

---

# SPLETNA APLIKACIJA ZA KATEGORIZACIJO IN SIMULACIJO DOKUMENTNIH TOKOV

---

NAVODILA ZA UPORABO



JUNIJ 2015



## Kazalo vsebine

<b>1 UVOD .....</b>	<b>2</b>
<b>2 DELO IN AKTIVNOSTI.....</b>	<b>2</b>
2.1 1. FAZA: RAZISKAVA ZAJEMA SPLETNIH BESEDIL TER ZAPISA V BAZO .....	2
2.2 2. FAZA: RAZISKAVA SENTIMENTA .....	4
<b>3. PROBLEMI PRI IZVEDBI RAZISKAVE .....</b>	<b>6</b>
<b>4. REZULTATI RAZISKAVE .....</b>	<b>7</b>
<b>5 NAVODILA ZA UPORABO SPLETNEGA MODULA.....</b>	<b>9</b>
<b>6 NAVODILA ZA DOLOČITEV SENTIMENTA NOVIH SPLETNIH BESEDIL V PROGRAMU WEKA.....</b>	<b>12</b>

## Kazalo slik

Slika 1: Prijava v spletni modul .....	2
Slika 2: Podatki o shranjenih člankih v bazi podatkov.....	3
Slika 3: Delež negativnih, nevtralnih, pozitivnih spletnih besedil glede na spletni medij.....	4
Slika 4: Oblak besed označenega korpusa s pozitivnim in negativnim sentimentom .....	5
Slika 5: Dostop do statistike analiziranih spletnih besedil.....	5
Slika 6: Statistika analiziranih spletnih besedil .....	5
Slika 7: Označevanje spletnih besedil .....	7
Slika 8: Prijava v spletni modul .....	9
Slika 9: Nastavitve iskanja in iskanje glede na iskane parametre.....	9
Slika 10: Pregledovanje in zajem zadetkov.....	10
Slika 11: Urejanje in shranjevanje zadetkov .....	10
Slika 12: Izvoz spletnih besedil za določitev sentimenta novih spletnih besedil .....	11
Slika 13: Zagon programa Weka (za WIN OS).....	12
Slika 14: Izbira osnovnega modula programa Weka – Explorer .....	13
Slika 15: Uvoz vhodne datoteke (določitev poti do vhodne datoteke joze_train.arff) .....	13
Slika 16: Izbira ustreznega filtra .....	14
Slika 17: Izbira ustreznega atributa kot ciljnega razreda v postopku določitve sentimenta novih spletnih besedil .....	15
Slika 18: Shranjevanje modela (model_train.arff) .....	16
Slika 19: Napovedna točnost modela določena z metodo Večrazsežnostni naivni bayes (Naive Bayes Multinomial) .....	16
Slika 20: Določitev sentimenta novih spletnih besedil.....	18

## 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 besedilih. **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:

- spletna stran – uporabnik lahko izbira med vgrajenim naborom spletnih strani, lahko pa se iskanje izvede splošno po spletnih straneh ali po posamezni spletni strani;
- 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;
- ključna beseda – uporabnik lahko vnaša ključne besede ali niz ključnih besed na podlagi katerih se iskanje izvrši;
- 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;

The screenshot shows the phpMyAdmin interface with a table named 'news' selected. The table contains 10 rows of data, each representing a saved article. The columns include id, timestamp\_create, main\_url, url, title, keywords, content, date, author, search\_words, user, timestamp\_arhiv, and user\_arhiv. The data is sorted by 'id' in descending order.

id	timestamp_create	main_url	url	title	keywords	content	date	author	search_words	user	timestamp_arhiv	user_arhiv
8893	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/srebrno-priznanje-za-fb-cla...	"Srebrno priznanje" za Zurnalov FB	Facebook, zurnal24.si	Zurnalova Facebook stran je med tiskanimi mediji d...	2012-02-27	P. M.		jose	2014-08-08 13:26:42	jose
6584	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/danes-je-zurnalov-dan-clan...	Začrtana Zurnalova prihodnost	Zurnalov dan	Zurnalov dan se je zaključil. Zdej je na vrsti ur...	2012-01-27	Polona Movrin		jose	2014-08-08 13:26:42	jose
6585	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/prvi-izlov-cipjev-clanek-1...	Prvi izlov cipjev		Zurnal je danes zjutraj na glavnem portorškem pom...	2011-01-27	Suzana Kos		jose	2014-08-08 13:26:42	jose
4781	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/zupani-v-parlament-clanek-1...	Zupani v parlament	Gibanje za Slovenjo, zupani	Zupansko gibanje. Radi bi uveljavitejšo državo...	2011-06-27	Ma. B.		jose	2014-08-08 13:26:42	jose
5385	2014-08-08 13:26:42	www.financ.si	http://www.financ.si/313763/Berlusconi-pora%C5%BE...	Berlusconi poraženec lokalnih volitev		Zupanski kandidati italijanske desnice v večjih, t...	2011-05-30	STA		jose	2014-08-08 13:26:42	jose
4885	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/skofjeloanci-spet-na-volitv...	Skofjeločani spet na volitve		Zupanske volitve na treh voliščih v Škofji Loki bo...	2007-11-20	A. Z.		jose	2014-08-08 13:26:42	jose
6284	2014-08-08 13:26:42	www.zurnal24.si	http://www.zurnal24.si/primestri-promet-do-bozica...	Primestri promet do božice		Zupani ne želijo subvencionirati izgube primestneg...	2008-04-21	Iztok Patokar, časnik Zurnal24		jose	2014-08-08 13:26:42	jose

