Zbornik radova Fakulteta tehničkih nauka, Novi Sad

**UDK: 004.9**
**DOI: https://doi.org/10.24867/17BE23Zuzic**

# DETEKCIJA OBLIGACIJA U UGOVORIMA NA ENGLESKOM JEZIKU

# OBLIGATION DETECTION IN ENGLISH CONTRACTS

Marko Žužić, *Fakultet tehničkih nauka, Novi Sad*

**Oblast – ELEKTROTEHNIKA I RAČUNARSTVO**

**Kratak sadržaj –** *U ovom radu predstavljen je sistem za detekciju obligacija u okviru ugovora napisanih na engleskom jeziku. Klasifikator obligacija kao ulaz prima rečenice iz ugovora, a kao izlaz daje informaciju da li su rečenice obligacije, ili ne.*

**Ključne reči:** *Detekcija obligacija, Pravni domen, Jezički modeli, NLP, Klasifikacija teksta, BERT*

**Abstract** – *This paper presents a system for obligation detection in english contracts. The classifier trained to solve the problem accepts sentences from contracts as an input and outputs information on whether they are considered obligations, or not.*

**Keywords:** *Obligation detection, Legal domain, Language models, NLP, Text classification, BERT*

## 1. UVOD

Pisani ugovori su sastavni deo današnjeg poslovnog sveta. Prilikom stupanja u bilo kakav poslovni odnos sklapa se ugovor kojim se definiše odnos između ugovornih strana. Pregledanje ugovora treba da obavi pravno obučeno lice i ono iziskuje dosta vremena, a samim tim i novca. Jasna je motivacija za pravljenje sistema koji bi delimično, ili u potpunosti, automatizovao, a samim tim i olakšao i ubrzao, proces analize pravnih dokumenata. Obligacije, kao osnova svakog pravnog dokumenta, predstavljaju bitan segment ovakvog sistema. Stoga je u ovom radu ponuđeno rešenje problema automatske detekcije obligacija.

Skup podataka koji je korišćen u radu sačinjen je od ručno anotiranih rečenica ("jeste obligacija", ili "nije obligacija") iz javno dostupnih ugovora.

Algoritmi koji su korišćeni za rešavanje problema detekcije obligacija birani su na osnovu dostupnih rezultata iz literature koja je pisana na istu ili sličnu temu. Isprobani su algoritmi za vektorizaciju rečenica, poput *bag of words* i *TFIDF*, kao i složeniji, poput *FastText*-a. Nakon vektorizacije, trenirano je nekoliko različitih binarnih klasifikatora. Kao osnovni klasifikator isproban je *Naive Bayes*, a potom i napredniji klasifikatori, poput *SVM* i *XGBoost*-a. Isprobane su i tehnike dubokog učenja, poput konvolutivnih neuronskih mreža i *BERT* transformer modela.

Dobijeni rezultati potvrdili su performanse algoritama vektorizacije teksta i klasifikatora koje su dobijene u prethodnim radovima. Najbolje rezultate dale su tehnike dubokog učenja (88% F1 skor prilikom korišćenja *LegalBERT* modela), dok su kontekstualne tehnike vektorizacije rečenica u kombinaciji sa jednostavnijim klasifikatorima dale slabije rezultate (82% F1 skor za *XGBoost* klasifikator).

U narednom poglavlju biće analizirana relevantna literatura za problem detekcije obligacija koji se analizira. U trećem poglavlju biće opisana predložena metodologija za rešavanje problema, dok četvrto poglavlje nudi uvid u rezultate i njihovu diskusiju. U petom poglavlju dat je zaključak o samom radu i problemu koji je rešavan.

## 2. PRETHODNA REŠENJA

Klasifikacija delova ugovora, poput klauzula, ili obligacija, nije obimno istražena tema u literaturi.

U radu [1] vršena je analiza tehnika za klasifikaciju sudskih procesa na različite pravne domene. Autori su implementirali ansambl sistem, sastavljen od nekoliko *SVM* klasifikatora sa linearnim kernelom i dostigli F1 skor od 96% nad skupom automatski označenih pravnih dokumenata kojih je bilo 126,865. Skup obeležja koji su iskoristili bio je jednostavan: *bag of words* sa unigramima i *bag of bigrams*, gde su gledali samo broj pojavljivanja svakog unigrama, ili bigrama u celom korpusu reči.

