Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/22BE11Lepar

MODELOVANJE TEMA U TEKSTU NA OSNOVU NASLOVA DOKUMENATA

TOPIC MODELING IN TEXT BASED ON TITLES OF DOCUMENTS

Minja Lepar, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U radu je predstavljen pristup za modelovanje tema i klasifikaciju tekstualnih dokumenata. Konkretno, vršena je 1) primena LDA (Latent Dirichlet Allocation) nad tekstom zarad dobijanja tema, pri čemu je evaluacija rađena kvalitativno, kroz semantiku pronađenih tema; 2) klasifikacija dokumenta primenom reprezentacije teksta dobijene kombinacijom tf-idf obeležja i tema izvučenih pomoću LSA (Latent Semantic Analysis); nad ovom reprezentacijom treniran je Naive Bayes klasifikator, a evaluacija je vršena računanjem F-mere, 3) klasifikacija dokumenta primenom tf-idf reprezentacije teksta, gde je eksperimentisano sa treniranjem SVM (Support Vector Machines) i RF (Random Fores) modela; I u ovom slučaju evaluacija je vršena računanjem F-mere.*

Ključne reči: *modelovanje tema, LDA, klasifikacija, SVM, Naive Bayes, Random Forest*

Abstract – *The paper presents an approach for topic modeling and document classification. Concretely, the paper explores 1) the application of LDA (Latent Dirichlet Allocation) to obtain topics from the text; This approach was evaluated qualitatively, relying on the semantics of the found topics. 2) document classification, where documents were represented using tf-idf features and topics extracted by applying LSA (Latent Semantic Analysis); Naive Bayes classifier was trained on the obtained representation, and F-measure was used for evaluation, 3) document classification, where the tf-idf representation was used to train a classification model, where we experimented with using SVM (Support Vector Machines) and RF (Random Forest) models; In this case, F-measure was used for evaluation.*

Keywords: *Topic modeling, LDA, classification, SVM, Naive Bayes, Random Forest*

1. UVOD

Potreba za organizacijom teksta, sve je veća i veća. Internet je preplavljen tekstom u elektronskoj formi. Informacija se potapa u nestruktuiranom formatu, u smislu da niko ne zna da je tu, i niko ne zna šta da radi s njom, te prisutna informacija biva neiskorišćena. Za potrebe ljudskog napretka, funkcionisanja i kvaliteta života, skrivena informacija može biti neprocenjiv izvor mnogih koristi, inovacija i razvoja.

Analiza teksta je tehnika koja se koristi za izvlačenje skrivene informacije iz teksta u nestruktuiranoj formi [4][6][7][8]. Zadaci modelovanja teksta su:

- organizacija i filtriranje vesti [8],
- rezimiranje dokumenata [1][5][6][7],
- praćenje zdravstvene nege [8],
- primena modelovanja tema u geografiji za ekstrahovanje tema iz geografskih dokumenata [4],
- primena modelovanja tema u političkim naukama za identifikovanje tema u političkim govorima [4], itd.

Modelovanje tema otkriva skrivene teme u tekstu [4]. Modelovanje tema može da unapredi postupak automatske klasifikacije dokumenata time što, umesto da tekst reprezentujemo putem pojedinačnih reči, reprezentujemo ga kroz ekstrahovane teme [8]. Ovim postupkom se u znatnoj meri redukuje dimenzionalnost reprezentacije teksta. U ovom cilju, rešenja primenjuju LSA (*Latent Semantic Analysis*) za modelovanje tema, a kao modele mašinskog učenja (*machine learning*, ML) tipično primenjuju SVM (*Support Vector Machines*) i *Naive Bayes* (NB) modele [8]. U ovom radu isproban je opisani postupak, uz dodatnu primenu *Random Forest* (RF) modela koji se pokazao kao jedan od superiornijih modela za klasifikaciju teksta zbog svojih odlika u klasifikaciji, odabiru atributa i dodavanja težina [9].