**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 zavijku, 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.

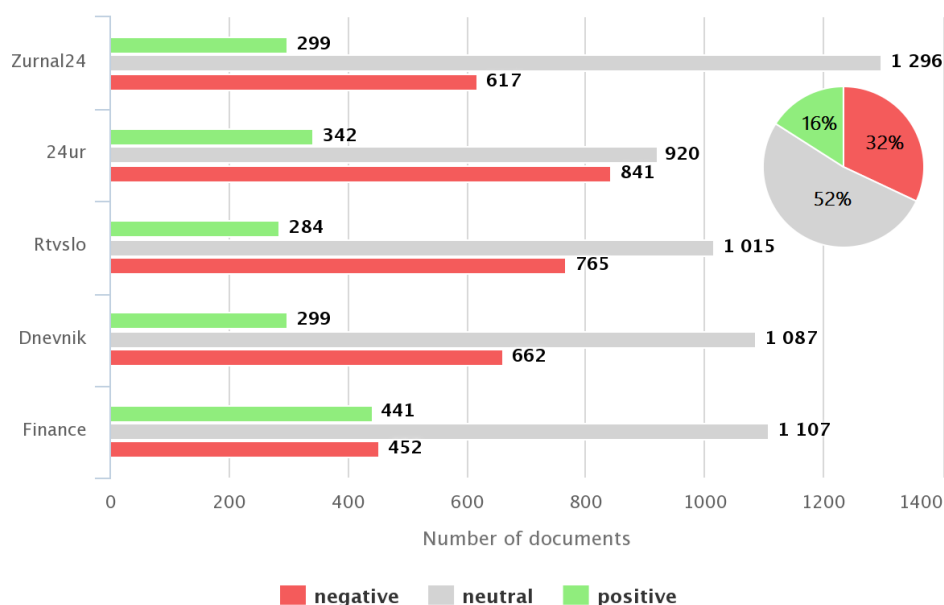
Razvili smo tudi paket funkcij v okviru Weka, ki se poveže z rezultati spletne aplikacije, in ugotavlja pozitiven oziroma negativen sentiment v novih spletnih besedilih v slovenskem jeziku.

## 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, Rtv slo, Žurnal24;
- 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 obnese metoda 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%).



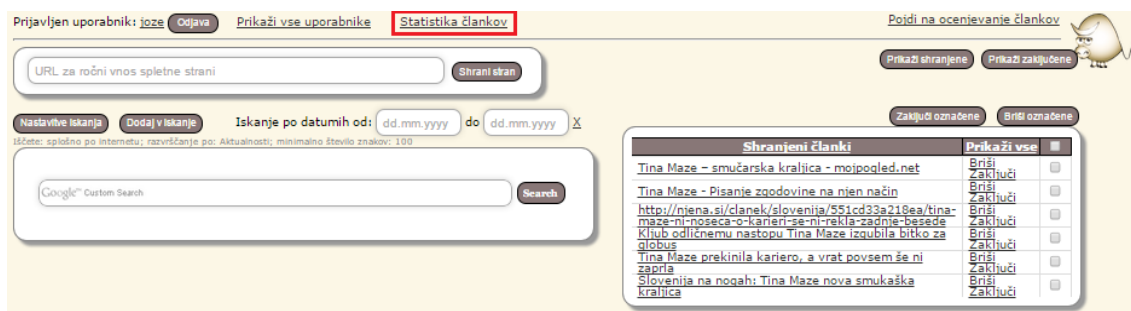
**Slika 3: Delež negativnih, nevtralnih, pozitivnih spletnih besedil glede na spletni medij**

- 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%.

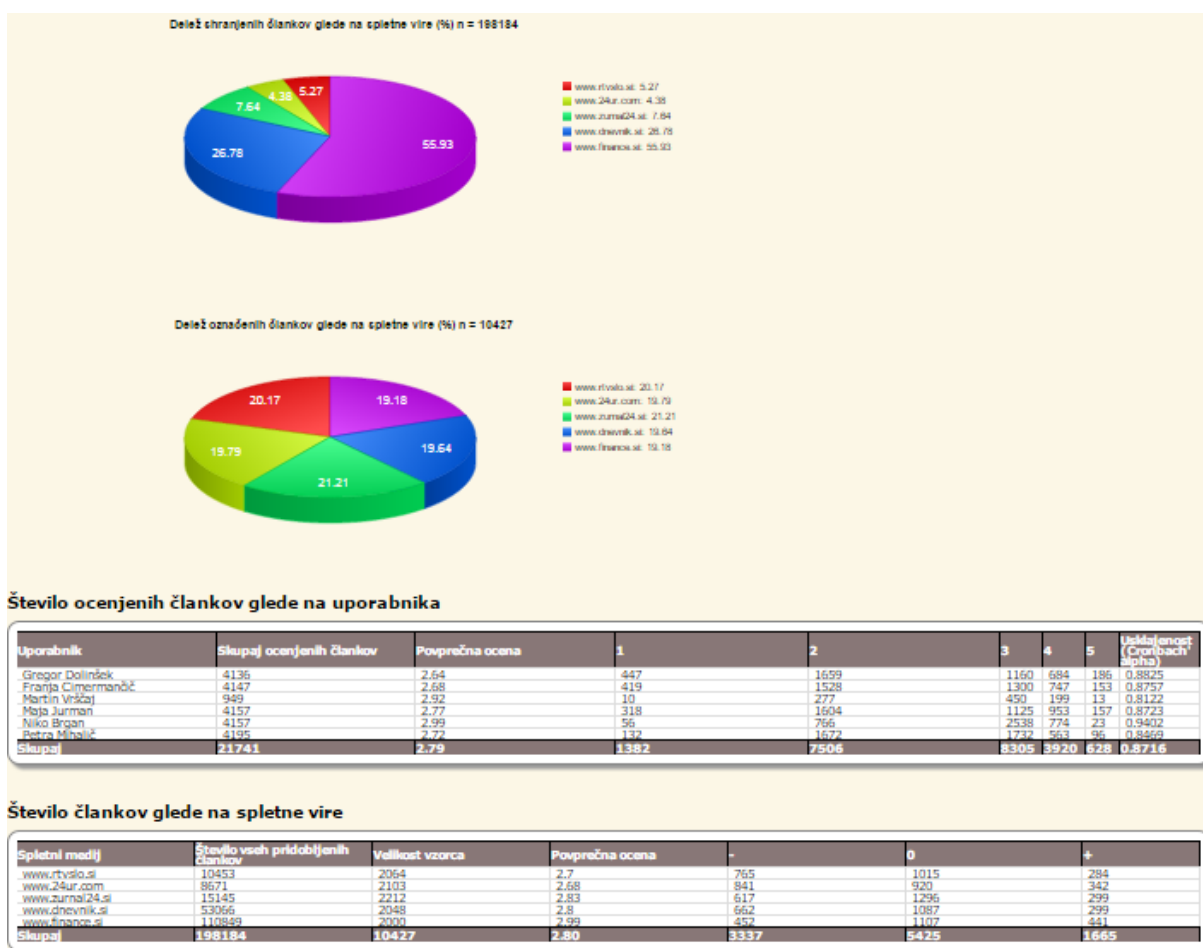
- raziskovanje možnosti za grafično predstavitev rezultatov in analiz (tabelarična in grafična predstavitev statističnih podatkov pridobljenih in označenih besedil, grafična predstavitev označenega korpusa z oblaki besed).



