SPLETNA APLIKACIJA ZA KATEGORIZACIJO IN SIMULACIJO DOKUMENTNIH TOKOV

NAVODILA ZA UPORABO

JUNIJ 2015

Kazalo vsebine

Kazalo slik

1 UVOD

Navodilo za uporabo opisuje dve fazi raziskave in razvoja spletnega modula, ki omogoča zajem, shranjevanje in označevanje spletnih besedil ter ugotavljanje sentimenta v spletnih besedil. **V poglavju 5 se nahaja navodila za uporabo spletnega modula. V poglavju 6 se nahaja postopek določitve sentimenta novih spletnih besedil.**

Dostop do spletnega modula je mogoč s klikom na povezavo:

http://dejan.amadej.si/test

uporabniško ime: xxxxxx geslo: xxxxxx

Slika 1: Prijava v spletni modul

2 DELO IN AKTIVNOSTI

2.1 1. FAZA: RAZISKAVA ZAJEMA SPLETNIH BESEDIL TER ZAPISA V BAZO

V začetni fazi smo raziskali, kateri izmed že obstoječih iskalnikov je primeren za rešitev problema. Izbrali smo spletni iskalnik Google, in sicer storitev Google Custom Search Engine, ki omogoča prilagajanje iskanih parametrov za potrebe raziskave.

Raziskali smo rešitev za spletni modul oziroma spletni portal, ki bi uporabniku omogočal iskanje zadetkov glede na naslednje iskane kriterije:

a) spletna stran – uporabnik lahko izbira med vgrajenim naborom spletnih strani, lahko pa se iskanje izvede splošno po spletnih straneh ali po posamezni spletni strani;

b) časovni okvir – določitev časa: možnost iskanja po spletnih vsebinah objavljenih med datumoma (od-do) ali na splošno; uporabnik bo lahko izbral časovni okvir z ročnim vnosom ali s klikanjem po vgrajenem koledarju;

c) ključna beseda – uporabnik lahko vnaša ključne besede ali niz ključnih besed na podlagi katerih se iskanje izvrši;

d) uporabniško dodajanje spletnih strani – rešitev dodajanja novih spletnih strani v iskane parametre.

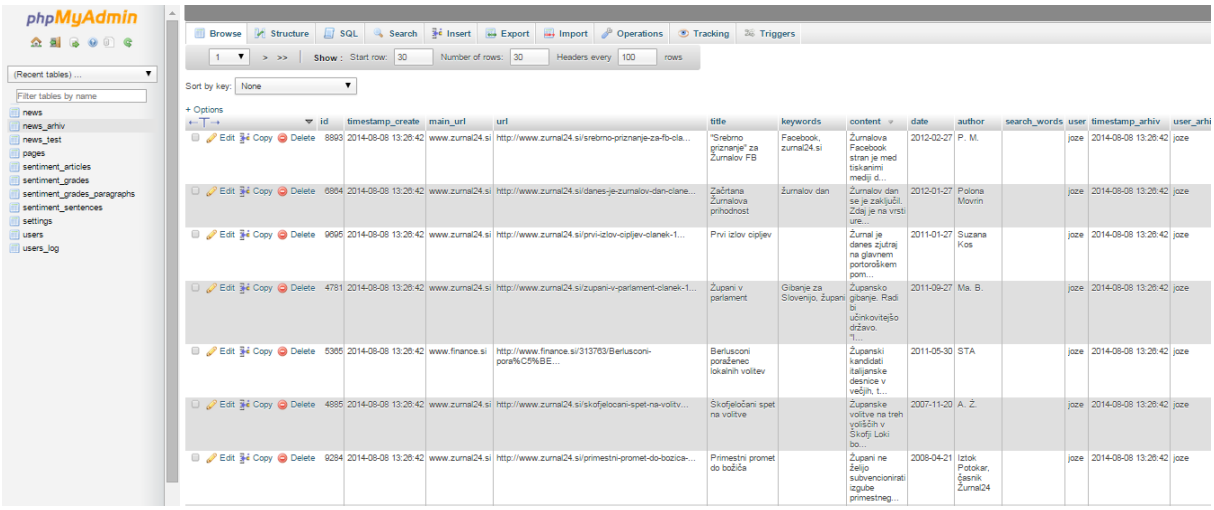
Raziskali smo:

- razvoj modula, ki omogoča avtomatsko zaznavanje in zajemanje spletnega besedila s spletnih virov ter ga zapiše v bazo (zajem HTML kode):

- težava je v striktnem določanju vsebine same objave, saj se spletne strani med seboj močno razlikujejo;
- v večini primerov modul pravilno zajema vsebino, uporabnik pa ima možnost pridobiti celotno vsebino; vsebino lahko tudi ročno popravi in shrani v bazo;
- raziskana je bila funkcija za optimalno prepoznavanje in zajem vsebine;

- ustreznost (relacijske) podatkovne baze MYSQL:

- glede na obsežnost raziskave je potrebno v podatkovno bazo shraniti veliko količino podatkov, v MySQL bazo se shranjujejo podatki o zadetkih (datum vnosa v bazo, URL matične spletne strani, URL naslov objave, naslov objave, ključne besede objave, ocene pomembnosti objave);
- zaradi napak v HTML kodi lahko včasih pride do težav, zaradi česar ima uporabnik možnost ročnega vnosa oziroma urejanja;



Slika 2: Podatki o shranjenih člankih v bazi podatkov

- rešitve za dodajanje vpisnega in registracijskega obrazca v spletni modul; izvedba raziskave za dvo-nivojski dostop (uporabnik in administrator);

- razvoj spletnega modula, ki omogoča:

- pregled najdenih zadetkov v svojem zavihku, shranjevanje zadetkov v bazo posamično ali glede na izbor označenih zadetkov;
- arhiviranje in brisanje zadetkov z možnostjo tabelarnega pregledovanja, iskanja po shranjenih zadetkih, urejanja shranjenih zadetkov in njihovih vsebin;
- ugotavljanje sentimenta v povezavi z metodami strojnega učenja.

Razvili smo tudi paket funkcij v okviru Weka, ki se poveže z rezultati spletne aplikacije, in ugotavlja pozitiven oziroma negativen sentiment v novih spletnih besedilih v slovenskem jeziku.

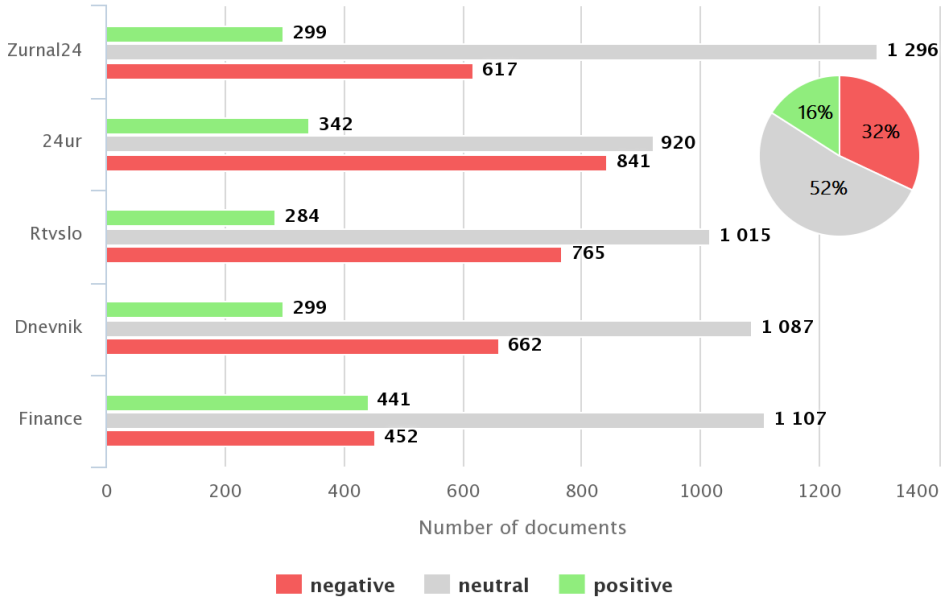
2.2 2. FAZA: RAZISKAVA SENTIMENTA

V drugi fazi je sledilo:

- raziskovanje filtriranja, čiščenja, predprocesiranja podatkov, oblikovanja označevalnika za slovenski jezik in izdelave korpusa (na podlagi preteklih člankov in novic), ki naj bo pripravljen za uporabo in testiranje metod klasifikacije;

- raziskovanje metod klasifikacije na označenem korpusu, testiranje in analiza učinkovitosti metod klasifikacije:

- označen korpus 10427 spletnih besedil v slovenskem jeziku z gospodarsko, finančno in politično vsebino med 1.9.2007 in 31.12.2013, spletna besedila zajeta z arhiva naslednjih slovenskih spletnih medijev: 24ur, Dnevnik, Finance, Rtvslo, Žurnal;
- sentiment: 52% nevtralen, 32% negativen in 16% pozitiven;
- Izmed petih pogosto uporabljenih klasifikacijskih metod (Naivni bayes, Večrazsežnostni naivni bayes, Metoda podpornih vektorjev, Metoda najbližjih sosedov ter Metoda naključnih gozdo se najbolje obneseta metodi Večrazsežnostni naivni bayes (Naive Bayes Multinomial) z napovedno točnostjo nekaj več kot 92% ter Metoda podpornih vektorjev (Support Vector Machines) z napovedno točnostjo več kot 85%).