Cilj ovog rada je:

1. Pronalaženje tema u skupu ulaznih tekstova primenom nenadgledanog učenja, odnosno, bez korišćenja anotiranih tema dostupnih u korišćenom skupa podataka.
2. Korišćenje anotiranih tema u skupu podataka radi klasifikacije dokumenata po temi.

U radu je korišćen je skup podataka dostupan na *Kaggle* stranici [10]. Kao ulaz u opisani postupak koristi se sirovi tekst, konkretno, naslov iz kolekcije naslova dokumenata dostupnih u skupu podataka. Dokumenti su u skupu podataka anotirani pripadnošću jednom od osam tema ('*TECHNOLOGY*', '*HEALTH*', '*WORLD*', '*ENTERTAINMENT*', '*SPORTS*', '*BUSINESS*', '*NATION*', '*SCIENCE*'), a cilj ovog rada je da se obučim model koji, na osnovu naslova dokumenta, automatski može da zaključi kojoj temi dokument pripada.

U cilju 1. zadatka (pronalaženja tema), u radu se primenjuje LDA (*Latent Dirichlet Allocation*), nenadgledani model. Eksperimentisano je sa odabirom broja tema, gde su isprobane vrednosti 8, što odgovara broju tema u skupu podataka, 10 i 12. U slučaju 8 tema,

---

NAPOMENA:

Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, vanr. prof.

model je evaluiran poređenjem rezultujućih tema sa anotiranim temama. U slučaju 10 i 12 tema, model je evaluiran kvalitativno. Kvalitativnom evaluacijom zaključeno je da se dobijaju semantički kvalitetne teme. Za veći broj tema dobijaju se specifičnije i rafiniranije teme. Međutim, u svim slučajevima je primećeno određeno preklapanje tema. Treba naglasiti da je rezultate teško evaluirati bez uključivanja domenskog eksperta jer se, na primer, u slučaju teme zdravlja javljaju podoblasti poput patologije, genetike i sl. koje zahtevaju dublje poznavanje domena.

U cilju 2. zadatka (klasifikacija dokumenata), trenirani su nadgledani modeli za klasifikaciju u postojeće teme. Izvršeni su sledeći eksperimenti:

a) naslov se pretvara u vektor obeležja (reči naslova) *tf-idf* (*term frequency - inverse document frequency*) transformacijom. Dalje, nad *tf-idf* reprezentacijom se primenjuje LSA radi ekstrakcije tema. Nad tako dobijenom reprezentacijom se primenjuje NB model za klasifikaciju u jednu od osam predefinisanih tema.
b) naslov se pretvara u vektor obeležja (reči naslova) *tf-idf* transformacijom. Zatim se primenjuje nadgledani model, gde je eksperimentisano sa: SVM i RF.

Cilj ovog 2. eksperimenta je da se utvrdi koliko primena LSA modela kao koraka pretprocesiranja doprinosi klasifikaciji.

Radi evaluacije oba zadatka, skup podataka je podeljen na trening i test skup. Za potrebe 1. zadatka za test skup je uzeto 2 300 primera, a model je evaluiran kvalitativno. Za potrebe 2. zadatka, skup podataka je podeljen 2:1, a kao glavna metrika evaluacije je korišćena *F*-mera. Kao najbolji model pokazao se SVM, čija *F*-mera iznosi 81%.

U literaturi ne postoji rad koji je primenio modelovanje tema i klasifikaciju teksta baš na skupu podataka [10]. Tipično se koristi skup podataka *Reuters-21578* i najbolje performanse u literaturi za ovaj skup podataka su iz rada koji vršio klasifikaciju dokumenata nad nižedimenzionom reprezentacijom reči indukovanom od LDA (uz 50 tema) i primenio SVM model [5]. Ovaj postupa postiže i do 97% tačnosti. U radu je pokazano da ovaj pristup značajno unapređuje performanse u odnosu na nižedimenzionu LDA reprezentaciju. Shodno tome, rad [5] predlaže tematsko-baziranu LDA reprezentaciju, kao filtrirajući algoritam za odabir reči u klasifikaciji teksta. *Reuters-21578* skup podataka korišćen u [5] sadrži 8 000 dokumenata i 15 818 reči, gde broj klasa nije naveden. To čini manji broj dokumenata, ali približno isti broj reči, u odnosu na skup [10] korišćen u ovom radu.