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



Slika 5: Dostop do statistike analiziranih spletnih besedil



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 Engine 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.



Prijavljen uporabnik: joze [Odjava] Prikaži vse uporabnike Statistika člankov [Pojdi na ocenjevanje člankov]

Nazaj na seznam sentimenta [Prikaži shranjene] [Prikaži zaključene] [Prikaži izbrisane] [Pojdi na iskanje]

### Podatki o članku - ocenjevanje sentimenta

ID članka	Naslov	Ključne besede	Datum članka	Avtor
36192	Kljub neprepičljivi sliki na trgu dela Wall Street dosegel nov rekord :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	Gospodarstvo, Borzni komentar	04.08.2013	T. O.

Vsebina

Kljub neprepičljivi sliki na trgu dela Wall Street dosegel nov rekord

Tedenski pregled dogajanja na finančnih trgih

4. avgust 2013 ob 07:01

Ljubljana - MMC RTV SLO/Reuters/STA

Čeprav je stopnja brezposelnosti v ZDA pri 7,4 odstotka dosegla štirinapolletno dno, je zadnje poročilo s trga dela na borzah pustilo malce grenak prokus, ni pa preprečilo novega rekorda indeksa Dow Jones. Vagatelje je na eni strani strah, da bo Fed kmalu začel zmanjševati obseg likvidnostnih ukrepov, na drugi strani pa jih je strah upočasnjenega gospodarstva.

Pravzaprav ne vedo točno, česa jih je strah, vedo pa, da so prestrašeni.

Poleg tega je zelo malo takšnih, ki bi želeli kupovati delnice ravno v trenutku, ko so na rekordno visokih vrednostih, je petkovo dogajanje na Wall Streetu komentiral eden od newyorških analitikov.

Vodilni indeksi so sprva sestopili z rekordnih vrednosti, na katerih so bili v četrtek, a na koncu dneva so šli spet v rekordno območje.

Elitni Dow Jones se je zvišal petino odstotka in mejnik premaknil na 15.658 točk.

Nova delovna mesta kljub nizki rasti BDP-ja Po podatkih ameriškega ministrstva za delo je bilo julija v ZDA 162 tisoč služb več kot mesec prej. kar je približno 22 tisoč manj od pričakovanih.

Ocena celotnega članka:

1 2 3 4 5

[Shrani oceno]

Prijavljen uporabnik: joze [Odjava] Prikaži vse uporabnike Statistika člankov [Pojdi na ocenjevanje člankov]

Nazaj na iskanje [Prikaži shranjene] [Prikaži zaključene] [Prikaži izbrisane] [Pojdi na iskanje]

### Seznam člankov - ocenjevanje sentimenta

[Nesortirani članki] [Ocenjeni članki]

Prikaži: 10 vnosov

Iskanje: [ ]

ID članka	Naslov	Datum članka	Avtor	Ocenjevalec	ID iskanja
36192	Kljub neprepičljivi sliki na trgu dela Wall Street dosegel nov rekord :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	04.08.2013	T. O.	joze	st39#st40#
36213	Palatki Himanta dobili odgovore, zanje naj bi se že zanimali drugi :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	12.08.2013	A. S.	joze	st39#st40#
36219	Anglor ša preizkuša vanilja konkurence :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	10.08.2013	A. Č., Lidija Pak, TV Dnevnik	joze	st39#st40#
36223	Dela v srednjem Makovcu so spet ustavljena :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	08.08.2013	G. C.	joze	st39#st40#
36231	Maher ne more biti minister, lahko pa je prvi nadzornik Elasa :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	20.08.2013	A. Č.	joze	st39#st40#
36232	Junija je gradbeništvo najbolj raslo v Sloveniji :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	20.08.2013	A. S.	joze	st39#st40#
36240	Povprečna plaša že drugi mesec navzdol, za energetiko ni krize :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	16.08.2013	A. Č.	joze	st39#st40#
36244	Apple naj bi leten mesec predstavil novo različico iPhone :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	14.08.2013	T. O.	joze	st39#st40#
36250	Tri Luke Koper se je naskladil ball dimi orisali je Galagar Miksi :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	27.08.2013	Mi. Li., @Tvitopis	joze	st39#st40#
36253	Imenovanje za Bratjskovo nedopustno, za Jankovića posumno :: Prvi interaktivni multimedijški portal, MMC RTV Slovenija	26.08.2013	Al. Ma., G. C.	joze	st39#st40#