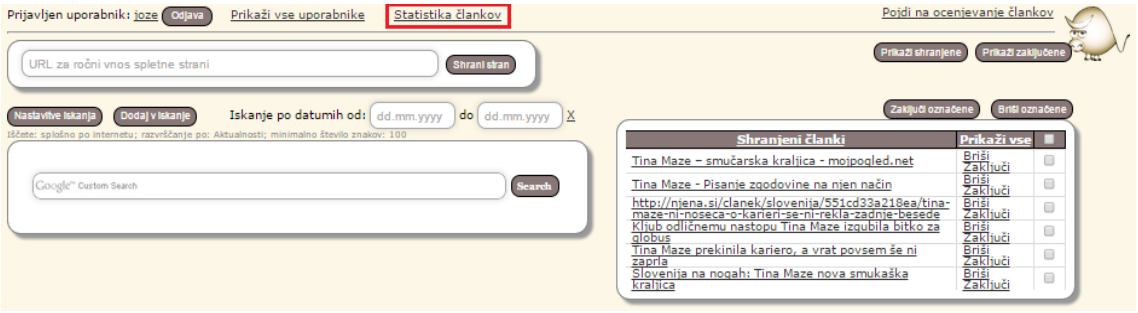


Slika 3: Delež negativnih, nevtralnih, pozitivnih spletnih besedil glede na spletni medij

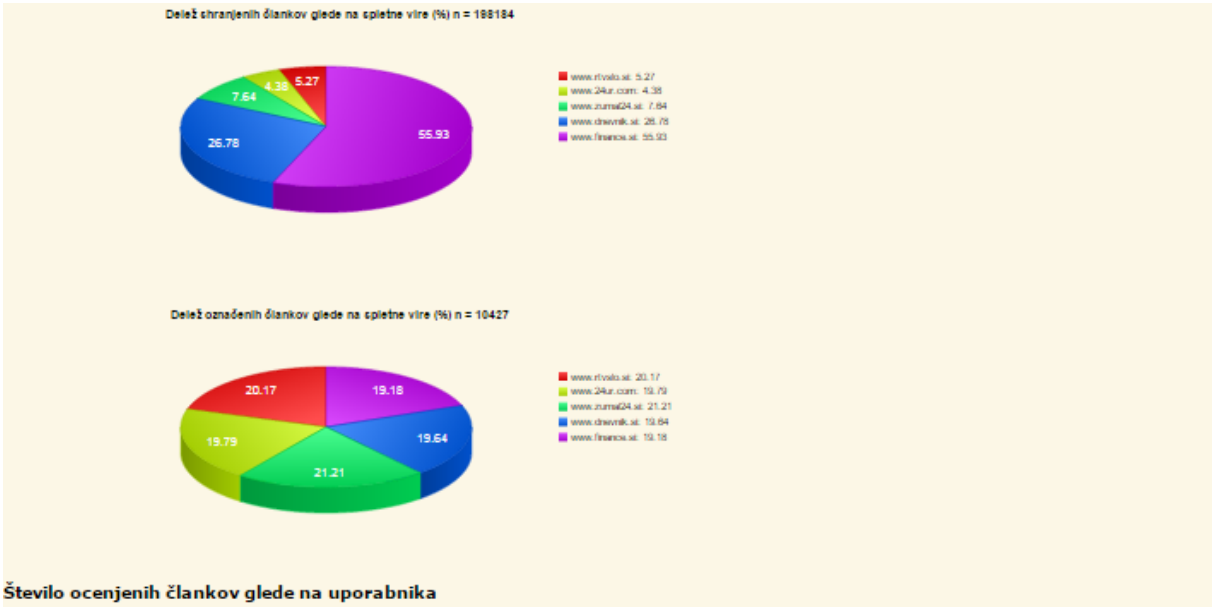
- ugotavljanje sentimenta novih spletnih besedil v povezavi z metodami strojnega učenja (klasifikacija besedil na podlagi metode Večrazsežnosti naivni bayes (Naive Bayes Multinomial), ki se med vsemi testiranimi metodami izkaže tako iz vidika napovedne točnosti modela kot tudi časovne zahtevnosti kot najprimernejša. Modul za ugotavljanje sentimenta novim besedilom je narejen s programskim paketom Weka, ki je pogosto uporabljeno in uveljavljeno orodje med strokovnjaki s področja podatkovnega rudarjenja in strojnega učenja. Zaradi boljše uporabniške izkušnje (predvsem zaradi časovne zahtevnosti algoritmov) smo za potrebe določitve sentimenta novih spletnih besedil implementirali metodo Večrazsežnosti naivni bayes (Naive Bayes Multinomial), ki omogoča točnost napovednega modela 83,2467%.

- raziskovanje možnosti za grafično predstavitev rezultatov in analiz (tabelarična in grafična predstavitev statističnih podatkov pridobljenih in označenih besedil, grafična predstavitev označenega korpusa z oblaki besed).

Slika 4: Oblak besed označenega korpusa s pozitivnim in negativnim sentimentom



Slika 5: Dostop do statistike analiziranih spletnih besedil



Število ocenjenih člankov glede na uporabnika

Uporabnik	Skupaj ocenjenih člankov	Povprečna ocena	1	2	3	4	5	Usklajenost (Cronbach' Alpha)
Gregor Dolinšek	4136	2.64	447	1659	1160	684	186	0.8825
Franja Cimermančič	4147	2.68	419	1528	1300	747	153	0.8757
Martin Vrščaj	949	2.92	10	277	450	199	13	0.8122
Maja Jurman	4157	2.77	318	1604	1125	953	157	0.8723
Niko Brgan	4157	2.99	56	766	2538	774	23	0.9402
Petra Mihalič	4195	2.72	132	1672	1732	563	96	0.8469
Skupaj	21741	2.79	1382	7506	8305	3920	628	0.8716

Število člankov glede na spletne vire

Spletni medij	Število vseh pridobljenih člankov	Velikost vzorca	Povprečna ocena	-	0	+
www.rtvslo.si	10453	2064	2.7	765	1015	284
www.24ur.com	8671	2103	2.68	841	920	342
www.zurnal24.si	15145	2212	2.83	617	1296	299
www.dnevnik.si	53066	2048	2.8	662	1087	299
www.finance.si	110849	2000	2.99	452	1107	441
Skupaj	198184	10427	2.80	3337	5425	1665

Slika 6: Statistika analiziranih spletnih besedil

3. PROBLEMI PRI IZVEDBI RAZISKAVE

Problemi pri izvedbi raziskave:

- prilagajanje in implementacija rešitev v Google Custom Search Engine glede na potrebe raziskave:
  - vključitev iskanja znotraj časovnega okna (tako vključitev polj kot vgradnja koledarja);
  - problematika z Google Custom Search Engine, saj določeni zadetki, ki so bili v določenem trenutku najdeni s strani Google Custom Search Engine, so že lahko zastarali (niso več aktualni oziroma čez čas preprosto niso več razpoložljivi na spletu, čeprav jih je Google zaznal), → rešitev: raziskovanje vključitve iskanja znotraj časovnega okna;
  - dodajanje parametrov v nastavitvah iskanja;
  - dodajanje možnosti shranjevanja vsakega posameznega zadetka znotraj okna z vrnjenimi zadetki;
  - skripte Google Custom Search Engina se občasno spremenijo, zato je potrebno včasih programsko kodo temu prilagoditi;
  - kompromisi glede želenih rešitev in omejenostjo podpornih funkcionalnosti;

- težave pri razvoju ustrezne rešitve za pridobivanje vsebine objav iz HTML kode, ker se spletne strani med seboj močno razlikujejo, je bilo potrebno dodatno raziskati možnosti za razvoj ustreznih funkcij za optimalno prepoznavanje in posledično za zajem vsebine;

- izkazalo se je, da napake v HTML kodi lahko povzročijo težave pri pridobivanju in zajemu vsebine, zato je bilo potrebno razviti rešitev, ki bi omogočila možnost ročnega vnosa oziroma urejanja s strani uporabnika;

- razčlenjevanje vsebine objav na stavke in shranjevanje teh segmentov v bazo (razčlenitev vsebine na stavke je ključna, ocenjevanje sentimenta segmentov (stavkov) nam omogoča bolj realno in verodostojno oceno sentimenta posameznim besedam, s čimer pa eksponentno raste kapaciteta podatkov v bazi);

