Klasifikacija novic na podlagi naslovov z uporabo metode SVM

Urban Ferlinc
urban.ferlinc@student.um.si
Fakulteta za Elektrotehniko,
Računalništvo in Informatiko,
Univerza v Mariboru
Koroška cesta 46
SI-2000 Maribor, Slovenia

Marija Jovanova
marija.jovanova@student.um.si
Fakulteta za Elektrotehniko,
Računalništvo in Informatiko,
Univerza v Mariboru
Koroška cesta 46
SI-2000 Maribor, Slovenia

Anđela Mihaljević
andjela.mihaljevic@student.um.si
Fakulteta za Elektrotehniko,
Računalništvo in Informatiko,
Univerza v Mariboru
Koroška cesta 46
SI-2000 Maribor, Slovenia

Cvetanka Pasinechka
cvetanka.pasinechka@student.um.si
Fakulteta za Elektrotehniko,
Računalništvo in Informatiko,
Univerza v Mariboru
Koroška cesta 46
SI-2000 Maribor, Slovenia

Stefan Srnjakov
stefan.srnjakov@student.um.si
Fakulteta za Elektrotehniko
Računalništvo in Informatiko,
Univerza v Mariboru
Koroška cesta 46
SI-2000 Maribor, Slovenia

POVZETEK

V tem članku bomo raziskali, kako lahko z uporabo metode podpornih vektorjev avtomatiziramo klasifikacijo novic glede na njihove naslove. Model bomo naučili prepoznavati ključne vzorce v naslovih in jih razvrščati v različne kategorije, kot so šport, kultura, poslovne novice, pri čemer bomo uporabili ustrezen nabor podatkov, ki je javno dostopen za učno množico. Namen takšnega pristopa je izboljšati organizacijo novic in omogočiti hitrejše iskanje ter analizo informacij.

KLJUČNE BESEDE

klasifikacija novic, podporni vektorski stroji (SVM), obdelava naravnega jezika (NLP), strojno učenje, razvrščanje besedil

1 UVOD

V dobi digitalnih medijev se vsakodnevno objavi na tisoče novic, ki pokrivajo različna področja – od športa in kulture do poslovnih novic ter črne kronike. Ob tako veliki količini informacij je učinkovita organizacija vsebin ključnega pomena za bralce in urednike. Ročno razvrščanje novic v ustrezne kategorije je zamudno, nepraktično in lahko podvrženo subjektivnosti, zato se v sodobni analizi podatkov vse bolj poslužujemo metod strojnega učenja.

Ena izmed učinkovitih metod za klasifikacijo besedil je metoda podpornih vektorjev (SVM), ki se je izkazala za zanesljivo pri različnih nalogah obdelave naravnega jezika. SVM omogoča natančno ločevanje med kategorijami na podlagi ključnih vzorcev v podatkih, pri čemer model učimo na izbranem naboru naslovov novic. Tak pristop ne le pospeši proces organizacije novic, temveč tudi izboljša uporabniško izkušnjo, saj zagotavlja bolj relevantne in natančno filtrirane informacije.

2 SORODNA DELA

Področje klasifikacije besedil je že dolgo raziskano, vendar ostaja aktualno zaradi napredka metod strojnega učenja. Klasifikacija besedil vključuje razvrščanje dokumentov v vnaprej določene kategorije, pri čemer so se kot učinkovite metode izkazali algoritmi, kot je metoda podpornih vektorjev. SVM temelji na iskanju optimalne ločilne hiperplošče, ki maksimizira razdaljo med različnimi razredi, s čimer zagotavlja visoko natančnost klasifikacije [4]. Uporablja se na številnih področjih, vključno s klasifikacijo novic, prepoznavanjem vzorcev in analizo medicinskih podatkov [8]. Pri klasifikaciji besedil je ključen tudi proces predobdelave, ki vključuje leksikalno analizo, odstranjevanje nepomembnih besed in izračun TF-IDF, kar izboljša učinkovitost modela.

Raziskave so pokazale, da naprednejše različice SVM, kot so LS-SVM, TWSVM in LS-TWSVM, omogočajo hitrejše in natančnejše klasifikacije. To se je izkazalo predvsem pri večrazrednem razvrščanju novic, kjer izboljšane metode zagotavljajo boljše rezultate [8]. Poleg klasifikacije novic se SVM uporablja tudi za razvrščanje besedil glede na posamezna čustva. Študija [2] obravnava razvrščanje čustev v naslovih novic, kjer avtorja analizirata naslove iz podatkovne zbirke in jih razvrščata v šest čustvenih kategorij: jeza, gnus, strah, veselje, žalost in presenečenje. Njuna metoda je pokazala boljšo natančnost v primerjavi z alternativnimi pristopi, kot so ročno določena pravila [1] in metode, ki temeljijo na iskanju spletnih virov [7]. Kljub visoki uspešnosti pa študija poudarja, da je za izboljšanje interpretacije potrebno vključiti širši kontekst besedila.