[Vrni po ID-ju] [Vrni po naslovu] [Vrni po datumu] [Vrni po avtorju] [Vrni po ocenjevalcu] [Vrni po ID iskanja]

Prikaz vnosov od 1 do 10. Vseh vnosov: 10427

[Prejeto] [Prejeto] [Prejeto] [Prejeto] [Prejeto]

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 iskanjih 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:

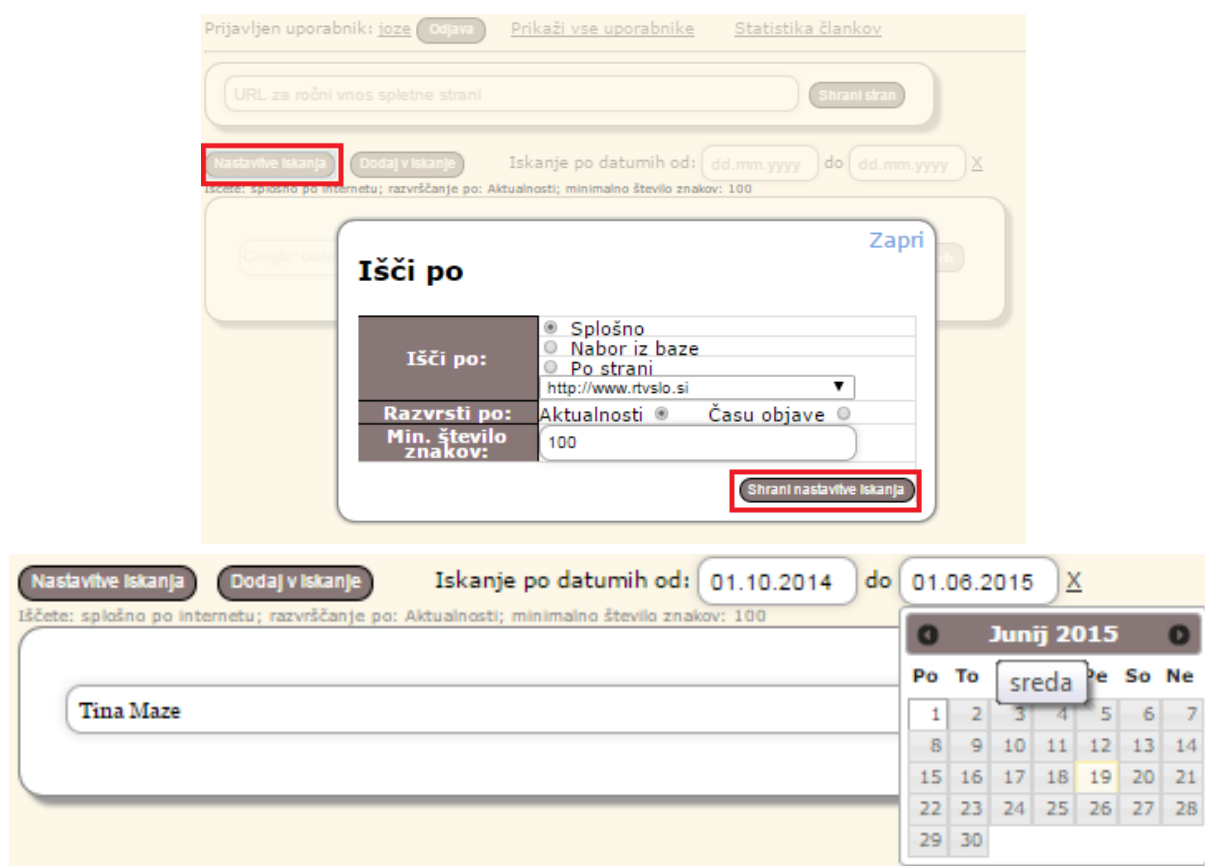


- razvoj univerzalnega »parserja« za vse spletne strani;
  - razvoj prilagojenih »parserjev« za določen nabor spletnih strani za bolj izpopolnjeno pridobivanje vsebine objav;
  - možnost pregleda in izbora enega ali več zadetkov ter možnost shranjevanja vsakega posameznega zadetka znotraj okna z vrnjenimi 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](http://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 dnevniških 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;
  - iskanjih zadetkov;
  - sentimenta: sentimenta celotnega članka, sentiment razčlenitve po stavkih;
  - uporabnikov: podatki o uporabnikih, podatki o času vpisa v portal;
  - nastavitve;
- 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 - nevtrarno: 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



Iskanje po datumih od: 1.10.2014 do: 1.6.2015

Tina Maze

About 10,700 results (0.38 seconds)

**Tina Maze (@TinaMaze) | Twitter**  
The latest Tweets from Tina Maze (@TinaMaze). My way is my decision. SLOVENIJA.  
<https://twitter.com/tinamaze>

**Tina Maze je "zamrzla" smučarsko kariero | Alpsko smučanje ...**  
7 maj 2015 ... Tina Maze: Enomačnega odgovora o (ne)načrtovanju smučarske kariere Tina Maze ni dala, vseeno pa je napovedala vrst začen enoletni ...  
[www.slo.si/sport/zimski-sporti/tina-maze-je-zamrzla-smucarsko-kariero](http://www.slo.si/sport/zimski-sporti/tina-maze-je-zamrzla-smucarsko-kariero)