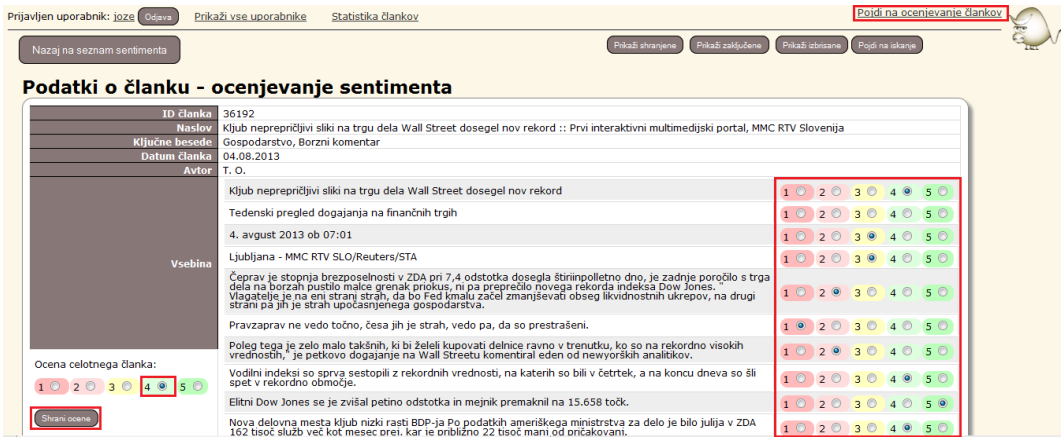
- težave pri pridobitvi korpusa besedil (za uporabo in testiranje klasifikacijskih metod in njihovo analizo učinkovitosti je potrebno pridobiti ocene vsaj nekaj tisoč člankov za verodostojne rezultate, časovno izjemno zamudno);

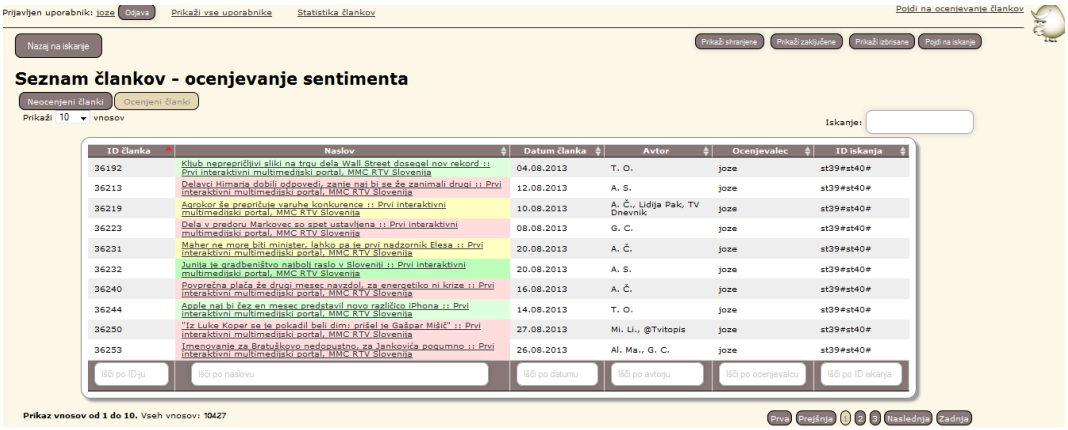
- podpora portala v različnih spletnih brskalnikih;

- težave z napadom na strežnik, omejitev funkcionalnosti, s časom smo izboljšali varnostne rešitve s čimer nam je uspelo zaščititi izvajanje določenih skript na strežniku;

- kompleksnost ureditve in gradnje MySQL baze, za kar je bilo potrebno izvesti nadaljnje raziskave o možnostih izbrane podatkovne baze.

Vsi problemi pri izvedbi raziskave so bili ves čas pod nadzorom, za vse probleme smo našli točno določene in konkretne rešitve.





Slika 7: Označevanje spletnih besedil

4. REZULTATI RAZISKAVE

Dosedanji rezultati so v veliki meri razvidni že v točki 1 (DELO IN AKTIVNOSTI). Uspelo nam je uspelo razviti načrtovan modul, ki omogoča zajem in shranjevanje spletnih besedil v MySQL bazo na podlagi iskanih kriterijev.

Razvili / izdelali smo:

- modul za avtomatsko prepoznavanje in zajem besedila s spletnih virov ter zapis v bazo:

- iskanje se izvrši glede na iskane kriterije (Slika 9):
  - spletne strani (izbor med vgrajenim naborom spletnih strani ali vnos novega nabora spletnih strani ter izvedba iskanja:
    - splošno po spletnih straneh;
    - po naboru iz baze;
    - po posamezni spletni strani;
    - možnost razvrščanja zadetkov po aktualnosti ali času objave;
    - določitev minimalnega števila znakov, znotraj katerega se izvede iskanje iskanega niza);
  - časovni okvir:
    - splošno iskanje brez določitve časovnega okvirja;
    - ročni vnos datuma iskanja preko tipkovnice;
    - vnos datuma iskanja s klikanjem in izbiro v vgrajenem koledarju;
  - ključne besede (vnos ključnih besed ali nizov ključnih besed na podlagi katerih se iskanje izvrši);
- razvoj funkcij za optimalno prepoznavanje in zajem iz HTML kode:

  - razvoj univerzalnega »parserja« za vse spletne strani;
  - razvoj prilagojenih »parserjev« za določen nabor spletnih strani za bolj izpopolnjeno pridobivanje vsebine objav;
- možnost pregleda in izbora enega ali več zadetkov ter možnost shranjevanja vsakega posameznega zadetka znotraj okna z vrnjenimi zadetki, čemur sledi zajem zadetkov ter zapis vsebin v bazo (Slika 10);
- možnost urejanja shranjenih zadetkov (Slika 11);

- v MSQL bazo se shranjujejo podatki o zadetkih (ID članka, datum vnosa v bazo, matična stran, povezava, ID iskanja, nastavitve iskanja, iskani niz, naslov, ključne besede, datum članka, avtor, pomembnost objave (TR rank in GL rank, vir: www.alexa.com), ime vnašalca vsebine) (Slika 2):

- pomembnost objave je določena na podlagi dveh meril:
  - TR rank – ocena spletne strani po priljubljenosti v državi (kombinacija povprečnih dnevnih obiskovalcev na tej spletni strani uporabnikov iz te države v preteklem mesecu, spletna stran z najvišjo kombinacijo obiskovalcev in ogledov strani v državi je uvrščena na 1. mesto);
  - GL rank – podobno kot TR rank, samo da je na globalni ravni, ocena posamezne spletne strani po (globalni) priljubljenosti v zadnjih 3 mesecih;
- možnost urejanja, shranjevanja in zaključevanja sprememb v bazo vnesenih podatkov;

- baze podatkov:

- objav: člankov, arhiviranih člankov, zbrisanih člankov;
- spletnih strani: spletnih strani in zbrisanih spletnih strani;
- iskanih zadetkov;
- sentimenta: sentimenta celotnega članka, sentiment razčlenitve po stavkih;
- uporabnikov: podatki o uporabnikih, podatki o času vpisa v portal;
- nastavitev;

- vpisni (angl. »login«) in registracijski obrazec;

- dvo-nivojski dostop (uporabnik in administrator), administrator ima vpogled v seznam uporabnikov ter njihovo zgodovino in statistiko, lahko arhivira in ureja objave, …;

- gumb za odjavo uporabnika ter avtomatska odjava po daljši odsotnosti;

- napredno iskanje po bazi (splošno iskanje, iskanje po kategorijah), tabelarično pregledovanje, izbor števila prikazanih vnosov, sortiranje glede na podatke o zadetkih;

- razčlenjevanje vsebine na stavke in shranjevanje teh segmentov v bazo;

- filtriranje, čiščenje, predprocesiranje podatkov, oblikovanja označevalnika za slovenski jezik, trenutno smo v fazi ocenjevanja člankov ter izdelave korpusa (na podlagi preteklih člankov in novic), ki lahko služi za uporabo in testiranje metod klasifikacije ter za analizo učinkovitosti metod klasifikacije;

- modul ocenjevanje člankov, 2 zavihka (Slika 7):

- neocenjeni članki;
- ocenjeni članki;

- pet-nivojsko ocenjevanje, grafično opremljeno z barvami, vsak v svoji barvi:

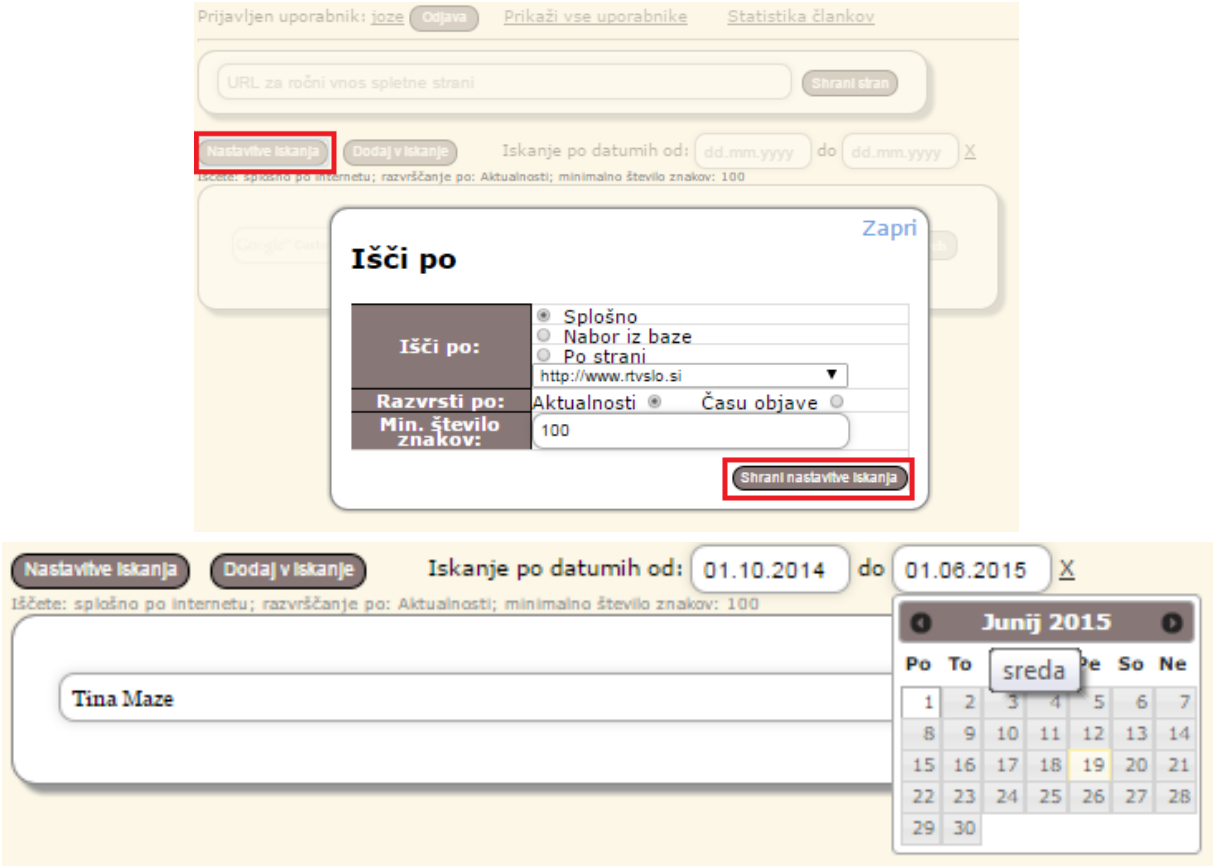
- 1 - zelo negativno: intenzivno rdeča;
- 2 - negativno: rdeča;
- 3 - nevtralno: rumena;
- 4 - pozitivno: zelena;
- 5 - zelo pozitivno: intenzivno zelena;

- statistika člankov ter njihova grafična predstavitev (Slika 6).

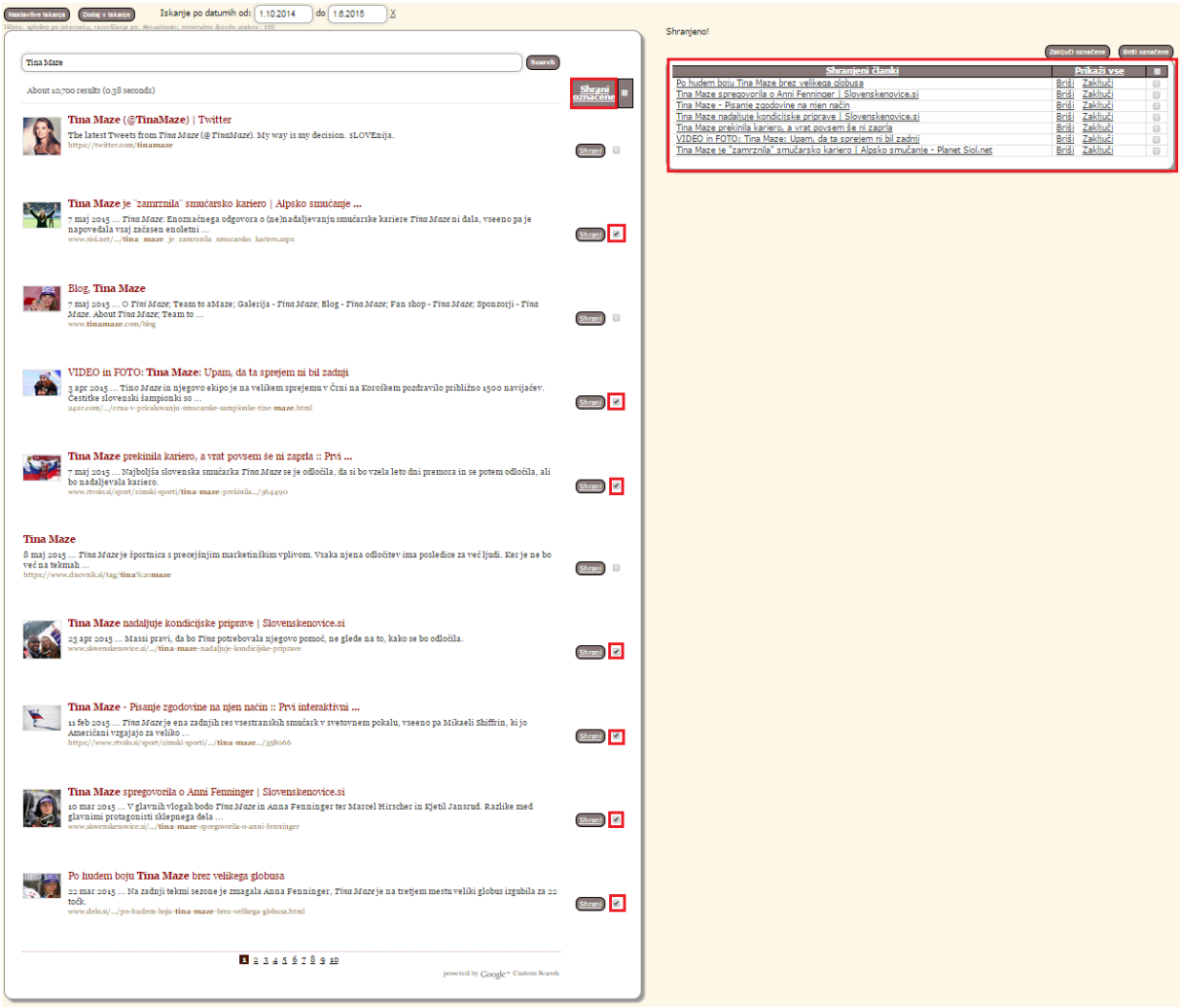
Spletni modul je na podlagi raziskav in izvedenih rešitev mogoče zasnovati tako, da ga je možno izpopolnjevati ter dograjevati.

5 NAVODILA ZA UPORABO SPLETNEGA MODULA

Slika 8: Prijava v spletni modul

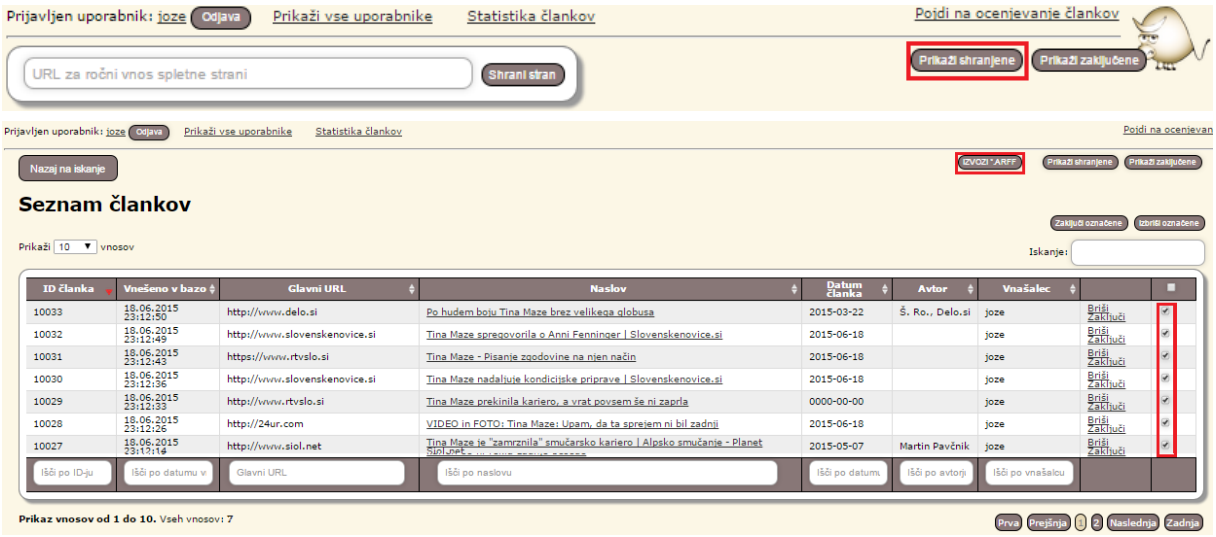


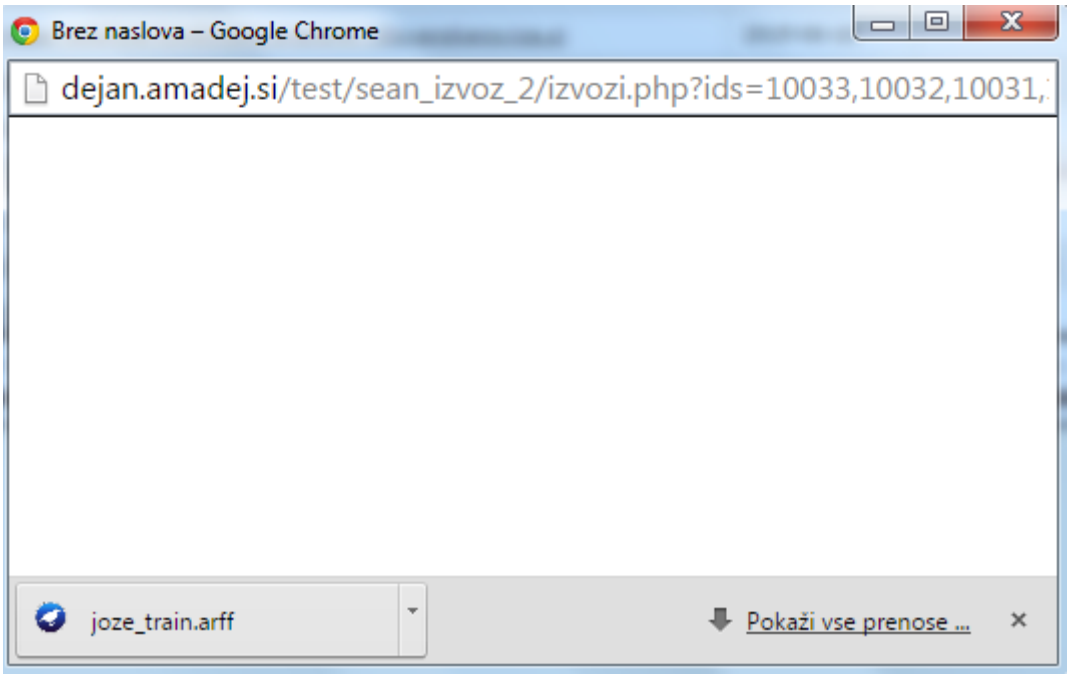
Slika 9: Nastavitve iskanja in iskanje glede na iskane parametre