Ovaj rad je organizovan na sledeći način. Poglavlje 2 prikazuje prethodne radove i koncepte na koje se oslanja ovaj rad. Poglavlje 3 daje teorijske osnove primenjenog metoda. Poglavlje 4 definiše metodologiju, a poglavlje 5 opisuje eksperiment i ukazuje na hiperparametre koji su upotrebljeni u modelima. U poglavlju 6 dati su rezultati, dok poglavlje 7 analizira greške modela. Poglavlje 8 zaključuje rad.

2. PRETHODNI RADOVI

U ovom poglavlju predstavljeni su srodna rešenja, vezana za koncepte koji se implementiraju u ovom radu.

2.1. Modelovanje tema

U radu [3] evaluira se primena LDA za preporuku tagova (oznaka), imajući resurse i dodeljene tagove. U radu [5] LDA je poređen sa drugim modelima. U radu je opisan LDA model i kako se vrši njegova primena nad skupom podataka.

U radu [6] predstavljen je pristup klasifikaciji dokumenata, gde su poređeni sledeći pristupi za modelovanje tema: treniranje Naive Bayes modela, LSA i LDA modelovanje tema. Pokazalo se da LSA i LDA postižu znatno bolje performanse pri modelovanju tema.

2.2. Klasifikacija dokumenata po temama sa i bez primene LSA

U radovima [1] i [2] opisana je primena LSA modela, što je analiza koja se primenjuje i u ovom radu.

U radu [4] pomenute su ekstenzije LDA pristupa koje pokazuju bolje performanse. U radu koji će se pisati, primenila se ekstenzija LSA na model NB.

U radu [7] opisana je strategija za klasifikaciju tekstualnih dokumenata. Navedeni su osnovni problemi vezani za odabir atributa iz ogromnog broja atributa, i odabir ML tehnika za klasifikaciju teksta. U ovom radu, primenjen je *tf-idf* statistički metod za uticaj na važnije tokene i postizanje boljih rezultata, sa ML tehnikom SVM za klasifikaciju dokumenata.

U radu [8] je primenjen pristup klasifikacije dokumenata primenom NB modela uz redukciju dimenzionalnosti reprezentacije teksta primenom LSA. Metodom je pokazano da tematsko kategorisanje dokumenata može dati tačnije klasifikovanje dokumenata u predefinisane teme. LSA je rezultirala redukcijom dimenzionalnosti. Ista ova ideja primenjena je u ovom radu.

U radu [9] opisan je RF model, njegove prednosti i nedostaci i razne sugestije uz koje se tehnika može poboljšati. Ovaj rad primeniće RF za klasifikaciju, pri čemu je rad [9] koristan za razumevanje i dublje analiziranje RF-a.

3. TEORIJSKE OSNOVE

LDA je generativni probabilistički model koji dokumente reprezentuje kao slučajne mešavine nad latentnim temama, gde je svaka tema karakterizovana kao distribucija verovatnoća nad rečima [5]. Reči sa najvećim verovatnoćama daju ideju o tome šta je tema [4].

SVM pronalazi hiper-ravan koja razdvaja klase, sa maksimalnom marginom (hiper-ravan koja je najdalja od svih tačaka skupa podataka).

NB je probabilistički klasifikator koji primenjuje *Bayes*-ovu teoremu [8]. Pretpostavka modela je da je u dokumentu verovatnoća pojavljivanja svake reči nezavisna od pojavljivanja ostalih reči [8]. Redukcija dimenzionalnosti se implementira primenom SVD (*Singular Value Decomposition*) [1]. Primenom LSA, procesirane reči separatišu se u značajne grupe [6].

RF je ansambl model koji koristi stablo odlučivanja kao bazni klasifikator koji određuje klasnu oznaku instance koja nema oznaku [9].

