TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1154

Postupci odabira značajki i prikaza dokumenta za klasifikaciju teksta

Mihael Šafarić

Zagreb, srpanj 2015.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 6. ožujka 2015.

Predmet: **Analiza i pretraživanje teksta**

DIPLOMSKI ZADATAK br. 1154

Pristupnik: **Mihael Šafarić (0036460167)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Postupci odabira značajki i prikaza dokumenta za klasifikaciju teksta**

Opis zadatka:

Klasifikacija teksta jest postupak pridjeljivanja oznaka tekstnim dokumentima na temelju njihovog sadržaja. U tu se svrhu najčešće koriste modeli strojnog učenja, primijenjeni na vektorsku reprezentaciju dokumenata kao vreće riječi. Premda su ti modeli razmjerno učinkoviti i robusni, istraživanja su pokazala da uspješnost klasifikacije može uvelike ovisiti o tome koje se značake koriste kao riječi. Predložen je niz postupaka za automatski odabir značajki kojima se nastoji poboljšati uspješnost klasifikacije. S druge strane, u novije vrijeme predložene su reprezentacije dokumenata temeljene na neuronskim mrežama, koje su se pokazale vrlo uspješnima, a ne iziskuju manipulaciju sa značajkama.
U okviru diplomskoga rada potrebno je proučiti postupke za odabir značajki u klasifikaciji teksta te novije modele prikaza riječi i dokumenata temeljene na neuronskim mrežama. Razviti programsku implementaciju postupaka odabira značajki te ga primijeniti na referentne zbirke tekstova na engleskome (Reuters Corpus RVC1) i hrvatskome jeziku (zbirka novinskih članaka Vjesnik i zbirka pravnih dokumenata NN13205). Primijeniti nekoliko klasifikacijskih modela te usporediti rezultate modela koji koriste odabir značajaki i modela koji koriste neuronske reprezentacije riječi. Razmotriti model prikaza koji bi kombinirao prednosti obaju pristupa. Provesti iscrpno vrednovanje, statističku obradu rezultata te analizu pogrešaka. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 30. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr. sc. Siniša Srbljić

SADRŽAJ

1. Uvod **1**

2. Klasifikacija teksta **3**

2.1. Klasifikatori 3

2.1.1. Naivan Bayesov klasifikator 3

2.1.2. Logistička regresija 4

2.1.3. Stroj potpornih vektora 5

2.2. Metode odabira značajki 6

2.2.1. Metoda frekvencija dokumenata 7

2.2.2. Metoda uzajamne informacije 7

2.2.3. Metoda korisnosti informacije 9

2.2.4. Metoda χ^2 (CHI) 10

2.3. Word2vec 11

2.4. Doc2vec 13

3. Eksperimenti **15**

3.1. Zbirke tekstova 15

3.1.1. NN13205 15

3.1.2. Vjesnik 15

3.1.3. Reuters RCV1 16

3.2. Eksperimentalne postavke 17

3.2.1. Tradicionalni pristup klasifikaciji teksta 18

3.2.2. Moderni pristupi klasifikaciji teksta 19

3.3. Metode vrednovanja klasifikatora 20

4. Rezultati **21**

4.1. Zbirka tekstova *Vjesnik* 21

4.2. Zbirka tekstova *NN13205* 25

4.3. Zbirka tekstova *Reuters RCV1* 26
4.4. Diskusija rezultata 30

5. Zaključak **31**

Literatura **32**

1. Uvod

Sve više tekstnih informacija dostupno je u digitalnom obliku, a učinkovitu ekstrakciju informacija teško je postići bez dobrog indeksiranja i sažimanja sadržaja dokumenata. Klasifikacija teksta jedno je od rješenja tog problema. Klasifikacija teksta jedan je od zadataka obrade prirodnog jezika, a u ovom radu obrađene su metode temeljene na strojnom učenju. Klasifikacija teksta je problem pridjeljivanja unaprijed zadanih kategorija dokumentima. Formalno, ako je d_i dokument iz kolekcije dokumenata $D = \{d_1, d_2, ..., d_n\}$ i $C = \{c_1, c_2, ..., c_m\}$ skup kategorija, tada je zadatak klasifikacije teksta pridjeljivanje kategorija c_j dokumentu d_i. Ako se dokumentu pridjeljuje točno jedna kategorija iz C, onda govorimo o klasifikaciji s jednom oznakom (engl. *single-label classification*), a ako je dokumentu moguće pridjeljeliti više od jedne ili nijednu kategoriju, onda govorimo o klasifikaciji s više oznaka (engl. *multi-label classification*). Glavna značajka problema klasifikacije teksta je velika dimenzionalnost prostora značajki koji se sastoji od jedinstvenih pojmova koji se pojavljuju u zbirci tekstova. Za smanjivanje veličine prostora značajki koriste se metode odabira značajki koje na temelju statističkih podataka iz primjera za učenje odabiru značajke koje ih najbolje opisuju.

Klasifikacija teksta subjektivan je zadatak te ako dvije osobe pridjeljuju jednu ili više kategorija iz C dokumentu d_i, njihovi odgovori vrlo često neće biti jednaki. Na primjer, članak koji govori o prodaji dionica nogometnog kluba može biti smješten u kategoriju *sport*, u kategoriju *financije*, u kategoriju *politika* ili u neku od kombinacija te tri kategorije. Također, izazov kod klasifikacije teksta je velik broj primjera za učenje kao i velik broj kategorija koje su često jako neujednačene.

Klasifikacija teksta primjenjuje se u mnogim zadacima kao što su automatsko indeksiranje dokumenata, filtriranje dokumenata, automatsko generiranje metapodataka, razrješavanje višeznačnosti riječi (engl. *word sense disambiguation*) i općenito u aplikacijama koje zahtijevaju organizaciju dokumenata.

U ovom radu obrađena je tema klasifikacije dokumenata na hrvatskom i engleskom jeziku koristeći različite metode odabira značajki i prikaza dokumenta s ciljem dobi-

vanja uvida o utjecaju tih parametara na učinkovitost klasifikacije. Eksperimenti su provedeni nad tri zbirke tekstova od kojih dvije sadrže dokumente na hrvatskom jeziku, dok jedna sadrži dokumente na engleskom jeziku. Dobiveni rezultati pomoći će u gradnji učinkovitijih modela za klasifikaciju dokumenata.

Ekperimenti su provedeni koristeći četiri različite metode odabira značajki: frekvencija dokumenata, metoda uzajamne informacije, korisnost informacije te metoda χ^2. Dobivene značajke korištene su kao ulaz za nekoliko različitih klasifikatora. Uz taj tradicionalni pristup klasifikaciji koji koristi reprezentaciju dokumenata kao vreće riječi, provedeni su i eksperimenti s modernim metodama temeljenim na neuronskim mrežama kao što su metode *word2vec* i *doc2vec* koje kao ulaz primaju dio teksta te vraćaju njegovu vektorsku reprezentaciju. Od prije spomenutih zbirki tekstova, dvije su primjer klasifikacije s više oznaka, dok je jedna primjer klasifikacije s jednom oznakom.

U idućem poglavlju opisani su korišteni klasifikatori, metode odabira značajki općenito, te detaljno svaka korištena metoda odabira značajki. Također, dana je teorijska podloga metoda koje iz čistog teksta grade njegovu vektorsku reprezentaciju. U trećem poglavlju opisana je struktura pojedine zbirke tekstova te mjere vrednovanja klasifikatora. U tom poglavlju navedeni su i modeli klasifikacije dokumenata isprobani u sklopu ovog rada, njihovi parametri te implementacijski detalji. U četvrtom poglavlju izneseni su dobiveni rezultati za pojedinu zbirku tekstova te je dana usporedba rezultata dobivenih korištenjem pojedinog modela.

2. Klasifikacija teksta

2.1. Klasifikatori

2.1.1. Naivan Bayesov klasifikator

Naivan Bayesov klasifikator je parametarski klasifikator opisan u (Jan Šnajder, 2012) i (McCallum et al., 1998), kod kojeg se klasifikacija primjera ostvaruje na temelju aposteriorne vjerojatnosti. Ta vjerojatnost izračunava se primjenom Bayesovog pravila (formula 2.1) koje za pojedinu klasu daje vjerojatnost da primjer pripada toj klasi koristeći pri tome pretpostavku o uvjetnoj nezavisnosti varijabli koja kaže da su varijable međusobno nezavisne za zadanu klasu.

$$P(C_j|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_j)P(C_j)}{\sum_{k=1}^{K} p(\boldsymbol{x}|C_k)P(C_k)} \tag{2.1}$$

Pretpostavka o uvjetnoj nezavisnosti varijabli omogućuje da parametri klasifikatora budu trenirani odvojeno, što uvelike pojednostavljuje učenje modela. Iako u praksi ta pretpostavka obično nije točna, naivan Bayesov klasifikator funkcionira vrlo dobro.

Bernoullijev model

Kod Bernoullijevog modela dokument je prikazan vektorom binarnih značajki koje označavaju prisutnost, odnosno odsutnost pojedine riječi u dokumentu. Ako x_i predstavlja i-tu značajku, a p_{ki} vjerojatnost pojave i-te značajke u klasi c_k, tada je vjerojatnost da dokument d_i pripada klasi c_k jednaka produktu vjerojatnosti pojavljivanja pojedine značajke u dokumentu, što je prikazano formulom 2.2.

$$p(d_i|c_k) = \prod_{i=1}^{n} p_{ki}^{x_i}(1 - p_{ki})^{(1-x_i)} \tag{2.2}$$

Poredak riječi u dokumentu i broj ponavljanja pojedine riječi ne uzimaju se u obzir, a vjerojatnosti značajki koje se ne pojavljuju u dokumentu eksplicitno se uzimaju u

obzir kod određivanja $p(d_i|C_k)$.

