

Razrješavanje višeznačnosti riječi

Anonimni autori

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

Sažetak

Ovaj rad razmatra uporabu metoda strojnog učenja u razrješavanju višeznačnosti riječi (engl. *word sense disambiguation*). Ispituje se uspješnost SVM-a i ansambla naivnih Bayesovih klasifikatora (ANBK) na jednom uravnoteženom skupu uzoraka i na jednom neuravnoteženom skupu uzorka, te se ocjenjuje uspješnost tih postupaka na takvim skupovima. Nenormalizirane riječi teksta (u originalnom leksičkom obliku) koriste se kao značajke. Rezultati pokazuju kako SVM ima visoku uspješnost točne klasifikacije, dok ANBK ima lošu uspješnost na neuravnoteženom skupu. U radu se provodi postupak pronalaska dobrih značajki za klasifikaciju, tj. nastoji se ustanoviti koliko riječi, koje se nalaze u neposrednoj blizini višeznačne riječi u samom tekstu, uzimati prilikom gradnje klasifikatora i klasifikacije. Rezultati pokazuju kako se izlučivanjem 15-ak riječi oko višeznačne riječi postiže dobra točnost klasifikacije.

1. Uvod

Riječ kao sastavnica svakog jezika, opisana je leksičkim izrazom i semantičkim značenjem. Jedan leksički izraz može imati više semantičkih značenja, pri čemu okolni kontekst sugerira semantičko značenje leksičkog izraza. Primjerice leksički izraz *jaguar* ima više semantičkih značenja (životinja iz porodice mačaka, britanski proizvođač automobila, britanska glazbena grupa, parfem), gdje se semantičko značenje riječi *jaguar* doznaje iz konteksta. Dakle svaka riječ u pisanom tekstu jedinstveno je označena leksičkim izrazom i semantičkim značenjem. Upravo iz razloga što odnos leksičkog izraza i semantičkog značenja nije injektivan, automatizirani postupci kao što je npr. pretraživanje teksta ili postavljanje upita nad bazom podataka i sl. vrlo često korisnicima daju neželjene rezultate. Danas lingvistička znanost i računarska znanost nastoje pronaći tehnike koje bi dovoljno dobro rješavale navedeni problem.

Tehnike strojnog učenja pružaju dovoljno dobru podlogu za realizaciju sustava koji bi bio sposoban razriješiti višeznačnosti riječi, tj. identificirati pravo značenje riječi ovisno o okolnom kontekstu. Stoga su u ovom radu istraženi postupci strojnog učenja za razrješavanje višeznačnosti riječi, a korišteni postupci su: stroj s potpornim vektorima (SVM) i ansambl naivnih Bayesovih klasifikatora (ANBK).

Slijedeći problem koji se pojavljuje je kako izlučiti značajke iz konteksta. U većini slučajeva za značajke se uzima n riječi koje se nalaze u neposrednoj blizini oko višeznačne riječi, gdje je $n \in N$ neka fiksna konstanta. Pogrešno odabran n može pogoršati uspješnost klasifikacije, također intuitivno se može naslutiti kako nije svejedno koliko od tih n riječi se pojavljuje u tekstu s lijeve strane višeznačne riječi, a koliko pak s desne strane. Upravo ovaj rad nastoji ustanoviti te zavisnosti, te zajedno s priloženim rezultatima provedenih eksperimenata iznosi zaključke kako odabrati kvalitetne značajke.

U idućem poglavlju dan je pregled srodnih radova. Treće poglavlje opisuje bazu podataka, ekstrakciju značajki te reprezentiranje uzoraka. Četvrto poglavlje daje uvid u provođenje eksperimenta, s kratkim opisom korištenih postupaka strojnog učenja zajedno s objašnjenjem postupka

provođenja validacije i testiranja. U petom poglavlju izneseni su rezultati eksperimenata te se raspravlja o dobivenim rezultatima. U zadnjem poglavlju izneseni su zaključci i smjernice za daljnji rad.

