Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/16BE22Raseta

KATEGORIZACIJA NOVINSKIH ČLANAKA POMOĆU MAŠINSKOG UČENJA

NEWS CATEGORIZATION USING MACHINE LEARNING

Marko Rašeta, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu korišćeno je više modela za klasifikaciju novinskog članka na osnovu njegovog kratkog sažetka, koji se najčešće sastoji iz jedne ili dve rečenice, radi utvrđivanja kojoj kategoriji članak pripada (sport, politika, zabava…). Svakom od tih modela prosleđen je kratki sažetak koji je prethodno obrađen nekom od metoda za vektorsku reprezentaciju teksta. Od modela korišćeni su: logistička regresija, naivni Bajes, Support Vector Machine, neuronska mreža, konvolutivna neuronska mreža i rekurentna neuronska mreža. Za vektorsku reprezentaciju teksta korišćeni su tf-idf, Word2vec i GloVe. Modeli su trenirani na skupu podataka koji sadrži članke iz Huffington Post novina iz perioda 2012-2018. godine, a evaluacija je rađena na tim podacima, kao i na novinskim člancima koji su dobijeni scrape-ovanjem sa njihove veb stranice. Preciznost je računata kao odnos broja tačno pogođenih kategorija i ukupnog broja pogađanja, a prikazana je i F-mera.*

Ključne reči: *klasifikacija teksta, logistička regresija, naivni Bajes, Support Vector Machine, neuronska mreža*

Abstract – *In this paper many different models are used in order to classify news articles using their short description, mostly consisting of one or two sentences, in order to determine article's category (sports, politics, entertainment…). Each of those models is given short description which is first transformed into its vector representation using different methods. Models used are: logistic regression, naïve Bayes, Support Vector Machine, artificial neural network, convolutional neural network and recurrent neural network. For representing text in vector form tf-idf, Word2Vec and GloVe are used. Models were trained on dataset containing articles from Huffington Post from 2012-2018, and evaluation was done using those articles, as well as articles obtained by scraping Huffington Post's webpage. Models' accuracies and F-measures are given.*

Keywords: *text classification, logistic regression, naïve Bayes, Support Vector Machine, neural network*

1. Uvod

U poslednjih nekoliko decenija svedoci smo naglog razvoja tehnologije u mnogim sferama zivota. Neke su od starta atraktivne, dok su neke vremenom postajale sve interesantnije i zastupljenije. U skladu sa tim pojavljuju se neke nove profesije, dok neke u prošlosti ne toliko razvijene doživljavaju tehnološku transformaciju i ekspanziju sto dovodi do popularizacije istih. Neke od njih su novinarstvo, komunikologija, marketing i odnosi sa javnošću. To sve dovodi do pitanja kako unaprediti i olakšati takve poslove, jer ako potpuna automatizacija nije moguća bilo kakav njen vid je poželjan i neophodan. Klasifikacija teksta je prvi odgovor. Njene primene su posebno bitne za organizaciju digitalnih dokumenata i drugih digitalnih podataka kojih je sve više. Klasifikacija ili kategorizacija teksta predstavlja svrstavanje delova teksta, odnosno delova teksta u jednu ili više prethodno definisanih kategorija . Na primer, članci u novinama su uglavnom razvrstani po rubirkama kojima pripadaju, naučni radovi su klasifikovani po oblastima koje obrađuju, medicinski kartoni pacijenata su indeksirani po istorijatu bolesti, brojevima osiguranja, itd.

Svaki od navedenih primera može se predstaviti nekim skupom osobina. Na taj način se dodeljuju klase. Jedna od prvih primena klasifikacije odnosila se na određivanje autora datog teskta, dok danas najveću primenu imamo kada je reč o informativnim portalima, društvenim mrežama itd.

Konkretno kada je reč o novinarstvu, novinarima je uvođenjem e-portala posao već u mnogome olakšan. Međutim poboljšanjem modela novinari nece morati tekstovima dodeljivati i kategorije kako bi čitaoci mogli vršiti lakšu i bržu pretragu, i na efikasniji način pronalaziti njima interesantne informacije, već će ceo taj proces biti automatizovan.