**Blog Tina Maze**  
7 maj 2015 ... O Tina Maze: Team to alpske; Galerija - Tina Maze; Blog - Tina Maze; Fan shop - Tina Maze; Sponzorji - Tina Maze; About Tina Maze; Team to ...  
[www.tinamaze.com/blog](http://www.tinamaze.com/blog)

**VIDEO in FOTO: Tina Maze: Upam, da ta sprejem ni bil zadnji**  
3 apr 2015 ... Tina Maze in njegovo ekipo je na velikem sprejemu v Črni na Koroškem podpravo približno 1500 navijačev. Certifika slovenski šampioni so ...  
[www.slo.si/sport/zimski-sporti/tina-maze-prekinila-smucarsko-kariero](http://www.slo.si/sport/zimski-sporti/tina-maze-prekinila-smucarsko-kariero)

**Tina Maze prekinila kariero, a vrst posvem še ni zaprla :: Prvi ...**  
7 maj 2015 ... Najboljša slovenska smučarka Tina Maze se je odločila, da si bo vzela leto dni premora in se potem odločila, ali bo nadaljevala kariero.  
[www.slo.si/sport/zimski-sporti/tina-maze-prekinila-smucarsko-kariero](http://www.slo.si/sport/zimski-sporti/tina-maze-prekinila-smucarsko-kariero)

**Tina Maze**  
8 maj 2015 ... Tina Maze je športnica s precejšnjim marketinškim vplivom. Vsaka njena odločitev ima posledice za več ljudi. Kar je ne bo vti na telesa ...  
<http://www.daroval.si/tag/tina-maze>

**Tina Maze nadaljuje kondicijske priprave | Slovenskenovice.si**  
23 apr 2015 ... Mami pravi, da bo Tina potrebovala njegovo pomoč, ne glede na to, kako se bo odločila.  
[www.slovenskenovice.si/tina-maze-nadaljuje-kondicijske-priprave](http://www.slovenskenovice.si/tina-maze-nadaljuje-kondicijske-priprave)

**Tina Maze - Pisanje zgodovine na njen način :: Prvi interaktivni ...**  
11 feb 2015 ... Tina Maze je ena zadnjih res vsestranskih smučark v svetovnem pokalu, vseeno pa Mikaeli Shiffin, ki jo Američani vzgajajo za veliko ...  
<http://www.slo.si/sport/zimski-sporti/tina-maze-prekinila-smucarsko-kariero>

**Tina Maze spregovorila o Anni Fenninger | Slovenskenovice.si**  
10 mar 2015 ... V glavni vlogi bodo Tina Maze in Anna Fenninger ter Marcel Hirscher in kjeti Jansrud. Razlike med glavnimi protagonistmi nilegga dela ...  
[www.slovenskenovice.si/tina-maze-spregovorila-o-anni-fenninger](http://www.slovenskenovice.si/tina-maze-spregovorila-o-anni-fenninger)

**Po hudem boju Tina Maze brez velikega globusa**  
22 mar 2015 ... Na zadnji tekmi sezone je zmagala Anna Fenninger, Tina Maze je na tretjem mestu veliki globus izgubila za 22 točk.  
[www.delo.si/sport/zimski-sporti/po-hudem-boju-tina-maze-brez-velikega-globusa.html](http://www.delo.si/sport/zimski-sporti/po-hudem-boju-tina-maze-brez-velikega-globusa.html)

powered by Google™ Custom Search

Shranjeni!

Shranjeni članki	Prikaži vse
Po hudem boju Tina Maze brez velikega globusa	Bridi Zaključ
Tina Maze spregovorila o Anni Fenninger   Slovenskenovice.si	Bridi Zaključ
Tina Maze - Pisanje zgodovine na njen način	Bridi Zaključ
Tina Maze nadaljuje kondicijske priprave   Slovenskenovice.si	Bridi Zaključ
Tina Maze prekinila kariero, a vrst posvem še ni zaprla	Bridi Zaključ
VIDEO in FOTO: Tina Maze: Upam, da ta sprejem ni bil zadnji	Bridi Zaključ
Tina Maze je "zamrzla" smučarsko kariero   Alpsko smučanje - Planet Sici.net	Bridi Zaključ

Slika 10: Pregledovanje in zajem zadetkov

Prijavljen uporabnik: joze | Odjavi | Prikaži vse uporabnike | Statistika člankov | Poidi na ocenjevanje člankov

URL za ročni vnos spletne strani | Shrani stran

Prikaži shranjene | Prikaži zaključene

Prijavljen uporabnik: joze | Odjavi | Prikaži vse uporabnike | Statistika člankov | Poidi na ocenjevanje člankov

Nazaj na seznam | Prikaži shranjene | Prikaži zaključene

**Podatki o članku**

ID članka	10026
Vnešeno v bazo	18.06.2015 22:32:34
Matična stran	<a href="http://www.delo.si">http://www.delo.si</a>
Povezava	<a href="http://www.delo.si/sport/zimski/po-hudem-boju-tina-maze-brez-velikega-globusa.html">http://www.delo.si/sport/zimski/po-hudem-boju-tina-maze-brez-velikega-globusa.html</a>
Naslov	Po hudem boju Tina Maze brez velikega globusa
Ključne besede	alpsko smučanje, Meribel, veleslalom, Tina Maze, Anna Fenninger, #foto
Datum članka	22.03.2015
Avtor	Š. Ro., Delo.si
Vnašalec	joze
Vsebina	Po hudem boju Tina Maze brez velikega globusa Na zadnji tekmi sezone je zmagala Anna Fenninger, Tina Maze je na tretjem mestu veliki globus izgubila za 22 točk. Š. Ro., Delo.si od, 22.03.2015, 09:30; spremenjen: 13:25 Ključne besede: alpsko smučanje, Meribel, veleslalom, Tina Maze, Anna Fenninger