Multinomijalni model

Kod multinomijalnog modela dokument je prikazan vektorom čiji elementi predstavljaju broj pojavljivanja pojedine značajke u dokumentu. Kao i u Bernoullijevom modelu, informacija o poretku riječi je izgubljena. Ako x_i predstavlja i-tu značajku u vektorskom prikazu dokumenta, a p_{ki} vjerojatnost pojave te značajke u klasi c_k, tada je vjerojatnost da dokument d_i pripada klasi c_k jednaka multinomijalnoj razdiobi, što je prikazano formulom 2.3.

$$p(d_i|c_k) = \frac{\left(\sum_i x_i\right)!}{\prod_i x_i!} \prod_i {p_{ki}}^{x_i} \tag{2.3}$$

2.1.2. Logistička regresija

Logistička regresija je probabilistički diskriminativni model koji izravno modelira aposteriornu vjerojatnost $P(C_j|x)$ prema 2.4, a opisan je u (Jan Šnajder, 2012).

$$P(C_j|x) = \sigma(\mathbf{w}^T\mathbf{x} + \omega_0) = \sigma(\tilde{\mathbf{w}}^T\tilde{\mathbf{x}}) \tag{2.4}$$

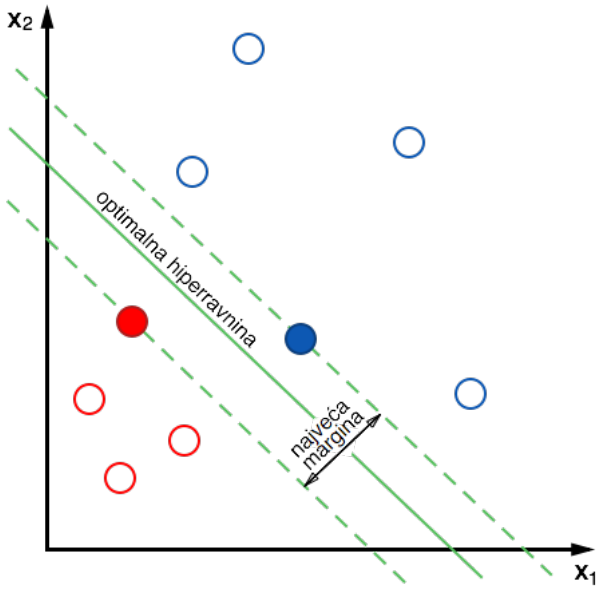
Funkcija $\sigma(\alpha)$ je logistička funkcija koja prema 2.5 preslikava sve realne brojeve na konačni interval (0, 1).

$$\sigma(\alpha) = \frac{1}{1 + exp(-\alpha)} \tag{2.5}$$

Učenje modela logističke regresije svodi se na određivanje parametara $\tilde{\mathbf{w}}$ minimizacijom funkcije pogreške na skupu za učenje. U ovom radu učenje modela logističke regresije vrši se pomoću stohastičkog gradijentnog spusta (engl. *stochastic gradient decent*) kod kojeg se ugađanje težina gradijenta obavlja na temelju svakog primjera pojedinačno.

2.1.3. Stroj potpornih vektora

Stroj potpornih vektora (engl. *support vector machine, SVM*), opisan u (Jan Šnajder, 2012), je diskriminativan model koji pronalazi optimalnu hiperravninu, odnosno onu koja daje maksimalnu marginu između primjera za učenje dviju klasa, to je prikazano slikom 2.1. Tu ravninu moguće je prikazati kombinacijom odabranih vektora primjera iz skupa za učenje koji se nazivaju potporni vektori.



Slika 2.1: Prikaz optimalne hiperravnine i pripadajuće maksimalne margine

SVM je vrlo učinkovit u radu s visokodimenzijskim prostorom značajki, a dodatnu učinkovitost postiže primjenom tzv. jezgrenih funkcija. Jezgrena funkcija mjeri sličnost dvaju vektora u nekom prostoru značajki. U ovom radu je uz linearni SVM isproban i nelinearni SVM s radijalnim baznim funkcijama (engl. *radial basis functions, RBF*) kao jezgrenim funkcijama. Radijalne bazne funkcije ovise samo o udaljenosti između primjera, a poseban slučaj radijalne bazne funkcije je Gaussova jezgra:

$$\kappa(x, x') = exp\Big\{ - \frac{\| x - x' \|^2}{2\sigma^2} \Big\} = exp\{-\gamma \| x - x' \|^2\} \tag{2.6}$$

Parametar $\gamma = \frac{1}{2\sigma^2}$ naziva se preciznost i kontrolira kojom brzinom $\kappa(x - x')$ teži k nuli u ovisnosti o udaljenosti. Gaussova jezgra mjeri sličnost dvaju primjera temeljem njihove udaljenosti u ulaznom prostoru. Za slične primjere vrijedi $\kappa(x, x') \rightarrow 1$, pa su ti primjeri i blizu u prostoru značajki. Za potpuno različite primjere vrijedi $\kappa(x, x') \rightarrow 0$, pa su ti primjeri ortogonalni u prostoru značajki.

2.2. Metode odabira značajki

Odabir značajki postupak je pretraživanja skupa značajki kojim se odabire podskup najrelevantnijih značajki za promatrani problem. Osnovna pretpostavka kod odabira značajki je da podaci opisani s puno značajki sadrže značajke koje su ili suvišne ili nebitne te stoga mogu biti uklonjene bez većeg gubitka informacija. Suvišne i nebitne značajke su dva različita pojma jer bitna značajka može biti suvišna u prisutstvu neke druge bitne značajke.

Korištenjem samo najbitnijeg skupa značajki smanjuje se prenaučenost modela jer manje suvišnih podataka dovodi do manje vjerojatnosti da će se klasifikacijska odluka donijeti na temelju šumova (engl. *noise data*). Značajka se klasificira kao šum ako se njezinim dodavanjem u skup bitnih značajki smanjuje učinkovitost modela. Učinkovitost se također povećava i uklanjanjem varljivih (engl. *misleading*) podataka, a manja količina podataka dovodi i do bržeg učenja modela. Smanjenjem broja značajki model se pojednostavljuje čime vizualizacija i razumijevanje podataka postaju lakši.

Metode za odabir značajki mogu se promatrati kao kombinacija tehnike pretraživanja za odabir podskupa značajki i mjere evaluacije kojom se ocjenjuju odabrani podskupovi značajki. Tako se metode za odabir značajki dijele na metode koje djeluju kao omotač (engl. *wrapper methods*), metode filtra i metode ugrađene u prediktivni model (engl. *embedded methods*).

Metode koje djeluju kao omotač za ocjenjivanje podskupa značajki koriste prediktivni model. Za pojedini podskup značajki vrši se učenje i evaluacija modela s tim značajkama te se na kraju uzima onaj podskup značajki koji daje najbolje rezultate.

Metode filtra koriste statističke mjere za određivanje ocjene pojedine značajke te se uzima skup značajki s najboljim ocjenama. Metode filtra su manje računalno zahtjevne od metoda koje djeluju kao omotač, ali rezultiraju s općenitijim skupom značajki koji nije podešen za specifičan prediktivni model i ne sadrži pretpostavke tog modela te je zbog toga prikladniji za određivanje veza između značajki.

Metode ugrađene u prediktivni model kombiniraju prednosti metoda filtra i metoda koje djeluju kao omotač te u procesu učenja modela određuju koje značajke najbolje pridonose poboljšanju učinkovitosti modela.

U ovom radu korištene su metode filtra jer ih je zbog njihove računalne složenosti moguće primjeniti na veće zbirke tekstova.

Kod opisa svih metoda za obir značajki korištena je sljedeća notacija:

- N – ukupan broj dokumenata
- $N_{\bullet 0}$ – broj dokumenata koji ne pripadaju klasi c_i

- $N_{\bullet 1}$ – broj dokumenata koji pripadaju klasi c_i
- $N_{0\bullet}$ – broj dokumenata koji ne sadrže pojam t
- $N_{1\bullet}$ – broj dokumenata koji sadrže pojam t
- N_{00} – broj dokumenata koji ne pripadaju klasi c_i i ne sadrže pojam t
- N_{01} – broj dokumenata koji pripadaju klasi c_i i ne sadrže pojam t
- N_{10} – broj dokumenata koji ne pripadaju klasi c_i i sadrže pojam t
- N_{11} – broj dokumenata koji pripadaju klasi c_i i sadrže pojam t

2.2.1. Metoda frekvencija dokumenata

Metoda frekvencije dokumenata (engl. *document frequency*) je najjednostavnija metoda smanjivanja dimenzionalnosti rječnika koja kao mjeru važnosti pojedinog pojma koristi broj dokumenata u kojima se promatrani pojam pojavljuje. Pojam je više rangiran ako se pojavljuje u više dokumenata. Osnovna pretpostavka je da pojmovi koji se rijetko pojavljuju ili sadrže malo informacija korisnih za predviđanje kategorije ili imaju malo utjecaja unutar cijele zbirke tekstova. U nekim područjima obrade prirodnog jezika smatra se da su pojmovi koji se rijetko pojavljuju relativno korisni pa se zbog toga ova metoda obično ne koristi za agresivno uklanjanje pojmova. Složenost ove metode je linearna te je stoga primjenjiva na velike zbirke tekstova.