2. Srodni radovi

Čitajući dostupne radove privukle su nas metode koje pokušavaju pristupiti tom problemu nama poznatim metodama, konkretnije pomoću naivnog Bayesovog klasifikatora i SVM-a. Pristup pomoću stroja s potpornim vektorima pokazao se dosta uspješan, ispravno klasificirajući značenje riječi u skoro 80% slučajeva, ali kako je opisano u (Lee et al., 2004), da bi se takav rezultat postigao potrebno je imati dodatne informacije o gramatičkim svojstvima teksta. Drugi pristup pomoću SVM-a je korištenje značenja riječi iz leksičke baze WordNet (Buscaldi et al., 2006) te usporedba riječi, koje opisuju značenje višeznačne riječi u rječniku, s riječima iz konteksta u kojem se ta riječ pojavljuje. Nažalost takav pristup nije se pokazao posebno uspješnim. Prema radu (Pedersen, 2000), naivni Bayesov klasifikator pokazao se vrlo robusnim rješenjem za ovaj problem. Njegova ideja temelji se na izradi više klasifikatora s različitim širinama lijevog i desnog konteksta, koji bi se koristili za klasifikaciju po sistemu nadglašavanja. Točnost klasifikacije ovakvog pristupa vrlo je bliska postotku od 80%.

3. Opis uzoraka

Označavanje vlastite baze podataka s višeznačnim riječima izuzetno je skup proces. Naime za relevantnu klasifikaciju potrebno je barem oko tisuću uzoraka za jednu višeznačnu riječ, pri tom označivač za svaki uzorak mora pročitati okolni kontekst i shvatiti semantičko značenje riječi te pridodati uzorku oznaku razreda, što predstavlja dugotrajan proces. Stoga je pribavljena besplatna baza podataka¹ s već označenim semantičkim značenjima za engleske riječi *interest* i *line*. Shodno tome u ovom radu bit će prikazani postupci razrješavanja višeznačnosti engleskih riječi *interest* i *line*. Tekstovi u bazi podataka su na engleskom jeziku i prikupljeni su iz *ACL/DCI Wall Street Journal* novina.

¹<http://www.senseval.org/data.html>

Kao značajke uzorka uzimaju se riječi koje se nalaze unutar prozora konteksta promatranog uzorka. Prozor konteksta (l, r) (engl. *window of context*) sastoji se od l riječi koji se nalaze lijevo od višeznačne riječi (lijevi prozor konteksta) i r riječi koji se nalaze desno od višeznačne riječi (desni prozor konteksta), pri čemu se interpunkcijski i pravopisni znakovi ignoriraju. Sve riječi koje su uključene u prozor konteksta ostavljene su originalnoj leksičkoj formi, tj. nisu normalizirane, osim što su sva velika slova pretvorena u mala slova. Ova metoda reprezentiranja skupa značajki, opisana u (Pedersen, 2000), varijanta je metode “vreća riječi” (engl. *bag-of-words*) prvi put opisane u (Gale et al., 1992), a razlikuje se po tome što razlikuje riječi koje se nalazi lijevo i desno od višeznačne riječi. Slika 1 opisuje izlučivanje značajki s definiranim prozorom konteksta (5,3). Naime sve značajke (riječi) unutar prozora konteksta (5,3) imat će vrijednost jedan, dok će ostale poprimiti vrijednost nula.

Forward supplies can largely be stored in these same areas, and land forces are best held in reserve on our own soil. Drawing a line between military aid and military involvement may be harder, but it can be done if we keep the distinction clearly in mind.

Slika 1: Primjer uzorka višeznačne riječi *line* s prozorom konteksta (5, 3).

Baza podataka sadrži 2368 primjera za riječ *interest* i 4148 primjera za riječ *line*, gdje je svaki primjer zapravo tekst u kojem se nalazi višeznačna riječ. Taj tekst smatramo kontekstom višeznačne riječi. Također svaki primjer, uz tekst, sadrži i atribut koji definira semantičko značenje višeznačne riječi u primjeru.

Riječ *interest* ima šest različitih semantičkih značenja (vidi tablicu 1), baš kao i riječ *line*, međutim za riječ *line* izdvojeni su samo oni primjeri kod kojih višeznačna riječ ima jedno od semantičkih značenja opisanih u tablici 2. Naime distribucija primjera, po semantičkim značenjima višeznačne riječi, u originalnoj bazi podataka za *interest* i *line* je vrlo neujednačen, stoga je stvoren umjetno uravnotežen skup primjera za riječ *line* kako bi mogli istražiti koliko koja metoda strojnog učenja kvalitetno klasificira na neuravnoteženom skupu podataka, a koliko na uravnoteženom skupu podataka.