Većinskim glasanjem svaki klasifikator daje jedan glas (predviđa klasnu labelu), a labela sa najviše glasova klasifikuje instancu [9].

4. METODOLOGIJA

4.1. Eksplorativna analiza

Primenom eksplorativne analize ovde je analiziran skup podataka [10]. Zaključeno je da skup podataka ne sadrži nedostajuće vrednosti. Postoje duplikati naslova, ali su oni zadržani jer pripadaju različitim temama. Skup ima 111 444 naslova i osam tema kojima ti naslovi mogu pripasti.

U skupu podataka postoje dve kolone: tekst kolona, koja sadrži naslove dokumenata i služi kao ulaz u modele prikazane u ovom radu i kolona sa temama koja služi kao ciljno obeležje. Jedinstvene vrednosti ciljnog obeležja su '*NATION*', '*SPORTS*', '*ENTERTAINMENT*', '*BUSINESS*', '*WORLD*', '*TECHNOLOGY*', '*HEALTH*' i '*SCIENCE*' i one su relativno balansirane u skupu podataka.

Kao priprema podataka za vektorizaciju teksta i pronalaženje rečnika, izvršena je tokenizacija i lematizacija (ekstrakcija gramatičkog korena reči). Takođe su uklonjene stop reči (skup reči koje će biti isključene iz rečnika).

4.2. *Latent Dirichlet Allocation*

U ovom poglavlju je opisana primena LDA modela za ekstrakciju tema. Dostupni podaci se dele na trening i test skup. Naslovi dokumenata se reprezentuju putem *count vectorizer*-a (svaki element vektora je broj pojave određene reči u dokumentu). Nad ovom reprezentacijom treniran je LDA model. Isprobano je više varijanti LDA modela, gde je kao broj tema (*n*) korišćeno 8, 10 i 12 tema. Osam tema je izabrano u skladu sa brojem anotiranih tema u skupu podataka, ali je isprobano i 10 i 12 tema u cilju ispitivanja da li će ovo rezultovati čistijom podelom na teme.

Nakon treniranja LDA modela, vizualizovane su naučene teme da bi se utvrdila njihova semantika. LDA transformiše trening i test skup, da bi se dobila raspodela tema za trening i test skup (2 300 naslova). Predefinisane kategorije skupa podataka mapiraju se u naučene kategorije LDA metoda za trening i test skup. Poređenjem trening i test skupa dobijaju se slični rezultati. Nad test skupom, poređenjem stvarne kategorije sa naučenom kategorijom se može vršiti analiza grešaka modela. Za sva tri modela kvalitet se ocenjuje semantikom pronađenih tema (kvalitativna evaluacija).

4.3. Klasifikacija

Da bi se izvršila klasifikacija dokumenata, najpre se vrši priprema podataka opisana u potpoglavlju 4.1. Konvertor oznaka konvertuje oznake iz tekstualne forme u numeričku. Za sve algoritme isti su koraci, *'counts'* (*CountVectorizer()*), *'tfidf'* (*TfidfTransformer()*), i primena klasifikatora. U slučaju NB-a i LSA proces iza *'tfidf'* je *'SVD'* (*TruncatedSVD()*) i *'normalize'* (*Normalizer()*). LSA je kombinacija SVD i *Normalizer*-a, prvi redukuje veličinu matrice, drugi normalizuje vrednosti u SVD matrici.

4.4. Korišćeni alati

U cilju implementacije postupka prikazanog u radu, korišćene su sledeće biblioteke i alati: *pandas*[1] (za manipulaciju skupom podataka), *sklearn*[2] (priprema podataka, konstrukcija modela, korišćenje metrika), *pyLDAvis*[3] i *wordcloud*[4] (vizualizacija tema), *numpy*[5] (manipulacija nizova), *matplotlib*[6] (vizualizacija uz grafikone), *seaborn*[7] (vizualizacija uz *heatmap*-e, toplotne mape), *spacy*[8], *nltk*[9] i *gensim*[10] (manipulacija teksta), *Anaconda*[11] (manipulacija *Jupyter Notebook*-a).

