Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Leon Noe Jovan

Prepoznavanje žaljivih objav z metodami strojnega učenja

diplomsko delo

univerzitetni študijski program prve stopnje
računalništvo in informatika

Mentor: doc. dr. Zoran Bosnić

Ljubljana 2013

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00083/2013
Datum: 09.04.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **LEON NOE JOVAN**

Naslov: **PREPOZNAVANJE ŽALJIVIH OBJAV Z METODAMI STROJNEGA UČENJA**

DETECTING OFFENSIVE POSTS WITH MACHINE LEARNING METHODS

Vrsta naloge: Diplomsko delo univerzitetnega študija prve stopnje

Tematika naloge:

V sodobnem času postaja splet popularen informacijski in družabni medij, ki omogoča preproste načine anonimnega objavljanja besedil. Etični in moralni problem na tem področju predstavljajo žaljive objave, ki so prisotne zlasti na družabnih forumih in za katere je zaželen obstoj algoritma, ki jih bo samodejno filtriral glede na neprimernost.

Kandidat naj v diplomski nalogi testira delovanje različnih algoritmovza rudarjenje po besedilih (text mining) na izbrani množici zajetih komentarjev s forumov. Testira naj različne predstavitve podatkov in parametre sistema (število atributov, parametre algoritma). Algoritme naj medseboj primerja z relevantnimi merami za ocenjevanje učenja.

Mentor:

doc. dr. Zoran Bosnić

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:

prof. dr. Nikolaj Zimic

Izjava o avtorstvu diplomskega dela

Spodaj podpisani Leon Noe Jovan, z vpisno številko **63090029**, sem avtor diplomskega dela z naslovom:

Prepoznavanje žaljivih objav z metodami strojnega učenja

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom doc. dr. Zorana Bosnića,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki "Dela FRI".

V Ljubljani, dne 11. septembra 2013 Podpis avtorja:

Zahvaljujem se mentorju doc. dr. Zoranu Bosniću za nasvete in pomoč pri izdelavi diplomske naloge.

Posebna zahvala gre tudi družini in punci Maji za potrpežljivost, razumevanje in podporo tekom študija.

Kazalo

Seznam uporabljenih kratic in simbolov

Kratica	Angleški izraz	Slovenski izraz
SVM	support vector machines	metoda podpornih vektorjev
NB	naive Bayes classifier	naivni Bayesov klasifikator
BSF	Bayesian spam filtering	Bayesovo filtriranje neželene pošte
DT	decision tree	odločitveno drevo
kNN	k-nearest neigbors	k-najbližjih sosedov
RF	random forests	naključni gozd
STACK	stacking	naučeno kombiniranje z meta-učenjem
IG	information gain	informacijski prispevek
MI	mutual information	medsebojna informacija
GINI	Gini index	Gini-indeks
χ^2	chi-square	hi kvadrat
GSS	Galavotti-Sebastiani-Simi Coefficient	Galavotti-Sebastiani-Simijev koeficient
CA	classification accuracy	klasifikacijska točnost
AUC	area under the ROC curve	ploščina pod krivuljo ROC
0/1	binary term weighting	binarna predstavitev

Povzetek

Cilj diplomske naloge je razviti sistem za prepoznavanje žaljivih komentarjev, s katerimi se srečujemo na spletu. Za boljše razumevanje so podana teoretična ozadja osnov strojnega učenja, tekstovnega rudarjenja in razvrščanja besedilnih dokumentov. V nalogi je predstavljen celoten potek dela, od predobdelave besedila, ocenjevanja in izbire najboljših atributov, določanja pomena atributov, do izbire najboljših klasifikatorjev. Rezultate smo preizkusili tudi na tekmovanju na spletnem portalu Kaggle. Za namen diplomskega dela je bila zgrajena tudi baza slovenskih komentarjev, ki služi kot učna in testna množica za preverjanje uspešnosti klasifikacije komentarjev v slovenskem jeziku.

Abstract

The main goal of this thesis was to develop a recognition system for offensive posts on the web. Theoretical backgrounds of machine learning, text mining and text categorization approaches are given for better understanding of this field of computer science. We present a framework of such a system, from text pre-processing, feature selection, term weighting to selection of best classifiers. The results are tested using the data obtained from a related competition on Kaggle. For the purpose of the thesis a database of Slovenian comments was built, which serves as a data set to verify the success of the classification of offensive comments in Slovenian language.

Poglavje 1

Uvod

1.1 Opis problema

Anonimnost in doseg svetovnega spleta sta povzročila večji obseg žaljivega, zlonamernega in celo kaznivega komentiranja ali govora. Spletni forumi, portali z novicami in ostali socialni mediji so poplavljeni s sovražnim govorom, vulgarnim izražanjem in grožnjami, kar onemogoča in odvrača uporabnike od konstruktivnih pogovorov. Razlogov in vrst takšnega izražanja je zelo veliko, vsem pa je skupno, da škodijo tako uporabnikom kot tudi spletnim stranem, na katerih se vse to dogaja [15].

Ročno pregledovanje vsebine je zelo zamudno, na večjih portalih, kjer uporabniki objavijo tudi več tisoč komentarjev dnevno, že skoraj nesmiselno. Pojavlja se potreba po avtomatiziranem načinu pregledovanja komentarjev.

Cilj te diplomske naloge je razviti sistem za prepoznavanje žaljivih komentarjev, preizkusiti različne pristope in metode strojnega učenja, na koncu pa ta sistem preizkusiti tudi na vzorčnih primerih komentarjev v slovenskem jeziku. S pomočjo takšnega sistema bi bilo pregledovanje spletnih vsebin precej hitrejše.

1.2 Definicija žaljivega sporočila

Sovražno vedenje se na svetovnem spletu nahaja v več oblikah, kot na primer sovražni govor, žaljenje, uporaba kletvic, podžigajoča sporočila, spolno nadlegovanje in zalezovanje. V angleščini obstaja nekakšen splošen izraz "flaming", ki si ga lahko razlagamo kot namerno agresivno dejanje, ki se odvija preko računalniško posredovanih kanalov. Omenjeno vedenje se lahko zdi precej neoprijemljiva oblika nasilja, saj se dogaja v virtualnem okolju, vendar ima na prejemnike lahko precejšnje psihološke učinke.

Če povzamemo, lahko žaljivo sporočilo vidimo kot:

- neposredno kritiko ali brezobzirno sporočilo,
- sporočilo, ki vsebuje sovražen jezik ali psovke,
- sporočilo, ki je nekonformno ali provokativno [4].

1.3 Sorodna dela

Na temo prepoznavanja žaljivih komentarjev je že objavljenih kar nekaj člankov, iz katerih lahko črpamo ideje in nasvete. Podoben problem je tudi prepoznavanje neželene elektronske pošte, o katerem je bilo objavljenih zelo veliko člankov, ki so nam lahko prav tako v pomoč. Po pregledu področja sem izbral štiri članke, ki so zanimivi in se problema lotevajo na različne načine:

1. Smokey: Automatic Recognition of Hostile Messages [8]
2. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety [1]
3. Comparative Study on Email Spam Classifier using Data Mining Techniques [3]
4. Optimally Combining Positive and Negative Features for Text Categorization [3]

Smokey [8] združuje obdelavo naravnega jezika in sociolingvistična opazovanja za prepoznavanje sporočil, ki ne le da vsebujejo žaljive besede, ampak jih tudi uporabijo na žaljiv način. Z uporabo različnih pravil določi 47 atributov, ki temeljijo na sintaksi kot tudi semantiki posameznih sporočil, nato pa z uporabo uveljavljenega algoritma C4.5 odloči, ali je besedilo žaljivo ali ne. V tem delu so zanimiva predvsem pravila, saj iščejo značilnosti, tako za žaljiva kot tudi za nežaljiva sporočila. Pravila prepoznavajo ukazovanje, žaljive besede, ki se nanašajo na določene ljudi (t.i. zlobneže), besedne zveze v drugi osebi, pohvale, vljudnostne fraze in ostale vzorce, ki so značilni za žaljiva ali nežaljiva sporočila. Model pravilno napove 64% žaljivih sporočil in 98% nevtralnih sporočil.

Drugi članek [1] se ukvarja z detekcijo uporabnikov, ki pišejo žaljiva sporočila. V ta namen so razvili ogrodje Lexical Syntactic Feature (LSF). Zgrajena sta bila dva slovarja žaljivih in manj žaljivih besed, s katerimi so s pomočjo stavčne analize ocenjevali, če so stavki žaljivi. Poudarek v tem članku je tudi na hitrosti detekcije. Preizkusili so naivni Bayesov klasifikator in metodo podpornih vektorjev. Ugotovili so, da je naivni Bayesov klasifikator hitrejši, vendar daje slabše rezultate, metoda podpornih vektorjev pa počasnejša, vendar bolj točna. Z metodo podpornih vektorjev so dosegli preciznost 77.8% in priklic 77.9%.

Tretji članek [3] obravnava filtriranje neželene elektronske pošte s pomočjo tehnik iz strojnega učenja. Zanimiv je zato, ker nam ponuja splošen pregled metod in učinkovitosti le-teh. Poleg tega je v članku obravnavan problem podoben prepoznavanju žaljivih komentarjev. Opisuje predvsem učinkovitost metod za izbiro najboljših atributov in učinkovitost vseh znanih metod za klasifikacijo.

Besedilo so opisali z 58 preprostimi atributi, ki so opisani v tabeli 1.1. Za najboljšo metodo so izbrali naključni gozd. Pred tem so izbrali atribute s Fisherjevim filtriranjem. S to rešitvijo so dosegli več kot 99% natančnost klasifikacije.

atribut	razlaga
A1-A48	frekvence posameznih črk v sporočila
A49-A54	povprečna dolžina neprekinjenega zaporedja velikih črk
A55	najdaljša dolžina neprekinjenega zaporedja velikih črk
A56	najdaljša dolžina neprekinjenega zaporedja velikih črk
A57	število velikih črk
A58	razred (spam/ham)

Tabela 1.1: Opis atributov, uporabljenih v članku [3]

Četrti članek [11] se ukvarja z izbiro atributov v razvrščanju besedilnih dokumentov. Tu se pogosto pojavlja problem ogromnega števila atributov. V članku za povišanje skalabilnosti predlagajo mere za ocenjevanje atributov, ki jih preprosto izračunajo le na podlagi štetja pojavitev izrazov v posamezni kategoriji. Preizkusili so več različnih mer za ocenjevanje atributov, kot so χ^2, korelacijski koeficient, razmerje obetov in GSS koeficient.

Poglavje 2

Podatki

2.1 Baza komentarjev v slovenskem jeziku

Za potrebe preverjanja uspešnosti klasifikacije žaljivih komentarjev smo izdelali bazo komentarjev v slovenskem jeziku. Baza slovenskih komentarjev vsebuje 584 komentarjev; od tega je 213 komentarjev žaljivih, ostalih 371 nežaljivih. 174 krajših in preprostih komentarjev smo zbrali na portalu Youtube, ostale pa sestavljajo komentarji, ki jih je posredovala organizacija Spletno Oko. To je spletna prijavna točka, ki sprejema anonimne prijave sovražnega govora na internetu. Največ teh komentarjev prihaja iz znanih slovenskih portalov, npr. siol.net, zurnal24.si, dnevnik.si, 24ur.si, rtvslo.si. Vse komentarje smo ročno pregledali in ocenili njihovo žaljivost.

Struktura podatkov je preprosta in je sestavljena iz dveh stolpcev, ki sta ločena s tabulatorjem. Prvi stolpec nam pove, ali je komentar žaljiv (1) oziroma nevtralen (0). Slika 2.1 prikazuje izsek iz baze komentarjev v slovenskem jeziku.

2.2 Podatki s spletnega portala Kaggle

Kaggle je spletni portal, kjer različna podjetja in ustanove predstavijo problem skupaj z zbirko podatkov, ljudje s celotnega sveta pa tekmujejo za

```
0   To ni noben nogomet, to je ragbi za šleve.
1   A si prepričan da nisi punčka?
1   Dejstvo je, da so črnci po konstrukciji zelo podobni opicam.
0   Eden pametnejših komentarjev zadnje čase in tudi precej resničen.
```

Slika 2.1: Izsek podatkov iz baze komentarjev v slovenskem jeziku

```
0   You must have great abs by now.
1   Shouldn't you be at your ballet lesson little girl?
0   Fucking Red Neck Bitch!!
1   Keep whining like a little pussy Mack. You are making my day.
0   You been goofin' w/ the bees?
0   Um are you planning on getting married again?
```

Slika 2.2: Izsek podatkov iz tekmovanja na portalu Kaggle

najboljšo rešitev problema. Po zaključku tekmovanja prireditelj podeli denarno nagrado zmagovalcu v zameno za intelektualno lastnino zmagovalnega modela.