Kako je cilj ovog rada ustanoviti koji prozor konteksta (l, r) odabrati ovisno o algoritmu strojnog učenja (razmatrani su SVM i ANBK) kako bi klasifikacija bila što uspješnija, stvoreno je 81 skupova uzoraka. Svaki skup uzoraka dobiven je drugačijim izlučivanjem značajki iz skupa podataka, pri čemu se skupovi uzoraka razlikuju po dimenzijama prozora konteksta (l, r), odnosno po kombinaciji koliko je riječi izlučeno s lijeve strane višeznačne riječi (varijabla l u definiciji prozora konteksta), a koliko s desne strane višeznačne riječi (varijabla r u definiciji prozora konteksta). Veličina lijevog prozora konteksta i desnog prozora konteksta, tj. varijable l i r poprimaju vrijednosti iz skupa

Tablica 1: Distribucija primjera po semantičkim značenjima za riječ *interest*

	Semantičko značenje	Broj primjera
I_1	interes	361
I_2	prouzročiti zainteresiranost drugih	11
I_3	pokazati zainteresiranost	66
I_4	prednost ili korist	178
I_5	udjel dionica u tvrtci	500
I_6	kamate	1252

Tablica 2: Distribucija primjera po semantičkim značenjima za riječ *line*

	Semantičko značenje	Broj primjera
L_1	tanak oblik, crta	373
L_2	umjetna podjela, granica	374
L_3	formacija ljudi ili stvari	349

$\{0, 1, 2, 3, 4, 5, 10, 25, 50\}$. Ne postoji poseban razlog zašto su odabrane baš te vrijednosti, koje mogu poprimiti varijable l i r , već se slijedila preporuka iz rada (Pedersen, 2000), gdje su korištene iste vrijednosti. Nad svakim od tih skupova uzoraka provedeni su algoritmi SVM i naivni Bayesov (NB) klasifikator, te su izračunate F1 mjere za svaku kombinaciju *skup uzoraka - algoritam*. Tablica 3 i tablica 4 prikazuju ukupan broj izlučenih značajki ovisno o kombinaciju (l, r), tj. ovisno o skupu uzoraka. Primjećuje se kako skupovi uzoraka koji su dobiveni izlučivanjem s većim prozorom konteksta imaju više značajki, što je i očekivano. Naime prozor konteksta s većim dimenzijama pohvatat će više različitih riječi pa će samim time i ukupan broj značajki biti veći.

4. Provođenje eksperimenata

4.1. Ansambl naivnih Bayesovih klasifikatora (ANBK)

ANBK funkcionira po principu nadglasavanja, u našem radu ANBK sadrži devet NB klasifikatora. Predočavanjem uzorka, svaki klasifikator glasa za jedno značenje višeznačne riječi, ovisno o “maximum a posteriori” (MAP) vjerojatnosti. Uzorak se zatim klasificira u razred (semantičko značenje) s najviše glasova.

Odluka koji će NB klasifikatori ući u ANBK donosi se ovisno o rezultatima pojedinačnih NB klasifikatora na skupu za validaciju. Za svaki prozor konteksta odgovoran je jedan NB klasifikator, što čini ukupno 81 NB klasifikatora. Nakon provedene validacije odabire se najuspješniji NB klasifikatora iz svake kategorije, koji zatim ulazi u ANBK. Ako više NB klasifikatora iz iste kategorije ima jednaku F1 mjeru, tada se odabire onaj kod kojeg je zbroj dimenzija prozora konteksta ($l + r$) manji. Razlikujemo devet kategorija ovisno u veličini prozora konteksta. Varijable l ili r mogu poprimiti vrijednosti iz raspona $\{0, 1, 2, 3, 4, 5, 10, 25, 50\}$, pri čemu se razlikuje uski raspon $\{0, 1, 2\}$,

Tablica 3: Broj značajki za riječ *interest* ovisno o veličini prozora konteksta (l, r)

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	1	219	845	1489	1918	2269	3387	4357	4524
1	495	656	1165	1726	2118	2445	3509	4451	4617
2	1034	1151	1580	2071	2423	2719	3720	4624	4785
3	1515	1604	1980	2415	2716	2988	3915	4781	4940
4	1944	2017	2345	2733	2996	3245	4125	4953	5109
5	2301	2364	2662	3015	3261	3494	4332	5129	5282
10	3546	3587	3817	4075	4254	4442	5121	5813	5955
25	4906	4933	5111	5313	5455	5604	6168	6763	6885
50	5137	5162	5335	5530	5665	5810	6361	6938	7060