Poleg SVM se pri klasifikaciji besedil pogosto uporablja tudi zmanjševanje dimenzionalnosti, saj lahko optimizacija števila funkcij izboljša učinkovitost modela. Članek [9] preučuje uporabo tehnik izbire funkcij v kombinaciji z vektorizacijo Naivni Bayes, pri čemer poudarja, da zmanjšanje števila značilnosti lahko izboljša učinkovitost klasifikacije, hkrati pa ohrani visoko natančnost. Nadaljnje raziskave [10] obravnavajo različne tehnike glajenja, kot sta glajenje po Laplaceu in Lidstonu, ki dodatno izboljšujejo generalizacijo modela. Ti pristopi so posebej relevantni za naloge, kot so analiza razpoloženja in ekstrakcija ključnih besed, kjer je uravnoteženje med učinkovitostjo in natančnostjo ključno.

Podobno se je metoda podpornih vektorjev izkazala kot učinkovita tudi pri personalizaciji digitalnih časopisov. Članek [3] opisuje uporabo SVM za avtomatsko razvrščanje naslovov novic v kategorije, kot so politika, šport in tehnologija, pri čemer sistem uporabnikom prikazuje novice na podlagi njihovih preferenc. Avtorji predlagajo model, ki uporablja značilnosti besedila in uporabniške

preference za izboljšanje razvrščanja vsebin. Rezultati eksperimenta, pridobljeni na podatkovnem naboru 20 Newsgroup in realnih podatkih s spletne strani Times of India, potrjujejo visoko natančnost klasifikacije [5]. Podobne ugotovitve so bile predstavljene [6], kjer so analizirali pomen metod strojnega učenja pri avtomatski klasifikaciji in distribuciji novic.

Vse omenjene raziskave potrjujejo, da metoda podpornih vektorjev ostaja ena najučinkovitejših tehnik za klasifikacijo besedil. Poleg tega pa nadaljnje izboljšave, kot so optimizacija predobdelave podatkov, izbira funkcij in glajenje, še dodatno povečujejo njeno uporabnost na različnih področjih obdelave naravnega jezika.

3 NAŠA IDEJA

V tej raziskavi preučujemo uporabo metode podpornih vektorjev za samodejno klasifikacijo novic na podlagi njihovih naslovov. Cilj je razviti model, ki bo sposoben natančno razvrščati naslove novic v vnaprej določene kategorije, kot so šport, kultura in poslovne novice. Pričakujemo, da bo takšen sistem omogočil boljšo organizacijo novic ter olajšal iskanje in analizo informacij.

Za učenje in vrednotenje modela nameravamo uporabiti javno dostopen nabor podatkov, ki vsebuje naslove novic s pripadajočimi oznakami kategorij. Pri obdelavi podatkov bomo uporabili ustrezne tehnike predobdelave, kot so leksikalna analiza, odstranjevanje nepomembnih besed in uporaba TF-IDF, s čimer želimo izboljšati kakovost vhodnih podatkov za model. Prav tako bomo preučili vpliv zmanjšanja dimenzionalnosti in izbire značilnosti na učinkovitost klasifikacije.

Pričakujemo, da bo metoda podpornih vektorjev dosegla visoko natančnost pri razvrščanju naslovov novic in s tem potrdila svojo uporabnost pri obdelavi besedilnih podatkov. Če bo model uspešen, bi lahko takšen pristop prispeval k avtomatizaciji klasifikacije novic, kar bi bilo koristno za digitalne medijske platforme in druge sisteme, ki se ukvarjajo z upravljanjem vsebin. Poleg tega želimo ugotoviti, kako različne tehnike predobdelave podatkov vplivajo na končne rezultate klasifikacije in poiskati optimalne strategije za izboljšanje zmogljivosti modela.