Pred časom je bilo izvedeno tekmovanje na temo prepoznavanja sovražnega govora in sicer *Detecting Insults in Social Commentary* [13], ki ga je organiziralo podjetje Impermium. Cilj tekmovanja je bil ustvariti sistem za prepoznavanje žaljenja med uporabniki, ki deluje v realnem času.

Na voljo so dali učno množico, ki vsebuje 3749 komentarjev, po končanem tekmovanju pa so objavili še testno množico z rešitvami, na podlagi katere se ocenjuje uspešnost modela na tekmovanju. Tekmovanje se je sicer zaključilo že 21. 9. 2012, vendar lahko še vedno preizkušamo uspešnost klasifikacije in se primerjamo s takratnimi tekmovalci. Zmagovalec je dosegel ploščino pod krivuljo ROC 0.84249 in s tem prejel denarno nagrado.

Poglavje 3

Uporabljene metode

3.1 Predstavitev besedilnih dokumentov

Pred uporabo metod strojnega učenja je potrebno besedilne dokumente ustrezno opisati, kar pomeni, da dokument, predstavljen kot niz znakov, prevedemo v model, ki bi bil primeren za učni algoritem. To storimo s pristopi, ki so opisani v tem poglavju.

3.1.1 Vreča besed

Vreča besed je osnoven način za transformacijo besedila. Besedilo predstavljajo besede, katerih vrstni red ni pomemben. Besedilo predstavimo kot vektor (3.1), kjer vsaka različna beseda predstavlja dimenzijo vektorskega prostora [10]:

$$d_i = (w_{i1}, w_{i2}, ..., w_{iT}), \tag{3.1}$$

kjer je w_{ij} (i = 1, 2, ... , N in j = 1, 2, ... , T) utež besede j v besedilu i. T predstavlja celotno število besed, N pa celotno število dokumentov.

3.1.2 Vreča besednih zvez

Vrečo besed lahko dopolnimo tako, da upoštevamo več besed, ki se nahajajo skupaj in s tem zaznamo različne besedne zveze in fraze, ki lahko povedo veliko več kot pa posamezne besede. Težava pri tem je, da je možnih kombinacij besed lahko ogromno, zato moramo izbrati le tiste, ki v sebi nosijo največ informacij.

3.1.3 N-terke znakov

Besedilo lahko opišemo tudi z znaki oziroma črkami, ki si sledijo v zaporedju. Če, na primer, uporabljamo pare znakov, lahko besedo "klasifikator" predstavimo s pari "kl", "la", "as", "si", ..., "or". Z večanjem števila n zelo hitro raste tudi število različnih n-terk. Da ta problem omilimo, lahko ne upoštevamo posebnih znakov v besedilu, ali pa velike in male črke obravnavamo enako.

3.1.4 Zmanjševanje števila atributov

Pri razvrščanju besedilnih dokumentov imamo običajno opravka z velikim številom različnih besed in s tem številom atributov reda 10^5 ali več, zato skušamo število uporabljenih besed zmanjšati.

Eden od načinov, kako zmanjšati število besed, je odstranjevanje splošnih besed, ki ne nosijo informacije o vsebini besedila. To so tako imenovane funkcijske besede: vezniki, predlogi, pomožni glagoli in podobno. Takšne besede je potrebno izločiti iz proučevanja. Izločanje besed sloni na frekvenci besed, kjer se kot funkcijske besede štejejo besede z visoko frekvenco. Pri tem lahko pride do napak in lahko izločimo tudi besede s pomenom. Namesto takšnega pristopa se pogosto uporablja seznam besed za odstranjevanje (*stopwords*).

Poleg odstranjevanja funkcijskih besed lahko besede nadomestimo z osnovnimi oblikami besede, ki jih imenujemo leme. To opravimo z računalniško podprtim postopkom, ki ga imenujemo lematizacija. Tabela 3.1 prikazuje primer lematiziranega besedila.

beseda	Mislim	,	torej	sem
lema	misliti	,	torej	biti

Tabela 3.1: Primer lematiziranega besedila

Omenjena postopka nista univerzalna in morata biti prilagojena vsakemu predstavljenemu jeziku. V diplomski nalogi smo za predobdelavo besedil uporabljali odprtokodno knjižnico Orange, ki podpira lematizacijo in odstranjevanje funkcijskih besed v več različnih jezikih.

3.2 Mere za ocenjevanje atributov

V tem poglavju so opisane ocene za ocenjevanje atributov. S pomočjo teh mer atribute ocenimo in nato izberemo in uporabimo le najboljše. Pametna izbira atributov ima veliko koristi. Bistvo tega postopka je, da odstranimo neinformativne atribute in tako poenostavimo problem in pohitrimo delovanje, brez velikega vpliva na kakovost klasifikacije. Pogosto tako celo povišamo klasifikacijsko točnost, ker s tem postopkom tudi zmanjšamo šum v podatkih.

Pred ocenjevanjem atributov moramo prej za vsak izraz prešteti število pojavitev v pozitivnem in negativnem razredu. Na podlagi tega štetja nato ocenimo atribut z eno od v tem poglavju opisanih mer. Za razlago formul smo uporabili črke A, B, C, D in N, kjer je:

A - število komentarjev, ki so žaljivi in vsebujejo izraz t,

B - število komentarjev, ki so nevtralni in vsebujejo izraz t,

C - število komentarjev, ki so žaljivi in ne vsebujejo izraza t,

D - število komentarjev, ki so nevtralni in ne vsebujejo izraza t,

N - število vseh komentarjev.

3.2.1 Informacijski prispevek

Informacijski prispevek je klasična mera za ocenjevanje pomembnosti atributa. Definiran je kot prispevek informacije atributa k določitvi vrednosti razreda (3.2):

$$Gain(a) = H_R + H_A - H_{RA} = H_R - H_{R|A} \tag{3.2}$$

Informacijski prispevek smo izračunali po formuli (3.2.1), ki smo jo našli v članku [2], ki govori o ocenjevanju in izbiri atributov v rudarjenju tekstovnih dokumentov:

$$Gain = \left| -log(\frac{A+C}{N}) - log(\frac{B+D}{N}) \right| - \tag{3.3}$$
$$- \left(\frac{A+B}{N} * \left| -log(\frac{A}{N}) - log(\frac{B}{N}) \right| + (1 - \frac{A+B}{N}) * \left| -log(\frac{C}{N}) - log(\frac{D}{N}) \right| \right)$$

3.2.2 Medsebojna informacija

Medsebojna informacija (*mutual information*) predvideva, da so bolj učinkovite besede, ki se pogosteje pojavljajo v žaljivih komentarjih. Medsebojna informacija se lahko preprosto izračuna po formuli (3.4):

$$MI = \frac{A * N}{(A+C)(A+B)}. \tag{3.4}$$

3.2.3 Gini-indeks

Ginijev koeficient je leta 1912 uvedel italijanski statistik Corrado Gini. Definiran je na intervalu med 0 in 1, pri čemer velja, da nižji kot je koeficient, bolj enakomerna je porazdelitev, in višji kot je koeficient, bolj neenakomerna je porazdelitev. Pomembnost atributa *Gini(A)* je definirana kot razlika med apriornim in pričakovanim aposteriornim Gini-indeksom:

$$Gini(A) = \sum_j p_{.j} \sum_k p^2_{k|j} - \sum_k p^2_{k.} \tag{3.5}$$

$p_{k.} = n_{k.}/n$,
$p_{.j} = n_{j.}/n$,

$p_{k|j} = n_{kj}/n_{.j}$.
n - število učnih primerov,
$n_{k.}$ - število učnih primerov iz razreda r_k,
$n_{.j}$ - število učnih primerov z j-to vrednostjo danega atributa A,
n_{kj} - število učnih primerov iz razreda r_k in z j-to vrednostjo danega atributa A.

3.2.4 Statistika χ^2

Statistika χ^2 je približno porazdeljena po zakonu hi-kvadrat z $(n_i - 1)(n_0 - 1)$ prostorskimi stopnjami, kjer je n_i število vrednosti atributa A_i, ki ga ocenjujemo, in n_0 število razredov. Statistika χ^2 je podana s:

$$\chi^2 = \sum_k \sum_j \frac{(e_{kj} - n_{kj})^2}{e_{kj}}, \tag{3.6}$$

pri tem je e_{kj} (3.7) pričakovano število primerov iz k-tega razreda in z j-to vrednostjo atributa, če bi bila atribut in razred neodvisna.

$$e_{kj} = \frac{n_{.j} n_{k.}}{n} \tag{3.7}$$

Za izračun statistike χ^2 smo uporabili formulo (3.8), ki je prilagojena za probleme klasifikacije besedilnih dokumentov [11].

$$\chi^2 = \frac{N * (AD - BC)^2}{(A + C)(B + D)(A + B)(C + D)}. \tag{3.8}$$

3.2.5 Koeficient GSS

GSS (Galavotti-Sebastiani-Simi) koeficient je poenostavljena varianta statistike χ^2. Avtorji so iz statistike χ^2 odstranili faktor N, saj naj ne bi bil potreben. Odstranili so tudi imenovalec, ker naj bi favoriziral redke besede in redke kategorije. GSS koeficient se izračuna po formuli:

$$GSS = AD - BC \tag{3.9}$$

3.3 Pomen atributov

To poglavje opisuje pristope, ki smo jih uporabili za izračun atributov, ki opisujejo posamezni komentar. Ideja je, da niso vsi izrazi enako pomembni v besedilu. Že v prejšnjih poglavjih smo omenili, da razne funkcijske besede kot so *in, ker, ali, tudi* zelo malo povedo o besedilu, zato jih smatramo kot manj pomembne. Po drugi strani pa razne kletvice nakazujejo, da gre lahko za žaljiv komentar, zato takšne besede bolj utežimo in jih dojemamo kot bolj pomembne. To poglavje opisuje različne pristope za način izračuna atributov.

3.3.1 Binarna predstavitev

Ta način ne ločuje besed po pomembnosti. Atributi imajo le dve možni vrednosti, in sicer 0 ali 1 (3.10). V našem primeru 1 pomeni, da se določena beseda nahaja v besedilu, 0 pa nam pove, da te besede ni v besedilu.

$$w(t,d) = \begin{cases} 1 & \text{if } t \in d \\ 0 & \text{if } t \notin d \end{cases} \tag{3.10}$$

3.3.2 Tf-idf

Tf-idf (*term frequency–inverse document frequency*) je mera, ki nam pove, kako pomembna je določena beseda v zbirki besedilnih dokumentov. Na splošno imamo raje besede, ki se nahajajo v manj dokumentih, saj so bolj specifične. Če bi uporabljali pogoste besede, ki se nahajajo v večini dokumentov, bi te besedilne dokumente med seboj težko razlikovali. Zato uvedemo mero inverzne frekvence (3.11) v dokumentih, ki meri pomembnost besede v nekem korpusu besedilnih dokumentov.

$$idf(t,D) = log\frac{|D|}{|d \in D : t \in d|} \tag{3.11}$$

Število $|D|$ je število vseh dokumentov v korpusu D, število $|d : t \in d|$ pa je število dokumentov, ki vsebujejo besedo t. Utež elementa je t v določenem

dokumentu d je zmnožek frekvence v tem dokumentu in pomembnosti tega elementa (3.12).

$$tf\text{-}idf(t,d) = tf(t,d) * idf(t) \tag{3.12}$$

3.3.3 Tf-žaljivost

Slabost mere tf-idf je, da ne upošteva razreda, v katerem se besedila nahajajo. Zato smo razvili novo mero, ki upošteva razred in uteži izraz glede na to, kako značilen je za določen razred.

Najprej izračunamo verjetnost (3.13), da je komentar žaljiv, če vsebuje izraz t:

$$žaljivost(t) = P(Z|T) = \frac{P(T|Z)}{P(T|Z) + P(T|N)} \tag{3.13}$$

Da dobimo končno vrednost atributa (3.14), moramo to verjetnost pomnožiti še s frekvenco izraza t v dokumentu.

$$tf\text{-}žaljivost = tf(t,d) * žaljivost(t) \tag{3.14}$$

3.4 Klasifikacija

3.4.1 Naivni Bayesov klasifikator

Naloga Bayesovega klasifikatorja je izračunati pogojne verjetnosti za vsak razred pri danih vrednostih atributov za novi primer, ki ga želimo klasificirati. Naivni Bayesov klasifkator predpostavlja pogojno neodvisnost vrednosti atributov pri danem razredu:

$$p(v_1, v_2, ...v_n|c) = \prod_i p(c|v_i) \tag{3.15}$$

Naivna Bayesova formula:

$$p(c|v_1, v_2, ...v_n) = p(c) * \prod_i \frac{p(c|v_i)}{p(c)} \tag{3.16}$$

Naivni Bayesov klasifikator nov primer klasificira tako, da za vsak možni razred c_i izračuna po naivni Bayesovi formuli verjetnost, da primer $(v_1, v_2, ...v_n)$ pripada razredu c_i, kar zapišemo $p(c_i|v_1, v_2, ...v_n)$. Primer klasificira v razred z največjo verjetnostjo.