2.2.2. Metoda uzajamne informacije

Metoda uzajamne informacije (engl. *mutual information*) temelji se na izračunu uzajamne informacije između pojma t i klase c. Izračunata vrijednost interpretira se kao količina informacije kojom prisutnost, odnosno odsutnost promatranog pojma pridonosi donošenju točne klasifikacijske odluke za klasu c. Drugim riječima, pokazuje koliko informacije o klasi c sadrži promatrani pojam.

$$I(T,C) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} P(T = e_t, C = e_c) log_2 \frac{P(T = e_t, C = e_c)}{P(T = e_t) \times P(C = e_c)} \tag{2.7}$$

gdje je T slučajna varijabla koja poprima vrijednost $e_t = 1$ ako dokument sadrži pojam t te vrijednost $e_t = 0$ ako dokument ne sadrži pojam t. Slučajna varijabla C poprima vrijednost $e_c = 1$ ako dokument pripada klasi c te $e_c = 0$ ako dokument ne pripada klasi c. Tada se uzajamna informacija između pojma t i klase c definira kao:

$$I(t,c) = p(t,c)\log_2 \frac{p(t,c)}{p(t) \times p(c)} \tag{2.8}$$

Formula 2.7 se tada može zapisati kao:

$$\begin{aligned} I(t,c) = \frac{N_{11}}{N}\log_2 \frac{N \times N_{11}}{N_{1\bullet} \times N_{\bullet 1}} + \frac{N_{01}}{N}\log_2 \frac{N \times N_{01}}{N_{0\bullet} \times N_{\bullet 1}} + \\ \frac{N_{10}}{N}\log_2 \frac{N \times N_{10}}{N_{t\bullet} \times N_{\bullet 0}} + \frac{N_{00}}{N}\log_2 \frac{N \times N_{00}}{N_{0\bullet} \times N_{\bullet 0}} \end{aligned} \tag{2.9}$$

Formula 2.9 je točkasta procjena uzajamne informacije jer se računa za dvije konkretne vrijednosti slučajnih varijabli T i C, tj. za fiksirani pojam i fiksiranu kategoriju. Vrijednost uzajamne informacije jednaka je nuli ako su pojam t i klasa c međusobno nezavisni. Kod određivanja ocjene pojma korištena su dva načina kombiniranja uzajamnih informacija pojma i pojedine kategorije koji su opisani u (Yang i Pedersen, 1997) te prikazani formulama 2.10 i 2.11.

$$I_{avg}(t) = \sum_{i=1}^{m} P(c_i) \times I(t,c_i) \tag{2.10}$$

$$I_{max}(t) = \max_{i=1}^{m}\{I(t,c_i)\} \tag{2.11}$$

Nedostatak metode uzajamne informacije je da vrijednost uzajamne informacije klase c i pojma t značajno ovisi o marginalnoj vjerojatnosti pojma.

U slučaju da dva pojma imaju jednaku uvjetnu vjerojatnost $P(t|c)$, rijeđi pojam imat će višu ocjenu od češće pojavljivanog pojma te zbog toga ocjene pojmova nisu usporedive ako se njihove frekvencije pojavljivanja značajno razlikuju.

2.2.3. Metoda korisnosti informacije

Metoda korisnosti informacije (engl. *information gain*) korštena je kod postupka klasifikacije dokumenata u radu (Yang i Pedersen, 1997), a često se koristi za rangiranje pojmova u području strojnog učenja. Pokazuje broj bitova informacije za predviđanje kategorije dobivene iz pretpostavke da se promatrani pojam pojavljuje, odnosno ne pojavljuje u dokumentu. Korisnost informacije *G(t)* određuje se prema:

$$\begin{aligned} G(t) = -\sum_{i=1}^{m} P(c_i) log P(c_i) + \\ P(t) \sum_{i=1}^{m} P(c_i|t) log P(c_i|t) + \\ P(\bar{t}) \sum_{i=1}^{m} P(c_i|\bar{t}) log P(c_i|\bar{t}) \end{aligned} \tag{2.12}$$

gdje je m broj kategorija.
Vjerojatnost klase $P(c_i)$ opisana je formulom:

$$P(c_i) = \frac{N_{\bullet i}}{N} \tag{2.13}$$

dok su vjerojatnosti pristutnosti i odsutnosti pojma opisane s:

$$P(t) = \frac{N_{1\bullet}}{N} \tag{2.14}$$

$$P(\bar{t}) = \frac{N_{0\bullet}}{N} \tag{2.15}$$

Uvjetne vjerojatnosti $P(c_i|t)$ i $P(c_i|\bar{t})$ dobiju se iz:

$$P(c_i|t) = \frac{N_{11}}{N_{1\bullet}} \tag{2.16}$$

$$P(c_i|\bar{t}) = \frac{N_{01}}{N_{0\bullet}} \tag{2.17}$$

Formule 2.14 - 2.17 su procjene najveće izglednosti koje pretpostavljaju da su primjeri međusobno nezavisni i da potječu od identične razdiobe. Za svaki jedinstveni pojam iz rječnika određuje se korisnost informacije i na kraju se uzima *k* najbolje rangiranih pojmova.

2.2.4. Metoda χ^2 (CHI)

U statistici χ^2 test koristi se za ispitivanje nezavisnosti dvaju događaja gdje se dva događaja A i B definiraju kao nezavisni ako vrijedi $P(AB) = P(A)P(B)$. Kod odabira značajki ta dva događaja su pojava pojma i pojava klase. Pojmovi se tada mogu rangirati prema:

$$\chi^2(D,t,c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{2.18}$$

gdje je N promatrana frekvencija u D, a E očekivana frekvencija u D. Na primjer E_{11} je očekivana frekvencija dokumenata koji pripadaju klasi c i sadrže pojam t uz pretpostavku da su pojam i klasa nezavisni. Formula 2.18 može se zapisati i kao:

$$\chi^2(D,t,c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})} \tag{2.19}$$

χ^2 mjera pokazuje koliko očekivana frekvencija E i promatrana frekvencija N međusobno odstupaju. Pojam t i klasa c su međusobno nezavisni ako je ta vrijednost jednaka nuli dok veća vrijednost te mjere ukazuje na to da je hipoteza o nezavisnosti, koja kaže da su frekvencije E i N slične, pogrešna.

Kod pridjeljivanja konačne ocjene pojmu, kao i kod metode uzajamne informacije, koriste se dva načina kombiniranja χ^2 mjere pojma i pojedine klase:

$$\chi^2_{avg}(t) = \sum_{i=1}^{m} P(c_i)\chi^2(t,c_i) \tag{2.20}$$

$$\chi^2_{max}(t) = \max_{i=1}^{m}\{\chi^2(t,c_i)\} \tag{2.21}$$

Glavna razlika metode χ^2 u odnosu na metodu uzajamne informacije je što je χ^2 normalizirana vrijednost te su stoga χ^2 vrijednosti usporedive između pojmova pojedine kategorije. Metoda χ^2 nije pouzdana za pojmove s niskom frekvencijom pojavljivanja jer se u tom slučaju normalizirane vrijednosti ne mogu točno usporediti s χ^2 razdiobom.

2.3. Word2vec

Algoritam *word2vec*[1] koristi se za određivanje vektorske reprezentacije riječi, a predstavljen je u (Mikolov et al., 2013). Algoritam kao ulaz prima skup tekstova, iz njega gradi rječnik te kao rezultat daje vektore koji predstavljaju riječi iz tog rječnika. Na početku je svaka riječ iz rječnika slučajni vektor dimenzije N, a tijekom učenja algoritam pronalazi optimalan vektor za pojedinu riječ. U postupku učenja koristi se jedna od dviju metoda:

- kontinuirana vreća riječi (engl. *continuous bag-of-words, CBOW*)
- kontinuirani skip-gram (engl. *continuous skip-gram*)

Jezični model neuronske mreže (engl. *neural network language model, NNLM*) predstavljen u (Bengio et al., 2003) omogućuje dobru reprezentaciju riječi, ali ima veliku računalnu složenost. U radu (Mikolov et al., 2009) donesen je zaključak da se *NNLM* može uspješno trenirati u dva koraka:

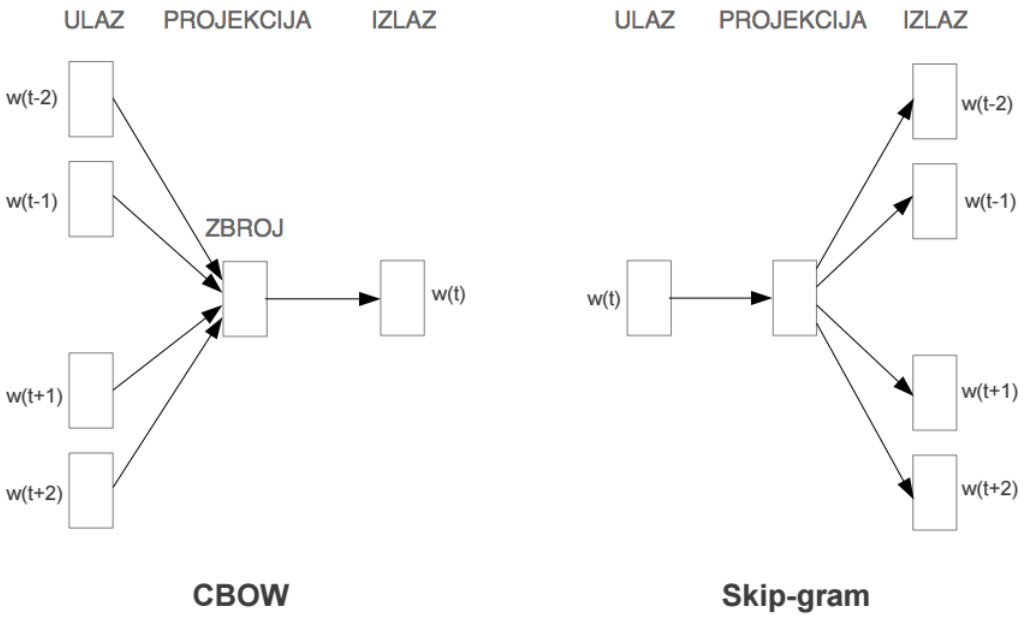
1. kontinuirani vektori riječi nauče se na jednostavnom modelu
2. N-gram *NNLM* uči se s vektorima iz prošlog koraka

Na temelju tog zaključka nastale su *CBOW* i *skip-gram* metode koje nisu u mogućnosti reprezentirati riječi s preciznošću kao *NNLM*, ali je njihova računalna složenost manja što omogućuje da se učenje modela izvede nad većom količinom podataka.

