavni su razlozi neuspeha ovog pristupa prilikom vršenja eksperimenata.

TF-IDF & ML-KNN

Ovaj pristup podrazumeva sličnu strukturu kao i prvi navedeni pristup sa običnom neuronskom mrežom i vektorima generisanim pomoću TF-IDF vrednosti. Neuronska mreža je zamenjena prethodno opisanim ML-KNN klasifikatorom, dok su ostali procesi pripreme i vektorizacije dokumenata ostali isti.

Iako u osnovi vrlo jednostavan model, KNN adaptiran za potrebe multi-label klasifikacije se u pogledu rezultata pokazao uporedivim sa pristupima gde je korišćena neuronska mreža. Međutim, obučavanje ML-KNN modela, kao i njegovo pokretanje nad nepoznatim podacima oduzima drastično više vremena u odnosu na najuspešnije metode (najmanje sat vremena potrebno za ML-KNN pristup u odnosu na približno 3 minute za neuronsku mrežu).

TF-IDF & Gradient boosting

Ovaj pristup je u osnovi dosta istražen i u proces implementacije se ušlo sa pretpostavkom o mogućoj dobroj uspešnosti. Inicijalna ideja je obuhvatala predprocesiranje i vektorizaciju dokumenata na sličan način kao u pristupu sa neuronskom mrežom, dok bi ona bila zamenjena implementacijom Gradient boosting algoritma iz XGBoost biblioteke.

Nažalost, usled nedostatka resursa, a prvenstveno radne memorije, treniranje ovakvog modela nije bilo moguće. Pretpostavka je da je zbog prirode problema koji se rešava, a posebno zbog velikog broja oznaka koje se određuju, bilo potrebno kreiranje velikog broja internih „slabih“ modela što je dovelo do popunjavanja raspoložive memorije.

Pregled i analiza rezultata

Prilikom evaluacije eksperimenata korišćene su sledeće mere:

- Micro average precision
- Micro average recall
- Micro average F score
- Macro average precision
- Macro average recall
- Macro average F score

Micro average metod – korišćenjem ove metode za evaluiranje rezultata veći značaj se daje primerima čije su oznake više zastupljene u skupu podataka u odnosu na prosek.

Macro average metod – suprotno od prethodnog, veći značaj se daje primerima čije oznake se pojavljuju manje puta u odnosu na prosek.

Ukoliko su vrednosti mera računatih *micro average* metodom značajno manje od onih dobijenih *macro average* metodom može se zaključiti da model pogrešno vrši klasifikaciju nad primerima sa najzastupljenijim oznakama. U obrnutom slučaju se može zaključiti da model pogrešno klasifikuje primere sa najmanje zastupljenim oznakama.

	MICRO AVG PRECISION	MICRO AVG RECALL	MICRO AVG F SCORE	MACRO AVG PRECISION	MACRO AVG RECALL	MACRO AVG F SCORE
LAW2VEC PRISTUP 1 & TF-IDF POND. & NN	0.54	0.61	0.57	0.46	0.44	0.40
LAW2VEC PRISTUP 2 & TF-IDF POND. & NN	0.60	0.57	0.59	0.47	0.40	0.40
TF-IDF & ML-KNN – K=3 & NN	0.74	0.59	0.66	0.55	0.42	0.45
TF-IDF & NN (OZNAKE U MIN 10 DOC)	**0.73**	**0.62**	**0.67**	**0.57**	**0.44**	**0.47**
TF-IDF & NN (SVE OZNAKE)	0.73	0.58	0.65	0.25	0.19	0.20
TF-IDF & NN (OZNAKE U MIN 50 DOC)	**0.73**	**0.66**	**0.69**	**0.69**	**0.56**	**0.60**

1 Pregled najboljih rezultata

Na slici 1 je dat pregled rezultata dobijenih u odgovarajućim eksperimentima. Može se zaključiti da su rezultati evaluirani metodama koje koriste *micro average* znatno veći u odnosu na *macro average* pristup, što dalje znači da su modeli generalno bolje obučeni za rad sa dokumentima čije su oznake među najzastupljenijima. Ovakvo ponašanje modela se najviše može pripisati nebalansiranosti samog skupa podataka.

Najbolji rezultati su postignuti korišćenjem kombinacije pristupa sa vektorizacijom dokumenata uz pomoć TF-IDF vrednosti i neuronskom mrežom i to na smanjenom skupu podataka, gde su u obzir uzimane samo oznake koje se nalaze u najmanje 10 (1289 oznaka) odnosno 50 (648 oznaka) dokumenata. Za dobijanje ovih rezultata korišćena je *weighted binary crossentropy* loss funkcija koju je moguće parametrizovati za dobijanje željenog odnosa

preciznosti i odziva modela. Vrednost parametra koji je korišćen je 1.7, čijim smanjivanjem bi se dobila veća preciznost, a povećavanjem bolji odziv. Ovo može biti korisno za konkretnu primenu te predstavlja parametar vredan pažnje u produkcionom sistemu.