U radu [2] rađena je *multilabel* klasifikacija nad 57,000 pravnih legislativa Evropske Unije na 4,300 različitih klasa. Autori ovog rada su kao osnovnu arhitekturu koristili algoritam logističke regresije sa *TF-IDF* vektorima ngrama reči u korpusu. Ova tehnika vektorizacije isprobana je i u ovom radu.

Kao napredne arhitekture u radu [2] korišćene su kompleksne neuronske mreže, poput *CNN-LWAN* i *BIGRU*. *CNN-LWAN* mreža podrazumeva neuronsku mrežu, gde su obeležja dokumenta (rečenice) su enkodovane preko *CNN* sloja mreže. Druga arhitektura koju su isprobali bila je *BIGRU* mreža, u okviru koje je dokument predstavljen kao niz vektora pojedinačnih reči u njemu, koji dalje prolazi kroz *BIGRU* sloj (specijalna vrsta rekurentne neuronske mreže). Arhitektura koja je dala najbolji rezultat u ovom radu bila je kombinacija *CNN-LWAN* i *BIGRU* arhitekture. Stoga je odlučeno da se i u ovom radu isprobaju neke specijalne tehnike kreiranja vektora, poput konvolutivne neuronske mreže koju koristi biblioteka *spaCy*. Pored pomenutih arhitektura, autori su isprobali da koriste i *BERT* transformer algoritam. On je dao dobre rezultate, te je stoga isproban kao jedan od algoritama u ovom radu.

**NAPOMENA:**
**Ovaj rad proistekao je iz master rada čiji je mentor bio prof. dr Aleksandar Kovačević.**

Rad [3] i drugi radovi na temu klasifikacije rečenica često isprobavaju kako druge tehnike vektorizacije utiču na rezultate. Jedna od tehnika vektorizacije je kreiranje *FastText* vektora, koja je korišćena u radovima [1], [2] i [3]. U radu [1] autori su pomoću jednostavnog linearnog klasifikatora i *FastText* vektora postigli rezultate slične onima dobijenim korišćenjem tehnika dubokog učenja (konvolutivne mreže). *FastText* vektorizacija je istražena i u ovom radu.

Pored *FastText* vektorizacije, autori radova na sličnu temu koristili su i druge algoritme koji uzimaju u obzir kontekst oko pojedinačnih reči u rečenici, poput *GloVe* i *Word2Vec* algoritma [1], [3]. Ove tehnike isprobane su i u ovom radu.

## 3. METODOLOGIJA I ALATI

Ovo poglavlje posvećeno je prikazu primenjene metodologije za detekciju obligacija: korišćenog skupa podataka, tehnika vektorizacije teksta i odabranih klasifikatora.

### 3.1. Formiranje skupa podataka

Skup podataka je dobijen manuelnim anotiranjem rečenica koja su izvučene iz javno dostupnih ugovora. Rečenice su anotirala pravna lica. Svaka rečenica označavana je jednom od dve klase: "jeste obligacija" (1), ili "nije obligacija" (0). Klase su prilično izbalansirane, imajući 7944 pozitivnih primera i 7001 negativnih primera. Trening i test skup su dobijeni deljenjem skupa podataka u odnosu 80:20, prilikom čega je očuvana distribucija pozitivnih i negativnih primera.

Skup podataka je formiran iterativno. Početni, manji skup podataka, formiran je korišćenjem jednostavnih pravila (korišćenjem frazi *"should", "should not", "is obliged", "is not obliged", "is obligated", "is not obligated"*). On je dat pravnicima, koji su formirali skup od 200 rečenica koje su zaista bile obligacije. Od ostatka skupa uzeto je još 200 rečenica koje nisu obligacije, kako bi se dobio balansiran skup podataka za treniranje početnog algoritma.