3.4.2 Bayesovo filtriranje nezaželene pošte

Ta metoda se pogosto uporablja za filtriranje neželene elektronske pošte z uporabo Bayesovega teorema. Zaradi podobnosti problema jo lahko z lahkoto prenesemo na problem žaljivih komentarjev.

Iz učne množice se nauči oziroma za vse besede v korpusu izračuna verjetnosti (3.17), da je besedilo žaljivo, če vsebuje določeno besedo.

$$P(S|W) = \frac{P(W|S)}{P(W|S) + P(W|H)} \tag{3.17}$$

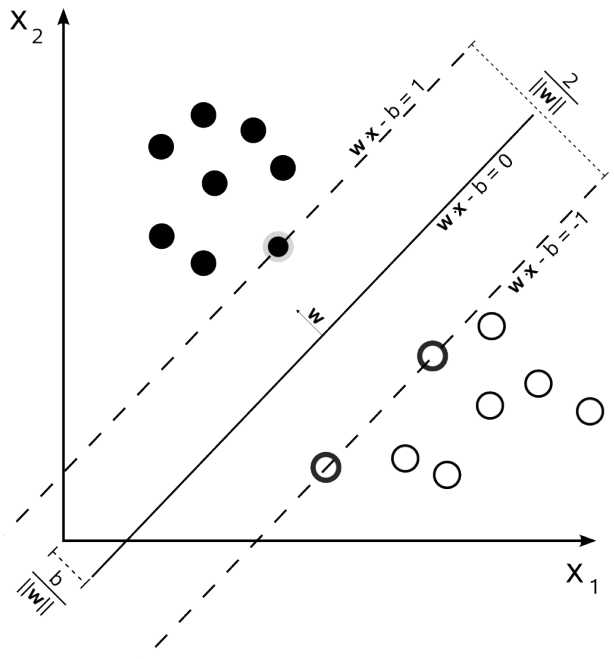
Za računanje verjetnosti, ali je sporočilo žaljivo ali ne, metoda naivno predpostavlja, da so pojavitve besed med seboj neodvisni dogodki. S to predpostavko lahko izpeljemo formulo (3.18) iz Bayesovega teorema.

$$p = \frac{p_1 p_2 ... p_N}{p_1 p_2 ... p_N + (1 - p_1)(1 - p_2)...(1 - p_N)} \tag{3.18}$$

Število p se ponavadi primerja z nekim mejnim pragom. Na osnovi rezultata nato odločimo, ali je besedilo žaljivo ali ne. Metoda je zelo podobna naivnemu Bayesovemu klasifikatorju, le da ta metoda ni pristranska in ne predvideva vnaprej, ali bo prišla neželena pošta.

3.4.3 Metoda podpornih vektorjev (SVM)

Metoda podpornih vektorjev je ena najbolj uspešnih metod za klasifikacijo in regresijo. Večina algoritmov strojnega učenja teži k minimalnemu številu atributov ter poišče ustrezno podmnožico pomembnih atributov, nad katerimi zgradi model. Pri metodi SVM uporabimo vse razpoložljive atribute, tudi manj pomembne, in jih z linearno kombinacijo uporabimo za napovedovanje odvisne spremenljivke. Pri SVM je pomemben predvsem način



Slika 3.1: Metoda podpornih vektorjev

kombiniranja atributov. Izbira atributov je manj pomembna, saj bo sama metoda z ustrezno kombinacijo izluščila želeno informacijo.

Metoda SVM je primerna za učenje na velikih množicah primerov z velikim številom bolj ali manj pomembnih atributov. Dosega visoko točnost napovedi. Njihova slaba stran je težka interpretacija naučenega, prav tako tudi razlaga posamezne odločitve.

Osnovna ideja metode je v danem prostoru atributov postaviti optimalno hiperravnino. Optimalna hiperravnina je tista, ki je enako in najbolj oddaljena od najbližjih primerov dveh razredov. Če imamo razredov več, postopek ponovimo za vsak razred, ki ga skušamo ločiti od ostalih. Najbližjim primerom optimalne hiperravnine pravimo podporni vektorji, razdalji hiperravnine od podpornih vektorjev pa rob. Torej je optimalna hiperravnina tista, ki ima maksimalni rob. Slika 3.1 prikazuje osnovno idejo metode podpornih vektorjev.

3.4.4 Metoda *k*-najbližjih sosedov

Gre za t.i. leno metodo, ki si v fazi učenja samo zapomni vse učne primere, in celotno delo prenese na fazo uvrščanja. Ko želimo napovedati razred novemu primeru, poiščemo med učnimi primeri k najbližjih primerov in pri klasifikaciji napovemo razred, kateremu pripada največ izmed k najbližjih sosedov.

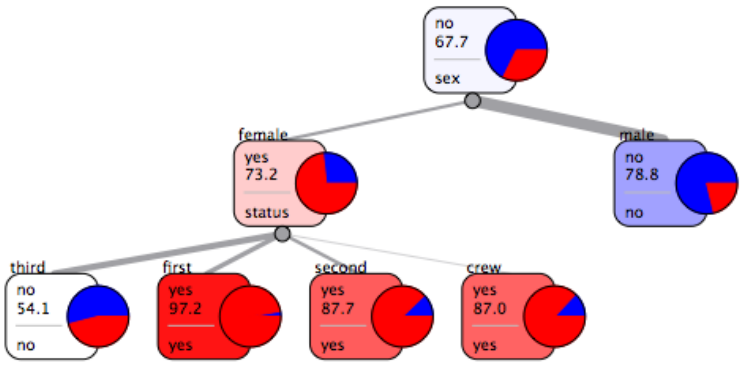
Algoritem k-najbližjih sosedov (*k-nearest neighbours*) je občutljiv na izbrano metriko pri računanju razdalj med novim primerom in učnimi primeri. Za klasifikacijo smo v tem delu uporabljali evklidsko razdaljo, število najbližjih sosedov k pa smo nastavili na koren števila vseh učnih primerov.

3.4.5 Klasifikacijsko drevo

Klasifikacijska drevesa so ena najbolj osnovnih in enostavnih orodij za gradnjo napovednih modelov pri problemih z diskretnim razredom. Osnovna ideja algoritma je razbitje začetne množice podatkov na čim bolj razredno čiste podmnožice. Za razbitje uporabimo en sam atribut, razbitje pa izvedemo na podlagi njegovih vrednosti. Ker tipično podatki vsebujejo več atributov, izberemo tistega, ki vodi k najbolj čistim podmnožicam, torej tistega, katerega informativnost je za dano množico primerov največja. Na sliki 3.2 je primer klasifikacijskega drevesa, ki je zgrajeno iz podatkov o potnikih na ladji Titanic.

3.4.6 Naključni gozd

Metoda naključnih gozdov je namenjena izboljšanju napovedne točnosti drevesnih modelov. Ideja je generirati zaporedje odločitvenih dreves, tako da se pri izbiri najboljšega atributa v vsakem vozlišču naključno izbere majhno število atributov, ki vstopajo v izbor za najboljši atribut. Breiman je predlagal naključno izbiro toliko atributov v vsakem vozlišču, kolikor znaša logaritem števila atributov plus 1. To število je lahko tudi 1, kar pomeni popolnoma naključno izbiro atributa v vsakem vozlišču vsakega drevesa.



Slika 3.2: Klasifikacijsko drevo, ki je zgrajeno iz podatkov o potnikih na ladji Titanic.

Vsako drevo se uporabi za klasifikacijo novega primera po metodi glasovanja - vsako drevo ima en glas, ki ga nameni razredu, v katerega bi klasificiralo nov primer. Število zgrajenih dreves je ponavadi 100 ali več. Iz vseh glasov dobimo verjetnostno porazdelitev po vseh razredih.

Metoda kljub robustnosti ponavadi dosega točnost, primerljivo z najboljšimi algoritmi, slaba stran pa je otežena razlaga odločitve, saj je množica 100 ali več dreves nepregledna in zato nerazumljiva za uporabnika.

Za klasifikacijo smo v tem delu naključni gozd sestavili iz 100 dreves. Število naključno izbranih atributov za vsako drevo pa je koren števila vseh atributov.

3.4.7 Naučeno kombiniranje z meta-učenjem (stacking)

Ta algoritem služi za kombiniranje napovedi različnih klasifikatorjev. Na validacijski množici vsak klasifikator vrne napoved za vsak primer. Te napovedi se uporabijo kot atributi za meta učenje. Naloga učnega algoritma je iz napovedi posameznih klasifikatorjev napovedati pravi razred. Kombinacija klasifikatorjev se zgenerira s pomočjo meta klasifikatorja. V praksi se najbolje obnesejo preprosti meta-učni algoritmi, zato da ne pride do prevelikega

		napovedana vrednost	
		+	-
dejanska	+	TP	FP
vrednost	-	FN	TN

Tabela 3.2: Kontingenčna matrika

prileganja učnim primerom.

3.5 Mere za ocenjevanje učenja

Ko uporabljamo algoritem za strojno učenje, nas zanima, kako uspešno bo reševal nove probleme. Če imamo klasifikacijski problem, želimo vedeti, kako uspešna bo klasifikacija z avtomatsko zgrajeno teorijo. V tem poglavju opisujemo mere za ocenjevanje učenja, ki smo jih uporabljali pri merjenjih in se pogosto uporabljajo pri problemih kategorizacije besedil.

Pri razlagi smo uporabljali oznake TP, FP, FN in TN, ki so prikazane v tabeli 3.2, pomenijo pa naslednje:

TP - zadetek (true positive),

TN - pravilna zavrnitev (false negative),

FP - lažni alarm, napaka I. reda (false positive),

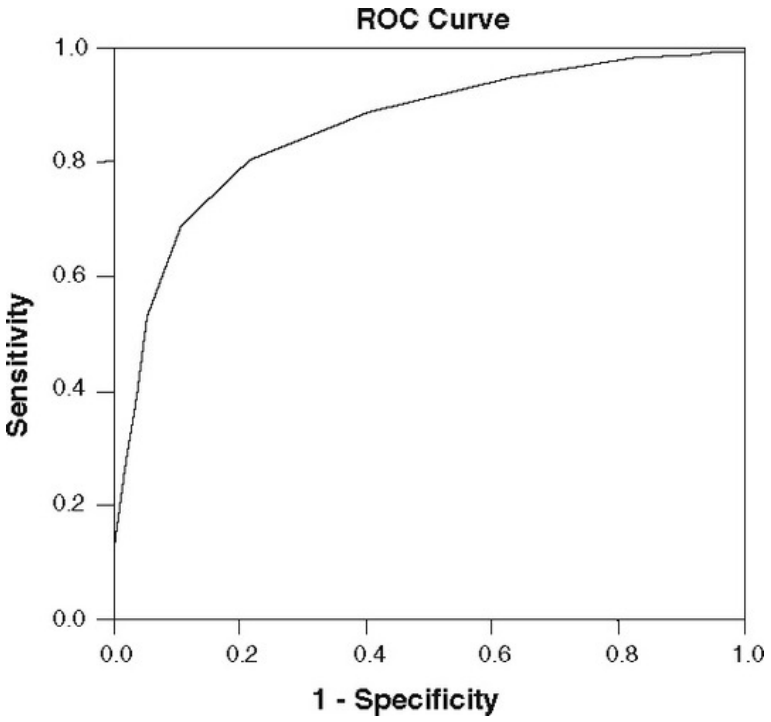
FN - pogrešek, napaka II. reda (false negative),

.