2. PREGLED TRENUTNOG STANJA U OBLASTI

U radu [1] korišćen je samo Naive Bayes model za predviđanje kategorije novinskog članka. Ovaj rad objavljen je 2003. godine i kao takav spada među radove sa začetka istraživanja u ovoj oblasti. Skup podataka preuzet je sa indijskih novina Eenaadu iz perioda 2003-2004. godine. Sastoji se od 9870 članaka i 27.5 miliona reči, ali je u ovom radu odabrano 794 članka iz četiri kategorije (politika, sport, biznis i film) kako bi se predvidela jedna od ove 4 kategorije sa velikom preciznošću. Za pretprocesiranje podataka rađen je stemming, uklanjanje stop reči i identifikacija fraza i izraza. Za validaciju modela korišćeno je nasumično odabranih 20% članaka iz skupa podataka i na njima je dobijena preciznost od čak 94.72%.

U radu [2] korišćena su tri različita modela. Korišćeni modeli su: konvolutivna neuronska mreža, naivni Bajes i

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

SVM. Reči su se svodile na osnovni oblik (stem) i korišćena je metoda gde se od svake reči uzme samo njenih prvih 4-7 slova. Ta dva pristupa korišćena su samostalno, uz stem tag i uz word tag. Najbolje rezultate za KNN dala je metoda gde se uzimalo prvih pet slova reči i to 92.50%. Za naivnog Bajesa takođe najbolji rezultat je za prvih pet slova i on je ujedno i najprecizniji model sa preciznošću od 94.37%. SVM se nalazi u sredini po preciznosti i to kada se uzme prvih pet ili šest slova, jer je preciznost u ta dva slučaja ista – 93.12%.

U radu [3] korišćen je SVM model kako bi se klasifikovali indonežanski novinski članci. Delili su se samo na tri kategorije: biznis vesti, političke vesti i sportske vesti. Svrha ovog rada bila je da uporedi pristup bez selekcije feature i sa selekcijom. Za selekciju feature korišćen je Information Gain pristup i on je doveo do poboljšanja, jer je preciznost SVM modela bez selekcije iznosila 95.11%, dok je sa Information Gain preciznost čak 98.057%.

U radu [4] automatsko dodeljivanje kategorije novinskom članku vršeno je uz pomoć modela Support Vector Machine i Naive Bayes. Skup podataka sastoji se od 130000 novinskih članaka od kojih svaki pripada tačno jednoj od osam disjunktnih kategorija. Isprobane su dve vrste vektorizacije: Count Vectorization i TF-IDF Vectorization. Poređenjem rezultata ustanovljeno je da tf-idf daje bolje rezultate. Nakon vektorizacije uklonjene su stop reči što je dodatno poboljšalo rezultate. Korišćene su dve tehnike za izbor feateure-a koji se prosleđuju modelima, jer modeli rade bolje ukoliko prime manje, ali važnije feature-e. Prva korišćena tehnika je Chi-Squared koja je za SVM model dala čak 88.6% preciznost, a za Naive Bayes preciznost od 79.6%. Druga tehnika je LASSO i ona je za SVM model dala gotovo identičnu preciznost, 88.5%, dok je za Naivnog Bajesa preciznost bila 76.9%.

U radu [5] opisano je rešenje problema klasifikacije novinskih članaka primenom ant colony optimization algoritma. U obzir je uzeto pet mogućih kategorija. Skup podataka sastojao se iz 1000 novinskih članaka iz svake kategorije, odnosno 5000 članaka ukupno. Nad podacima je prvo rađen stemming, odnosno sve reči svođene su na svoj osnovni oblik. Zatim je rađena normalizacija podataka tako što su uklanjani svi znakovi interpunkcije i sva slova su prebačena u mala slova. Nakon toga izvršena je tokenizacija i uklonjene su sve takozvane stop reči, odnosno reči koje se često pojavljuju a ne nose gotovo nikakvo značenje. Model je evaluiran na tri skupa podataka: Berita10, Berita50 i Berita500 koji sadrže redom 10, 50 i 500 novinskih članaka od kojih svaki pripada jednoj od pet kategorija. Purity vrednosti za ova tri skupa bile su redom: 70%, 46% i 28.2%.

3. METODOLOGIJE

U ovom poglavlju biće opisane korišćene metode za pretprocesiranje podataka, kao i skup podataka koji je korišćen za obučavanje i testiranje modela.