Za dalje proširivanje skupa podataka najpre je istreniran početni algoritam nad 400 obligacija. Ovaj model je dalje korišćen za inicijalno obeležavanje preostalog skupa podataka koji je sakupljen pomoću pravila, pomenutih u pasusima iznad.

Ovako je skup formiran iterativno. Pretpostavka je bila da će algoritam davati bolje rezultate sa većom količinom podataka.

### 3.2. Tehnike preprocesiranja teksta

Iz formalnih rečenica koje su sačinjavale skup podataka su izbačeni znakovi interpunkcije, specijalni znakovi i stop reči (engl. *stop words*), dok je ostatak teksta prebačen u mala slova [7].

### 3.3. Vektorizacija teksta

Odabrano je i isprobano nekoliko tehnika vektorizacije teksta, po ugledu na prethodnu literaturu na sličnu temu. Sve tehnike opisane su u nastavku teksta.

*Bag of words* je pojednostavljena reprezentacija teksta u okviru koje se delovi teksta predstavljaju frekvencijom reči koje taj tekst čine [8]. Ova tehnika je istražena u ovom radu kao osnovna tehnika (engl. *baseline*), kao i u radu [1].

Kod *TF-IDF* tehnike broj pojavljivanja termina deli se ukupnim brojem termina koji su prisutni u tom dokumentu. Kako bi se termini razlikovali po važnosti, uvodi se "*Inverse Document Frequency*", koji dodeljuje veću težinu terminima koji su retki. Što je veći broj pojavljivanja termina u dokumentu, veći je njegov *term frequency*, a što se manje pojavljuje u drugim dokumentima, raste njegova važnost [9]. Očekivano je da ova tehnika vektorizacije da bolje rezultate kada se koristi za klasifikaciju teksta [1] [2].

*GloVe* (engl. *Global Vectors*) metod pretpostavlja da se veza između reči može zaključiti proučavanjem verovatnoće njihovog zajedničkog pojavljivanja. Stoga, *GloVe* reprezentacijom moguće je iskazati i semantičke informacije [10]. U slučaju ovog rada, vektorizovala se cela rečenica, tako što su se uzimali *GloVe* vektori pojedinačnih reči u njoj i na kraju računala srednja vrednost vektora svih vektora reči u rečenici [1] [2] [3].

*FastText* algoritam vektorizacije uzima u obzir internu strukturu reči i njenu okolinu [11]. Ovaj algoritam pretpostavlja da se svaka reč sastoji iz n-grama karaktera, gde dužina n može varirati od 1 do dužine reči [10]. Slično kao i kod *GloVe* reprezentacije, vektorizuje se cela rečenica tako što se uzima *FastText* reprezentacija svih pojedinačnih reči u rečenici i na kraju računa srednja vrednost vektora za celu rečenicu [1] [2] [3].

### 3.4. Klasifikacija teksta

U ovom radu isprobani su klasični algoritmi mašinskog učenja, kao i algoritmi dubokog učenja. Svi korišćeni algoritmi opisani su u nastavku teksta.

*Naive Bayes* je algoritam mašinskog učenja zasnovan na Bajesovoj teoremi [13]. Ovaj algoritam, međutim, podrazumeva međusobnu uslovnu nezavisnost atributa u okviru jedne klase. U prethodnoj literaturi je pokazao prilično dobre rezultate za probleme binarne klasifikacije teksta [1].

*SVM* je algoritam mašinskog učenja razvijen za binarnu klasifikaciju [12]. Ovaj algoritam je zasnovan na ideji pronalaženja hiperravni (engl. *hyperplane*) koja najbolje deli skup podataka u dve kategorije. U prethodnim radovima na sličnu temu, *SVM* klasifikator davao je bolje rezultate nego *Naïve Bayes*, te je stoga odlučeno da se isproba i u ovom radu [1] [3].

*Extreme Gradient Boosting* (*XGBoost*) je algoritam iz grupe *boosting* algoritama, koji kombinuju slabe "učenike", tj. jednostavna stabla odlučivanja, koja nemaju veliku preciznost prilikom klasifikacije, u precizniji algoritam, iterativno. *XGBoost* obično daje najbolje metrike od svih klasičnih algoritama, koji ne spadaju u kategoriju neuronskih mreža [14]. Stoga je odlučeno da se isproba i u ovom radu.