Vektori riječi dobiveni algoritmom *word2vec* sadrže informacije o jezičnim pravilima i o kontekstu okolinih riječi. Na primjer vektorske operacije *vector('king') - vector('man') + vector('woman')* rezultiraju vektorom koji je blizu vektora *vector('queen')*. Da bi se postiglo takvo ponašanje modela $word2vec$, on mora biti naučen s velikom količinom podataka uz dostatnu dimenzionalnost vektora.

CBOW metoda uzima k riječi koje prethode i k riječi koje slijede nakon trenutne riječi te s njima trenira klasifikator i pokušava predvidjeti trenutnu riječ, dok *skip-gram* metoda za svaku riječ predviđa k riječi koje prethode i k riječi koje slijede nakon trenutne riječi. Arhitektura tih dviju metoda prikazana je na slici 2.2 gdje je $w(t)$ trenutna riječ.

[1] https://code.google.com/p/word2vec



Slika 2.2: Arhitektura metoda *CBOW* i *skip-gram*, preuzeto iz (Bengio et al., 2003)

2.4. Doc2vec

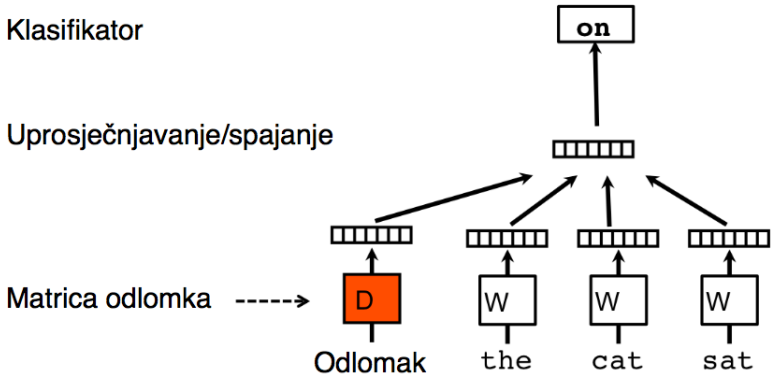
Metoda *Doc2vec* koristi algoritam vektora odlomka (engl. *Paragraph Vector*) predložen u (Le i Mikolov, 2014), koji kao ulaz prima tekst proizvoljne duljine, a kao rezultat daje njegovu vektorsku reprezentaciju. Pojedini odlomak preslikan je na jedinstveni vektor prikazan stupcem u matrici D, a pojedina riječ preslikana je na jedinstven vektor prikazan stupcem u matrici W. Vektori odlomaka nisu dijeljeni između odlomaka, dok je matrica riječi W zajednička svim odlomcima. Vektori odlomaka i vektori riječi trenirani su stohastičkim gradijentnim spustom, a za dobivanje gradijenta korišten je algoritam propagacije pogreške unazad (engl. *backpropagation algorithm*).

Ako je pojedini odlomak preslikan na p dimenzija, a pojedina riječ na q dimenzija, tada je broj parametara modela jednak $N \times p + M \times q$, gdje je N broj odlomaka u zbrici tekstova, a M veličina rječnika. U slučaju velikog broja odlomaka N, broj parametara modela je velik, ali je dodavanje novih informacija tijekom učenja rijetko (engl. *sparse*) te je stoga učinkovito.

Postoje dva različita modela koja se koriste u postupku učenja:

- model vektora odlomka s distribuiranom memorijom (engl. *distributed memory model of paragraph vector, PV-DM*)
- model vektora odlomka s distribuiranom vrećom riječi (engl. *distributed bag of words version of paragraph vector, PV-DBOW*)

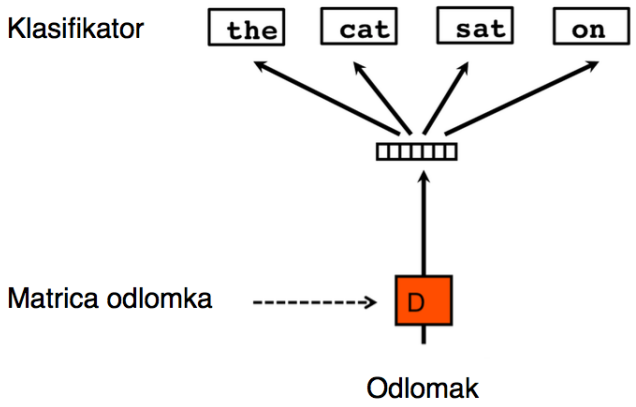
Model *PV-DM* (2.3) kod predviđanja sljedeće riječi u trenutnom kontekstu koristi vektor dobiven spajanjem ili uprosječnjavanjem vektora odlomka i vektora pojedinih riječi. Kontekst je fiksne duljine i dobiven je pomicanjem prozora po poragrafu.



Slika 2.3: Arhitektura PV-DM metode, preuzeto iz (Le i Mikolov, 2014)

Model *PV-DBOW* (2.4), za razliku od modela *PV-DM*, zanemaruje kontekst riječi

te na temelju vektora odlomka predviđa nove riječi.



Slika 2.4: Arhitektura PV-DBOW metode, preuzeto iz (Le i Mikolov, 2014)

Metoda *doc2vec* može se primjeniti na zadatke koji nemaju dovoljno označenih podataka jer se njezino učenje provodi s neoznačenim podacima. Također, metoda nasljeđuje dobra svojstva vektora riječi kao što su značenje riječi i kontekst u kojem se ona pojavljuje.

3. Eksperimenti

3.1. Zbirke tekstova

Svi eksperimenti provedeni su nad tri zbirke tekstova: *NN13205*, *Vjesnik* i *Reuters RCV1* od kojih *NN13205* i *Vjesnik* sadrže dokumente na hrvatskom jeziku, a *Reuters RCV1* sadrži dokumente na engleskom jeziku.

3.1.1. NN13205

Zbirka tekstova *NN13205* napravljena je u sklopu projekta *CADIAL*[1] i sastoji se od 13205 zakonodavnih dokumenata Republike Hrvatske na hrvatskom jeziku objavljenih prije 2009. godine u službenom glasniku Republike Hrvatske (Narodne Novine Republike Hrvatske). Zbirka tekstova sastoji se od 1.2M jedistvenih riječi i 39M riječi ukupno, dok je prosječna veličina dokumenta oko 3K riječi. Dokumenti iz zbirke označeni su kategorijama iz pojmovnika *EuroVoc* i *CroVoc*. *EuroVoc* pojmovnik sadrži 6797 kategorija podijeljenih u 21 različitih područja. Pojmovnik *EuroVoc* organiziran je u 8 razina od kojih je u ovom radu korišteno njih 6. Kombinacijom ta dva pojmovnika dobilo se ukupno 14571 kategorija od kojih su neke nadređene drugim kategorijama, a neke nemaju pridjeljen niti jedan dokument. Kad se takve kategorije uklone ostaje njih 3951 s prosjekom od 3.6 kategorija po dokumentu.

3.1.2. Vjesnik

Zbirka tekstova *Vjesnik* sastoji se od 258869 članaka objavljenih u Vjesniku, političkom dnevnom listu koji je izlazio u Republici Hrvatskoj, od listopada 1999. godine do rujna 2009. godine. Svi članci su na hrvatskom jeziku. Dokumenti su svrstani u 13 kategorija čija je zastupljenost prikazana u tablici 3.1. Svakom dokumentu pridjeljena je točno jedna kategorija. Ukupan broj riječi u zbirci je 92M od kojih je 2.3M

[1] http://www.cadial.org

jedinstvenih, dok je prosječna veličina dokumenta oko 350 riječi.

Tablica 3.1: Zastupljenost kategorija u zbirci tekstova *Vjesnik*

kategorija	# dokumenata	udio dokumenata u kategoriji
sta	5268	2.04 %
gle	5285	2.04 %
kom	8291	3.2 %
pis	11904	4.6 %
tem	14966	5.78 %
gos	20829	8.05 %
sss	20943	8.09 %
zag	21685	8.38 %
crn	21956	8.48 %
kul	21961	8.48 %
van	28450	10.99 %
spo	34872	13.47 %
unu	42459	16.4 %

3.1.3. Reuters RCV1

Zbirka tekstova *Reuters RCV1* sadrži 806791 dokumenata koji sadrže članke na engleskom jeziku koje je emitirala televizijska agencija Reuters, a struktura te zbirke opisana je u (Lewis et al., 2004). Sastoji se od tri skupa kategorija: teme (engl. *topics*), ekonomija (engl. *industries*) i regije (engl. *regions*), a u ovom radu korišten je samo skup kategorija s temama koji klasificira dokument s obzirom o temi koju obrađuje. Taj skup kategorija, nakon izbacivanja kategorija u kojima nema niti jednog dokumenta, sastoji se od 103 kategorije. Zastupljenost kategorija je vrlo neujednačena pa se tako u kategoriji *MILLENNIUM ISSUES* nalazi 5 dokumenata dok se u kategoriji *CCAT* nalazi 374316 dokumenata, a prosječno su dokumentu pridjeljene 3 kategorije. Zbirka se sastoji od 185M riječi od kojih je 2.3M jedinstveno, dok duljina dokumenata varira od nekoliko stotina do nekoliko tisuća riječi s prosječnom duljinom od 230 riječi.