3.5.1 Klasifikacijska točnost

Klasifikacijsko točnost (3.19) interpretiramo kot verjetnost, da bo naključno izbran primer pravilno klasificiran.

$$CA = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.19}$$



Slika 3.3: Primer krivulje ROC

3.5.2 AUC - Ploščina pod krivuljo ROC

Pri iskanju optimalnega klasifikatorja nas zanima predvsem razmerje med senzitivnostjo in specifičnostjo. Klasična metoda je krivulja ROC (*Recievер Operating Characteristic*) 3.3, ki omogoča analizo razmerja med tema dvema merama. Na vodoravni X-osi je prikazano relativno število napačno klasificiranih negativnih primerov (FP), t.j. 1 - specifičnost. Na navpični Y-osi pa je prikazano relativno število pravilno klasificiranih pozitivnih primerov (TP), t.j. senzitivnost. Vsak klasifikator, ki napove samo en razred, je na sliki predstavljen s točko, ki ustreza njegovi senzitivnosti in specifičnosti. Za klasifikatorje, ki vrnejo za vsak primer verjetnostno distribucijo po razredih, dobimo za vsak prag, ki ga izberemo za odločitveno pravilo, drugo točko na grafu. Običajni prag je 0.5 - torej klasificiramo primer v tistih od dveh razredov, ki ima večjo verjetnost.

Če prag spreminjamo, se gibljemo po krivulji ROC. Prag 0.0 pomeni, da vse primere klasificiramo kot negativne (senzitivnost = 0.0 in specifičnost = 1.0), prag 1.0 pa pomeni, da so vsi pozitivni primeri pravilno klasificirani in vsi negativni napačno (senzitivnost = 1.0 in specifičnost = 0.0). Ta dva ekstremna primera sta prikazana kot točki v levem spodnjem kotu in desnem zgornjem kotu. Diagonala, ki povezuje enkstremna primera, ustreza naključnim klasifikatorjem, ki uporabljajo različne pragove. Idealni klasifikator, ki ima senzitivnost in specifičnost enako 1.0, se nahaja v levem zgornjem kotu grafa. Bolj kot se klasifikator približa tej točki, tem boljši je.

Če narišemo celotno krivuljo ROC tako, da zvezno spreminjamo prag odločitve od 0.0 do 1.0, nam o kvaliteti klasifikatorja govori ploščina pod krivuljo ROC (Area Under the ROC Curve, AUC). Izkaže se, da je AUC enaka verjetnosti, da bo klasifikator pravilno razločil med pozitivnim in negativnim primerom.

3.5.3 Preciznost in priklic

Ti dve meri se uporablja na področju iskanja pomembnih dokumentov (*information retrieval*) in ne upoštevata števila pravilno klasificiranih negativnih primerov (TN), saj nas pri iskanju relevantnih primerov le-ti ne zanimajo. Preciznost (3.20) ocenjuje odstotek pravilno klasificiranih primerov, ki so bili klasificirani kot pozitivni. Priklic (3.21) ocenjuje odstotek pomembnih odkritih primerov glede na vse pomembne primere.

$$Preciznost = \frac{TP}{TP + FP} \tag{3.20}$$

$$Priklic = \frac{TP}{TP + FN} \tag{3.21}$$

Poglavje 4

Evalvacija pristopov na zbirki Kaggle

Za prepoznavanje žaljivih komentarjev smo za predstavitev besedil preizkusili tri različne pristope:

1. vreča besed,
2. vreča besednih zvez,
3. n-terke črk in znakov.

Za vsak pristop smo poskusili izbrati parametre, ki bi zagotovili čimvečjo uspešnost klasifikacije. Ti parametri so mera za ocenjevanje atributov, način izračuna atributov, število atributov in klasifikator. Pred analizo smo besedila lematizirali, odstranili funkcijske besede in pretvorili vse črke v majhne. Za vse tri pristope smo najprej preizkusili vse kombinacije sedmih metod za klasifikacijo in treh načinov izračuna atributov. Ta dva parametra sta najbolj vplivala na rezultate. Po izbrani najboljši kombinaciji teh dveh parametrov, smo poskusili izbrati tudi čimboljše ostale parametre. Na koncu smo vse parametre, ki so se izkazali kot najboljši, preizkusili na testnih podatkih s tekmovanja na spletnem portalu Kaggle in tako lahko primerjali rezultate tudi z drugimi ljudmi, ki so se ukvarjali s tem problemom.

Za merjenje uspešnosti klasifikacije smo uporabili 10-kratno prečno preverjanje. Kot učne primere smo uporabili učno množico s tekmovanja na portalu Kaggle, ki vsebuje 3947 komentarjev v angleškem jeziku.

4.1 Vreča besed

To poglavje obravnava prepoznavanje komentarjev, če kot atribute uporabimo le posamezne besede. Najprej smo preizkusili vse kombinacije različnih klasifikatorjev in različnih načinov izračuna atributov. Besedila smo predstavili s 100 najboljšimi atributi, izbranimi z Gini-indeksom. Podrobni rezultati vseh merjenj so v prilogi A.1.

4.1.1 Izbira klasifikatorja

Tabela 4.1 prikazuje povprečne rezultate vsakega od klasifikatorjev. Najbolje so se pričakovano obnesle metode naivni Bayes, metoda podpornih vektorjev in naključni gozd. Tudi metoda naučenega kombiniranja z meta-učenjem, ki združuje napovedi omenjenih metod, daje dobre rezultate, vendar je časovno precej potratna. Odločitveno drevo in metoda najbližjih sosedov dajeta solidne rezultate, slednja v kombinaciji z načinom izračuna atributov tf-žaljivost daje celo rezultate, primerljive z najboljšimi klasifikatorji. Metoda Bayesovo filtriranje neželene pošte se je obnesla najslabše. Klasifikacijska točnost je slabša, kot če bi napovedali večinski razred in tudi AUC nam pove, da je metoda približno tako dobra, kot če bi naključno klasificirali komentarje.

4.1.2 Način izračuna atributov

Tabela 4.2 prikazuje povprečno uspešnost klasifikacije glede na različne pomene atributov. Binarna predstavitev daje zelo dobre rezultate, kljub svoji preprostosti. Metoda tf-idf se je presenetljivo obnesla slabše od ostalih, čeprav se uporablja zelo pogosto pri podobnih problemih.

Na splošno smo najboljše rezultate dobili z mero tf-žaljivost, ki smo jo

klasifikator	CA	AUC	preciznost	priklic
NB	0.78	0.83	0.59	0.61
BSF	0.64	0.51	0.28	0.23
DT	0.75	0.69	0.53	0.52
SVM	0.81	0.84	0.71	0.47
kNN	0.77	0.79	0.76	0.23
RF	0.82	0.85	0.69	0.56
STACK	0.81	0.86	0.64	0.68
avg	0.77	0.77	0.60	0.47

Tabela 4.1: Povprečna uspešnost klasifikacije glede na različne klasifikatorje

razvili sami (glej poglavje 3.3.3). S to mero smo dobili tudi najboljši posamezni rezultat (CA: 0.84, AUC: 0.88, preciznost: 0.75, priklic: 0.59), in sicer v kombinaciji z metodo naključni gozd, s katero smo klasificirali komentarje.

način izračuna atributov	CA	AUC	preciznost	priklic
binarna predstavitev	0.78	0.78	0.63	0.48
tf-idf	0.74	0.73	0.55	0.43
tf-žaljivost	0.79	0.79	0.63	0.50
avg	0.77	0.77	0.60	0.47

Tabela 4.2: Povprečna uspešnost klasifikacije glede na način izračuna atributov

4.1.3 Mera za ocenjevanje atributov

Do sedaj smo kot mero za ocenjevanje atributov uporabljali privzeto mero Gini-indeks, za katero smo predvidevali, da daje dobre rezultate. V tem poglavju bomo obravnavali vpliv različnih mer za ocenjevanje atributov na rezultate klasifikacije. Preizkusili smo pet različnih mer za ocenjevanje atributov: informacijski prispevek, skupna informacija, Gini-indeks, χ^2 in mera

mera za ocenjevanje atributov	CA	AUC	preciznost	priklic
IG	0.84	0.88	0.76	0.60
MI	0.84	0.88	0.76	0.58
GINI	0.84	0.88	0.75	0.59
χ^2	0.84	0.88	0.74	0.59
GSS	0.83	0.87	0.74	0.58

Tabela 4.3: Uspešnost klasifikacije glede na izbiro mere za ocenjevanje atributov

GSS.

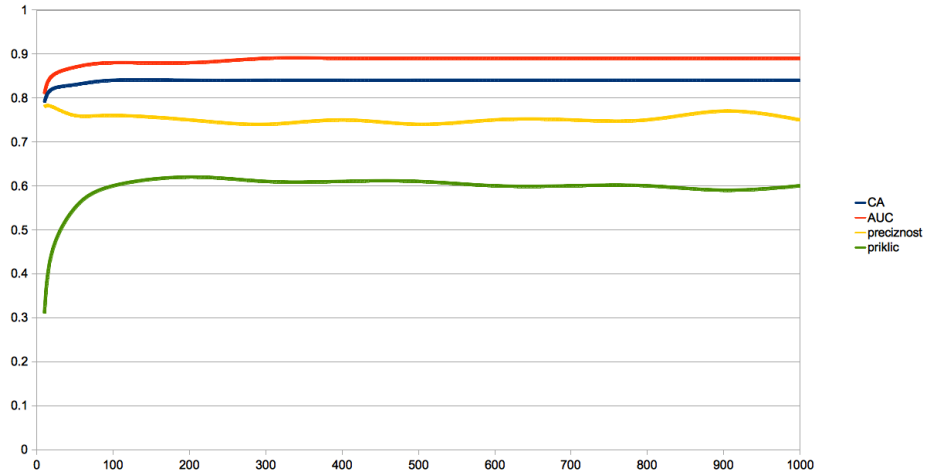
Tabela 4.3 prikazuje rezultate klasifikacije za vsako od mer za ocenjevanje atributov s katermi smo izbrali 100 najboljših atributov. Na podlagi merjenj v prejšnjih poglavjih smo izbrali najboljšo kombinacijo načina izračuna atributov in klasifikatorja. Za klasifikacijo smo uporabili metodo naključni gozd. Vrednosti atributov smo določili z mero tf-žaljivost.

Iz tabele je razvidno, da izbira mere za ocenjevanje atributov ne vpliva veliko na kvaliteto klasifikacije. Najbolje se je obnesel informacijski prispevek, vendar le po malo večjem priklicu. Le mera GSS daje nekaj slabše rezultate od ostalih mer, vendar tudi tu razlike niso velike.

4.1.4 Število atributov

V prejšnjem poglavju smo ugotovili, da najboljše atribute izberemo z mero informacijski dobitek. V tem poglavju pa nas zanima, kolikšno število najboljših atributov oziroma besed moramo izbrati, da dobimo čimboljše rezultate. Graf 4.3 prikazuje rezultate klasifikacije v odvisnosti od števila atributov, ki jih uporabimo za klasifikacijo. Podrobni rezultati so prikazani v prilogi A.2.

Iz grafa je razvidno, da število atributov do določene mere vpliva na točnost klasifikacije, potem pa je vpliv vedno manjši. Razlika je največja predvsem v priklicu, ki narašča nekje do sto atributov. Za dobro klasifikacijo



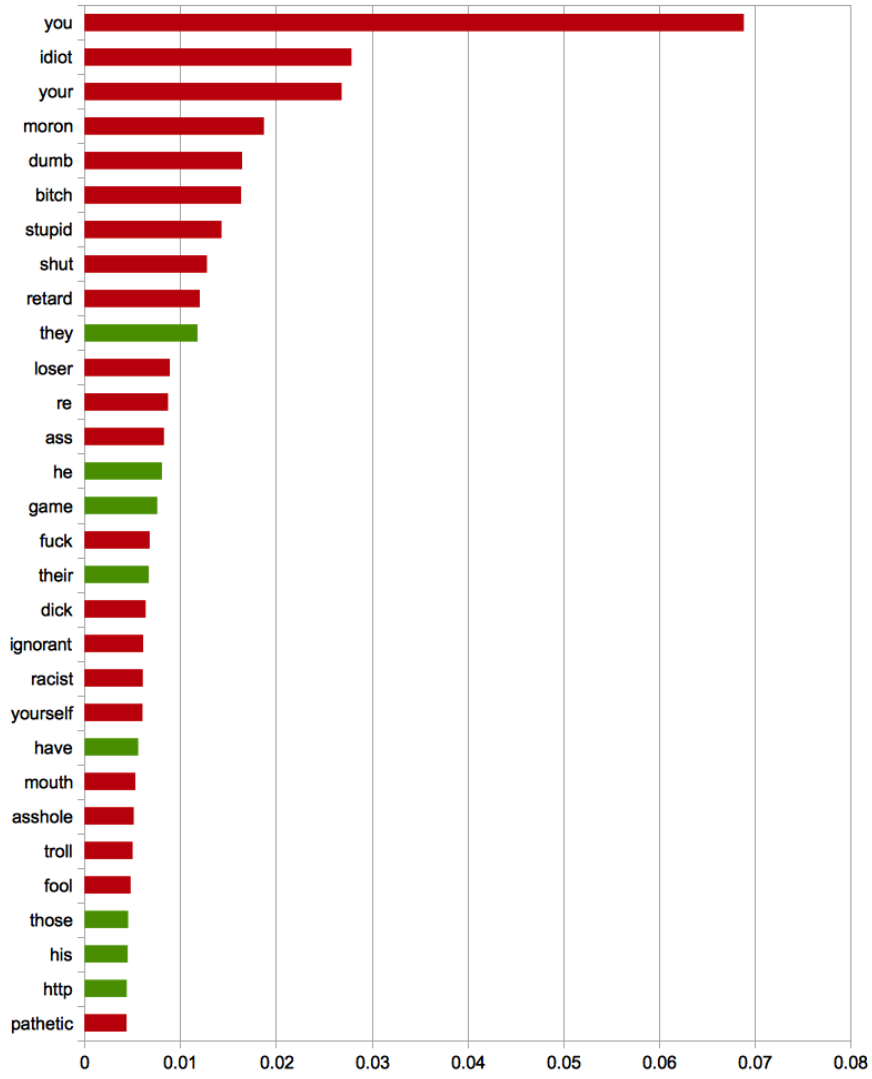
Slika 4.1: Uspešnost klasifikacije glede na število atributov

je dovolj, če uporabimo vsaj sto atributov. Najboljši rezultat smo dosegli z uporabo najboljših 400 atributov (CA: 0.84, AUC: 0.89, preciznost: 0.75, priklic: 0.61). Izbrali smo ga na podlagi najvišjega CA in AUC, poleg tega pa ima tudi visoko preciznost in priklic. Ostali rezultati z višjo preciznostjo imajo večinoma nižji priklic, zato smo jih smatrali kot slabše.