*SpaCy* je biblioteka koja nudi različite pretrenirane modele nad generalnim skupom podataka. Arhitektura

koja je u ovom radu korišćena kao osnova za treniranje tekstualnog klasifikatora bila je neuronska mreža, gde se vektori pojedinačnih tokena računaju koristeći jednostavnu konvolutivnu neuronsku mrežu (*CNN*). Više detalja o specifičnoj implementaciji može se pronaći na zvaničnoj stranici *spaCy* biblioteke [6]. Ovaj model je isproban zbog toga što je u literaturi pronađeno da vektori dobijeni kroz konvolutivne mreže doprinose dobrim metrikama prilikom tekstualne klasifikacije [1].

*BERT* (*Bidirectional Encoder Representations from Transformers*) je transformer model, koji koristi bidirekciono treniranje i *attention* [1]mehanizam kako bi napravio model jezika (engl. *Language model*) [4]. *BERT* koristi transformer i *attention* mehanizam, koji ima sposobnost da nauči kontekstualne odnose između reči (ili delova reči) u tekstu. U slučaju ovog rada korišćen je *BERT* specijalizovan za pravni domen, koji se zove *LegalBERT* [5]. Korpus koji je korišćen za treniranje ovakvog jezičkog modela je uključivao tekst iz različitih pravnih dokumenata (legislativa, sudskih dokumenata, presuda, ugovora).

## 4. EKSPERIMENTALNI REZULTATI I DISKUSIJA

Eksperimentalna evaluacija izvršena je radi određivanja i kvantifikovanja performansi svih pojedinačnih algoritama.

Svaki klasifikator obučen je pomoću skupa koji se sastojao od približno 10500 parova (rečenica, binarna labela), gde je binarna labela imala vrednost 1 ili 0, u zavisnosti od toga da li je data rečenica obligacija, ili ne. Za svaki klasifikator optimizovani su hiperparametri. Metrike klasifikatora su određene uz pomoć testnog skupa, koji je činilo 4.500 rečenica, na isti način anotiranih kao i kod obučavajućeg skupa.

Metrike koje su korišćene u radu za evaluaciju svih klasifikatora su preciznost (engl. *precision*), odziv (engl. *recall*) i F1 mera. Odziv je posebno bitna metrika u radu, jer je bitnije da se u ugovoru prepoznaju i označe sve obligacije.

### 4.1. Evaluacija sistema

Prvi model koji je istreniran i evaluiran bio je multinomijalni *Naive Bayes*. Kod ovog modela trenirane su tri varijante (spram vektorizacije): *bag of words* i *TFIDF* (sa tehnikama preprocesiranja i bez njih). Najbolje rezultate (F1 skor od 73%) dalo je korišćenje *TFIDF* i tehnika preprocesiranja teksta, što je u skladu sa prethodnom literaturom na sličnu temu [1].

Kod *SVM* modela istrenirano je nekoliko varijanti (spram vektorizacije): *bag of words*, *TFIDF*, *GloVe* i *FastText* nad generalnim domenom. Svaka varijanta modela trenirana je korišćenjem tehnika preprocesiranja. Najbolje rezultate dalo je korišćenje *TFIDF* i tehnika preprocesiranja teksta (F1 skor od 82%). Tehnike vektorizacije uz pomoć *FastText*-a i *GloVe*-a nisu dali očekivano visoke rezultate, jer su to vektori koji se formiraju nad generalnim domenom teksta, a ne nad domenom ugovora.

*XGBoost* model je treniran u nekoliko varijanti (spram vektorizacije): *bag of words*, *TFIDF*, *GloVe* i *FastText* vektora nad pravnim domenom. Najbolja varijanta ovog modela dobijena je korišćenjem *BoW* vektorizacije (F1 skor od 82%).

Poslednja dva modela koja su istrenirana i evaluirana bile su duboke neuronske mreže: konvolutivna neuronska mreža kroz biblioteku *SpaCy*, kao i transformer model *LegalBERT*.