Slika 10: Pregledovanje in zajem zadetkov

Slika 11: Urejanje in shranjevanje zadetkov





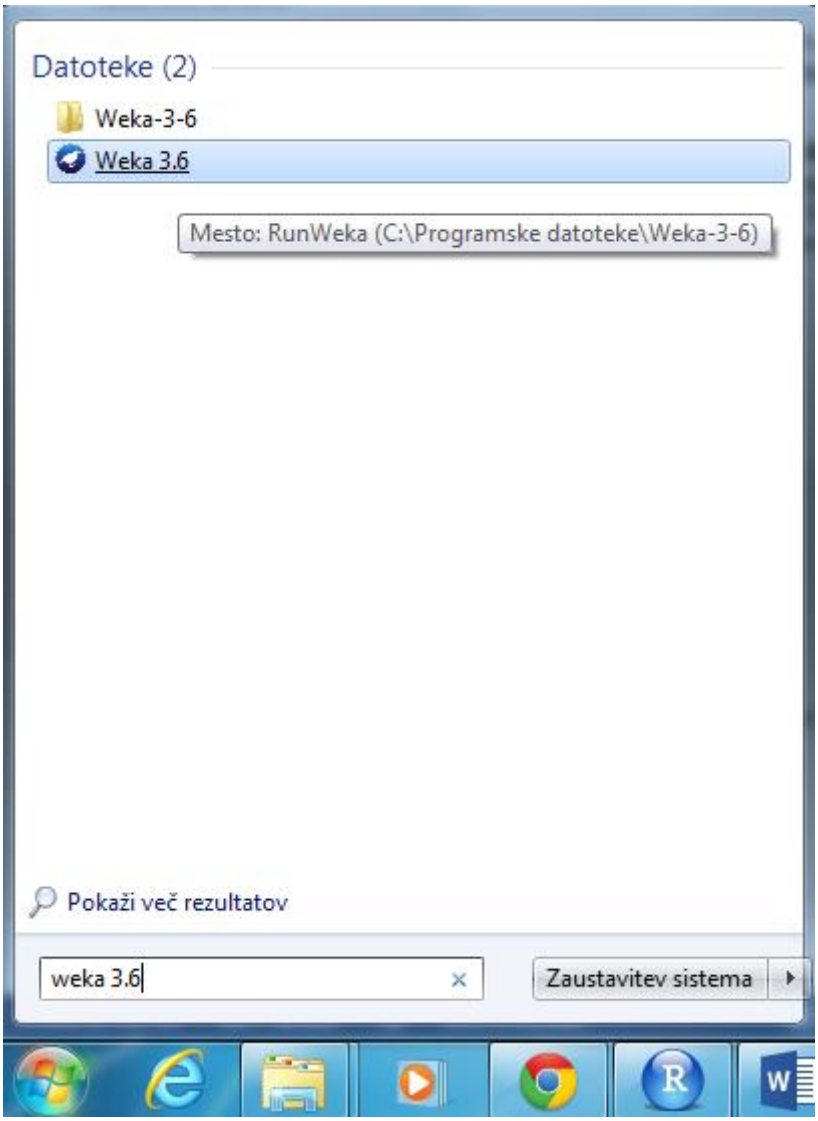
Slika 12: Izvoz spletnih besedil za določitev sentimenta novih spletnih besedil

6 NAVODILA ZA DOLOČITEV SENTIMENTA NOVIH SPLETNIH BESEDIL V PROGRAMU WEKA

Sistemske zahteve:

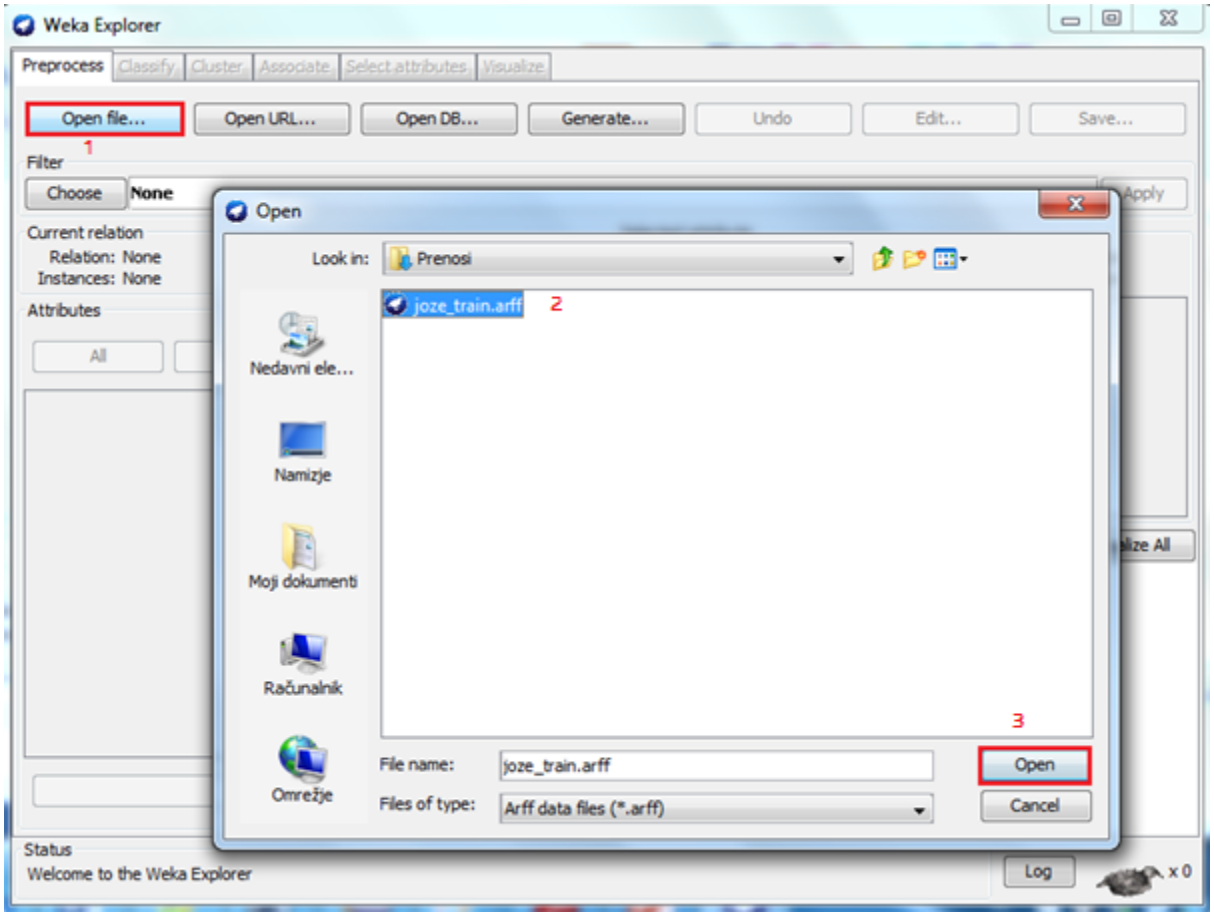
- potrebno je naložiti in inštalirati zadnjo verzijo programa Weka, ki ustreza vašemu operacijskemu sistemu, povezava do programske opreme, katero inštalirate na vaš računalnik: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

- v kolikor želite boljšo funkcionalnost programa, odprite datoteko **RunWeka.ini** kot administrator, datoteka se nahaja v delovnem prostoru, kamor ste inštalirali program (ponavadi je privzeta pot podobna C:\Program Files\Weka-3-6), in nastavite maxheap iz 512MB na vrednost, ki ustreza delovnemu spominu – RAM-u vašega računalnika, npr. **maxheap=4096MB** v kolikor imate 4GB RAM-a; nastavite tudi kodiranje na **fileEncoding=utf-8**)



Slika 13: Zagon programa Weka (za WIN OS)