3.2. Eksperimentalne postavke

U svim zbirkama tekstova najprije su uklonjeni interpunkcijski znakovi, a nakon toga su uklonjene riječi s manje od tri znaka te su sva velika slova zamijenjena pripadajućim malim slovima. Na kraju su izbačene riječi koje sadrže bilo koji znak osim crtice "–" i slova. Nakon tih promjena ukupan broj riječi u zbirkama tekstova kao i broj jedinstvenih riječi te veličina dokumenata smanjeni su na vrijednosti prikazane tablicom 3.2.

Tablica 3.2: Veličina zbirki tekstova nakon predobrade

zbirka tekstova	# riječi	# jedinstvenih riječi	prosječna duljina dokumenta
Vjesnik	51M	640K	200 riječi
NN13205	21M	252K	1600 riječi
Reuters RCV1	136M	469K	170 riječi

Svaka zbirka tekstova podijeljena je slučajnim odabirom na dva dijela – skup dokumenata za učenje i skup dokumenata za testiranje. Na skup dokumenata za učenje otpada 67% ukupnog broja dokumenata (osim kod zbirke tekstova *Reuters RCV1*), tj. učenje klasifikatora za zbirku tekstova *NN13205* provodi se s 8847 dokumenta, za zbirku tekstova *Vjesnik* s 173442 dokumenata dok se za zbirku tekstova *Reuters RCV1* zbog velikog broja dokumenata učenje vrši s njih 100000.

Kod klasifikacije s više oznaka korištena je metoda binarne relevantnosti (engl. *binary relevance method*), koja problem s više oznaka preslikava na problem s jednom oznakom pretpostavljajući da su oznake međusobno nezavisne. Problem s više oznaka rastavlja se na nekoliko binarnih klasifikacijskih problema i to po jedan za svaku oznaku koja sudjeluje u početnom problemu. U procesu pridjeljivanja oznaka nekom primjeru za svaku oznaku vrši se binarna klasifikacija te se u konačnici primjeru pridjeljuju sve oznake koje su klasificirane pozitivno.

Kod implementacije klasifikacije korištene su već gotove implementacije klasifikatora iz paketa *scikit-learn*[2], a za klasifikaciju s više oznaka korišten je *OneVsRestClassifier*, koji kao argument prima jedan od klasifikatora te se ponaša kao omotač oko njega kreirajući po jedan binarni klasifikator za svaku klasu. U ovom radu isprobani su klasifikatori Bernoullijev i multinomijalni naivan Bayes, linearni SVM, SVM s RBF jezgrenom funkcijom te logistička regresija.

Za dobivanje optimalnih hiperparametara klasifikatora korištena je 5-struka una-

[2]http://scikit-learn.org/stable/

krsna provjera (engl. *5-folded cross validation*), a zbog velikog broja klasifikatora u obzir je uzet samo malen opseg vrijednosti pojedinog hiperparametra. Za regularizacijski parametar C kod linearnog SVM-a i RBF SVM-a te za koeficijent jezgrene funkcije *gamma* kod RBF SVM-a u obzir su uzete vrijednosti iz skupa $\{0.1, 1, 10\}$.

3.2.1. Tradicionalni pristup klasifikaciji teksta

Pojedini dokument prikazan je vrećom riječi (engl. *bag-of-words*), tj. vektorom dimezije $|V|$ gdje je $|V|$ veličina rječnika. Elementi tog vektora su frekvencije pojavljivanja pojedine riječi u promatranom dokumentu. Na taj vektor primjenjena je jedna od metoda odabira značajki i novodobiveni vektor normaliziran je metodom *tf-idf* (engl. *term frequency-inverse document frequency*). Za odabir značajki korištene su lokalne implemetacije metoda opisanih u poglavlju 2.2. Metoda *tf-idf* za svaku riječ daje ocjenu u kojoj je mjeri ta riječ važna za promatrani dokument. Visoka ocjena dobiva se za riječi s visokom frekvencijom pojavljivanja u promatranom dokumentu, a niskom frekvencijom pojavljivanja gledajući sve dokumente u zbirci tekstova. Ako je t trenutna riječ, a d dokument iz skupa dokumenata D u kojem se nalazi promatrana riječ t, tada se ocjena *tf-idf* izračuna po formuli 3.1.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \tag{3.1}$$

Frekvencija riječi (engl. *term frequency, tf*) je broj poljavljivanja riječi t u dokumentu d, a u ovom radu korištena je modificirana verzija (3.2) kod koje se broj pojavljivanja riječi u promatranom dokumentu dijeli s ponavljanjem najučestalije riječi u tom dokumentu i time se sprječava pristranost metode prema duljim dokumentima.

$$tf(t, d) = 0.5 + \frac{0.5 \times f(t, d)}{max\{f(w, d) : w \in d\}} \tag{3.2}$$

Inverzna frekvencija dokumenta (engl. *inverse document frequency, idf*) pokazuje koliko informacije nosi promatrana riječ, tj. koliko često se promatrana riječ ponavlja u skupu dokumenata za učenje. *Idf* dobije se prema formuli 3.3, gdje je N broj dokumenata iz skupa za učenje.

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}| + 1} \tag{3.3}$$

3.2.2. Moderni pristupi klasifikaciji teksta

Kod metoda *word2vec* i *doc2vec* sadržaj svakog dokumenta se u procesu predobrade mijenja na način da se sva slova zamijene pripadajućim malim slovima, a simboli za novi redak zamijene se razmakom. Uz to, dodaje se razmak prije i nakon svakog interpunkcijskog znaka čime se postiže da se interpukcijski znakovi promatraju jednako kao i riječi. Za obje metode korištena je implementacije iz paketa *gensim*[3].

Učenje modela *word2vec* vrši se dodavanjem novih rečenica koje ažuriraju trenutni model. Za to je napravljen generator rečenica koji prolazi kroz prethodno obrađene dokumente iz skupa za učenje i iz njihovog sadržaja gradi rečenice. Krajem rečenice smatra se ili kraj dokumenta ili pronalazak jednog od znakova iz skupa $\{., !, ?\}$. Nakon što je model naučen, vektori dokumenata dobivaju se uprosječavanjem vektora pojedinih riječi iz promatranog dokumenta, tj. najprije se zbroje vektori pojedinih riječi iz promatranog dokumenta dobiveni kroz model *word2vec* te se nakon toga novodobiveni vektor podijeli s brojem riječi u promatranom dokumentu. Dobiveni vektori dokumenata koriste se kao podaci za učenje klasifikatora.

Kod učenja modela *doc2vec* cijeli sadržaj dokumenta promatra se kao jedna rečenica, a učenje se provodi dodavanjem označenih tekstova u model. Kao i kod metode *word2vec* napravljen je generator tekstova koji prolazi kroz dokumente iz skupa za učenje te vraća označene tekstove. Isprobane su dvije inačice gradnje vektora dokumenata koji se koriste kao ulaz u klasifikator. Prva je korištenje vektora dokumenta koji se dobije kao rezultat modela *DM* ili *DBOW*, dok je druga inačina korištenje rezultata oba modela na način da se vektori za promatrani dokument dobiveni od pojedinog modela spoje.

Isprobane su i hibridne metode koje kombiniraju metode odabira značajki s metodama za prikaz dokumenta temeljenim na neuronskim mrežama. Najprije se procesom odabira značajki dobije skup najznačajnijih riječi nakon čega se iz dokumenata izbacuju sve riječi koje se ne nalaze u tom skupu (interpunkcijski znakovi se ne izbacuju). Novonastali dokumenti koriste se kao ulaz u metode *word2vec* i *doc2vec*. Obje metode dale su znatno lošije rezultate od modela koji koriste samo tradicionalan ili samo moderan pristup, a razlog tome je što su novonastali dokumenti bili duljine od oko 10 riječi, pa su svi dokumenti bili međusobno slični. Također, količina teksta za učenje drastično je smanjena, a kontekst u kojem se pojedina riječ nalazi je izgubljen.

[3] https://radimrehurek.com/gensim/index.html

3.3. Metode vrednovanja klasifikatora

Vrednovanje klasifikatora provedeno je s nekoliko standardnih mjera vrednovanja – preciznost (P), odziv (R) i F_1-mjera, odnosno mikro-uprosječene (engl. *micro-averaged*) i makro-uprosječene (engl. *macro-averaged*) vrijednosti tih mjera. Pregled mjera u nastavku temelji se na inačicama spomenutih mjera opisanih u (McCallum et al., 1998), koje su prilagođene za vrednovanje modela za klasifikaciju teksta.