4 EKSPERIMENT

4.1 Pridobivanje podatkov

Za potrebe eksperimenta smo uporabili javno dostopen nabor podatkov, pridobljen s spletne platforme Kaggle, v formatu JSON. Vsak zapis v zbirki predstavlja eno novico in vključuje informacije, kot so povezava do izvornega članka, naslov novice (ang. *headline*), kategorija, kratek opis vsebine, avtorji in datum objave. Osrednji atribut, ki smo ga uporabili za klasifikacijo, je naslov novice. Primer posameznega zapisa v podatkovnem nizu je:

```
{
    "link": string,
    "headline": string,
    "category": string,
    "short_description": string,
    "authors": string,
    "date": string
}
```

4.2 Priprava okolja

Eksperiment je bil izveden v okolju Jupyter Notebook z uporabo programskega jezika Python. V ta namen smo uporabili več odprtokodnih knjižnic, kot so `pandas`, `numpy`, `scikit-learn` in `nltk`, ki omogočajo učinkovito obdelavo podatkov in gradnjo modelov strojnega učenja. Rezultati klasifikacije so bili analizirani s pomočjo standardnih metrik uspešnosti, kot jih ponuja knjižnica `sklearn.metrics`. Koda je bila strukturirana modularno, kar omogoča preprosto zamenjavo modelov in ponovljivost eksperimenta.

4.3 Postopek

Eksperimentalni postopek je bil usmerjen v rešitev naloge kategorizacije naslovov novic v vnaprej določene tematske razrede. Proces smo razdelili v naslednje ključne faze:

Uvoz in čiščenje podatkov. Najprej smo uvozili podatke iz JSON datoteke ter izločili zgolj polja, relevantna za klasifikacijsko nalogo, in sicer naslove (*headline*) in pripadajoče kategorije (*category*). Manjkajoče ali nepopolne zapise smo izločili iz analize.

Predobdelava besedila. Naslovi so bili predhodno obdelani z metodami čiščenja besedila: pretvorba v male črke, odstranjevanje posebnih znakov in številk ter tokenizacija. Nadalje smo odstranili angleške stop besede in uporabili algoritem za korenjenje (*stemming*), da bi zmanjšali variabilnost besed in izboljšali konsistenco vhodnega prostora.

Kodiranje izhodnih razredov. Kategorije novic smo pretvorili v numerične vrednosti z uporabo `LabelEncoder`, kar omogoča uporabo klasifikacijskih algoritmov.

Delitev podatkov na učni in testni del. Podatkovni niz smo razdelili na učno in testno množico v razmerju 80:20, pri čemer smo uporabili stratificirano delitev, da smo ohranili razmerje med razredi v obeh množicah.

Izgradnja in učenje klasifikacijskih modelov. Preizkusili smo več različnih modelov za večrazredno klasifikacijo:

- **LinearSVC z One-vs-Rest strategijo**: Klasičen linearen klasifikator, ki dobro deluje na visoko-dimenzionalnih podatkih, kot so TF-IDF vektorji.
- **Prilagojen model LS-TWSVM**: Implementirali smo lastno različico algoritma *Least Squares Twin Support Vector Machine*, prilagojeno za redke predstavitve podatkov in večrazredno klasifikacijo s strategijo One-vs-Rest.
- **RidgeClassifier**: Alternativni model s penalizacijo L2, prav tako uporabljen v kombinaciji z One-vs-Rest pristopom.

Evaluacija modelov. Uspešnost modelov smo ovrednotili s pomočjo standardnih metrik za večrazredno klasifikacijo: natančnost (*precision*), priklic (*recall*) in F1-mera. Metrike so bile izračunane za vsako posamezno kategorijo, kar omogoča podrobno primerjavo uspešnosti med različnimi modeli.

4.4 Avtorski prispevek

V okviru eksperimenta smo razvili in preizkusili lastno različico modela LS-TWSVM (*Least Squares Twin Support Vector Machine*), prilagojeno za uporabo nad redkimi TF-IDF predstavitvami. Model

temelji na ideji dveh ločitvenih ploskev, kjer se problema klasifikacije pristopi kot sistemu linearnih enačb z dodano regularizacijo. Reševanje sistema poteka z uporabo metode najmanjših kvadratov (LSQR), ki je primerna za redke matrike velike dimenzije.

Model smo vključili v `scikit-learn` združljiv `Cevovod`, kar omogoča integracijo v obstoječ eksperimentalni okvir in primerjavo z drugimi pristopi. Z implementacijo lastnega modela smo želeli raziskati njegove prednosti in slabosti v primerjavi z obstoječimi metodami ter prispevati k razumevanju učinkovitosti alternativnih klasifikatorjev na področju analize besedil.

V preglednici **??** predstavljamo primerjavo rezultatov naše implementacije z rezultati, predstavljenimi v znanstvenem članku. Naši rezultati kažejo zelo dobro uspešnost klasifikacije, kjer dosegamo povprečno natančnost 0.87, občutljivost 0.86 in F1-vrednost 0.86. Posebej dobri rezultati so vidni pri kategorijah DIVORCE (natančnost 0.92) in CRIME (natančnost 0.90), kar kaže na dobro sposobnost razločevanja teh dveh kategorij.