4.1.5 Najbolj informativne besede

Zanimivo je videti, katere besede so zelo značilne za žaljive oziroma nevtralne komentarje. Graf 4.2 prikazuje 30 najpomembnejših besed, ocenjenih z informacijskim prispevkom. Rdeče črte predstavljajo besede, ki so značilne za žaljive komentarje, zelene črte pa besede, ki so značilne za nevtralne komentarje.

Za žaljive komentarje so poleg žaljivih besed, najbolj pomembne besede, ki se nanašajo na drugo osebo. To so besede, kot so “you”, “your”, “yourself”. To je precej smiselno glede na to, da je cilj tekmovanja na portalu Kaggle prepoznavanje medsebojnega žaljenja, ki se nanaša na osebo, katera je vpletena v pogovor.



Slika 4.2: 30 najbolj informativnih besed izbranih z informacijskim prispevkom

preverjanje	CA	AUC	preciznost	priklic
prečno preverjanje	0.84	0.88	0.76	0.60
testni podatki	0.70	0.79	0.84	0.47

Tabela 4.4: Uspešnost klasifikacije z različnimi množicami podatkov

Na splošno se besede v žaljivih komentarjih veliko bolj ponavljajo, nevtralni komentarji pa so po besedah veliko bolj raznoliki, zato je verjetno med najbolj pomembnimi besedami le malo takšnih, ki bi bile izrazito nevtralne. Večinoma so to besede, ki se nanašajo na tretjo osebo, kot so “they”, “he”, “his”in “their”. Zanimiva je tudi beseda “http”, ki nakazuje, da se spletne povezave večinoma pojavljajo v nevtralnih komentarjih. Za nevtralne komentarje so predvsem značilne tudi besede, ki se uporabljajo v pogovorih o politiki, na primer “Obama”, “president”in “government”.

4.1.6 Preizkus na spletnem portalu Kaggle

Najboljši rezultat z vrečo besed smo torej dobili z uporabo 400 najboljših atributov, ki smo jih izbrali z informacijskim dobitkom. Vrednosti atributov smo določili z našo metodo tf-žaljivost in na koncu klasificirali s metodo naključni gozd.

To kombinacijo smo preizkuslili na testni množici podatkov, ki je služila za končni preizkus, na podlagi katerega so določili zmagovalca. Dosegli smo AUC 0.7948, s katerim bi se na tekmovanju uvrstili na 17. mesto. Tabela 4.4 prikazuje rezultate končne kombinacije parametrov, tako na učni množici s prečnim preverjanjem, kot tudi na testni množici.

4.2 Vreča besed in besednih zvez

Vrečo besed iz prejšnjega poglavja smo razširili z vrečo besednih zvez in tako poskusili ujeti nekaj več pomena med besedami. Kot v prejšnjem poglavju smo najprej preizkusili vse kombinacije različnih klasifikatorjev in različnih

načinov izračuna atributov. Besedne zveze so lahko dolge največ pet besed. Ostale nastavitve so nastavljene na privzete vrednosti. Podrobni rezultati vseh merjenj so v prilogi B.1.

4.2.1 Izbira klasifikatorja

Tabela 4.5 prikazuje povprečne rezultate vsakega od klasifikatorjev. Najbolje so se tako kot v prejšnjem poglavju obnesle metode naivni Bayes, metoda podpornih vektorjev, naključni gozd in metoda naučeno kombiniranje z meta-učenjem, ki združuje napovedi treh najboljših klasifikatorjev. Odločitveno drevo in metoda najbližjih sosedov dajeta povprečne rezultate. Metoda najbljižjih sosedov izstopa spet v kombinaciji z načinom izračuna atributov tf-žaljivost, kjer dosega zelo dobre rezultate. Metoda Bayesovo filtriranje neželene pošte se je obnesla najslabše. Na splošno so rezultati v primerjavi s pristopom vreča besed precej podobni, le malce se povišata preciznost in priklic.

klasifikator	CA	AUC	preciznost	priklic
NB	0.81	0.85	0.65	0.60
BSF	0.57	0.52	0.27	0.35
DT	0.77	0.68	0.59	0.51
SVM	0.80	0.83	0.68	0.47
kNN	0.77	0.80	0.75	0.22
RF	0.82	0.85	0.69	0.59
STACK	0.81	0.86	0.62	0.72
avg	0.76	0.77	0.61	0.49

Tabela 4.5: Povprečna uspešnost klasifikacije glede na različne klasifikatorje

4.2.2 Način izračuna atributov

Tabela 4.6 prikazuje povprečno uspešnost klasifikacije glede na različne pomene atributov. Približno enako dobre rezultate dajeta binarna predstavitev in metoda tf-žaljivost. Metoda tf-idf se je obnesla malo slabše. Najboljši rezultat smo ponovno dosegli s kombinacijo tf-žaljivost in metodo naključni gozd (CA: 0.84, AUC: 0.88, preciznost: 0.75, priklic: 0.59).

način izračuna atributov	CA	AUC	preciznost	priklic
binarna predstavitev	0.78	0.79	0.64	0.50
tf-idf	0.74	0.73	0.55	0.46
tf-žaljivost	0.78	0.79	0.62	0.52
avg	0.77	0.77	0.60	0.49

Tabela 4.6: Povprečna uspešnost klasifikacije glede na način izračuna atributov

4.2.3 Mere za ocenjevanje atributov

Tabela 4.7 prikazuje vpliv različnih mer za ocenjevanje atributov na klasifikacijo. Za klasifikacijo smo z vsako mero izbrali 100 najboljših atributov. Primere smo klasificirali z naključnim gozdom. Vrednosti atributov smo določili z metodo tf-žaljivost.

V primeru uporabe besednih zvez ima izbira mere za ocenjevanje atributov rahlo večji vpliv kot pri uporabi vreče besed, vendar še vedno precej majhen. Najboljše rezultate smo dobili z uporabo Gini-indeksa, ki daje najvišjo klasifikacijsko točnost.

4.2.4 Število atributov

V prejšnjem poglavju smo ugotovili, da najboljše atribute izberemo z Gini-indeksom. V tem poglavju pa nas zanima, kolikšno število najboljših atributov oziroma besed in besednih zvez moramo izbrati, da dobimo čimboljše

mera za ocenjevanje atributov	CA	AUC	preciznost	priklic
IG	0.83	0.88	0.72	0.59
MI	0.83	0.87	0.72	0.60
GINI	0.84	0.88	0.74	0.63
χ^2	0.83	0.88	0.72	0.61
GSS	0.83	0.87	0.74	0.58

Tabela 4.7: Uspešnost klasifikacije glede na izbiro mere za ocenjevanje atributov

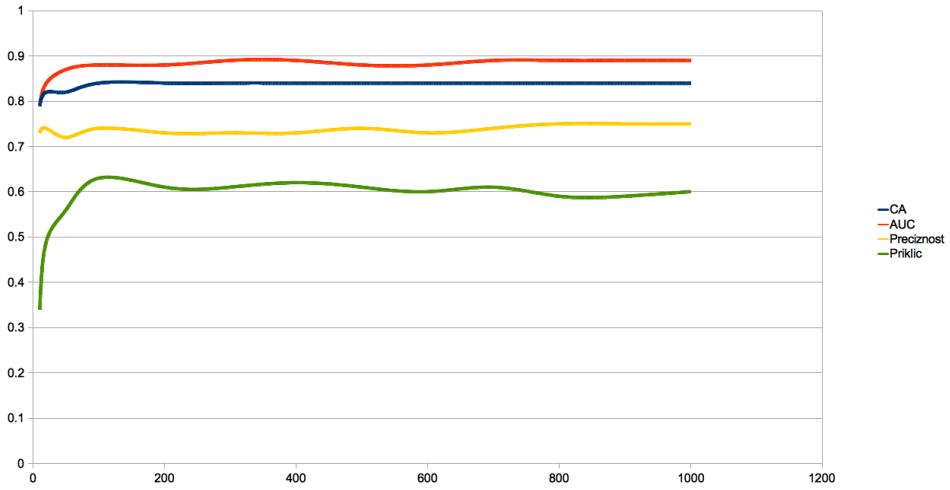
rezultate. Graf 4.3 prikazuje rezultate klasifikacije v odvisnosti od števila atributov, ki jih uporabimo za klasifikacijo. Atribute smo ocenili z Gini-indeksom, atribute izračunali s tf-žaljivost in nato primere klasificirali z metodo naključni gozd. Podrobni rezultati so prikazani v prilogi A.2.

Iz grafa je razvidno, da število atributov do neke mere vpliva na točnost klasifikacije, potem pa je vpliv vedno manjši. Sto atributov že zadostuje za dobro klasifikacijo, od tu naprej se rezultati le počasi izboljšujejo. Najboljše rezultate dobimo, če uporabimo 1000 atributov (CA: 0.84, AUC: 0.89, preciznost: 0.75, priklic: 0.60), vendar je razlika zelo majhna.

4.2.5 Širina okna

To poglavje obravnava vpliv širine okna na uspešnost klasifikacije. Širina okna je določena z največjim številom besed, ki lahko sestavljajo besedno zvezo. Tabela 4.8 prikazuje uspešnost klasifikacije glede na širino okna.

Razlike med rezultati so precej majhne, kar pomeni, da že same besede dovolj povejo o besedilu. Vseeno pa ob povečanju širine okna na pet besed dobimo sočasno največjo preciznost in priklic, zato ta rezultat smatramo kot najboljši.



Slika 4.3: Uspešnost klasifikacije glede na število atributov

širina okna	CA	AUC	preciznost	priklic
1	0.84	0.89	0.75	0.59
2	0.84	0.89	0.75	0.59
3	0.84	0.89	0.74	0.60
4	0.84	0.89	0.74	0.60
5	0.84	0.89	0.75	0.60

Tabela 4.8: Uspešnost klasifikacije glede na širino okna

4.2.6 Najbolj informativne besede in besedne zveze

Zanimivo je videti, katere besede in besedne zveze so zelo značilne za žaljive oziroma nevtralne komentarje. Graf 4.4 prikazuje 30 najpomembnejših besed in besednih zvez, ocenjenih z Gini-indeksom. Rdeče črte predstavljajo izraze, ki so značilne za žaljive komentarje, zelene črte pa izraze, ki so značilne za nevtralne komentarje.

Iz grafa vidimo, da so same besede lahko zelo veliko povedo o besedilu, saj večino najbolj informativnih atributov sestavlja samo ena beseda. Te besede so seveda zelo podobne tistim, kot smo jih opazili v prejšnjem poglavju, kjer smo uporabljali le vrečo besed. Zanima nas predvsem, katere besedne zveze so najbolj informativne.