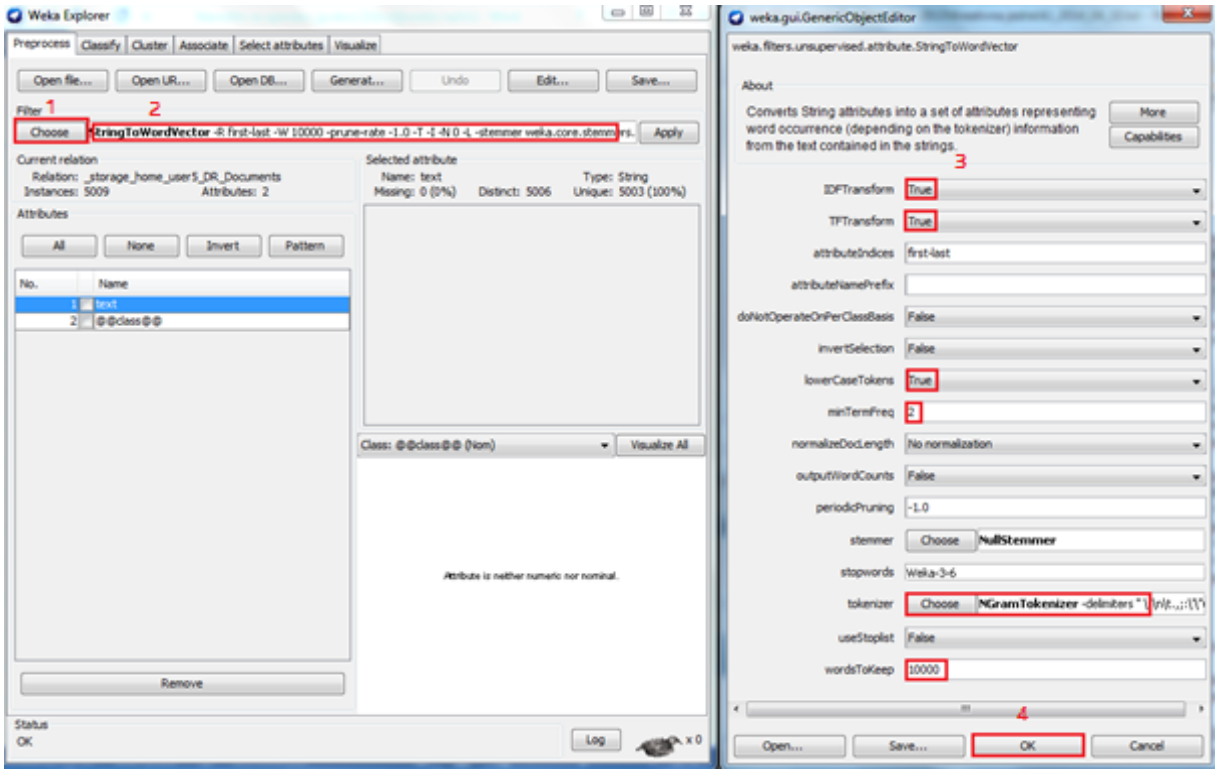
Slika 14: Izbira osnovnega modula programa Weka – Explorer

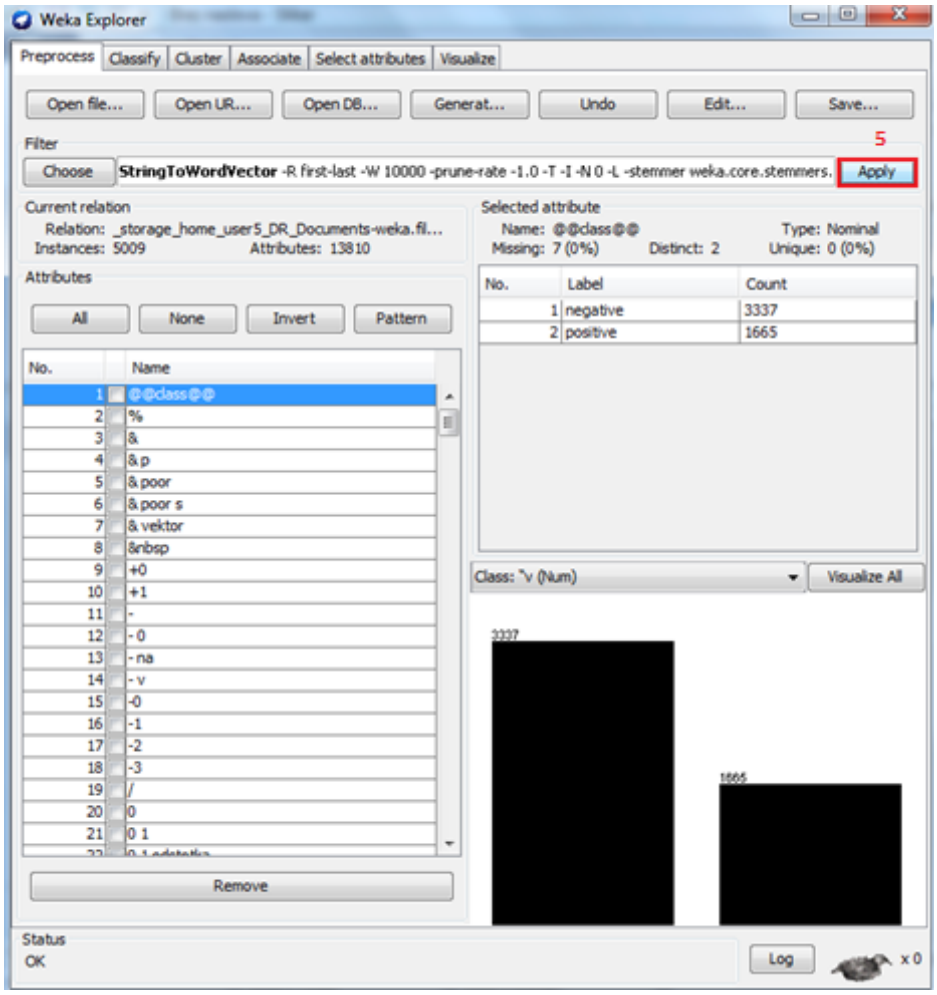


Slika 15: Uvoz vhodne datoteke (določitev poti do vhodne datoteke joze_train.arff)

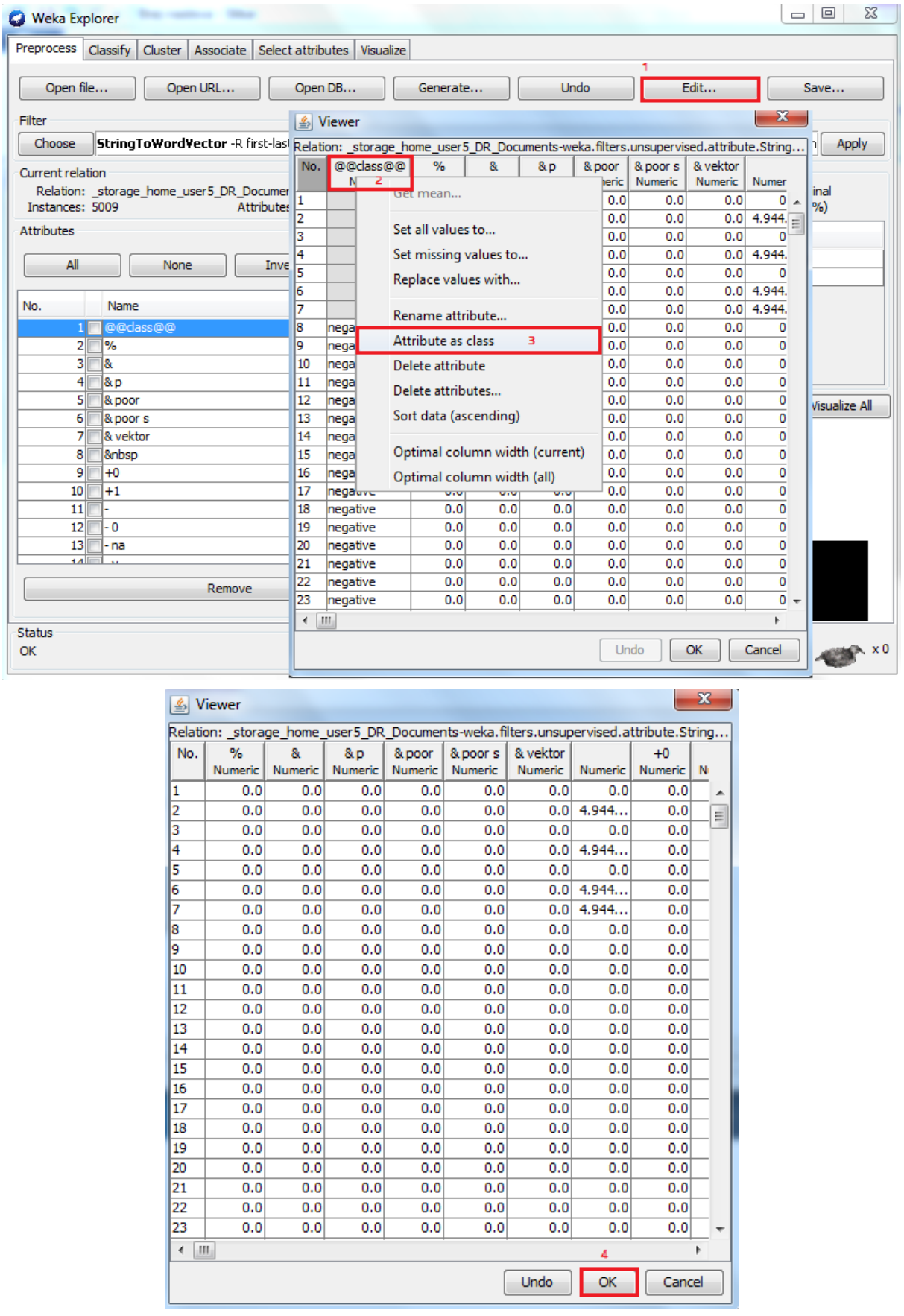
Izbira filtra

Filter-> filters -> unsupervised -> attribute -> StringToWordVector - > kliknete z levim gumbom na miški na filter da lahko urejate lastnosti filtra (IDFTransform=True, TFTransform=True, lowerCaseTokens=True,minTermFreq=2, tokenizer=NGramTokenizer (min=1,max=3), wordsToKeep=10000) (traja ~ 30s) -> Apply