Preciznost (engl. *precision*) je udio točno klasificiranih primjera u skupu pozitivno klasificiranih primjera, dok je odziv (engl. *recall*) udio točno klasificiranih primjera u skupu svih pozitivnih primjera. Primjenom tih definicija na zadatak klasifikacije teksta dobivaju se izrazi 3.4 i 3.5.

$$P = \frac{\textit{broj točno pridjeljenih kategorija}}{\textit{ukupan broj pronađenih kategorija}} = \frac{BT}{BP} \tag{3.4}$$

$$R = \frac{\textit{broj točno pridjeljenih kategorija}}{\textit{broj kategorija koji je trebao biti pridjeljen}} = \frac{BT}{BS} \tag{3.5}$$

Mikro vrijednosti preciznosti, odziva i F_1-mjere dobiveni su prema 3.6 i 3.7.

$$P^{micro} = \frac{\sum_{i=1}^{K} BT_i}{\sum_{i=1}^{K} BP_i} \qquad R^{micro} = \frac{\sum_{i=1}^{K} BT_i}{\sum_{i=1}^{K} BS_i} \tag{3.6}$$

$$F_1^{micro} = \frac{2 \times P^{micro} \times R^{micro}}{P^{micro} + R^{micro}} \tag{3.7}$$

Ako je F_1-mjera klase C_i označena s F_i, tada se F_1-makro dobije prema formuli 3.8.

$$F_1^{macro} = \frac{1}{K} \sum_{i=1}^{K} F_i \tag{3.8}$$

Makro-F_1 sve klase tretira jednako te zbog toga primjeri iz malih klasa imaju veći utjecaj na mjeru nego što bi imali kod mjere mikro-F_1. S druge strane, mikro-F_1 je najvećim dijelom određena brojem ispravno klasificiranih pozitivnih primjera (engl. *true positives*) te zbog toga kod te mjere donimiraju veće klase.

4. Rezultati

4.1. Zbirka tekstova *Vjesnik*

U tablici 4.1 prikazani su najbolje vrednovani modeli za zbirku tekstova *Vjesnik* uz vektor dokumenta dimenzije 500 značajki. Kako se kod zbirke tekstova *Vjesnik* svakom dokumentu pridjeljuje točno jedna kategorija, vrijednosti mikro P, mikro R i mikro F_1 su jednake pa se u tablicama prikazuje samo mjera mikro F_1. U tablici 4.2 prikazani su rezultati vrednovanja klasifikatora uz korištenje metode korisnosti informacije i vektor dokumenta dimenzije 60, odnosno, svi modeli trenirani su s istim podacima, a jedina razlika je klasifikator. RBF SVM pokazao se najboljim klasifikatorom za ovaj problem te najlošije vrednovan model koji koristi taj klasifikator ima bolje rezultate od najboljeg modela koji koristi naivan Bayesov klasifikator.

Tablica 4.1: Najbolje vrednovani modeli za zbirku tekstova *Vjesnik* uz vektor dokumenta dimenzije 500 značajki

klasifikator	metoda odabira značajki	mikro F1	makro F1
RBF SVM	korisnost informacije	71.07	63.94
RBF SVM	frekvencija dokumenata	70.44	63.36
linearni SVM	korisnost informacije	70.09	62.22
RBF SVM	χ^2	69.54	61.41
RBF SVM	frekvencija dokumenata	69.28	61.41
RBF SVM	uzajamna informacija	68.86	61.81
logistička regresija	korisnost informacije	68.24	59.36

Prikaz utjecaja broja značajki na rezultate klasifikacije nalazi se u tablici 4.3. Klasifikacija se vrši linearnim SVM-om uz korištenje metode korisnosti informacije kao metode za odabir značajki. Rezultati pokazuju da veći broj značajki korišten za prikaz dokumenata daje bolje rezultate te da se prelaskom s 2500 značajki na 5000 rezultat klasifikacije neznatno poboljšao.

Tablica 4.2: Rezultati vrednovanja klasifikatora uz korištenje metode korisnosti informacije za odabir značajka i vektor dokumenta dimenzije 60 nad zbirkom tekstova *Vjesnik*

klasifikator	mikro F1	makro F1
RBF SVM	55.05	47.06
logistička regresija	53.39	43.52
linearni SVM	52.64	41.26
Bernoullijev naivan Bayes	48.92	42.58
multinomijalni naivan Bayes	48.49	38.79

Tablica 4.3: Utjecaj broja značajki na rezultate klasifikacije tekstova iz zbirke tekstova *Vjesnik* korištenjem linearnog SVM-a i metode korisnosti informacije za odabir značajki

# značajki	mikro F1	makro F1
20	41.01	29.62
60	52.64	41.26
100	56.51	45.58
250	65.69	55.44
500	70.09	62.23
1000	73.99	65.74
2500	76.15	68.62
5000	76.43	68.9

Primjenom metoda *word2vec* i *doc2vec* umjesto metoda za odabir značajki ostvareni su bolji rezultati koji su prikazani u tablici 4.4. Rezultati su dobiveni uz vektor dokumenta dimenzije 100 značajki, a učenje modela *doc2vec* provedeno je kroz 10 iteracija. Utjecaj dimenzije vektora dokumenta prikazan je u tablici 4.5. Rezultati su dobiveni primjenom logističke regresije i metode *word2vec*. Dimenzija vektora dokumenta nema velik utjecaj na učinkovitost klasifikacije kao kod metoda koje koriste reprezentaciju dokumenta kao vreće riječi te se najbolji rezultat dobiva za vektor dokumenta dimenzije 250 značajki.

Tablica 4.4: Najbolje vrednovani modeli za zbirku tekstova *Vjesnik* uz korištenje metoda *word2vec* i *doc2vec* i vektor dokumenta dimenzije 100 značajki

metoda	klasifikator	mikro F1	makro F1
word2vec	RBF SVM	76.43	67.95
word2vec	linearni SVM	75.92	66.2
doc2vec (dm+dbow)	RBF SVM	75.38	65.26
doc2vec (dbow)	RBF SVM	74.91	66.91
doc2vec (dm+dbow)	Logistička regresija	73.02	63.39

Tablica 4.5: Utjecaj broja značajki na rezultate klasifikacije tekstova iz zbirke tekstova *Vjesnik* korištenjem logističke regresije i metode *word2vec*

# značajki	mikro F1	makro F1
20	63.41	47.73
60	66.79	51.8
100	67.72	52.24
250	**69.01**	**52.92**
500	68.69	52.14
1000	68.2	51.34

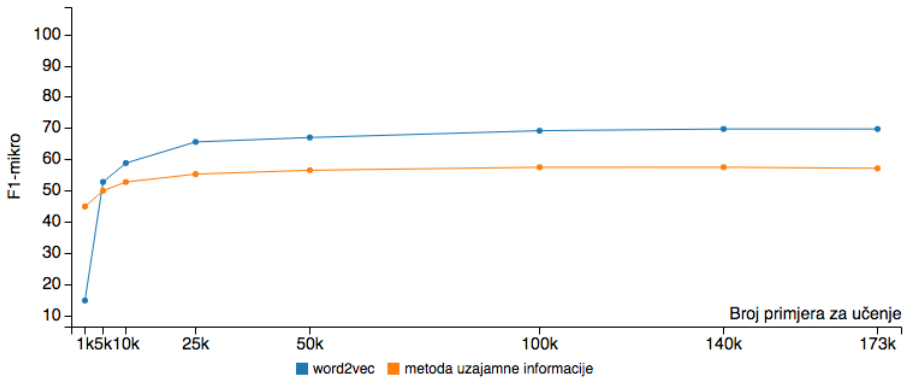
Povezanost duljine dokumenta s učinkovitošću klasifikacije prikazana je u tablici 4.6. Dokumenti su podijeljeni na duge i kratke pri čemu su dugi dokumenti oni koji su dulji od prosječne duljine dokumenta iz promatrane zbirke tekstova, a kratki oni koji su kraći od prosječne duljine dokumenta. Klasifikacijski modeli učeni su i dugim i kratkim dokumentima, a klasifikacija dugih i kratkih dokumenata vršila se odvojeno. Rezultati pokazuju da svi isprobani modeli učinkovitije klasificiraju kraće dokumente, a razlog tome leži u činjenici da dulji dokumenti sadrže više riječi koje bolje opisuju

neku drugu kategoriju, a ne kategoriju u koju dokument mora biti klasificiran pa zbog toga dolazi do pogrešnih klasifikacija.

Tablica 4.6: Utjecaj duljine dokumenta na učinkovitost klasifikacije tekstova iz zbirke tekstova *Vjesnik*

klasifikator	metoda	kratki dokumenti		dugi dokumenti	
		mikro F1	makro F1	mikro F1	makro F1
logistička reg.	korisnost info.	66.64	49.54	62.21	54.43
logistička reg.	word2vec	75.49	51.61	67.65	52.49
logistička reg.	doc2vec	78.17	54.35	72	62.53

Na slici 4.1 nalaze se krivulje učenja modela koji koristi metodu korisnosti informacije za odabir značajki te modela *word2vec* korištenjem logističke regresije za klasifikaciju s vektorom dokumenta dimenzije 100 značajki. Vidljivo je da je za obje metode dovoljno 100K primjera za učenje kako bi klasifikacija bila najučinkovitija.



Slika 4.1: Krivulje učenja modela *word2vec* i modela koji koristi metodu korisnosti informacije za odabir značajki s tekstovima iz *Vjesnik* zbirke tekstova uz vektor dokumenta dimenzije 100 značajki

4.2. Zbirka tekstova *NN13205*

Posebnost zbirke tekstova *NN13205* je velik broj kategorija koje imaju mali broj primjera, a kako se odabir primjera za učenje vrši slučajnim odabir, postoji mogućnost da su za neku kategoriju svi dokumenti iskorišteni za učenje, a jednako je tako moguće da binarni klasifikator za neku kategoriju uopće nije kreiran jer nema niti jednog dokumenta za učenje. Najbolje vrednovani modeli uz korištenje vektora dokumenta dimenzije 500 značajki prikazani su u tablici 4.7. Linearni SVM pokazao se najboljim klasifikatorom za ovu zbirku tekstova za razliku od RBF SVM-a koji ima znatno višu preciznost, ali vrlo loš odziv.