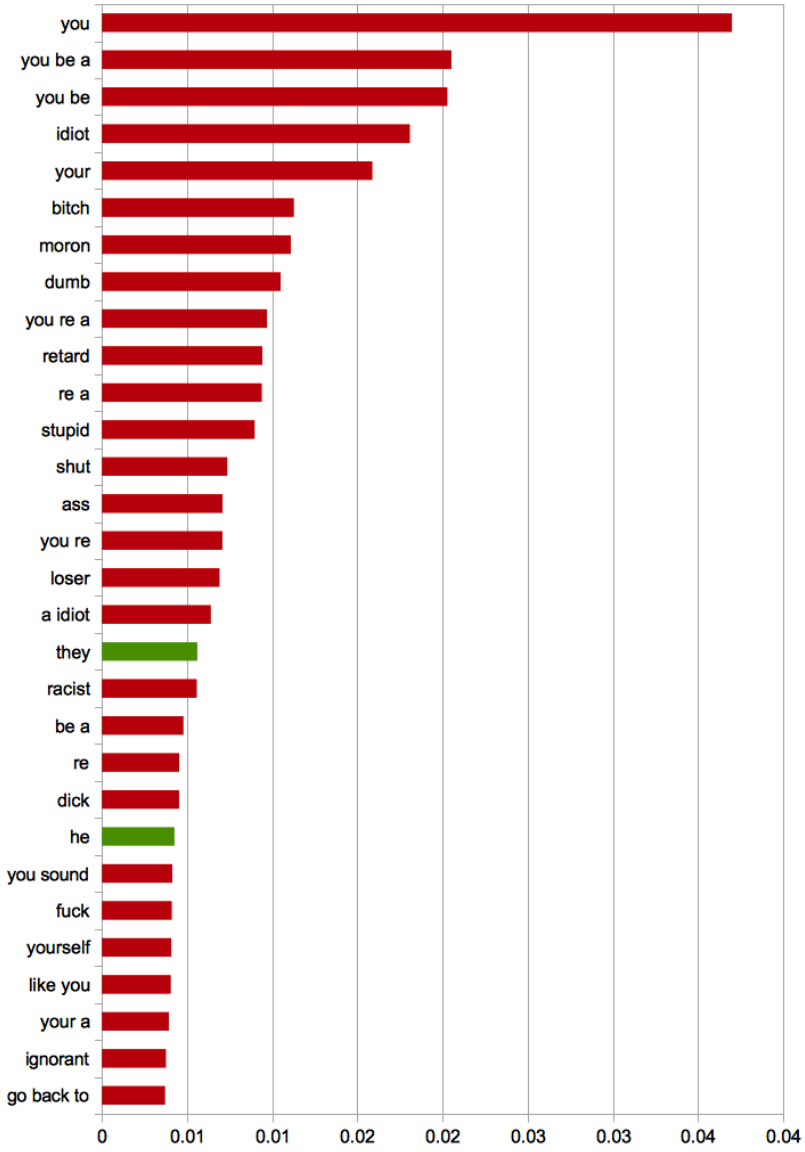
Zelo značilne besedne zveze za žaljive komentarje so tiste, ki se nanašajo na drugo osebo, na primer "you are a", "you re a", "you really are" in podobno. Tem besednim zvezam pogosto sledi ena od kletvic. Podobni primeri so tudi "you sound like", "you look like", "like you", "you little", "you piece of", "go back to your"in podobno. Pogosti so tudi izrazi kot so "shut up", "shut the f**k up", "get a life"in "fuck you". Zanimivo je, da beseda "fuck"na splošno niti ni tako značilna za žaljive komentarje, ampak postane bolj žaljiva šele v povezavi z drugimi besedami.

Za nevtralne komentarje so značilne besedne zveze kot so "they are", "i think", "would be"in "will be". Te besedne zveze se večinoma nanašajo na tretjo ali prvo osebo. Na splošno pa smo opazili, da so takšne besedne zveze precej manj informativne od tistih, ki so žaljive.

4.2.7 Preizkus na spletnem portalu Kaggle

Najboljši rezultat s pristopom vreča besed smo torej dobili z uporabo 1000 najboljših atributov, ki smo jih izbrali z Gini-indeksom. Vrednosti atributov smo določili z metodo tf-žaljivost in na koncu klasificirali s klasifikatorjem naključni gozd.

To kombinacijo smo preizkusili na testni množici podatkov, ki je služila



Slika 4.4: 30 najbolj informativnih besed in besednih zvez, izbranih z Gini-indeksom

preverjanje	CA	AUC	preciznost	priklic
prečno preverjanje	0.84	0.89	0.75	0.60
testni podatki	0.71	0.80	0.86	0.48

Tabela 4.9: Uspešnost klasifikacije z različnimi množicami podatkov

za končni preizkus na podlagi katerega so določili zmagovalca. Dosegli smo AUC 0.80436, s katerim bi se na tekmovanju uvrstili na 13. mesto. Z uporabo vreče besednih zvez smo torej izboljšali rezultate v primerjavi s pristopom vreča besed. Tabela 4.9 prikazuje rezultate končne kombinacije parametrov, tako na učni množici s prečnim preverjanjem, kot tudi na testni množici.

4.3 N-terke znakov

To poglavje obravnava prepoznavanje komentarjev, če kot atribute uporabimo n-terke znakov. Najprej smo preizkusili vse kombinacije različnih klasifikatorjev in različnih načinov izračuna atributov. Besedila smo predstavili s 100 najboljšimi atributi, izbranimi z Gini-indeksom. Atribute lahko predstavljajo terke dolge od 1 do 5 znakov. Podrobni rezultati vseh merjenj so v prilogi C.1.

4.3.1 Izbira klasifikatorja

Tabela 4.10 prikazuje povprečne rezultate vsakega od klasifikatorjev. Najvišjo uspešnost klasifikacije dobimo z metodo podpornih vektorjev, metodo naključni gozd in k-najbljižjih sosedov. Naučeno kombiniranje z meta-učenjem združuje napovedi teh treh najboljših klasifikatorjev in daje daje dobre rezultate, predvsem poviša priklic, vendar na račun preciznosti. Odločitveno drevo in naivni Bayesov klasifikator dajeta solidne rezultate. Bayesovo filtriranje neželene pošte se je obnesla najslabše.

klasifikator	CA	AUC	preciznost	priklic
NB	0.75	0.80	0.52	0.65
BSF	0.27	0.50	0.27	1.00
DT	0.77	0.70	0.62	0.43
SVM	0.80	0.83	0.74	0.35
kNN	0.79	0.81	0.74	0.33
RF	0.81	0.84	0.73	0.45
STACK	0.79	0.84	0.60	0.62
avg	0.71	0.76	0.61	0.55

Tabela 4.10: Povprečna uspešnost klasifikacije glede na različne klasifikatorje

4.3.2 Način izračuna atributov

Tabela 4.11 prikazuje povprečno uspešnost klasifikacije glede na različne pomene atributov. Razlike med različnimi načini izračuna atributov so majhne. Najboljše rezultate smo dobili, če smo pomen atributov določili z binarno predstavitvijo, ki daje rahlo višji priklic. Z binarno predstavitvijo v kombinaciji z naključnim gozdom smo dobili tudi najboljši posamezni rezultat (CA:0.79, AUC:0.85, preciznost: 0.7, priklic:0.47).

način izračuna atributov	CA	AUC	preciznost	priklic
binarna predstavitev	0.71	0.76	0.60	0.57
tf-idf	0.71	0.76	0.62	0.53
tf-žaljivost	0.71	0.76	0.60	0.53
avg	0.71	0.76	0.61	0.55

Tabela 4.11: Povprečna uspešnost klasifikacije glede na način izračuna atributov

mera za ocenjevanje atributov	CA	AUC	preciznost	priklic
IG	0.81	0.85	0.72	0.47
MI	0.81	0.85	0.73	0.47
GINI	0.81	0.85	0.76	0.47
χ^2	0.81	0.85	0.73	0.47
GSS	0.79	0.82	0.69	0.41

Tabela 4.12: Uspešnost klasifikacije glede na izbiro mere za ocenjevanje atributov

4.3.3 Mere za ocenjevanje atributov

Tabela 4.12 prikazuje vpliv različnih mer za ocenjevanje atributov na uspešnost klasifikacije. Za klasifikacijo smo z vsako mero izbrali 100 najboljših atributov. Primere smo klasificirali z metodo naključni gozd. Vrednosti atributov določa binarna predstavitev s številoma 0 in 1.

V primeru uporabe n-terk ima izbira mere za ocenjevanje atributov majhen vpliv na uspešnost klasifikacije. Razen koeficienta GSS so se vse mere obnesle približno enako dobro. Najboljše rezultate smo dobili z uporabo Gini-indeksa, ki rahlo poviša preciznost.

4.3.4 Širina okna

To poglavje obravnava vpliv širine okna na uspešnost klasifikacije. Širina okna pomeni, iz največ koliko črk je lahko sestavljena posamezna terka. Preizkusili smo 15 različnih kombinacij. Toplotna slika 4.5 prikazuje klasifikacijsko točnost (modra) in AUC (oranžna) glede na širino okna. Stolpci predstavljajo največjo dolžino posamezne terke, stolpci pa najmanjšo dolžino. Podrobni rezultati so zapisani v prilogi C.2.

Iz rezultatov je razvidno, da kratke terke povedo malo o besedilu in znižajo uspešnost klasifikacije. Najboljše rezultate smo dobili, če smo uporabili samo terke dolge 4 znake. Tako smo dosegli klasifikacijsko točnost 0.83 in AUC 0.86.

od/do	1	2	3	4	5
1	0.74	0.77	0.81	0.82	0.81
2		0.77	0.81	0.82	0.81
3			0.82	0.82	0.82
4				0.83	0.82
5					0.82

od/do	1	2	3	4	5
1	0.69	0.82	0.85	0.85	0.85
2		0.8	0.85	0.86	0.85
3			0.85	0.86	0.85
4				0.86	0.86
5					0.86

Slika 4.5: Toplotna slika, ki prikazuje klasifikacijsko točnost (modra barva) in AUC (oranžna barva) glede na izbiro dolžine terk znakov

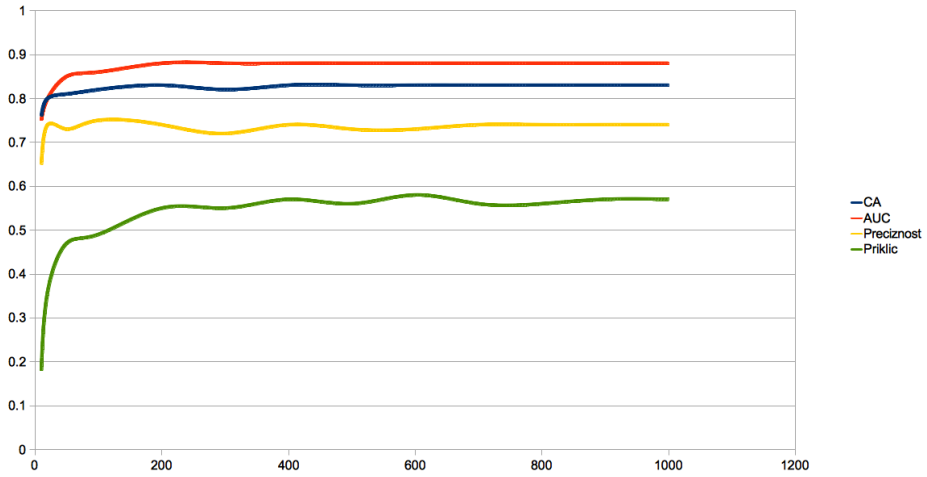
4.3.5 Število atributov

Zanima nas, kolikšno število najboljših atributov oziroma besed in besednih zvez moramo izbrati, da dobimo čimboljše rezultate. Graf 4.6 prikazuje rezultate klasifikacije v odvisnosti od števila atributov, ki jih uporabimo za klasifikacijo. Atribute smo ocenili z Gini-indeksom, atribute izračunali s tf-žaljivost in nato primere klasificirali z metodo naključni gozd. Podrobni rezultati so prikazani v prilogi C.3.

Iz rezultatov je razvidno, da uspešnost klasifikacije znatno raste nekje do 200 atributov, potem pa je vpliv števila atributov vedno manjši. Najboljše rezultate smo dobili, če smo izbrali 400, 900 ali tisoč atributov (CA: 0.83, AUC: 0.88, preciznost: 0.74, priklic: 0.57). Zaradi varčnosti smo izbrali 400 atributov kot najboljšo rešitev.