Pridobi celotno vsebino

Shrani spremembe

Shrani in Zaključ

Slika 11: Urejanje in shranjevanje zadetkov



Prijavljen uporabnik: joze [Odjava](#) [Prikaži vse uporabnike](#) [Statistika člankov](#) [Pojdi na ocenjevanje člankov](#)

URL za ročni vnos spletne strani [Shrani stran](#) [Prikaži shranjene](#) [Prikaži zaključene](#)

Prijavljen uporabnik: joze [Odjava](#) [Prikaži vse uporabnike](#) [Statistika člankov](#) [Pojdi na ocenjevanje](#)

[Nazaj na iskanje](#) [IZVOZI\\*ARFF](#) [Prikaži shranjene](#) [Prikaži zaključene](#)

### Seznam člankov

Prikaži 10 vnosov

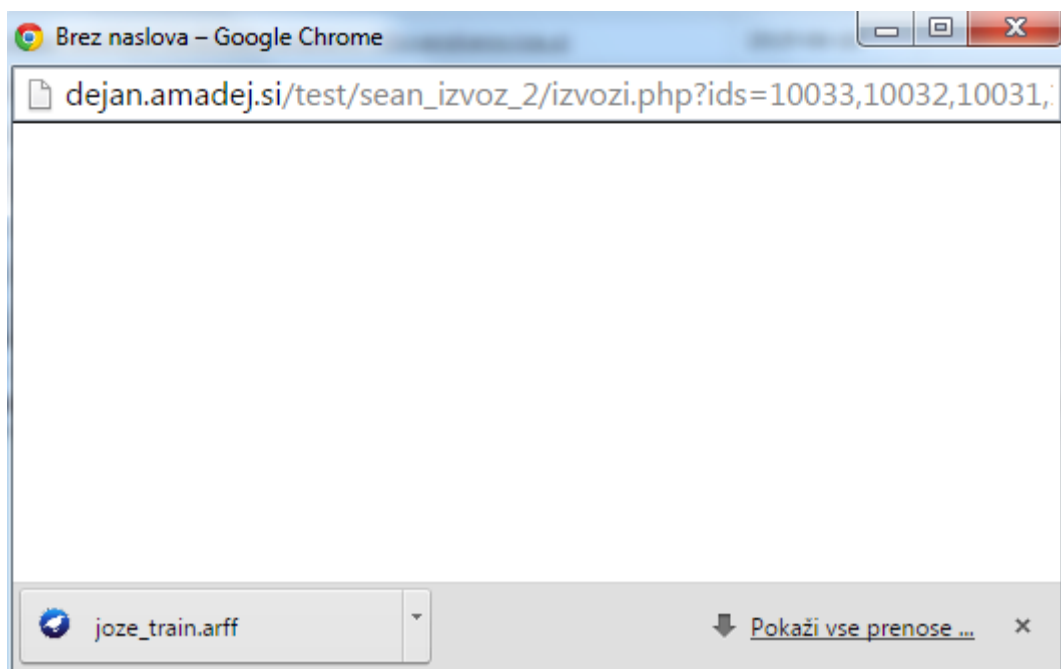
Iskanje:

ID članka	Vnešeno v bazo	Glavni URL	Naslov	Datum članka	Avtor	Vnašalec	
10033	18.06.2015 23:12:50	http://www.delo.si	Po hudem boju Tina Maze brez velikega globuru	2015-03-22	Š. Ro., Delo.si	joze	Briži zaključeni ✓
10032	18.06.2015 23:12:49	http://www.slovenskenovice.si	Tina Maze spregovorila o Anni Fenninger   Slovenskenovice.si	2015-06-18	joze	joze	Briži zaključeni ✓
10031	18.06.2015 23:12:43	https://www.rtvslo.si	Tina Maze - Pisanje zgodovine na njen način	2015-06-18	joze	joze	Briži zaključeni ✓
10030	18.06.2015 23:12:36	http://www.slovenskenovice.si	Tina Maze nadaljuje kondicijske priprave   Slovenskenovice.si	2015-06-18	joze	joze	Briži zaključeni ✓
10029	18.06.2015 23:12:33	http://www.rtvslo.si	Tina Maze prekinila kariero, a vrata povsem še ni zaprla	0000-00-00	joze	joze	Briži zaključeni ✓
10028	18.06.2015 23:12:26	http://24ur.com	VIDEO in FOTO: Tina Maze: Upam, da ta sprejem ni bil zadnji	2015-06-18	joze	joze	Briži zaključeni ✓
10027	18.06.2015 23:12:14	http://www.siol.net	Tina Maze je "zamrznila" smučarsko kariero   Alpsko smučanje - Planet	2015-05-07	Martin Pavčnik	joze	Briži zaključeni ✓

[Išči po ID-ju](#) [Išči po datumu v](#) [Glavni URL](#) [Išči po naslovu](#) [Išči po datumu](#) [Išči po avtorju](#) [Išči po vnašalcu](#)

Prikaz vnosov od 1 do 10. Vseh vnosov: 7

[Prva](#) [Prejšnja](#) [1](#) [2](#) [Naslednja](#) [Zadnja](#)