Tablica 4.7: Najbolje vrednovani modeli za zbirku tekstova *NN13205* uz vektor dokumenta dimenzije 500 značajki

klasifikator	odabir značajki	P mikro	R mikro	mikro F1	makro F1
linearni SVM	korisnost info.	63.26	43.87	51.81	46.79
linearni SVM	uzajamna info. (avg)	53.77	41.11	46.6	42.26
logistička reg.	uzajamna info. (avg)	50.27	39.29	44.11	40.04

Utjecaj dimenzije vektora dokumenta na učinkovitost klasifikacije prikazan je u tablici 4.8. Prema očekivanju, učinkovitost klasifikacije raste s porastom veličine vektora dokumenta.

Tablica 4.8: Utjecaj broja značajki na rezultate klasifikacije tekstova iz zbirke tekstova *NN13205* korištenjem linearnog SVM-a i metode korisnosti informacije za odabir značajki

# značajki	P mikro	R mikro	mikro F1	makro F1
20	77.23	4.88	9.19	5.82
60	81.49	13.97	23.85	16.47
100	71.51	22.93	34.73	26.74
250	55.14	39.40	45.96	41.39
500	63.26	43.87	51.81	46.79

Rezultati dobiveni primjenom metoda *word2vec* i *doc2vec* prikazani su u tablici 4.9 te su ti rezultati lošiji od rezultata dobivenih metodama koje koriste reprezentacija dokumenta kao vreće riječi.

Kao i kod zbirke tekstova *Vjesnik*, dokumenti su podijeljeni na duge i kratke te se klasifikacija dugih i kratkih dokumenata vršila odvojeno. Svi isprobani modeli učinkovitije su klasificirali kratke dokumente, što je prikazano u tablici 4.10.

Tablica 4.9: Rezultati klasifikacije dokumenata iz zbirke tekstova *NN13205* uz korištenje metoda *word2vec* i *doc2vec* i vektor dokumenta dimenzije 100 značajki

metoda	klasifikator	P mikro	R mikro	mikro F1	makro F1
doc2vec (dm + dbow)	linearni SVM	40.55	36.7	38.53	35.35
word2vec	linearni SVM	90.72	23.78	37.68	31.26
doc2vec (dm + dbow)	logistička reg.	36.18	36.17	36.18	33.35

Tablica 4.10: Utjecaj duljine dokumenta na učinkovitost klasifikacije tekstova iz zbirke tekstova *NN13205*

klasifikator	metoda	kratki dokumenti		dugi dokumenti	
		mikro F1	makro F1	mikro F1	makro F1
logistička reg.	korisnost info.	32.38	25.15	28.88	21.83
logistička reg.	word2vec	25.46	17.87	11.37	9.85
logistička reg.	doc2vec	38.89	33.45	32.78	31.41

4.3. Zbirka tekstova *Reuters RCV1*

Učinkovitost klasifikacije tekstova iz zbirke tekstova *Reuters RCV1* prikazana je u tablici 4.11. Rezultati su dobiveni korištenjem vektora dokumenta dimenzije 100 značajki. RBF SVM klasifikator nije isproban s ovom zbirkom tekstova zbog velikog broja kategorija i dokumenata za učenje dok logistička regresija i linearni SVM daju podjednako dobre rezultate.

Tablica 4.11: Najbolje vrednovani modeli za zbirku tekstova *Reuters RCV1* uz vektor dokumenta dimenzije 100 značajki

klasifikator	odabir značajki	P mikro	R mikro	mikro F1	makro F1
logistička reg.	uzajamna info.(max)	80.82	51.36	62.8	59.73
logistička reg.	korisnost info.	80.7	50.37	62.02	59.08
linearni SVM	uzajamna info.(max)	84.68	47.74	61.06	57.9
linearni SVM	korisnost info.	83.85	47.52	60.66	57.84
logistička reg.	χ^2 (avg)	80.35	47.69	59.86	56.89
multi. Bayes	χ^2 (max)	37.71	55.23	44.82	41.40

Ovisnost učinkovitosti klasifikacije o dimenziji vektora dokumenta prikazana je u tablici 4.12. Rezultati su dobiveni korštenjem logističke regresije kao klasifikatora i metode uzajamne informacije za odabir značajki. Za razliku od preciznosti, odziv vrlo

ovisi o dimenziji vektora dokumenta te je dvostruko veći s vektorom dokumenta dimenzije 250 nego s vektorom dokumenta dimezije 20 značajki. Uzrok tome je što s malim brojem značajki klasifikatori nemaju dovoljno informacija za pozitivnu klasifikaciju, pa je ukupan broj pronađenih kategorija malen što je dovodi do manjeg odziva.

Tablica 4.12: Utjecaj dimenzije vektora dokumenta na rezultate klasifikacije tekstova iz zbirke tekstova *Reuters RCV1* korištenjem logističke regresije i metode uzajamne informacije za odabir značajki

# značajki	Pmikro	R mikro	mikro F1	makro F1
20	79.49	29.38	42.9	39.38
60	79.95	43.75	56.56	53.42
100	80.82	52.36	62.8	59.73
250	83.3	61.71	70.9	69.14

Primjenom metoda temeljenih na neuronskim mrežama s tekstovima iz zbirke tekstova *Reuters RCV1* ostvareni su bolji rezultati od metoda koje koriste reprezentaciju dokumenta kao vreće riječi, što je prikazano u tablici 4.13. Modeli su učeni sa skupom od 100K slučajno odabranih dokumenta iz zbirke tekstova te se metoda *word2vec* pokazala boljom u odnosu na metodu *doc2vec*.

Tablica 4.13: Rezultati klasifikacije dokumenata iz zbirke tekstova *Reuters RCV1* uz korištenje metoda *word2vec* i *doc2vec* i vektor dokumenta dimenzije 100 značajki

metoda	klasifikator	P mikro	R mikro	mikro F1	makro F1
word2vec	linearni SVM	88.34	63.74	74.05	73.18
doc2vec (dm + dbow)	linearni SVM	79.63	54.59	64.77	62.73
doc2vec (dm + dbow)	logistička reg.	78.61	52.89	63.24	61.15

S povećanjem dimenzije vektora dokumenta raste i učinkovitost klasifikacije modela *word2vec*, ali ne u mjeri kao što je to kod metoda koje koriste reprezentaciju dokumenta kao vreće riječi. Kod modela *doc2vec* učinkovitost klasifikacije raste do dimenzije vektora dokumenta 100 značajki nakog čega počinje padati. To je prikazano tablicom 4.14.

Google je omogućio preuzimanje modela *word2vec* (*Google word2vec model*) koji je učen sa skupom tekstova iz zbirke tekstova *Google News*, odnosno sa oko 100M riječi. Model *Google word2vec* prepoznaje oko 3M različitih riječi i fraza koje prikazuje vektorom dimenzije 300 značajki.

Tablica 4.14: Utjecaj dimenzije vektora dokumenta na učinkovitost klasifikacije tekstova iz *Reuters RCV1* zbirke tekstova

# značajki	metoda word2vec		metoda doc2vec	
	mikro F1	makro F1	mikro F1	makro F1
20	61.14	60.14	59.52	58
60	70.65	69.67	63.3	61.41
100	74.05	73.18	63.24	61.15
300	78.59	78.05	61.26	58.97
500	79.72	79.31	59.03	56.87

Tablica 4.15: Usporedba modela *word2vec* učenog s tekstovima iz *Reuters RCV1* zbirke tekstova i modela učenog s tekstovima iz zbirke tekstova *Google News*

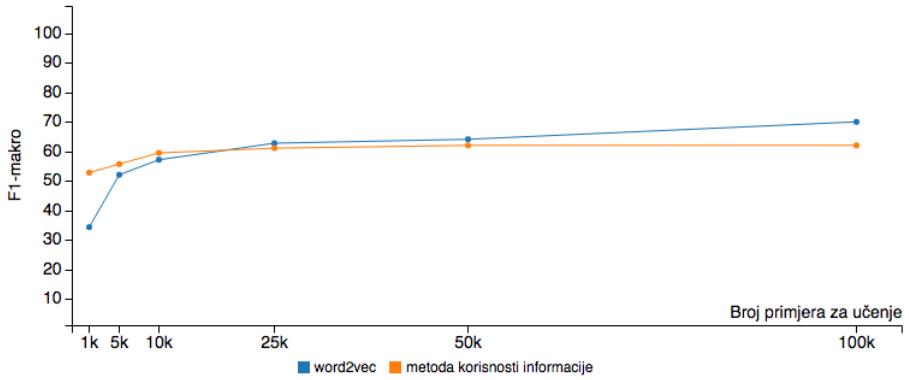
zbirka tekstova	klasifikator	P mikro	R mikro	mikro F1	makro F1
Reuters RCV1	linearni SVM	89.59	70	78.59	78.05
Google News	linearni SVM	90.01	63.85	74.71	73.47
Reuters RCV1	logistička regresija	89.96	57.79	70.38	69.32
Google News	logistička regresija	90.66	40.67	56.15	54.66

U tablici 4.15 dana je usporedba modela učenog s tekstovima iz *Reuters RCV1* zbirke tekstova i modela *Google word2vec*. Za klasifikaciju su korišteni linearni SVM i logistička regresija s dimenzijom vektora dokumenta 300 značajki. S modelom *Google word2vec* ostvareni su vrlo dobri rezultati, a to dovodi do zaključka da modeli *word2vec* mogu biti učeni na jednom skupu podataka te nakon toga uspješno primjenjeni na tekstove iz nekog drugog tematskog područja.