4.3.6 Najbolj informativni atributi

Zanimivo je videti, kateri atributi so najboljši če uporabimo pristop n-terk znakov. Graf 4.7 prikazuje 30 najpomembnejših terk dolgih 4 znake, ocenje-



Slika 4.6: Uspešnost klasifikacije glede na število atributov

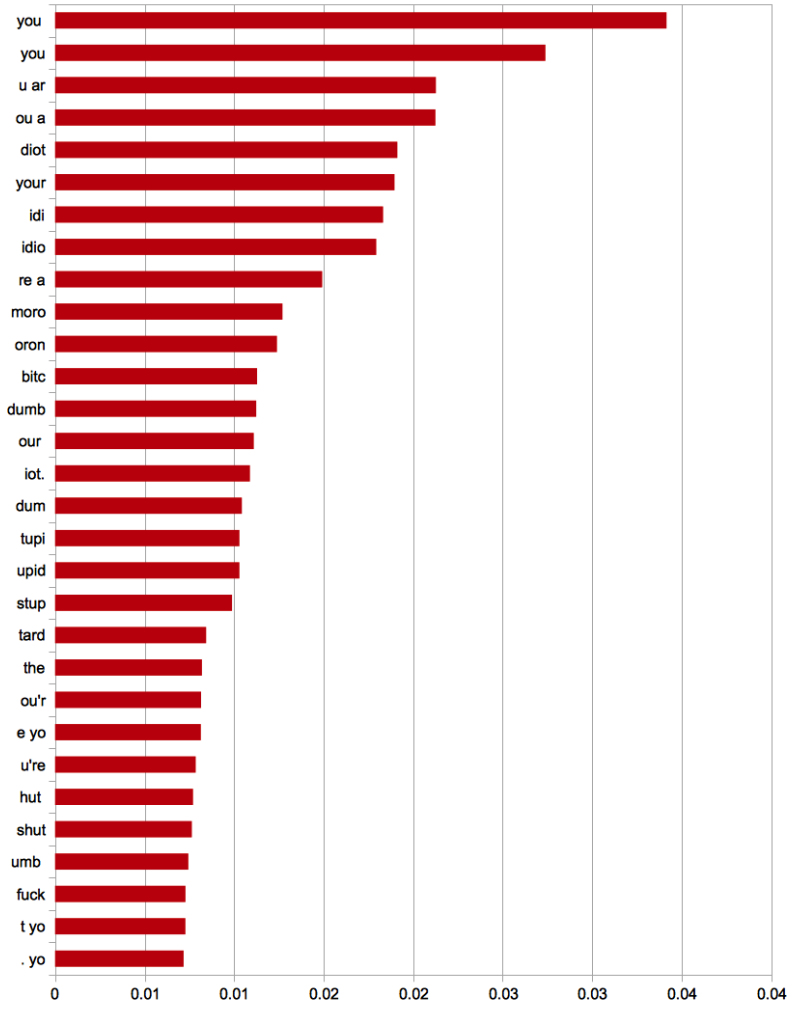
nih z Gini-indeksom. Nekatere terke izgledajo dolge samo 3 znake, vendar ne smemo spregledati raznih ločil in presledkov, ki smo jih prav tako upoštevali.

Najbolj informativni atributi so precej podobni besedam, oziroma iz njih hitro ugotovimo za katere besede gre. Najbolj značilne so terke, ki izhajajo iz besednih zvez "you are a", "you're a". Opazimo pa tudi terke znakov, ki izhajajo iz besed kot so "stupid", "idiot", "moron"in podobno. Zanimivo je tudi, da med 30 najpomembnejšimi atributi ni nobene terke, ki bi bila značilna za nevtralne komentarje.

4.3.7 Preizkus na spletnem portalu Kaggle

Najboljši rezultat z uporabo terk znakov smo torej dobili z uporabo 400 najboljših atributov, ki smo jih izbrali z Gini-indeksom. Vrednosti atributov smo predstavili z binarno predstavitvijo in na koncu klasificirali s klasifikatorjem naključni gozd.

To kombinacijo smo preizkusili na testni množici podatkov s portala Kaggle. Dosegli smo AUC 0.78638, s katerim bi se na tekmovanju uvrstili na 21. mesto. Pristop s terkami znakov v primerjavi z vrečo besed in vrečo



Slika 4.7: 30 najboljših atributov, izbranih z Gini-indeksom

preverjanje	CA	AUC	preciznost	priklic
prečno preverjanje	0.83	0.88	0.75	0.57
testni podatki	0.69	0.79	0.86	0.42

Tabela 4.13: Uspešnost klasifikacije z različnimi množicami podatkov

besednih zvez daje slabše rezultate, vendar je razlika majhna. Tabela 4.13 prikazuje rezultate končne kombinacije parametrov, tako na učni množici s prečnim preverjanjem, kot tudi na testni množici.

Poglavje 5

Prilagoditev za slovenski jezik

Sistem za prepoznavanje žaljivih komentarjev, ki smo ga oblikovali, je zelo splošen in lahko deluje načelno za vsak jezik, če le imamo na voljo bazo komentarjev v tem jeziku iz katerih se sistem lahko uči. Edina omejitev je lematizacija, ki je specifična za vsak jezik. V ta namen uporabljamo dodatek orange-text za knjižnico Orange, ki vsebuje lematizatorje za 13 različnih jezikov, med drugim tudi za slovenščino.

Za preizkus našega sistema smo zgradili bazo slovenskih spletnih komentarjev. Preizkusili smo pristop vreče besed in besednih zvez, ki je na angleških podatkih s tekomavnja na portalu Kaggle dal najboljše rezultate. To pomeni, da smo atribute ocenjevali z Gini-indeksom, atribute smo izračunali z metodo tf-žaljivost in na koncu klasificirali z naključnim gozdom. Edina razlika je v številu atributov, ki smo ga zmanjšali na 100, ker je prihajalo do napak pri merjenju, verjetno zaradi manjše množice podatkov. Tabela 5.1 prikazuje rezultate, ki smo jih dosegli z 10-kratnim prečnim preverjanjem.

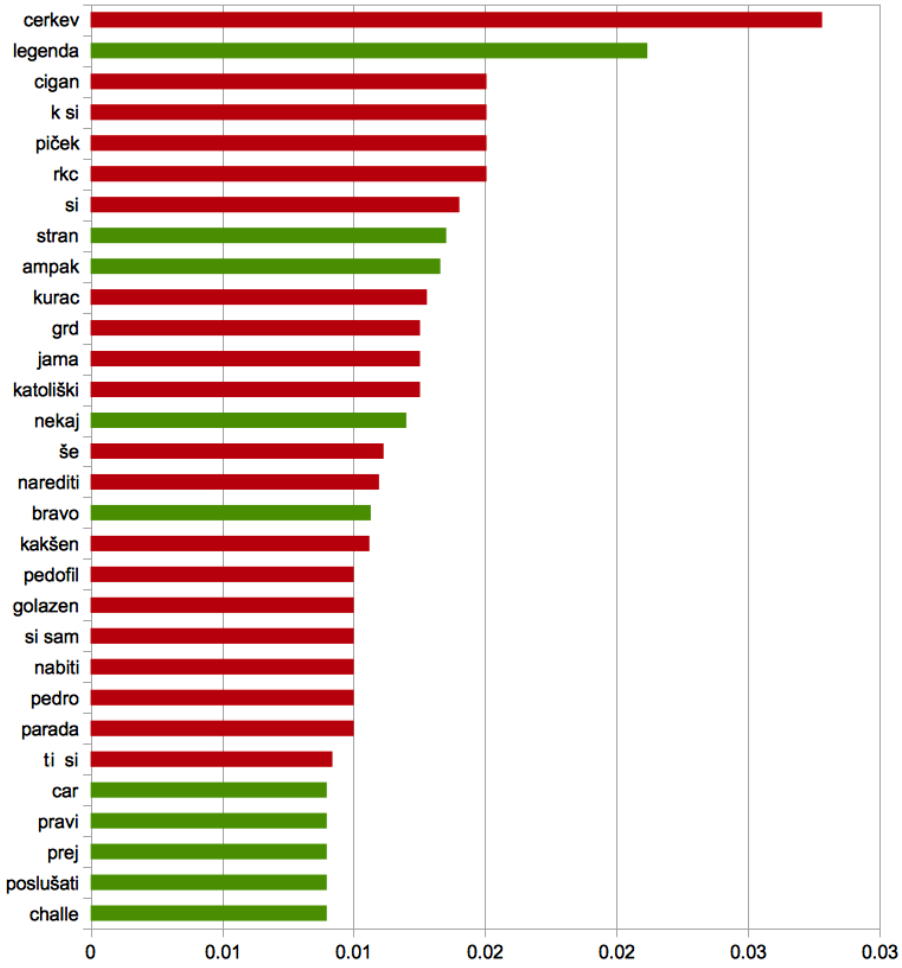
preverjanje	CA	AUC	preciznost	priklic
prečno preverjanje	0.60	0.70	0.86	0.14

Tabela 5.1: Uspešnost klasifikacije na množici komentarjev v slovenskem jeziku

Uspešnost klasifikacije je na množici podatkov v slovenskem jeziku precej drugačna kot na podatkih v angleščini. Sistem prepozna manj žaljivih komentarjev, vendar pa je tudi manj napak II. reda, ki so bolj nezaželene. Na splošno je množica slovenskih komentarjev težja za klasifikacijo, saj vsebuje več oblik žaljivih komentarjev, ne samo medosebno žaljenje. Če pogledamo 30 najbolj informativnih besed in besednih zvez, ki so prikazane na grafu 5.1, vidimo besede značilne, za različne vrste žaljivih komentarjev.

Besede kot so "cerkev", "rkc", "katoliški", "pedofil", "islam" se pogosto pojavljajo v komentarjih, ki so versko nestrpni. Za etnično nestrpne komentarje so značilne besede kot so "cigan", "golazen" in "na ič". Vzorci medsebojnega žaljenja so podobni kot v angleškem jeziku. Na primer "ti si" in v nadaljevanju ena od neprimernih besed ali žaljivih pridevnikov.

Za razliko od baze angleških komentarjev, lahku tu opazimo več besed, ki so značilne za nevtralne komentarje. Besede kot so "legenda", "bravo"in "car" se pojavljajo v komentarjih, kjer se hvali določeno osebo. Značilna nevtralna beseda je tudi "stran", ki pa se pojavlja v komentarjih na različne načine. V nekaterih primerih se ta beseda nanaša na spletno stran, drug primer pa je v povezavi z besedno zvezo "po eni strani" ali "po drugi strani", kjer uporabnik izraža svoje mnenje. Oba načina uporabe besede stran se pojavlja predvsem v nevtralnih komentarjih.



Slika 5.1: 30 najboljših atributov, izbranih z Gini-indeksom

Poglavje 6

Sklepne ugotovitve

6.1 Izboljšave

Pri razvoju klasifikacijskega modela je vedno prostor za izboljšave. Obstaja še veliko možnosti, ki v tem dokumentu niso omenjene, a so kljub temu vredne nadaljnjega raziskovanja. Seznam opisuje tiste izboljšave, ki si zaslužijo več pozornosti, vandar niso bile omenjene v tem dokumentu.

Uporaba pravil Klasifikacijsko točnost bi verjetno izboljšali z uporabo pravil, tako kot so omenjena v članku Smokey [8]. Takšna pravila prepoznajo kontekst znotraj sporočila. Vendar pa bi za tak pristop morali porabiti več časa za analiziranje žaljivih komentarjev, da bi našli zakonitosti in pojave, ki se pogosto pojavljajo v takšnem izražanju.

Uporaba označevalnika in razčlenjevalnika Označevalnik je računalniški program, s katerim poljubno besedilo razdelimo na enote in posameznim besedam pripišemo dodatne informacije, npr. besedno vrsto, v katero spada, in kakšne so njene lastnosti, ali njeno osnovno obliko, če gre za besedo, ki ima več pregibnih oblik [16].

Skladenjski razčlenjevalnik je računalniški program, s pomočjo katerega besedam v povedih lahko pripišemo skladenjska razmerja [17].

S tovrstnimi programomi lahko računalnik besedilo analizira veliko globje. Z uporabo takšnih orodij bi lahko povečali točnost napovedi.

Sodelovanje uporabnikov Če bi klasifikacijski model dejansko uporabili v praksi, bi sodelovanje uporabnikov verjetno precej pripomoglo k prepoznavanju žaljivih komentarjev. Danes ima večina spletnih portalov možnost "(ne)všečkanja" komentarjev oziroma možnost prijave sporne vsebine. Ta način se uporablja pri velikih spletnih portalih skupaj z avtomatskim filtriranjem sporočil kot tudi ročnim pregledovanjem.

6.2 Zaključek

Razvili smo sistem za prepoznavanje žaljivih komentarjev. Opisali in ovrednotili smo rezultate različnih metod, ki se pogosto pojavljajo pri klasifikaciji besedilnih dokumentov.

Za izračun atributov smo razvili metodo tf-žaljivost. Na podlagi uporabe metode tf-žaljivost na testnih podatkih ocenjujemo, da razvita metoda daje boljše rezultate od navadno uporabljenih metod za izračun atributov.