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

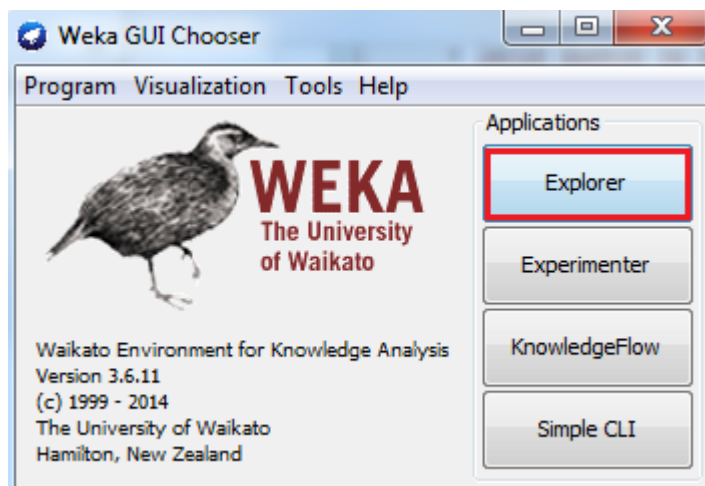
## 6 NAVODILA ZA DOLOČITEV SENTIMENTA NOVIH SPLETNIH BESEDIL V PROGRAMU WEKA

Sistemske zahteve:

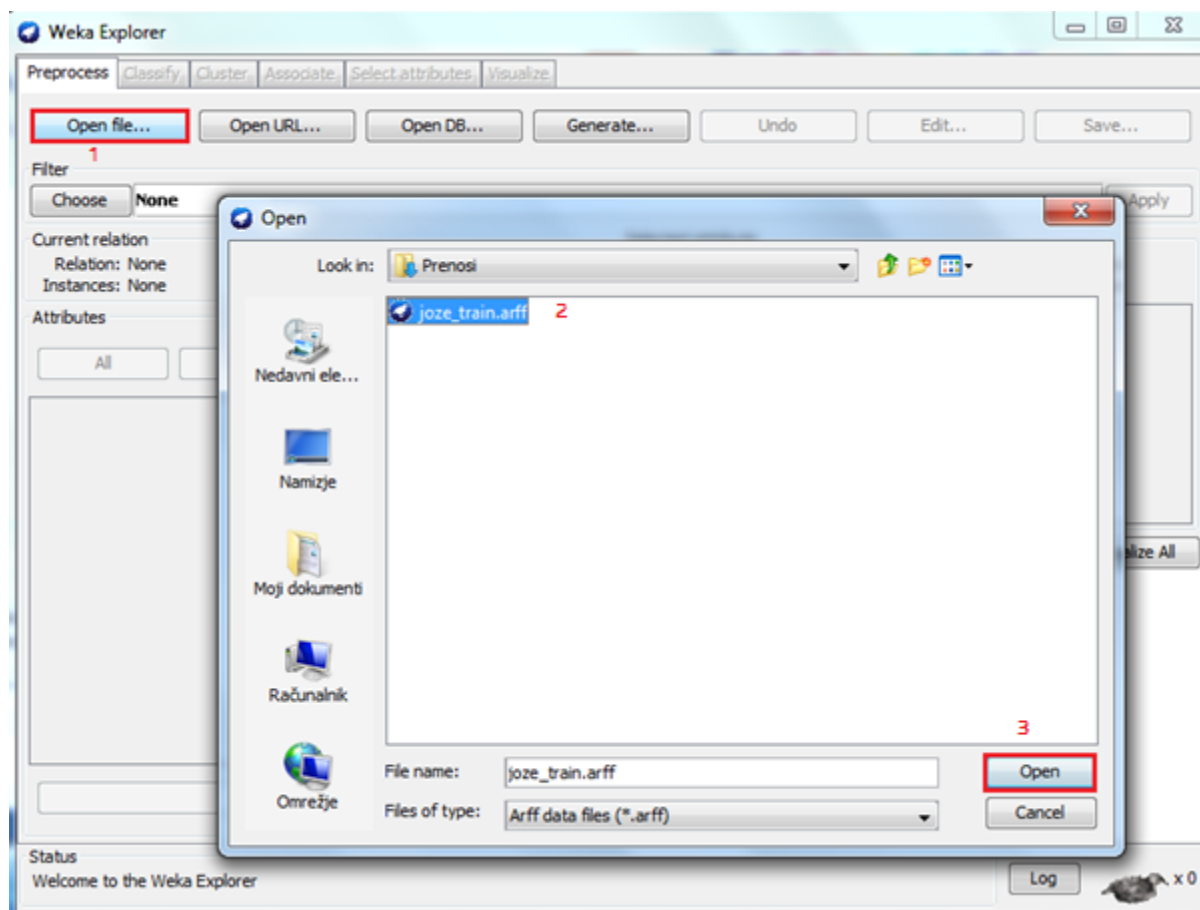
- potrebno je naložiti in inštalirati zadnjo verzijo programa Weka, ki ustreza vašemu operacijskemu sistemu, povezava do programske opreme, katero inštalirate na vaš računalnik: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>
- v kolikor želite boljšo funkcionalnost programa, odprite datoteko **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**