Tablica 4: Broj značajki za riječ *line* ovisno o veličini prozora konteksta (l, r)

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	1	224	663	1020	1299	1543	2433	3395	3647
1	235	422	825	1159	1427	1657	2531	3483	3731
2	554	691	1053	1363	1617	1836	2679	3609	3849
3	951	1058	1377	1657	1898	2109	2907	3803	4037
4	1326	1416	1703	1959	2186	2384	3143	4002	4231
5	1683	1766	2030	2267	2478	2659	3378	4198	4412
10	3185	3244	3447	3620	3783	3927	4494	5166	5346
25	5789	5828	5970	6094	6203	6296	6721	7205	7359
50	7046	7082	7205	7315	7408	7491	7854	8271	8406

srednji $\{3, 4, 5\}$ i široki $\{10, 25, 50\}$. Svaka kombinacija navedenih raspona čini jednu kategoriju. Primjerice naivni Bayesov (NB) klasifikator s prozorom konteksta $(2, 3)$ pripada kategoriji (*uski, srednji*).

Nakon što se iz svake kategorije odabere NB klasifikator koji ima najveću točnost, pristupa se postupku testiranja ansambla na skupu za testiranje.

4.2. Stroj s potpornim vektorima (SVM)

U ovom radu koristi se LibSVM implementacija SVM-a, odnosno C-SVM s linearnom jezgrenom funkcijom. C-SVM je verzija SVM-a s mekim marginama i ima jedan parametar (cijena koštanja C). Takav SVM omogućuje provlačenje hiperravnine čak i ako problem nije linearno odvojiv, u takvim slučajevima parametar C određuje koliko će se kazniti svaki uzorak koji se nalazi s pogrešne strane hiperravnine.

Vrijednost parametra C određuje se validacijom, baš kao i dimenzije prozora konteksta. Postupkom *pretraživanja mreže* (engl. *grid search*) pronalazi se optimalan prozor konteksta (raspon mogućih prozora konteksta isti je kao i za ANBK). Za svaki prozor konteksta računa se optimalna vrijednost parametra C i to postupkom *pretraživanja uzorka* (engl. *pattern search*), gdje početna vrijednost parametra C iznosi 2.5, a početni korak 2 (korak se dijeli s dva sve dok nije manji od 0.1). Testiranje se zatim provodi na SVM-u s onim parametrima (C , prozor kontekst) koji na skupu za validaciju imaju najmanju pogrešku klasifikacije.

4.3. Validacija i testiranje

Validacija i testiranje modela provedeno je unakrsnom provjerom. Poredak uzoraka u skupu uzoraka nasumično je ispremiššan prije unakrsne provjere. Skup uzoraka zatim je podijeljen na pet podskupova, četiri podskupa služe za učenje modela, dok se posljednji podskup dijeli popola na još dva podskupa, od kojih jedan služi za validaciju, a drugi za testiranje. Prije podjele popola, podskup je nasumično ispremiššan s namjerom sprječavanja nepravilne distribucije uzoraka u skupu za validaciju ili skupu za testiranje. Validacija se vrši na dobivenom skupu za validaciju, dok skup za testiranje ostaje neaktivan. Nakon završene validacije slijedi slijedeća iteracija unakrsne provjere, tj. podskup za validaciju i testiranje ubacuje se u podskupove za učenje, a jedan od 4 prijašnja podskupa za učenje postaje skup za validaciju i testiranje. Takvih iterativnih koraka ima ukupno pet. Valja naglasiti kako se testiranje ne provodi odmah nakon validacije u svakom iterativnom koraku unakrsne provjere, već se testiranje provodi nakon što je validacija provedena nad svakim od pet mogućih skupova za validaciju, nakon čega se računaju optimalni parametri modela, te se tek nakon toga vrši testiranje.