5. EKSPERIMENT

U ovom poglavlju opisano je kako su određeni hiperparametri za sve eksperimente i kako je vršena evaluacija rešenja.

5.1. *Latent Dirichlet Allocation*

Određuju se LDA hiperparametri, kako je broj tema u skupu podataka osam, ovde se vrši treniranje za osam tema, ali se vrši i eksperiment za 10 i 12 tema.

5.2. Klasifikacija

Određuju se hiperparametri za klasifikacione modele. SVM model radi linearno odvajanje klasa, odnosno, *kernel* (jezgro) je *linear*. Za NB i redukciju dimenzionalnosti *TruncatedSVD()* ima parametar 100 tema, prema preporuci dokumentacije. Za RF nisu definisani hiper-parametri, već su korišćene njihove predefinisane vrednosti.

5.3. Evaluacija

LDA model ocenjen je kvalitativno. Kroz rečničke reprezentacije naučenih tema, posmatra se semantika tema, i uviđa se kvalitet. Veličina skupa za testiranje postavljena je na 2 300 naslova.

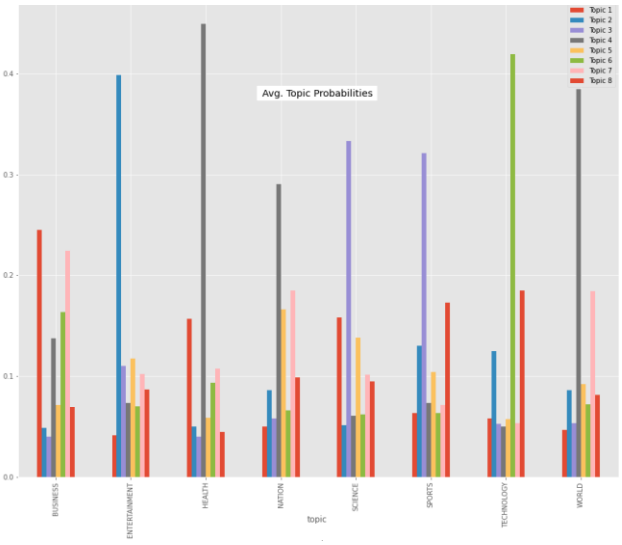
Klasifikacioni modeli evaluirani su *F*-merom, koja je glavna metrika evaluacije. Za sve klasifikacione modele urađen je jedan način podele na trening i test skup i on iznosi 2:1.

6. REZULTATI

Na slici 1 prikazan je rezultat primene LDA za 8 tema. Slika prikazuje vizualizaciju raspodele verovatnoća tema *Topic 1 – 8* za svaku od stvarnih tema (*topic*). Za sve odabire broja tema, naime n = 8, 10, 12, vizualizacije bez referenciranja istinske kategorije rezultiraju u semantički kvalitetne teme.

[1] https://pandas.pydata.org
[2] https://scikit-learn.org/stable
[3] https://pyldavis.readthedocs.io
[4] https://pypi.org/project/wordcloud/
[5] https://numpy.org
[6] https://matplotlib.org/stable
[7] https://seaborn.pydata.org
[8] https://spacy.io/api/doc
[9] https://www.nltk.org
[10] https://pypi.org/project/gensim/
[11] https://docs.anaconda.com

Izvršena je analiza grešaka LDA modela za 8 tema, gde znamo stvarnu kategoriju na osnovu skupa podataka. Za



10 i 12 tema ova analiza nije vršena. Razlog pogrešno klasifikovanih naslova je prisustvo reči koje su indikativne o nekoj naučenoj kategoriji koja nije istinska kategorija naslova.

Slika 1. *Vizualizacija raspodele verovatnoća tema Topic 1 – 8 za svaku od stvarnih tema (topic)*

Kod klasifikacije dokumenata, SVM model je postigao *F*-meru od 81%, RF je postigao *F*-meru od 75%, a NB uz LSA pretprocesiranje 52%. podataka. Utvrđeno je da LAS kao korak pretprocesiranja nije doprineo poboljšanju performansi klasifikacije dokumenata na ovom skupu podataka.