3.1. Pretprocesiranje podataka

Svakom kratkom opisu novinskog članka prvo budu uklonjene stop reči i budu sva slova prebačena u mala slova. To je vršeno funkcijom pod nazivom clean_text. Ona je implementirana tako što se prvo sva slova prebacuju u mala slova. Zatim se karakteri /, (,), {, }, [,], |, @, ; i zapeta zamenjuju razmakom. Kada se to izvrši svi karakteri koji nisu malo slovo, cifra, #, + ili _ se uklanjaju. Ovaj postupak se primenjuje kako bi se iz teksta uklonili svi karakteri koji nemaju veliku težinu, odnosno ne nose nikakav ili nose veoma mali značaj. Nakon toga prolazi se reč po reč kroz tako obrađen tekst i izbacuju se sve reči koje pripadaju skupu stop reči za engleski jezik. Stop reči su reči koje ne nose nikakvu semantiku, pa se iz tog razloga mogu ukloniti. Npr. To su reči a, the, is, are. Nakon toga se kratak opis prebacuje u svoju vektorsku reprezentaciju. To je urađeno tako što su reči prosleđivane modelima koji su zatim vršili transformaciju vraćajući vektore. Kako kratki opisi vesti najčešće nisu duži od dve rečenice, do vektorske reprezentacije celog teksta dolazilo se tako što se uzimala srednja vrednost vektora reči tog opisa.

Za tf-idf[4] transformaciju u vektor korišćen je TfidfVectorizer biblioteke sklearn. Word2Vec[5] model koji je korišćen preuzet je preko gensim downloader-a i zove se word2vec-google-news-300 i on prevodi svaku reč u vektor dimenzije 300. Konačno, GloVe[6] model koji je korišćen takođe je preuzet uz pomoć gensim i nosi naziv glove-wiki-gigaword50 i prevodi svaku reč u vektor dimenzije 50.

3.2. Skup podataka

Za rešavanje problema korišćen je skup podataka koji se sastoji od 200000 novinskih članaka koji pripadaju nekoj od čak 41 različitoj kategoriji. Kako je to prevelik broj kategorija da bi se napravili precizni modeli i neke kategorije nisu zastupljene u velikom broju, taj skup podataka sveden je na članke koji pripadaju nekoj od sledećih šest kategorija: sport, kriminal, politika, zdravlje, zabava i putovanje. Time je broj članaka sveden na 85000. Primer jednog JSON objekta koji se nalazi u skupu podataka:

{

"category": "CRIME",

"headline": "There Were 2 Mass Shootings In Texas Last Week, But Only 1 On TV",

"authors": "Melissa Jeltsen",

"link":"https://www.huffingtonpost.com/entry/texas-amanda-painter-mass-shooting_us_5b081ab4e4b0802d69caad89",

"short_description": "She left her husband. He killed their children. Just another day in America.",

"date": "2018-05-26"

}

4. EKSPERIMENTALNI REZULTATI I DISKUSIJA

U ovom poglavlju će biti izneti i prodiskutovani rezultati u zavisnosti od toga koji model i koji način pretprocesiranja podataka su korišćeni.

4.1 Evaluacija

Evaluacija modela rađena je na dva načina. Jedan je tako što je iz celog skupa podataka od oko 85000 članaka

uzeto nasumično 30% podataka, što iznosi oko 25000 vesti, i na tome se vršilo testiranje, dok su se na preostalih 70% podataka modeli obučavali. Drugi način je tako što je sa veb stranice novina čiji su članci i korišćeni za obuku modela prikupljeno 50 članaka koji pripadaju jednoj od šest kategorija na kojima su modeli i obučavani. Ovi rezultati su manje merodavni iz razloga što ih nema mnogo, ali su ipak zanimljivi kako bi se videlo kako se modeli pokazuju na vestima koje su objavljene nekoliko godina nakon vesti na kojima su modeli obučavani, jer su modeli obučavani na podacima koji su prikupljani u periodu 2012-2018. godine. Za prikupljanje je korišćen selenium. Na veb stranici Huffington Post novina nalazi se stranica na kojoj se nalaze vesti sortirane od najnovijih ka starijim. Nakon što je prosleđena putanja do te stranice, od svake vesti skriptom su prikupljene kategorije i kratak opis na osnovu kog modeli predviđaju kategoriju tog članka.