Program 1: Pseudokôd implementirane unakrsne provjere

```

1 unakrsnaProvjera(skup uzoraka, algoritam)
2 {
3   lista_skupova = podijeli skup uzoraka na
                     5 jednakih dijelova
4   za svaki podskup i iz lista_skupova{

```

```

5  skup_za_ucenje = 4 podskupa iz
    lista_skupova među kojima nije i;
6  promijesaj i;
7  podijeli i na dva dijela
8  skup_za_validaciju = prva polovica
    podijeljenog skupa i
9
10 nauci_model(skup_za_ucenje, algoritam)
11 validacija(skup_za_validaciju, algoritam
    )
12 }
13
14 izracunaj parametre za algoritam
15
16 za svaki podskup i iz lista_skupova{
17  skup_za_ucenje = 4 podskupa iz
    lista_skupova među kojima nije i;
18  podijeli i na dva dijela
19  skup_za_testiranje = druga polovica
    podijeljenog skupa i
20
21  nauci_model(skup_za_ucenje, algoritam)
22  testiraj(skup_za_testiranje, algoritam)
23 }
24 }

```

Testiranje se provodi baš kao i validacija, tj. unakrsnom provjerom kroz pet koraka, ali je sada skup za validaciju neaktivan, a skup za testiranje aktivan. Ovakav postupak učenja, validacije i testiranja modela odabran je s razlogom, naime želi se iskoristi što veći skup uzoraka, i za validaciju, i za testiranje. Naime ako bi se testiranje vršilo odmah nakon validacije u svakom iterativnom koraku unakrsne provjere, prilikom testiranja dobiveni rezultati ne bi bili potpuno objektivni, jer bi rezultati testiranja u tom koraku ovisili o parametrima izračunatim na trenutačnom skupu za validaciju.

5. Rezultati

5.1. Klasifikacija pomoću ansambla naivnih Bayesovih klasifikatora

5.1.1. Rezultati za riječ *interest*

Nakon provedene validacije odabire se devet najtočnijih klasifikatora iz svake kategorije kako bi se dobio ansambl naivnih Bayesovih klasifikatora (ANBK). U tablici 5 prikazani su rezultati validacije (F1 mjere) za svaki od 81 NB klasifikatora. NB klasifikatori koji su uključeni u ansambl su oni s prozorom konteksta: (1, 1), (3, 1), (10, 1), (1, 3), (4, 3), (10, 3), (2, 10), (5, 10), (10, 25).

Nakon provedenog testiranja F1 mjera ANBK-a iznosi

$$F1_{ANBK(interest)} = 0.7. \quad (1)$$

Valja primijetiti kako u ovom slučaju ansambl od devet NB klasifikatora ima manju točnost klasifikacije od svakog pojedinačnog NB klasifikatora od kojih je sastavljen, iz čega se može zaključiti da ANBK ne mora nužno poboljšati točnost klasifikacije. Naime u ovom slučaju bilo bi bolje, umjesto ansambla, klasificirati npr. s NB klasifikatorom s definiranim prozorom konteksta (1, 1), kao što se može iščitati iz tablice 5, njegova F1 mjera iznosi

$$F1_{(1,1)} = 0.81. \quad (2)$$

Doduše treba uzeti u obzir da je mjera $F1_{ANBK(interest)}$ izračunata na skupu za testiranje, dok je $F1_{(1,1)}$ izračunata na skupu za validaciju, tj. te dvije mjere nisu baš usporedive jer nisu dobivene na temelju istih uzoraka.

Tablica 6: Matrica zabune ANBK-a za riječ *interest*

I_1	I_2	I_3	I_4	I_5	I_6	
71	0	0	0	28	108	I_1
1	0	0	0	0	6	I_2
1	0	0	0	4	37	I_3
3	0	0	0	7	92	I_4
0	0	0	0	213	58	I_5
0	0	0	0	0	556	I_6

Ako se promotri tablica 6 lako se uočava kako niti jedan uzorak nije uspješno klasificiran u razrede I_2 , I_3 , I_4 . Naime ti razredi imaju barem za red veličine manje uzoraka od najzastupljenijeg razreda I_6 u skupu uzoraka (vidi tablicu 1). Ovakvu ne posve uspješnu, ali prihvatljivu klasifikaciju možemo opravdati neuravnoteženošću uzoraka između razreda u skupu uzoraka.

Ovakav postupak za riječ *interest* proveden je i u radu (Pedersen, 2000), pri čemu je korištena identična metoda i identični način validacije i testiranja. U tom radu točnost² klasifikacije iznosi 0.88. Razlog zbog čega je naš klasifikator neuspješniji, u odnosu na spomenuti rad, nije nam potpuno poznat. Pretpostavljamo da je razlog u različitom pristupu zaglađivanja apriornih vjerojatnosti s vjerojatnošću nula, tj. u različitim implementacijama Bayesovog klasifikatora (u ovom radu korištena je biblioteka sustava Weka).