Najboljši rezultat smo dosegli s pristopom vreče besednih zvez, dolgih največ pet besed. Izbrali smo 400 najboljših atributov, ocenjenih z Gini-indeksom. Atribute smo izračunali z metodo tf-žaljivost. Primere smo klasificirali z naključnim gozdom in dosegli AUC 0.80436. S takšnim rezultatom bi se na portalu Kaggle uvrstili na 13. mesto.

Uporabo sistema smo preizkusili tudi na bazi podatkov v slovenskem jeziku. Za ta namen smo sestavili bazo slovenskih komentarjev, zbranih iz znanih slovenskih spletnih portalov in spletne strani YouTube.

Sistem za prepoznavanje žaljivih komentarjev bi lahko uporabili na spletnem portalu, kjer imamo opravka z velikim številom sporočil. Čeprav točnost takšnih algoritmov še ni tako dobra, da bi lahko to opravilo zaupali le programu, lahko še vedno dobro služi za odkrivanje potencialnih žaljivih komentarjev, ki jih potem pregleda moderator in jih po potrebi izbriše. To bi lahko precej olajšalo delo moderatorjev.

Literatura

[1] Y. Chen, Y. Zhou, S. Zhu, H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety". Dostopno na: http://faculty.ist.psu.edu/xu/papers/Chen_etal_SocialCom_2012.pdf

[2] G. Forman, "Feature Selection for Text Classification". Published as a book chapter in Computational Methods of Feature Selection ,Copyright 2007 CRC Press/Taylor and Francis Group Dostopno na: http://www.hpl.hp.com/techreports/2007/HPL-2007-16R1.pdf

[3] R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, "Comparative Study on Email Spam Classifier using Data Mining Techniques". Proceedings of the International MultiConference of Engineers and Computer Sciencetists 2012, Vol 1, IMECS 2012, March 14-16, 2012, Hong Kong. Dotopno na: http://www.iaeng.org/publication/IMECS2012/IMECS2012_pp539-544.pdf

[4] V. Kočevar, "Verbalno naslije v interaktivnih forumih". Dostopno na: http://www.safe.si/uploadi/editor/1256762674Kocevar-Valentina.pdf

[5] I. Kononenko, R. Robnik Šikonja, "Inteligentni sistemi". Založba FE in FRI, 2010

[6] P. Peer, B. Cargo, I. Kononenko, "Razširitev algoritma ReliefF". Dostopno na:
http://www.lrv.fri.uni-lj.si/~peterp/publications/ev97.pdf

[7] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, "Offensive Language Detection Using Multi-level Classification". Dostopno na:
http://www.site.uottawa.ca/~diana/publications/Flame_Final.pdf

[8] E. Spertus, "Smokey: Automatic Recognition of Hostile Messages". Dostopno na: http://people.mills.edu/spertus/Smokey/smokey.pdf

[9] Z. Zheng, R. Srihari, "Optimally Combining Positive and Negative Features for Text Categorization". Dostopno na: http://www.site.uottawa.ca/~nat/Workshop2003/zheng.pdf

[10] M. Vončina, "Predstavitev spletnih novic iz več virov", Dostopno na: http://eprints.fri.uni-lj.si/1618/1/Voncina1.pdf

[11] Z. Zheng, R. Srihari, "Optimally Combining Positive and Negative Features for Text Categorization'. Dotopno na: http://www.site.uottawa.ca/~nat/Workshop2003/zheng.pdf

[12] Bayesian spam filtering. Dostopno na:
https://en.wikipedia.org/wiki/Bayesian_spam_filtering

[13] Detecting insults in social commentary. Dostopno na:
https://www.kaggle.com/c/detecting-insults-in-social-commentary

[14] Spletni portal Kaggle. Dostopno na:
http://www.kaggle.com/about

[15] Podjetje Impermium. Dostpono na:
https://www.impermium.com/what-we-do/icp

[16] Označevalnik. Dostopno na:
http://www.slovenscina.eu/tehnologije/oznacevalnik

[17] Razčlenjevalnik. Dostopno na:
`http://www.slovenscina.eu/tehnologije/razclenjevalnik`

Priloge

Dodatek A

Podrobni rezultati evalvacije pristopa vreča besed

A.1 Rezultati meritev različnih kombinacij klasifikatorjev in načinov izračuna atributov

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.79	0.78	0.78	0.78
BSF	0.64	0.64	0.64	0.64
DT	0.76	0.70	0.79	0.75
SVM	0.84	0.77	0.81	0.81
kNN	0.75	0.76	0.81	0.77
RF	0.84	0.77	0.84	0.82
STACK	0.82	0.78	0.84	0.81
avg	0.78	0.74	0.79	0.77

Tabela A.1: CA

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.85	0.81	0.83	0.83
BSF	0.51	0.51	0.51	0.51
DT	0.71	0.62	0.74	0.69
SVM	0.87	0.80	0.86	0.84
kNN	0.78	0.75	0.84	0.79
RF	0.88	0.78	0.88	0.85
STACK	0.88	0.83	0.88	0.86
avg	0.78	0.73	0.79	0.77

Tabela A.2: AUC

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.59	0.59	0.59	0.59
BSF	0.28	0.28	0.28	0.28
DT	0.55	0.43	0.62	0.53
SVM	0.77	0.59	0.76	0.71
kNN	0.79	0.77	0.73	0.76
RF	0.75	0.58	0.75	0.69
STACK	0.65	0.58	0.69	0.64
avg	0.63	0.55	0.63	0.60

Tabela A.3: Preciznost

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.66	0.56	0.60	0.61
BSF	0.23	0.23	0.23	0.23
DT	0.56	0.49	0.52	0.52
SVM	0.54	0.45	0.41	0.47
kNN	0.09	0.14	0.45	0.23
RF	0.57	0.53	0.59	0.56
STACK	0.70	0.64	0.69	0.68
avg	0.48	0.43	0.50	0.47

Tabela A.4: Priklic

A.2 Število atributov

število atributov	CA	AUC	preciznost	priklic
10	0.79	0.81	0.78	0.31
20	0.82	0.85	0.78	0.45
50	0.83	0.87	0.76	0.55
100	0.84	0.88	0.76	0.60
200	0.84	0.88	0.75	0.62
300	0.84	0.89	0.74	0.61
400	0.84	0.89	0.75	0.61
500	0.84	0.89	0.74	0.61
600	0.84	0.89	0.75	0.60
700	0.84	0.89	0.75	0.60
800	0.84	0.89	0.75	0.60
900	0.84	0.89	0.77	0.59
1000	0.84	0.89	0.75	0.60

Tabela A.5: Uspešnost klasifikacije glede na število atributov

Dodatek B

Podrobni rezultati evalvacije pristopa vreča besednih zvez

B.1 Rezultati meritev različnih kombinacij klasifikatorjev in načinov izračuna atributov

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.82	0.79	0.82	0.81
BSF	0.57	0.57	0.57	0.57
DT	0.78	0.75	0.79	0.77
SVM	0.83	0.76	0.80	0.80
kNN	0.75	0.76	0.80	0.77
RF	0.84	0.77	0.84	0.82
STACK	0.83	0.79	0.81	0.81
avg	0.78	0.74	0.78	0.77

Tabela B.1: CA

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.86	0.83	0.85	0.85
BSF	0.52	0.52	0.52	0.52
DT	0.70	0.63	0.71	0.68
SVM	0.87	0.78	0.85	0.83
kNN	0.80	0.74	0.85	0.80
RF	0.87	0.79	0.88	0.85
STACK	0.88	0.83	0.88	0.86
avg	0.79	0.73	0.79	0.77

Tabela B.2: AUC

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.66	0.62	0.68	0.65
BSF	0.27	0.27	0.27	0.27
DT	0.60	0.53	0.63	0.59
SVM	0.76	0.56	0.71	0.68
kNN	0.78	0.73	0.73	0.75
RF	0.74	0.58	0.74	0.69
STACK	0.67	0.58	0.61	0.62
avg	0.64	0.55	0.62	0.60

Tabela B.3: Preciznost

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.65	0.56	0.59	0.60
BSF	0.35	0.35	0.35	0.35
DT	0.53	0.49	0.51	0.51
SVM	0.55	0.43	0.43	0.47
kNN	0.09	0.18	0.39	0.22
RF	0.61	0.54	0.63	0.59
STACK	0.72	0.68	0.76	0.72
avg	0.50	0.46	0.52	0.49

Tabela B.4: Priklic

B.2 Število atributov

število atributov	CA	AUC	preciznost	priklic
10	0.79	0.79	0.73	0.34
20	0.82	0.84	0.74	0.49
50	0.82	0.87	0.72	0.56
100	0.84	0.88	0.74	0.63
200	0.84	0.88	0.73	0.61
300	0.84	0.89	0.73	0.61
400	0.84	0.89	0.73	0.62
500	0.84	0.88	0.74	0.61
600	0.84	0.88	0.73	0.60
700	0.84	0.89	0.74	0.61
800	0.84	0.89	0.75	0.59
900	0.84	0.89	0.75	0.59
1000	0.84	0.89	0.75	0.60

Tabela B.5: Uspešnost klasifikacije glede na število atributov

Dodatek C

Podrobni rezultati evalvacije pristopa n-terke znakov

C.1 Rezultati meritev različnih kombinacij klasifikatorjev in načinov izračuna atributov

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.74	0.75	0.75	0.75
BSF	0.27	0.27	0.27	0.27
DT	0.78	0.77	0.77	0.77
SVM	0.81	0.80	0.78	0.80
kNN	0.79	0.78	0.80	0.79
RF	0.81	0.81	0.81	0.81
STACK	0.79	0.79	0.80	0.79
avg	0.71	0.71	0.71	0.71

Tabela C.1: CA

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.81	0.81	0.78	0.80
BSF	0.50	0.50	0.50	0.50
DT	0.69	0.72	0.68	0.70
SVM	0.84	0.83	0.81	0.83
kNN	0.82	0.81	0.81	0.81
RF	0.85	0.84	0.84	0.84
STACK	0.84	0.84	0.83	0.84
avg	0.76	0.76	0.76	0.76

Tabela C.2: AUC

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.50	0.53	0.52	0.52
BSF	0.27	0.27	0.27	0.27
DT	0.63	0.61	0.61	0.62
SVM	0.71	0.79	0.73	0.74
kNN	0.76	0.75	0.70	0.74
RF	0.71	0.76	0.73	0.73
STACK	0.59	0.60	0.62	0.60
avg	0.60	0.62	0.60	0.61

Tabela C.3: Preciznost

klasifikator	0/1	TF-IDF	TF-Zaljivost	avg
NB	0.68	0.65	0.61	0.65
BSF	1.00	1.00	1.00	1.00
DT	0.47	0.41	0.42	0.43
SVM	0.45	0.33	0.27	0.35
kNN	0.29	0.28	0.41	0.33
RF	0.47	0.45	0.43	0.45
STACK	0.66	0.62	0.59	0.62
avg	0.57	0.53	0.53	0.55

Tabela C.4: Priklic

C.2 Širina okna

n	CA	AUC	preciznost	priklic
1-1	0.74	0.69	0.78	0.05
1-2	0.77	0.82	0.70	0.27
1-3	0.81	0.85	0.76	0.44
1-4	0.82	0.85	0.75	0.48
1-5	0.81	0.85	0.71	0.46
2-2	0.77	0.80	0.69	0.27
2-3	0.81	0.85	0.75	0.75
2-4	0.82	0.86	0.75	0.49
2-5	0.81	0.85	0.72	0.47
3-3	0.82	0.85	0.76	0.46
3-4	0.82	0.86	0.74	0.49
3-5	0.82	0.85	0.73	0.49
4-4	0.83	0.86	0.75	0.53
4-5	0.82	0.85	0.73	0.53
5-5	0.82	0.86	0.71	0.54

Tabela C.5: Uspešnost klasifikacije glede na širino okna

C.3 Število atributov

število atributov	CA	AUC	preciznost	priklic
10	0.76	0.75	0.65	0.18
20	0.80	0.80	0.74	0.36
50	0.81	0.85	0.73	0.47
100	0.82	0.86	0.75	0.49
200	0.83	0.88	0.74	0.55
300	0.82	0.88	0.72	0.55
400	0.83	0.88	0.74	0.57
500	0.83	0.88	0.73	0.56
600	0.83	0.88	0.73	0.58
700	0.83	0.88	0.74	0.56
800	0.83	0.88	0.74	0.56
900	0.83	0.88	0.74	0.57
1000	0.83	0.88	0.74	0.57

Tabela C.6: Uspešnost klasifikacije glede na število atributov