4.2. Diskusija

Modeli neuronskih mreža pokazali su bolje rezultate od modela naivnog Bajesa, SVM modela i logističke regresije. Od ta tri modela logistička regresija je za nijansu uspešnija od SVM-a, dok je naivni Bajes pokazao najslabije rezultate. Od neuronskih mreža konvolutivna se pokazala za malo boljom od rekurentne, dok je obična neuronska mreža dala malo lošije rezultate od prethodne dve. Što se tiče pretprocesiranja podataka, zajedničko svim modelima je da se Word2Vec vektorska reprezentacija pokazala najboljom, GloVe reprezentacija za neke modele nije bila mnogo lošija, a tf-idf reprezentacija je dala najlošije rezultate u svakom modelu. Modeli su, sa obzirom na veći broj kategorija, pokazali solidne rezultate, ali ipak nedovoljno dobre za neku praktičnu upotrebu, npr. Potpuna automatizacija procesa dodeljivanja kategorija. Takvi rezultati bi mogli da se postignu ako bi se skup podataka proširio tako da sve kategorije budu podjednako zastupljene, jer su F-mere ipak bile najmanje za one kategorije kojih ima manje u skupu podataka. Drugi način bi bio da se modeli vektorske reprezentacije pojedinačnih reči, zamene modelima pretprocesiranja podataka koji bi ceo tekst prebacivali u vektor (npr. Doc2Vec[15]).

5. ZAKLJUČAK

U ovom radu korišćeno je šest različitih modela za kategorizaciju vesti. Svaka vest pripada jednoj od šest kategorija. Korišćeni su naivni Bajes, support vector mačine, logistička regresija, veštačka neuronska mreža, konvolutivna neuronska mreža i rekurentna neuronska mreža. Za vektorizaciju ulaznog teksta korišćeni su tf-idf, GloVe i Word2Vec. Kao modeli najbolje su se pokazali KNM i RNM, dok se od metoda za vektorizaciju Word2Vec pokazao kao najbolji, ali nije mnogo lošiji bio ni GloVe. Iz svega navedenog jasno je da je problem rešiv i da je moguće napraviti modele koji će sa velikom preciznošću kategorisati vesti. Veća preciznost mogla bi biti ostvarena ukoliko bi se smanjio broj kategorija. Takođe moguće je isprobati i modele koji u ovom radu nisu korišćeni u cilju poređenja sa njima.

Ostaje prostora i za dalja istraživanja, jer je velik izazov napraviti model koji bi kategorisao ne samo vesti, već i neke druge tekstove i radove i to tako što bi dodeljivao jednu od 10, pa kasnije čak i više kategorija. To bi bilo moguće postići proširivanjem skupa podataka ili korišćenjem modela za vektorsku reprezentaciju dokumenata.

Konkretna primena takvog modela mogla bi biti u novinama, kako ne bi urednik morao da dodeljuje ručno kategoriju svakoj vesti, ali za tako nešto potrebno je imati model sa preciznošću od bar 90% koji dodeljuje jednu od bar 10 kategorija. Pored toga, ovakvi modeli bi mogli doprineti i boljoj pretrazi na Google-u, jer ako postoji neki tekst ili rad bez određenih oznaka ili direktno definisane oblasti/teme, on bi opet mogao da bude rezultat pretrage vezane za neku temu ukoliko model dodeli tu temu tom tekstu.

6. LITERATURA

[1] Kavi Narayana Murthy (2003), “Automatic Categorization of Telugu News Articles”

[2] Burak Kerim Akkus, Ruket Cakici (2013), “Categorization of Turkish News Documents with Morphological Analysis”

[3] Adhy Rizaldy, Heru Agus Santoso (2017), “Performance improvement of Support Vector Machine (SVM) With information gain on categorization of Indonesian news documents”

[4] Juan Ramos (2003), “Using TF-IDF to Determine Word Relevance in Document Queries”

[5] KW Church (2017), “Word2Vec”

[6] Jeffrey Pennington, Richard Socher, Christopher D. Manning (2014), “GloVe: Global Vectors for Word Representation”

Kratka biografija

Marko Rašeta rođen je 15. aprila 1997. godine u Novom Sadu, Srbija. Fakultet tehničkih nauka upisao je 2016. na studijskom programu Računarstvo i Automatika. Diplomski rad iz oblasti XML i veb servisi odbranio je 2020. godine. Master rad na usmerenju Elektronsko poslovanje odbranio je 2021. godine.