5.1.2. Rezultati za riječ *line*

Rezultati validacije mogu se vidjeti u tablici 7. ANBK je sastavljen od Bayesovih klasifikatora s prozorom konteksta: (2, 2), (5, 2), (10, 1), (1, 4), (5, 3), (10, 4), (2, 10), (5, 10) i (10, 10). F1 mjera ANBK-a u ovom slučaju iznosi

$$F1_{ANBK(line)} = 0.91, \quad (3)$$

što se može smatrati izuzetno uspješnom klasifikacijom. Također valja primijetiti kako u ovom slučaju ANBK ima veću točnost od bilo kojeg pojedinačnog NB klasifikatora, što se može pravdati uravnoteženošću skupa uzoraka po razredima (semantičkim značenjima).

Tablica 8: Matrica zabune ANBK-a za riječ *line*

L_1	L_2	L_3	
201	8	3	L_1
11	175	2	L_2
22	6	122	L_3

²Kod klasifikacije u više od dvije klase, točnost je jednaka F1 mjeri.

Tablica 5: Validacijom dobivene F1 mjere NB klasifikatora s prozorom konteksta (l, r) za riječ *interest*

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	0,61	0,75	0,71	0,73	0,7	0,7	0,69	0,71	0,69
1	0,76	0,81	0,8	0,78	0,76	0,75	0,72	0,73	0,71
2	0,72	0,79	0,78	0,77	0,76	0,75	0,74	0,74	0,72
3	0,74	0,78	0,77	0,76	0,76	0,77	0,75	0,73	0,72
4	0,74	0,77	0,76	0,78	0,76	0,77	0,75	0,73	0,73
5	0,73	0,77	0,77	0,78	0,77	0,76	0,76	0,74	0,74
10	0,68	0,73	0,73	0,73	0,72	0,73	0,74	0,75	0,74
25	0,68	0,69	0,69	0,69	0,7	0,7	0,72	0,72	0,72
50	0,66	0,68	0,68	0,69	0,69	0,68	0,7	0,72	0,72

Tablica 7: Validacijom dobivene F1 mjere NB klasifikatora s prozorom konteksta (l, r) za riječ *line*

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	0,32	0,55	0,63	0,67	0,66	0,63	0,66	0,66	0,65
1	0,67	0,68	0,73	0,76	0,8	0,79	0,77	0,73	0,72
2	0,72	0,73	0,79	0,78	0,79	0,79	0,82	0,77	0,76
3	0,77	0,74	0,78	0,75	0,78	0,8	0,83	0,79	0,81
4	0,78	0,77	0,81	0,8	0,79	0,8	0,83	0,83	0,85
5	0,8	0,81	0,83	0,83	0,82	0,83	0,86	0,84	0,85
10	0,81	0,86	0,83	0,86	0,87	0,87	0,9	0,9	0,9
25	0,76	0,77	0,81	0,81	0,82	0,84	0,85	0,83	0,83
50	0,72	0,76	0,76	0,76	0,79	0,81	0,82	0,82	0,81

Tablica 9: Pogreške i vrijednosti parametara C nakon provedene validacije SVM-a za riječ *interest*

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	0.46; 2.5	0.28; 2.5	0.27; 2.5	0.26; 0.75	0.24; 1.5	0.23; 0.625	0.24; 0.5	0.24; 0.75	0.24; 0.375
1	0.22; 2.5	0.15; 1.75	0.14; 0.5	0.13; 1.5	0.15; 0.75	0.14; 0.5	0.15; 2.5	0.15; 0.5	0.15; 0.375
2	0.20; 1	0.12; 1.25	0.12; 0.625	0.12; 0.75	0.13; 0.75	0.12; 1	0.12; 0.75	0.13; 0.5	0.13; 0.5
3	0.19; 1	0.11; 0.5	0.11; 1.5	0.11; 0.75	0.11; 1.5	0.11; 0.5	0.11; 0.25	0.13; 0.5	0.13; 0.5
4	0.18; 2	0.11; 0.75	0.11; 0.75	0.11; 2.5	0.10; 0.25	0.10; 2.5	0.10; 0.25	0.13; 2.5	0.13; 2.5
5	0.18; 0.5	0.12; 0.5	0.12; 2.5	0.11; 0.5	0.10; 0.75	0.10; 2.5	0.10; 0.5	0.12; 2.5	0.12; 2.5
10	0.20; 0.25	0.12; 0.25	0.12; 0.25	0.12; 2.5	0.12; 2.5	0.12; 2.5	0.11; 2.5	0.12; 2.5	0.12; 2.5
25	0.21; 0.5	0.14; 0.375	0.14; 0.5	0.14; 2.5	0.13; 2.5	0.12; 0.5	0.12; 2.5	0.13; 2.5	0.13; 0.25
50	0.20; 0.5	0.14; 0.5	0.14; 0.25	0.14; 2.5	0.13; 0.5	0.12; 0.25	0.12; 2.5	0.13; 0.125	0.14; 2.5