U tabeli 4.1 mogu se videti objedinjeni rezultati najboljih istreniranih varijanti svih isprobanih modela.

| **Model** | **Preciznost** | **Odziv** | **F1** |
|---|---|---|---|
| NB TFIDF pp | 76% | 75% | 74% |
| SVM TFIDF | 82% | 82% | 82% |
| XGB BoW | 82% | 82% | 82% |
| SpaCy CNN | 86% | 86% | 86% |
| LegalBERT | 88% | 88% | 88% |

Tabela 4.1: Rezultati najboljih varijanti svih modela

Može se primetiti da je *SVM* ostvario bolje rezultate od *Naive Bayes* modela. U prethodnoj literaturi je to takođe bio slučaj [1] [3].

Takođe, vidimo da *XGBoost* daje slične rezultate kao *SVM* model, ali bolje od *Naive Bayes*-a. Ovo potvrđuje pretpostavke iz literature [17].

Metrike *CNN* su za par procenata bolje od najboljeg "klasičnog" modela istreniranog odabranim tehnikama vektorizacije rečenica. Odziv nam je naročito bitan za detekciju obligacija, jer želimo da ih sve pronađemo i prikažemo pravnim licima koji analiziraju ugovore. Ovde je odziv iznosio 86%, što je za 4% bolje od najboljeg odziva "klasičnih" modela. Ovo je u skladu i sa prethodnim istraživanjima na sličnu temu [3].

*LegalBERT* daje daje najbolje rezultate do sada: prosečan odziv je 88%, a i F1 skor iznosi 88%. Pretreniranje *BERT*-a je veoma značajno, jer je u ovom slučaju korišćen transformer model treniran baš nad domenom nad kojim treniramo i sam klasifikator. Ovi rezultati su u skladu i sa rezultatima iz prethodnih radova [3] [5].

### 4.2. Analiza grešaka

Pored evaluacije, analizirane su greške osnovnog modela (*BoW* u kombinaciji sa multinomijalnim *Naive Bayes*-om), konvolutivne neuronske mreže, kao i modela baziranog na transformer arhitekturi, koji je dao najbolje rezultate (*LegalBERT*).

Baseline model, *Naive Bayes*, grešio je u dosta slučajeva gde su rečenice bile dugačke. Takve rečenice obično su sadržale reči koje nisu u kontekstu obligacija, nego "uobičajene" reči, poput "*software*", ili "*books*", i ne javljaju se mnogo u skupu podataka u sklopu obligacija.

[1] selektivno fokusiranje na određene delove vektora koji su relevantni

Konvolutivna neuronska mreža greši kod sličnih rečenica, jer nije uspela da dobro poveže kontekst određenih reči, koje se pojavljuju u dugačkim rečenicama.

*LegalBERT* model uglavnom greši kod klasifikacije nestandardnih struktura rečenica i prilikom korišćenja fraza koje se ne javljaju često u rečenicama koje nisu obligacije, kao npr. *"agrees to sign"*.

Pored ovoga, *LegalBERT* greši i u sledećim tipovima primera:

*"This Section 18.0 shall not delay or excuse Client's payment obligations."*

Ovakva rečenica nije obligacija. Međutim, u primeru se pojavljuje fraza *"Section 18.0",* koju je *LegalBERT* verovatno tretirao kao entitet u samom dokumentu (ugovornu stranu) i stoga pogrešno klasifikovao ovu rečenicu kao obligaciju.

## 5. ZAKLJUČAK

U okviru ovog rada implementiran je algoritam za detekciju obligacija u ugovorima napisanim na engleskom jeziku. Ulaz u algoritam je rečenica iz ugovora, a izlaz je odgovor da li je rečenica obligacija, ili ne.
Za potrebe rada trenirano je i evaluirano 5 različitih tipova klasifikatora: *Naive Bayes, SVM, XGBoost, SpaCy CNN* i *LegalBERT*. Isprobane su i različite tehnike vektorizacije teksta: od jednostavnijih, kao što su *bag of words i TF-IDF*, do naprednih tehnika koje se baziraju na vektorima koje enkoduju okolinu pojedinačnih reči u rečenici: *GloVe i FastText*.