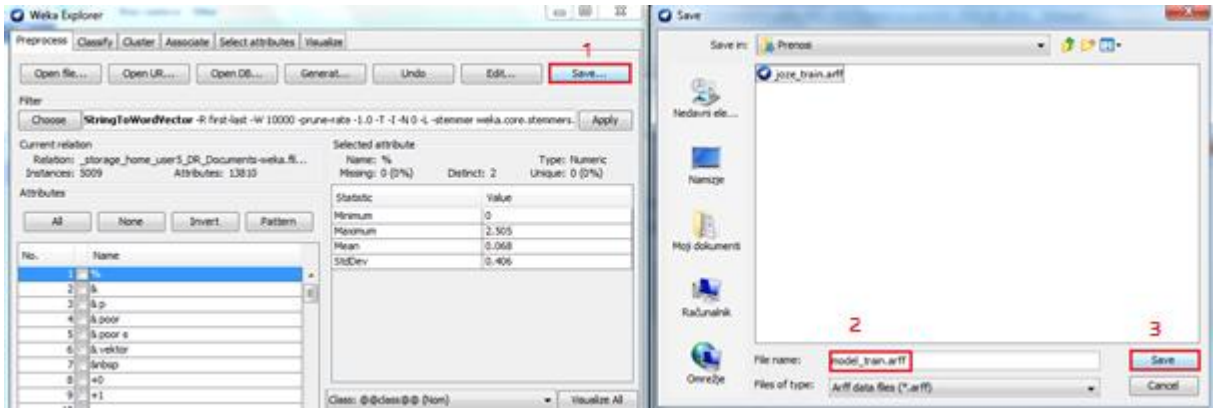




Slika 16: Izbira ustreznega filtra

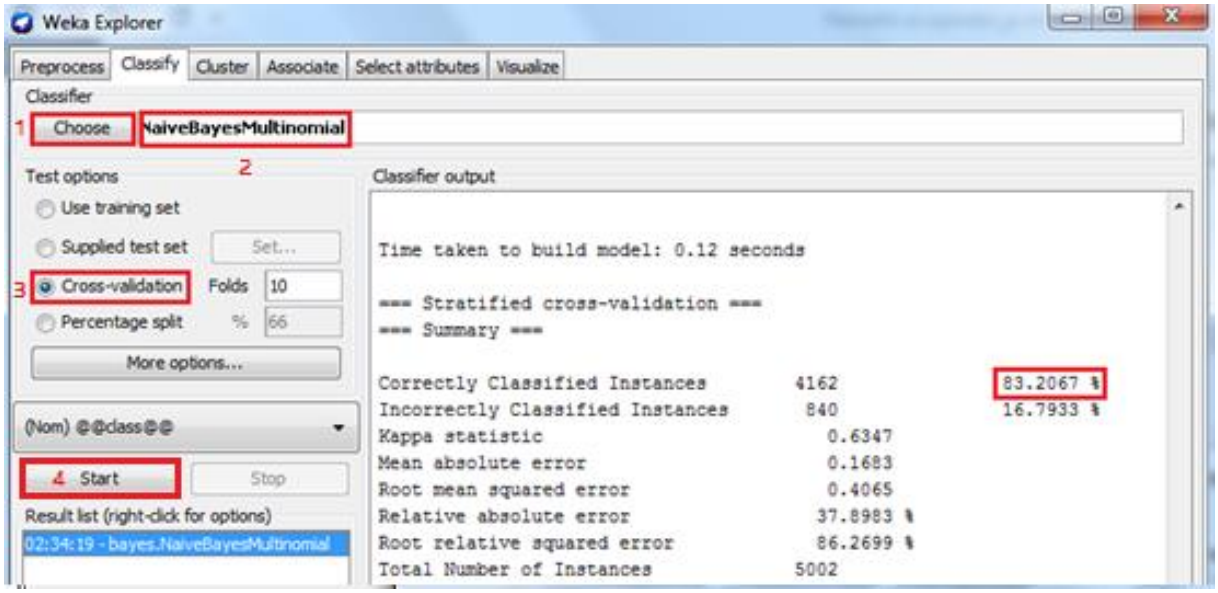


Slika 17: Izbira ustreznega atributa kot ciljnega razreda v postopku določitve sentimenta novih spletnih besedil



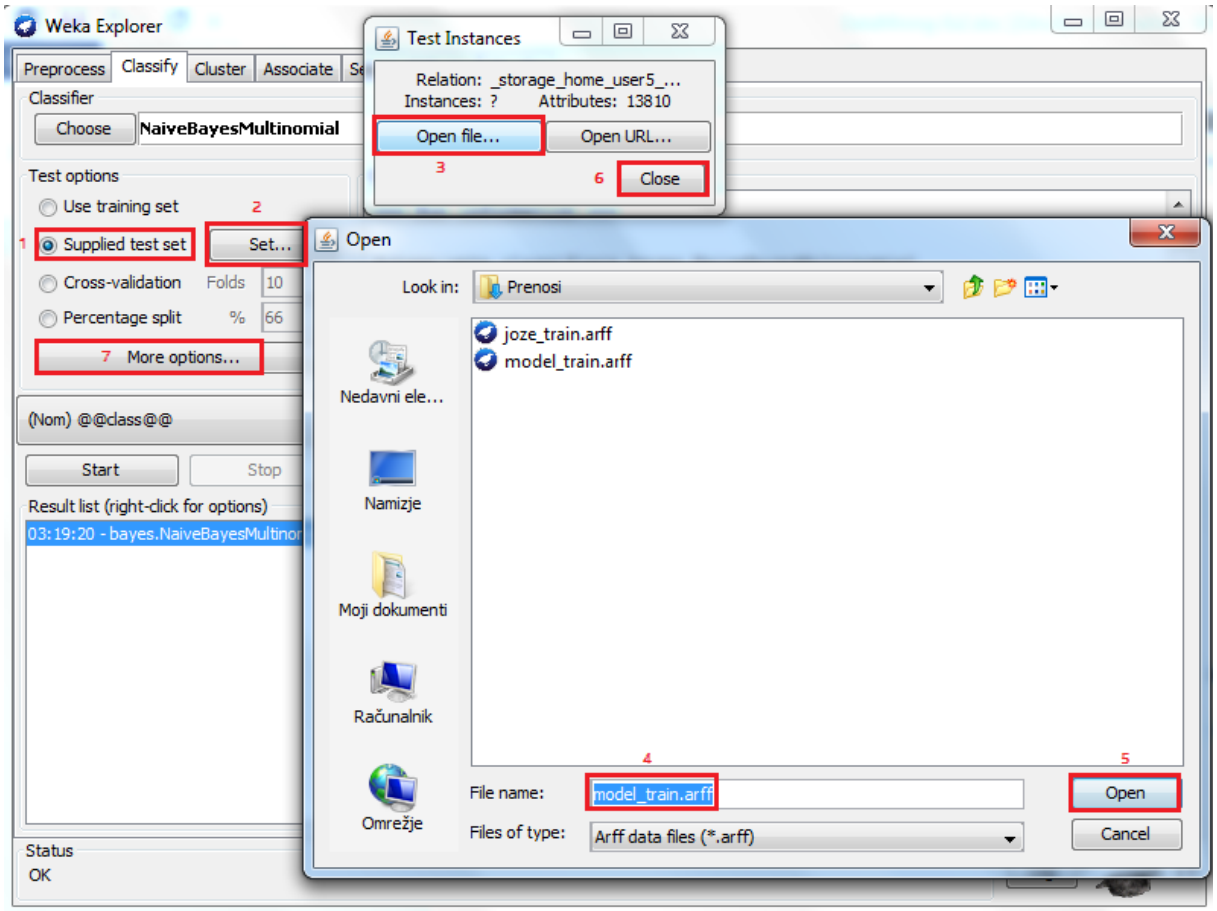
Slika 18: Shranjevanje modela (model_train.arff)