Tablica 10: Pogreške i vrijednosti parametara C nakon provedene validacije SVM-a za riječ *line*

$l \setminus r$	0	1	2	3	4	5	10	25	50
0	0.67; 2.5	0.40; 2.5	0.39; 3	0.36; 4.5	0.36; 2.75	0.36; 1	0.35; 0.375	0.36; 0.625	0.36; 0.5
1	0.34; 2.5	0.27; 2.5	0.27; 1.5	0.27; 2	0.27; 3	0.26; 1.5	0.27; 0.375	0.27; 0.375	0.25; 0.375
2	0.25; 1.5	0.20; 2.5	0.18; 0.5	0.19; 2.5	0.18; 1	0.17; 0.875	0.18; 1.375	0.19; 1.5	0.19; 0.875
3	0.22; 6	0.18; 3.75	0.17; 0.5	0.16; 0.25	0.18; 4.5	0.18; 0.5	0.16; 1.5	0.16; 0.5	0.16; 0.5
4	0.23; 2.5	0.18; 0.5	0.17; 0.5	0.16; 0.375	0.16; 0.5	0.15; 1.5	0.14; 0.5	0.15; 4.5	0.13; 0.25
5	0.21; 0.625	0.18; 0.5	0.17; 2.5	0.16; 0.5	0.16; 2.5	0.16; 2.5	0.16; 2.5	0.16; 0.5	0.16; 0.5
10	0.24; 0.5	0.21; 2.5	0.19; 2.5	0.17; 0.5	0.16; 2.5	0.15; 2.5	0.14; 2.5	0.14; 2.5	0.15; 2.5
25	0.24; 0.375	0.20; 0.5	0.19; 0.25	0.19; 2.5	0.19; 2.5	0.18; 2.5	0.16; 2.5	0.16; 2.5	0.15; 2.5
50	0.27; 2.5	0.19; 0.141	0.21; 2.5	0.21; 0.5	0.21; 0.5	0.20; 2.5	0.18; 2.5	0.17; 2.5	0.17; 2.5

5.2. Klasifikacija pomoću SVM-a

5.2.1. Rezultati za riječ *interest*

U tablici 9 nalaze se rezultati validacije. Za svaki prozor konteksta (l, r) validacijom je izračunat parametar C koji daje najmanju pogrešku na skupu za validaciju. Npr. SVM s prozorom konteksta $(2, 1)$ ima pogrešku klasifikacije 0.12, dok parametar C iznosi 1.25. Onaj SVM kod kojeg kombinacija dimenzija prozora konteksta i parametra C daje najmanju pogrešku na skupu za validaciju, odabire se za testiranje. U našem slučaju, kao što se može očitati iz tablice 9, to je SVM s prozorom konteksta $(5, 4)$ i parametrom $C = 0.75$. Rezultati testiranja vidljivi su u tablici 11, a F1 mjera iznosi

$$F1_{SVM(interest)} = 0.88. \quad (4)$$

Ako se ovaj rezultat usporedi s ANBK-om, vidljivo je kako je SVM klasifikator za riječ *interest* podosta uspješniji od ANBK. Također ako se pomnije promotri tablica 9, vidljivo je kako se dobre dimenzije prozora konteksta nalaze oko veličine $(5, 4)$, kako se udaljavamo od tih dimenzija u bilo kojem smjeru, pogreška klasifikacija raste. Upravo prozor konteksta s dimenzijama $(5, 4)$ daje najmanju pogrešku, naime takav prozor konteksta izlučuje najbolje značajke. Manji prozori konteksta (manje vrijednosti varijabli l i r) ne zahvaćaju karakteristične riječi (one koje dobro diskriminiraju između različitih semantičkih značenja) u dovoljnoj mjeri. Veliki prozor konteksta pak zahvaća i riječi koje ne pridonose diskriminaciji različitih semantičkih značenja riječi, tj. zahvaća previše šuma, pa je pogreška klasifikacije velika, što potvrđuju i rezultati iz tablice 9.