Motivacija za rešavanje problema detekcije obligacija leži u omogućavanju automatizacije procesa čitanja dugačkih ugovora na engleskom jeziku. Implementirani su algoritmi koji klasifikuju rečenice po tome da li predstavljaju obligaciju, ili ne.

Skup podataka je napravljen u saradnji sa advokatima koji govore engleski jezik. Najpre je sakupljena (engl. *scrape*[2]) velika količina ugovora sa sajta *Edgar*[3], koja je zatim data advokatima na anotaciju. Oni su najpre anotirali tipove klauzula u dokumentu, a zatim i pojedinačne rečenice iz klauzula u dva tipa: da li one jesu ili nisu obligacija. Skup je izbalansiran i podeljen na trening i test skup.

Za evaluaciju klasifikatora korišćene su metrike preciznosti, odziva i F1 mere. Jako bitan je bio odziv svakog klasifikatora, zato što želimo da pravniku koji čita ugovore prikažemo sve obligacije koje u njemu postoje. Dobijeni rezultati klasifikatora su u skladu sa prethodnim rešenjima na istu, ili sličnu temu. Najbolje rezultate dao je klasifikator baziran na pretreniranom *LegalBERT* transformer modelom [4][5].

Osim sakupljanja veće količine podataka, mogu se probati i druge arhitekture dubokih neuronskih mreža u kombinaciji sa složenijim vektorskim reprezentacijama teksta ili pretreniranje nekog drugog jezičkog modela, baziranog na transformerima, na pravnom domenu.

[2] korišćenje specijalnih programa za preuzimanje sadržaja sa nekog veb-sajta
[3] *https://www.sec.gov/edgar.shtml*, otvorena baza ugovora na engleskom jeziku

## 6. LITERATURA


[1] Octavia-Maria S¸ulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P. Dinu, Josef van Genabith, "Exploring the Use of Text Classification in the Legal Domain", ASAIL, oktobar 2017.
[2] Vladimir Zolotov, David Kung, "Analysis and Optimization of FastText Linear Text Classifier", IBM Watson Research

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, "Large-Scale Multi-Label Text Classification on EU Legislation", Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, jul 2019

[4] Asaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lilion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", NIPS 2017, Long Beach, CA, USA

[5] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, Ion Androutsopoulos, "LEGAL-BERT: The Muppets straight out of Law School", oktobar 2020.

[6] TextCat Ensemble Model, spaCy: https://spacy.io/api/architectures#TextCatEnsemble

[7] V. Srividhya, R. Anitha, "Evaluating Preprocessing Techniques in Text Categorization"

[8] Somuya George K, Shibily Joseph, "Text Classification by Augmenting Bag of Words (BOW) Representation with Co-Occurence Feature", IOSR/JCE, volume 16, Issue 1, Jan 2014.

[9] Shahzad Qaiser, Ramsha Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents", International Journal of Computer Applications, Volume 181, Jul 2018.

[10] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, "Enriching Word Vectors with Subword Information", Transactions of the Association for Computational Linguistics, 2017.

[11] Birol Kuyumcu, Cuneyt Aksakalli, Selman Delil, "An automated new approach in fast text classification (FastText): A case study for Turkish text classification without preprocessing", NLPIR 2019, jun 2019.

[12] Corinna Cortes, Vladimir Vapnik, "Support-Vector Networks", Machine Learning, 20, 273-297, 1995.

[13] I. Rish, "An Empirical Study of the Naive Bayes Classifier", jan. 2001.

[14] Ashish Chaturvedi, Santosh Yadav, Mohd. Abuzar Mohd. Haroon Ansari, Mahendra Kanojia, "Comparative Multinomial Text Classification Analysis of Naïve Bayes and XGBoost with SMOTE on Imbalanced Dataset", septembar 2021.


### Kratka biografija:


**Marko Žužić** rođen je u Subotici 28. aprila 1994. godine. Master rad na Fakultetu tehničkih nauka, oblast Elektrotehnika i računarstvo – Računarske nauke i informatika odbranio je 2021. godine.
Kontakt: markozuzic@nordnet.rs