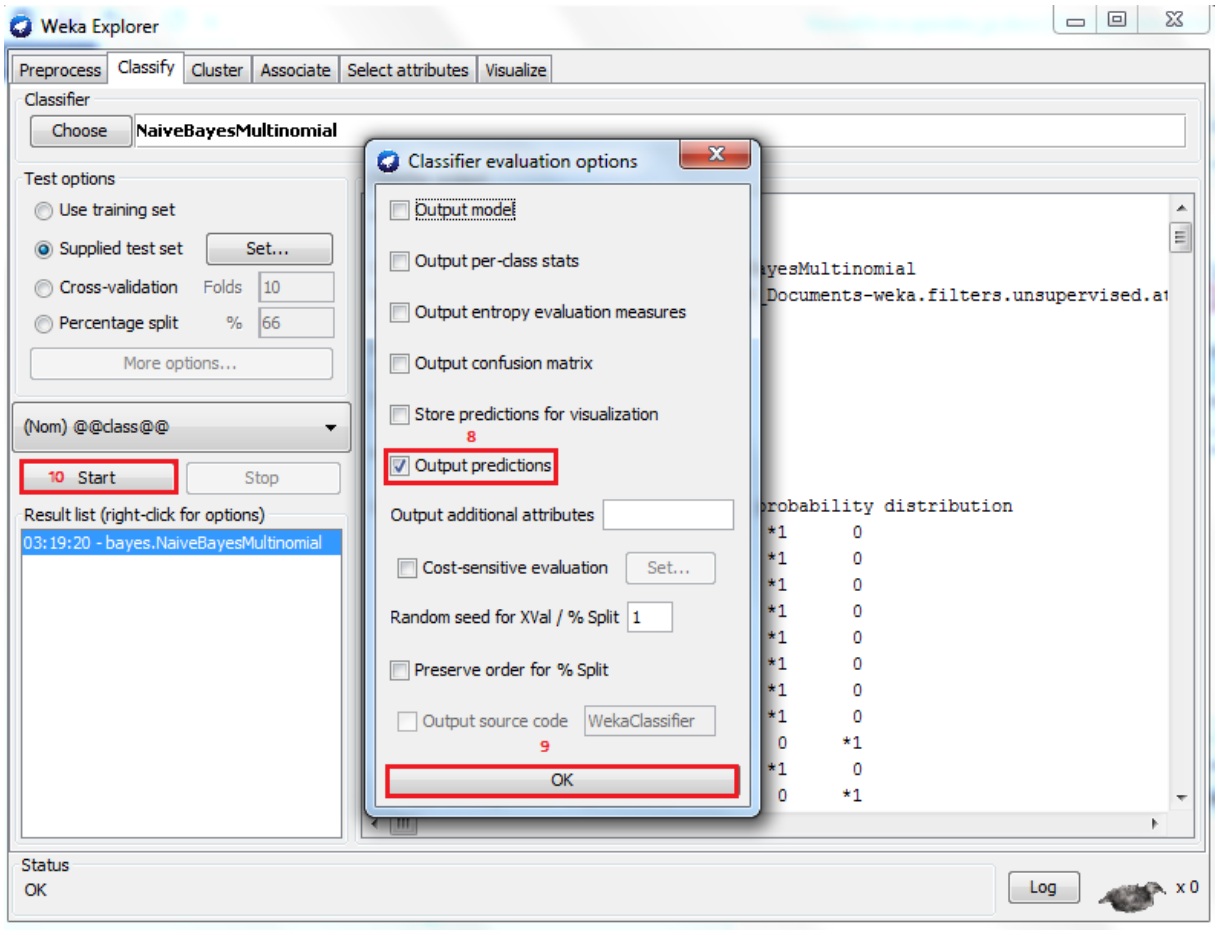
Če želite preveriti točnost napovednega modela, kliknete na zavihek Classify -> izberete klasifikator s klikom na gumb Choose -> classifiers -> bayes -> NaiveBayesMultinomial -> pod Test options izberete Cross-validation -> Start (traja ~5s) -> v oknu z rezultati (Classifier output) lahko opazite točnost napovednega modela 83.2467%

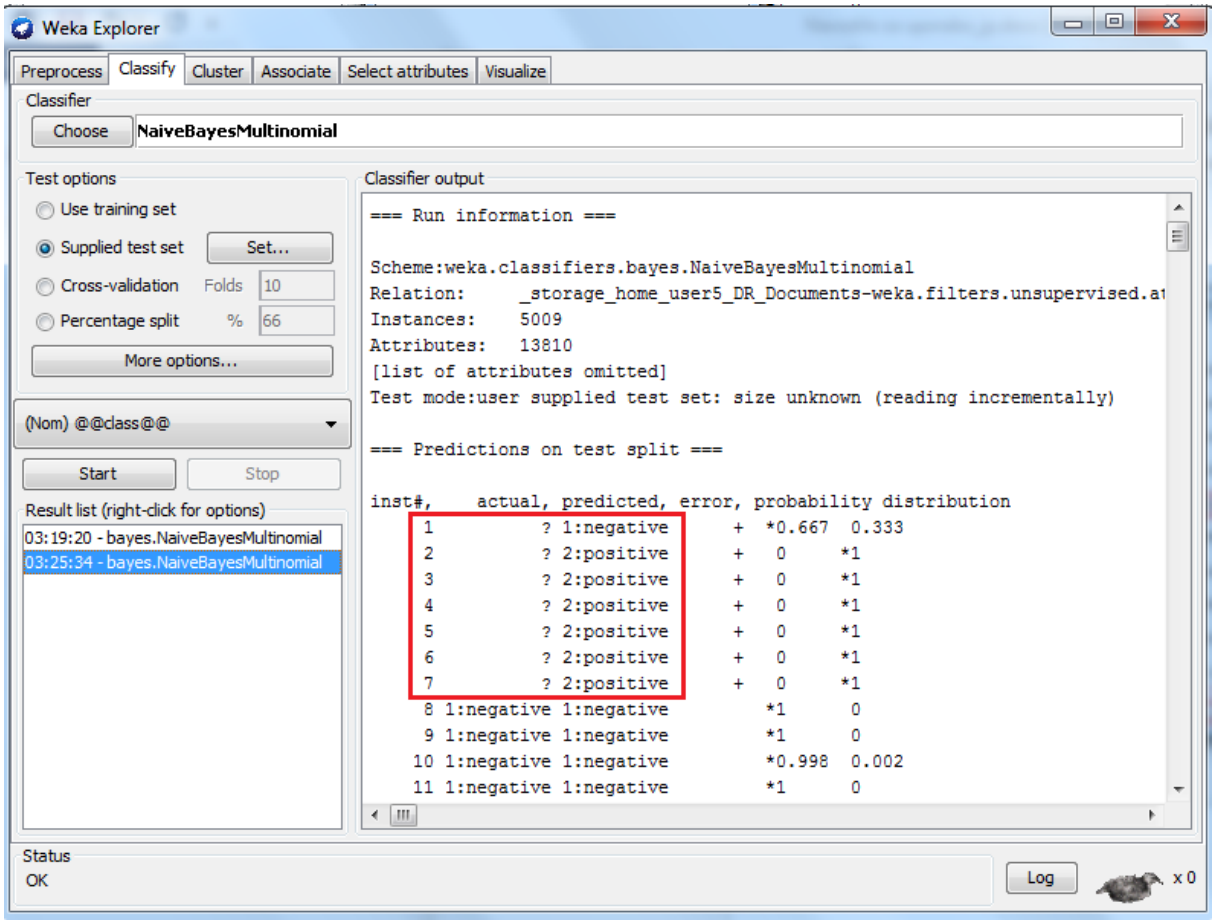


Slika 19: Napovedna točnost modela določena z metodo Večrazsežnostni naivni bayes (Naive Bayes Multinomial)

Za določanje sentimenta novim besedilom: kliknete na zavihek Classify -> izberete klasifikator s klikom na gumb Choose -> bayes -> NaiveBayesMultinomial -> pod Test options izberete Supplied test set -> Set -> Open file... -> izberete datoteko model_train.arff -> kliknete na gumb More options -> označite samo možnost Output predictions, vse ostale možnosti odznačite -> OK -> Start -> v oknu z rezultati (Classifier output) lahko pod === Predictions on test split ===, kjer so ? lahko opazite napoved ocene sentimenta izbranih besedil (besedila so nanizana v enakem vrstnem redu kot so shranjena v aplikaciji).


Weka Explorer
Preprocess
Classify
Cluster
Associate
Classifier
Choose
NaiveBayesMultinomial
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) @@class@@
Start
Stop
Result list (right-click for options)
03:19:20 - bayes.NaiveBayesMultinor
Status
OK
Test Instances
Relation: _storage_home_user5_...
Instances: ?
Attributes: 13810
Open file...
Open URL...
Close
Open
Look in:
Prenosi
joze_train.arff
model_train.arff
Nedavni ele...
Namizje
Moji dokumenti
Računalnik
Omrežje
File name:
model_train.arff
Files of type:
Arff data files (*.arff)
Open
Cancel


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
NaiveBayesMultinomial
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) @@class@@
Start
Stop
Result list (right-click for options)
03:19:20 - bayes.NaiveBayesMultinomial
Classifier evaluation options
Output model
Output per-class stats
Output entropy evaluation measures
Output confusion matrix
Store predictions for visualization
Output predictions
Output additional attributes
Cost-sensitive evaluation
Set...
Random seed for XVal / % Split
1
Preserve order for % Split
Output source code
WekaClassifier
OK
Status
OK
Log
x 0



Slika 20: Določitev sentimenta novih spletnih besedil