Primerjava z rezultati iz članka kaže, da je naša implementacija nekoliko manj natančna (razlika približno 0.11 točk), vendar še vedno dosega zelo konkurenčne rezultate. Razlika v rezultatih je deloma posledica različnih podatkovnih nizov in kategorij, ki smo jih uporabili v naši študiji. Medtem ko članek uporablja štiri kategorije (Acq, Earn, Money-fx, Grain), naša implementacija obravnava deset različnih kategorij, kar predstavlja kompleksnejši problem klasifikacije.

Kljub temu, da naši rezultati ne dosegajo popolnoma enake ravni natančnosti kot rezultati iz članka, je pomembno omeniti, da naša implementacija še vedno dosega zelo dobro uspešnost klasifikacije, kar potrjuje učinkovitost uporabljene metode. Posebej pomembno je, da naša implementacija ohranja visoko stopnjo konsistentnosti med natančnostjo, občutljivostjo in F1-vrednostjo, kar kaže na dobro uravnoteženost klasifikacijskega modela.

Tabela 1: Rezultati naše klasifikacije po kategorijah (delni izpis)

Kategorija	Natančnost	Občutljivost	F1-vrednost
ARTS	0.85	0.82	0.83
ARTS & CULTURE	0.83	0.81	0.82
BLACK VOICES	0.87	0.86	0.86
BUSINESS	0.88	0.89	0.88
COLLEGE	0.86	0.84	0.85
COMEDY	0.85	0.87	0.86
CRIME	0.90	0.88	0.89
CULTURE & ARTS	0.86	0.85	0.85
DIVORCE	0.92	0.90	0.91
EDUCATION	0.87	0.86	0.86
...	...	...	...
Povprečje (macro)	0.87	0.86	0.86
Povprečje (micro)	0.89	0.89	0.89

Tabela 2: Rezultati klasifikacije iz članka

Kategorija	Natančnost	Občutljivost	F1-vrednost
Acq	0.9897	0.9614	0.9754
Earn	0.9811	0.9909	0.9860
Money-fx	0.9707	1.0000	0.9851
Grain	0.9817	1.0000	0.9907
Povprečje	0.9808	0.9881	0.9843

LITERATURA

[1] Francois-Regis Chaumartin. 2007. UPAR7: A knowledge-based system for headline sentiment tagging. *Proceedings of SemEval-2007* (06 2007).

[2] Ratnadeep Deshmukh and D. Kirange. 2012. Emotion Classification of News Headlines Using SVM. (05 2012).

[3] R. R. Deshmukh and D. K. Kirange. 2013. Classifying News Headlines for Providing User Centered E-Newspaper Using SVM. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)* 2, 3 (May–June 2013). https://www.researchgate.net/publication/262793684_Classifying_News_Headlines_for_Providing_User_Centered_E-Newspaper_Using_SVM

[4] EITCA Academy. 2023. Kaj je stroj podpornih vektorjev (SVM)? https://sl.eitca.org/Umetna-inteligenca/eitc-ai-mlp-strojno-uÄŀenje-s-pythonom/podporni-vektorski-stroj/parametri-svm/kaj-je-podporni-vektorski-stroj-svm/ Dostopno: 16. marec 2025.

[5] Qi Gao, Fabian Abel, Geert-Jan Houben, and Ke Tao. 2011. Interweaving Trend and User Modeling for Personalized News Recommendation. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. IEEE. https://doi.org/10.1109/WI-IAT.2011.184

[6] Angel L. Garrido, Oscar Gómez, Sergio Ilarri, and Eduardo Men. 2011. NASS: News Annotation Semantic System. In *2011 23rd IEEE International Conference on Tools with Artificial Intelligence*. https://doi.org/10.1109/ICTAI.2011.109

[7] Zornitsa Kozareva, Borja Navarro-Colorado, Sonia Vázquez, and Andrés Montoyo. 2007. UA-ZBSA: A headline emotion classification through Web information. (01 2007), 334–337.

[8] Pooja Saigal and Vaibhav Khanna. 2020. Multi-category news classification using Support Vector Machine based classifiers. *SN Applied Sciences* 2 (2020), 458. https://doi.org/10.1007/s42452-020-2266-6

[9] Hajah T. Sueno, Bobby D. Gerardo, and Ruji P. Medina. 2020. Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique. *International Journal of Advanced Trends in Computer Science and Engineering* 9, 3 (May–June 2020), 3937–3944. https://doi.org/10.30534/ijatcse/2020/216932020

[10] Hajah T. Sueno, Bobby D. Gerardo, and Ruji P. Medina. 2020. Transforming Text Documents Into Numerical Format Using Enhanced Bayesian Vectorization For Multi-class Classification. 3 (2020), 18–22.