Izvršena je analiza pogrešno klasifikovanih naslova. Utvrđeno je da do grešaka dolazi jer su naslovi jesu previše kratki. Greške su se javljale i kod naslova koji sadrže reči koje su pripale nekoj temi, a koje ne postoje u temi koja je za taj naslov anotirana u skupu

8. ZAKLJUČAK

U radu su predstavljeni metodi za automatsku klasifikaciju tekstualnih dokumenata na osnovu njihovih naslova. Korišćen je skup podataka [10].

Prva grupa metoda modelovala je teme u tekstu nenadgledano, odnosno, bez korišćenja anotacija tema dostupnih u skupu podataka. Toj grupi metoda pripada LDA koji je treniran za 8, 10, i 12 tema. Modeli su ocenjeni kvalitativno i utvrđeno je da su dobijene semantički kvalitetne teme, iako postoje određena preklapanja. Sa porastom broja tema raste, teme postaju specifičnije za poddomene. Nad test skupom objašnjeno je pogrešno klasifikovanje naslova na osnovu reči iz kojih se naslov sastoji, poređenjem sa istinskom temom naslova. Predlaže se da je za detaljniju analizu tema potreban domenski ekspert.

Drugoj grupi metoda pripadaju metodi za klasifikaciju koji koriste anotirane teme naslova. Najvišu *F*-meru od 81% postigao je SVM model treniran nad *tf-idf* obeležjima. Pokazalo se da LSA kao korak pretprocesiranja nije doprineo poboljšanju performansi klasifikacije.

Predlog za unapređivanje modela je da se pronađe optimalan broj tema koji će jasno definisati teme, koje se ne preklapaju. Primena LDA umesto LSA modela, kao i implementacija boljih transformera podataka bi mogla poboljšati postupak ekstrakcije tema.

9. LITERATURA

[1] Evangelopoulos, N., Zhang, X. and Prybutok, V.R., 2012. Latent semantic analysis: five methodological recommendations. European Journal of Information Systems, 21(1), pp.70-86.

[2] Wiemer-Hastings, P., Wiemer-Hastings, K. and Graesser, A., 2004, November. Latent semantic analysis. In Proceedings of the 16th international joint conference on Artificial intelligence (pp. 1-14).

[3] Krestel, R., Fankhauser, P. and Nejdl, W., 2009. Latent dirichlet allocation for tag recommendation. In Proceedings of the third ACM conference on Recommender systems (pp. 61-68).

[4] Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y. and Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications, 78(11), pp.15169-15211.

[5] Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.

[6] Rajasundari, T., Subathra, P. and Kumar, P.N., 2017. Performance analysis of topic modeling algorithms for news articles. Journal of Advanced Research in Dynamical and Control Systems, 11, pp.175-183.

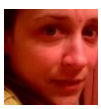
[7] Dalal, M.K. and Zaveri, M.A., 2011. Automatic text classification: a technical review. International Journal of Computer Applications, 28(2), pp.37-40.

[8] Sedghpour, A.S. and Sedghpour, M.R.S., 2020. Web Document Categorization Using Naive Bayes Classifier and Latent Semantic Analysis. arXiv preprint arXiv:2006.01715.

[9] Fawagreh, K., Gaber, M.M. and Elyan, E., 2014. Random forests: from early developments to recent advancements. Systems Science & Control Engineering: An Open Access Journal, 2(1), pp.602-609.

[10] Abdelsalam, K. topic_balanced_dataset, Version 1. Retrieved June 13, 2021 from https://www.kaggle.com/karimamd95/topic-balaned-dataset/version/1.

Kratka biografija:



Minja Lepar rođena je 1987. god. u Odžacima. Osnovne akademske studije završila je na Prirodno-matematičkom fakultetu, odseku za primenjenu matematiku, smer Inženjer matematike. Master rad na Fakultetu tehničkih nauka iz oblasti Softverskog inženjerstva i informacionih tehnologija – Inteligentni sistemi brani 2022.god.
kontakt: minja.lepar@gmail.com