Rezultati korišćenjem pristupa sa BERT i LSTM modelima, koji su ranije opisani, a najverovatnije zbog već navedenih razloga, su veoma loši.

Najbolji rezultati pristupa sa korišćenjem LegalBERT modela su dobijeni merama koje se zasnivaju na *micro average* metodi: 0.07 (precision), 0.11 (recall) i 0.09 (F score).

Takođe, pristup uz korišćenje LSTM modela dao je slične rezultate: 0.08 (precision), 0.15 (recall) i 0.09 (F score).

Problemi i ograničenja

Tokom istraživanja i sprovođenja eksperimenata autori su se susreli sa problemima:

- Nedostatka računarskih resursa,
- Izrazito nebalansiranog skupa podataka
- Potencijalno nedovoljnog broja primera u skupu podataka

Pretpostavka je da bi se rešavanjem barem prvog problema stekli uslovi za sprovođenje mnogo zahtevnijih eksperimenata i dobijanje boljih rezultata primenom najsavremenijih metoda.

Zaključak

U ovom radu prezentovane su i upoređene metode i pristupi različite kompleksnosti za klasifikaciju pravnih dokumenata na osnovu višestrukih oznaka. Izvršeni su mnogi eksperimenti kako sa različitim numeričkim reprezentacijama dokumenata tako i sa klasifikacionim modelima mašinskog učenja. Iako, nažalost, uslovi nisu bili adekvatni za dobijanje dobrih rezultata upotrebom najsavremenijih metoda, potvrđena je upotrebljivost tradicionalnijih i jednostavnijih pristupa. Metode ovde razmatrane, a zasnovane na upotrebi TF-IDF i klasične neuronske mreže su, iako najjednostavnije, dale najbolje rezultate. Takav pristup bi svakako trebalo razmotriti u uslovima ograničene procesne moći i neadekvatnog označenog skupa podataka. Utvrđeno je i da mogućnost korišćenja BERT modela drastično opada prilikom rada sa velikim tekstovima, kakvi su pravni dokumenti.

Važno je naglasiti značaj strukture pravnih dokumenata. Utvrđeno je da različiti delovi dokumenta nose različitu količinu korisnih informacija za izvršavanje konkretnog zadatka. Kao najuticajniji delovi istakli su se naslov, zaglavlje i reference dokumenta, iz čega se može zaključiti da njihovi autori vrlo verovatno i sami izdvajaju najznačajnije informacije. Reference pravnog dokumenta su najčešće drugi pravni akti, a značaj ovog dela za konkretni zadatak može biti objašnjen tako da slični dokumenti često referenciraju slične pravne akte. Mogućnost izdvajanja samo pojedinih delova pravnog dokumenta dodatno pojednostavljuje razvijanje i poboljšava performativnost prikazanih metoda prvenstveno smanjivanjem količine podataka koju je potrebno numerički predstaviti. Ova osobina bi bila posebno korisna u uslovima postojanja više (desetina) hiljada oznaka.

Uz postojanje odgovarajućeg skupa podataka iz željenog pravnog domena mogao bi biti razvijen softverski alat za ispomoć domenskim stručnjacima za označavanje, kategorizaciju, pretragu i pri drugim poslovima za rad sa dokumentima.

Literatura

[1] https://arxiv.org/pdf/2010.01653.pdf An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels

[2] https://archive.org/details/Law2Vec Law2Vec: Legal Word Embeddings

[3] https://yashuseth.blog/2019/06/12/bert-explained-faqs-understand-bert-working/ BERT explained

[4] https://arxiv.org/abs/1912.06905 Long-length Legal Document Classification

[5] https://tomaxent.com/2018/04/27/Micro-and-Macro-average-of-Precision-Recall-and-F-Score Micro- and Macro-average of Precision, Recall and F-Score

[6] https://www.researchgate.net/publication/4196695_A_k-nearest_neighbor_based_algorithm_for_multi-label_classification A k-nearest neighbor based algorithm for multi-label classification

[7] https://dl.acm.org/doi/abs/10.1145/3077136.3080834 Deep Learning for Extreme Multi-label Text Classification

[8] https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5 Multi Label Text Classification with Scikit-Learn

[9] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21 Guide to LSTM's and GRU's: A step by step explanation

[10] https://colah.github.io/posts/2015-08-Understanding-LSTMs/ Understanding LSTM Networks

[11] https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d Gradient Boosting from scratch

[12] https://arxiv.org/pdf/2010.02559.pdf LEGAL-BERT: The Muppets straight out of Law School

[13] https://machinelearningmastery.com/maximum-a-posteriori-estimation/ Introduction to Maximum a Posteriori (MAP) for Machine Learning

[14] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 Machine Learning Basics with the K-Nearest Neighbors Algorithm

[15] https://towardsdatascience.com/journey-to-the-center-of-multi-label-classification-384c40229bff Deep dive into multi-label classification..! (With detailed Case Study)

[16] https://arxiv.org/pdf/1905.10892.pdf Large-Scale Multi-Label Text Classification on EU Legislation