U tablici 4.16 prikazan je utjecaj duljine dokumenata na učinkovitost klasifikacije. Sve metode osim *doc2vec* daju bolje rezultate klasificirajući kratke dokumente.

Tablica 4.16: Utjecaj duljine dokumenta na učinkovitost klasifikacije dokumenata iz zbirke tekstova *Reuters RCV1*

klasifikator	metoda	kratki dokumenti		dugi dokumenti	
		mikro F1	makro F1	mikro F1	makro F1
logistička reg.	uzajamna info.	64.38	73.47	60.16	59.67
logistička reg.	word2vec	77.73	75.69	65.23	65.12
logistička reg.	doc2vec	63.14	60.05	64.96	64



Slika 4.2: Krivulje učenja modela koji koristi metodu korisnosti informacije za odabir značajki te modela *word2vec* s tekstovima iz zbirke tekstova *Reuters RCV1* uz vektor dokumenta dimenzije 100 značajki

Na slici 4.2 nalaze se krivulje učenja za model koji koristi metodu korisnosti informacije za odabir značajki te model *word2vec* uz korištenje logističke regresije kao klasifikatora s vektorom dokumenta dimenzije 100 značajki. Vidi se da je modelu koji koristi metodu korisnosti informacije dovoljno 50K primjera za učenje kako bi dosegnuo najbolje rezultate dok model *word2vec* najbolje rezultate daje kad je naučen sa 100K primjera.

4.4. Diskusija rezultata

Najbolji rezultati za zbirku tekstova *Vjesnik* ostvareni su modelom *word2vec* uz vektor dimenzije 100 značajki i korištenjem RBF SVM klasifikatora, čime su se dobili F_1 mikro 76.43% te F_1 makro 67.95%. Najbolji rezultati za zbirku tekstova *NN13205* dobili su se korištenjem linearnog SVM-a i metode korisnosti informacije za odabir značajki uz vektor dokumenta dimenzije 500 značajki. Taj model ostvario je preciznost od 63.26% s odzivom od 43.87% te mjerama F_1 mikro 51.81% i F_1 makro 46.70%. Najbolji rezultati za zbirku tekstova *Reuters RCV1* ostvareni su modelom *word2vec* korištenjem linearnog SVM-a uz dimenziju vektora dokumenta 300 značajki. Tim modelom ostvarena je preciznost od 89.59% i odziv od 70% što daje F_1 mikro 78.59% i F_1 makro 78.05%. Opisani rezultati prikazani su u tablici 4.17.

Tablica 4.17: Prikaz najboljih rezultata za pojedinu zbirku tekstova

zbirka tekstova	klasifikator	dimenzija vektora dokumenta	odabir značajki	mikro F1	makro F1
Vjesnik	RBF SVM	100	korisnost info.	71.07	63.94
Vjesnik	RBF SVM	5000	korisnost info.	76.43	68.9
Vjesnik	RBF SVM	100	word2vec	76.43	67.95
Vjesnik	RBF SVM	100	doc2vec	75.38	65.26
NN13205	lin. SVM	500	korisnost info.	51.81	46.79
NN13205	lin. SVM	100	doc2vec	38.53	35.35
NN13205	lin. SVM	100	word2vec	37.68	31.26
Reuters	log. reg.	100	uzajamna info.	62.8	59.73
Reuters	log. reg.	250	uzajamna info.	70.9	69.14
Reuters	lin. SVM	100	word2vec	74.05	73.18
Reuters	lin. SVM	300	word2vec kratki dok.	79.72	79.31
Reuters	lin. SVM	100	doc2vec	64.77	62.73

5. Zaključak

Primjena klasifikacije teksta raste s porastom digitalno dostupnih informacija, a time se javlja i potreba za metodama koje mogu obraditi veće količine podataka. U ovom radu isprobane su različite metode odabira značajki i prikaza dokumenta sa svrhom dobivanja što boljeg uvida o utjecaju tih parametara na učinkovitost klasifikacije.

Kod metoda koje koriste reprezentaciju dokumenta kao vreće riječi, problem velikog prostora značajki uspješno se rješava odabirom najznačajnijih značajki, ali se time gubi poredak i kontekst riječi. Najučinkovitija metoda odabira značajki pokazala se metoda korisnosti informacije, a dobri rezultati dobili su se i metodom uzajamne informacije i metodom χ^2. Metode temeljene na neuronskim mrežama kod stvaranja vektora dokumenta u obzir uzimaju cijeli dokument te gledaju okolinu pojedine riječi čime uzimaju u obzir kontekst u kojem se riječ pojavljuje. Te metode daju bolje rezultate od metoda koje koriste reprezentaciju dokumenta kao vreće riječi te rezultati tih metoda nisu toliko ovisni o dimenziji vektora dokumenta. Neovisno o korištenoj metodi, za dobre rezultate potrebna je zbirka tekstova sastavljena od ujednačenih klasa i dovoljno primjera za učenje. Oba pristupa učinkovitije klasificiraju kraće tekstove zbog manjeg broja riječi koje su specifičnije za neku drugu klasu. Sa stajališta klasifikatora, naivan Bayesov klasifikator ostvario je najlošije rezultate dok su linearni SVM i logistička regresija za sve isprobane metode ostvarili dobre rezultate. Nelinearni SVM s RBF jezgrenom funkcijom ostvario je bolje rezultate kod klasifikacije s jednom oznakom, dok je kod klasifikacije s više oznaka odziv modela s tim klasifikatorom bio jako nizak. Uzrok tome može biti premalen broj primjera za učenje i neuravnoteženost klasa u korištenoj zbirci tekstova.

U budućem radu može se provjeriti utjecaj zaustavnih riječi na učinkovitost klasifikacije na način da se zaustavne riječi izbace iz tekstova prije učenja modela. Također, model *word2vec* mogao bi se naučiti s velikim skupom tekstova na hrvatskom jeziku iz različitih tematskih područja te se nakon toga taj model može primijeniti na zbirke tekstova koje imaju mali broj dokumenata, kao što je zbirka tekstova *NN13205*.

LITERATURA

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, i Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Everton Alvares Cherman, Maria Carolina Monard, i Jean Metz. Multi-label problem transformation methods: a case study. *CLEI ELECTRONIC JOURNAL*, 14(1), 2011.

Isabelle Guyon i André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

Dalbelo Bašić Jan Šnajder. *Strojno učenje*. verzija 3.6 (2013-02-04) izdanju, 2012.

Quoc V Le i Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

David D Lewis, Yiming Yang, Tony G Rose, i Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, svezak 1. Cambridge university press Cambridge, 2008.

Andrew McCallum, Kamal Nigam, et al. A comparison of event models for naive bayes text classification. U *AAAI-98 workshop on learning for text categorization*, svezak 752, stranice 41–48. Citeseer, 1998.

Tomáš Mikolov, Jiří Kopeckỳ, Lukáš Burget, Ondřej Glembek, i Jan Honza Černockỳ. Neural network based language models for highly inflective languages. U *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, stranice 4725–4728. IEEE, 2009.

Tomas Mikolov, Kai Chen, Greg Corrado, i Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Frane Šaric, Bojana Dalbelo Bašic, Marie-Francine Moens, i Jan Šnajder. Multi-label classification of croatian legal documents using eurovoc thesaurus. U *Semantic Processing of Legal Texts (SPLeT-2014) Workshop Programme*, stranica 7, 2014.

Fabrizio Sebastiani. Text categorization., 2005.

Yiming Yang i Jan O Pedersen. A comparative study on feature selection in text categorization. U *ICML*, svezak 97, stranice 412–420, 1997.

Postupci odabira značajki i prikaza dokumenta za klasifikaciju teksta

Sažetak

S porastom količine digitalnih informacija raste i potreba za metodama klasifikacije teksta koje su učinkovitinije i u mogućnosti obraditi velike količine podataka. Iz tog razloga pozornost se sve više okreće metodama temeljenim na strojnom učenju. U ovom radu isprobane su metode koje koriste reprezentaciju dokumenta kao vreće riječi koje u prvom koraku rade odabir najznačajnijih značajki. Uz to, isprobane su i metode temeljene na neuronskim mrežama koje za pojedini dokument grade njegovu vektorsku reprezentaciju. Korištenjem tih metoda ostvarili su se bolji rezultati nego korištenjem metoda koje koriste reprezentaciju dokumenta kao vreće riječi. Svi eksperimenti provedeni su nad različitim zbirkama tekstova na hrvatskom i engleskom jeziku.

Ključne riječi: strojno učenje, obrada prirodnog jezika, klasifikacija teksta, odabir značajki, word2vec, doc2vec, reprezentacija dokumenta

Feature selection and document representation methods for text classification

Abstract

With growing amount of online information, there is a growing need for text classification methods that are more efficient and capable to process large amount of data. Therefore, there is increased attention to the methods based on machine learning. This paper experiments with methods that use bag-of-words document representation and feature selection methods. In addition, this paper experiments with a neural network based methods. These methods build vector representation of each document and the results achieved with these methods are better than the results achieved using methods that use bag-of-words document representation. All of the experiments are performed over a few different document collections in Croatian and English.

Keywords: machine learning, natural language processing, text classification, feature selection, word2vec, doc2vec, document representation