Tablica 11: Matrica zabune SVM-a za riječ *interest*

I_1	I_2	I_3	I_4	I_5	I_6	
164	0	1	12	27	7	I_1
0	0	0	5	0	0	I_2
6	0	18	7	3	1	I_3
13	0	3	63	11	6	I_4
16	0	0	4	204	4	I_5
7	0	0	2	3	598	I_6

5.2.2. Rezultati za riječ *line*

Na uravnoteženom skupu uzoraka za riječ *line* proveden je isti postupak kao i za riječ *interest*. Rezultati validacije SVM-a vidljivi su u tablici 10. Baš kao i za riječ *interest*, nakon validacije odabire se SVM s onim parametrima koji daju najmanju pogrešku na skupu za validaciju. Upravo SVM s prozorom konteksta $(4, 50)$ i parametrom $C = 0.25$ daje najmanju pogrešku na skupu validacije. Za takav SVM nakon testiranja na skupu za testiranje, dobivena je matrica zabune prikazana u tablici 12 iz koje se može izračunati:

$$F1_{SVM(line)} = 0.84. \quad (5)$$

Valja primijetiti kako su F1 mjere za riječ *interest* (4) i *line* (5) slične, ne postoji velika razlika kao u

Tablica 12: Matrica zabune SVM-a za riječ *line*

L_1	L_2	L_3	
170	10	29	L_1
15	159	9	L_2
23	0	135	L_3

slučaju ANBK-a. Dakle SVM bi bio bolji izbor za razrješavanje vižeznačnosti riječi jer nije toliko osjetljiv na neuravnotežene i uravnotežene skupove uzoraka kao što je ANBK. Također zamjećuje se manja uspješnost SVM-a na uravnoteženom skupu uzoraka (*line*) u odnosu na neuravnoteženi (*interest*), dok je kod ANBK-a obrnuti slučaj.

6. Zaključak

Rad rješava problem razrješavanja vižeznačnih riječi na postupcima strojnog učenja. Izneseni su rezultati za SVM postupak i za ansambl naivnih Bayesovih klasifikatora (ANBK) gdje se odluka donosi glasanjem devet naivnih Bayesovih klasifikatora.

Za ovakav nenormaliziran skup podataka, rezultati pokazuju veću uspješnost SVM algoritma. Eksperimenti su pokazali kako povezivanje više NB klasifikatora u ansambl (glasački stroj) ne poboljšava nužno uspješnost klasifikacije. ANBK pokazao se izrazito uspješnim na uravnoteženom skupu podataka, dok na neuravnoteženom skupu podataka daje loše rezultate (velika pogreška klasifikacije). Također proveden je eksperiment koji pronalazi najoptimalniji prozor konteksta za izlučivanje značajki iz skupa uzoraka. Zaključeno je kako prozor konteksta ne bi trebao biti niti prevelik, niti premali; svakako ne bi trebao obuhvaćati više od 15-ak riječi, u suprotnom uspješnost klasifikacije opada.

U daljnjim istraživanjima svakako bi valjalo provjeriti kako se ANBK i SVM ponašaju na normaliziranom skupu uzoraka. Naime takav skup bi dao veće težiste razlučivim značajkama, istovremeno bi se smanjio ukupan broj značajki, a posebice šum.

Literatura

- D. Buscaldi, P. Rosso, F. Pla, E. Segarra, and E. Sanchis Arnal. 2006. Verb sense disambiguation using support vector machines: Impact of wordnet-extracted features. *Int. Conf. on Comput. Linguistics and Intelligent Text Processing*.
- W. Gale, K. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*.
- Y. K. Lee, H.T. Ng, and T.K. Chia. 2004. Disambiguation with support vector machines and multiple knowledge sources. *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.
- T. Pedersen. 2000. A simple approach to building ensembles of naive bayesian classifiers for word sense disambiguation. *Proceeding Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*.