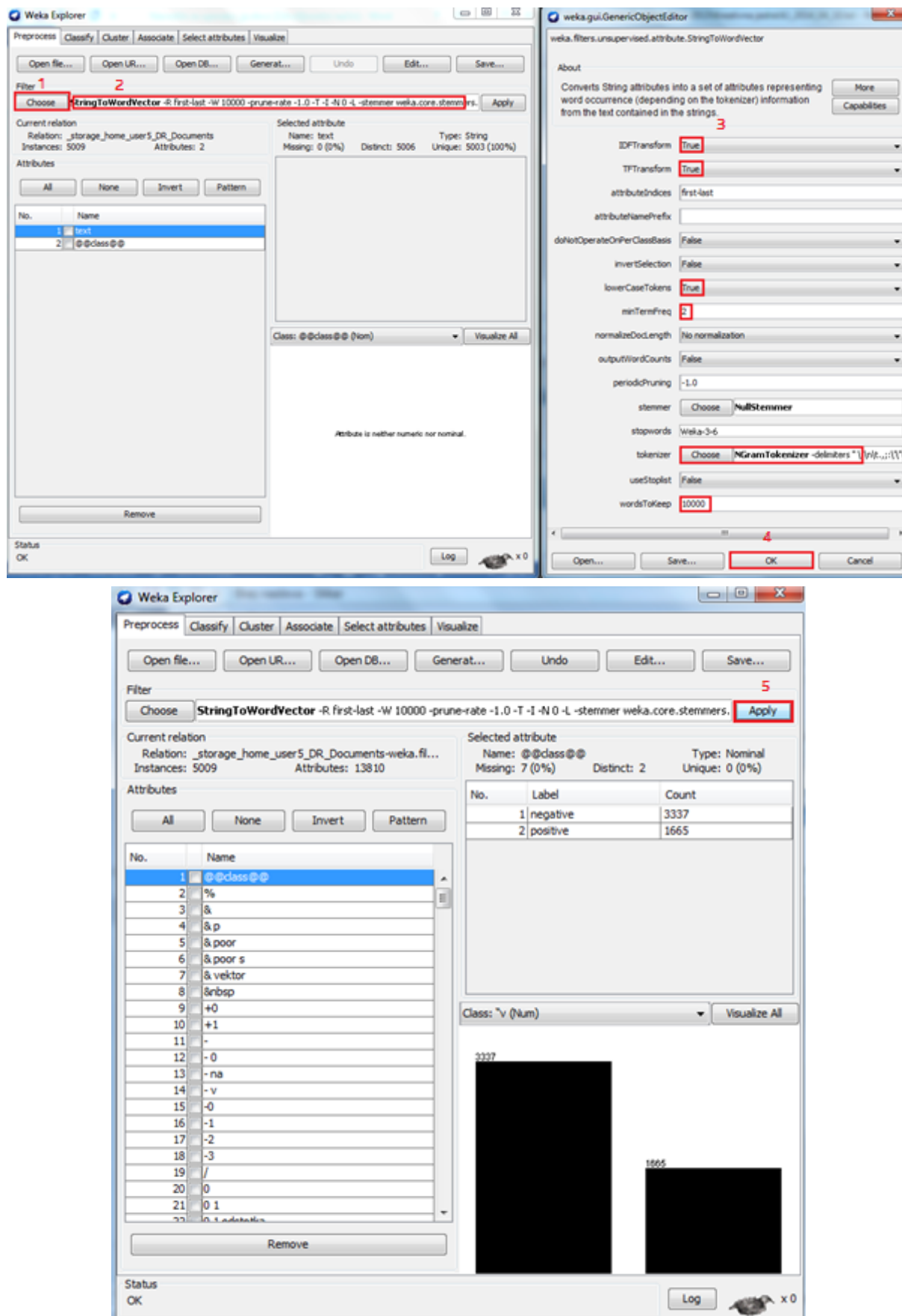


**Slika 15: Uvoz vhodne datoteke (določitev poti do vhodne datoteke joze\_train.arff)**

#### Izbira filtra

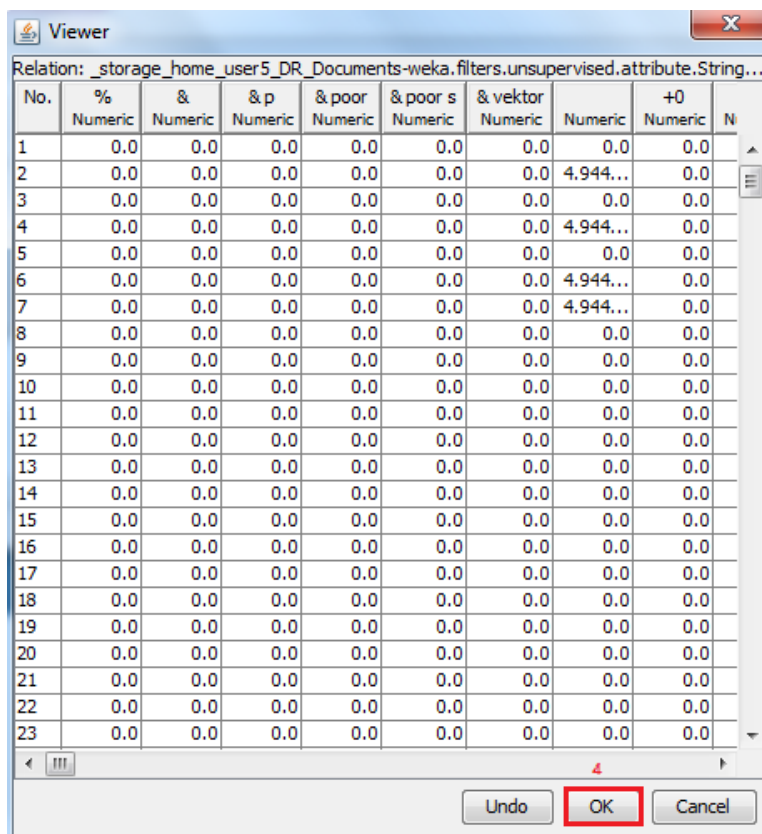
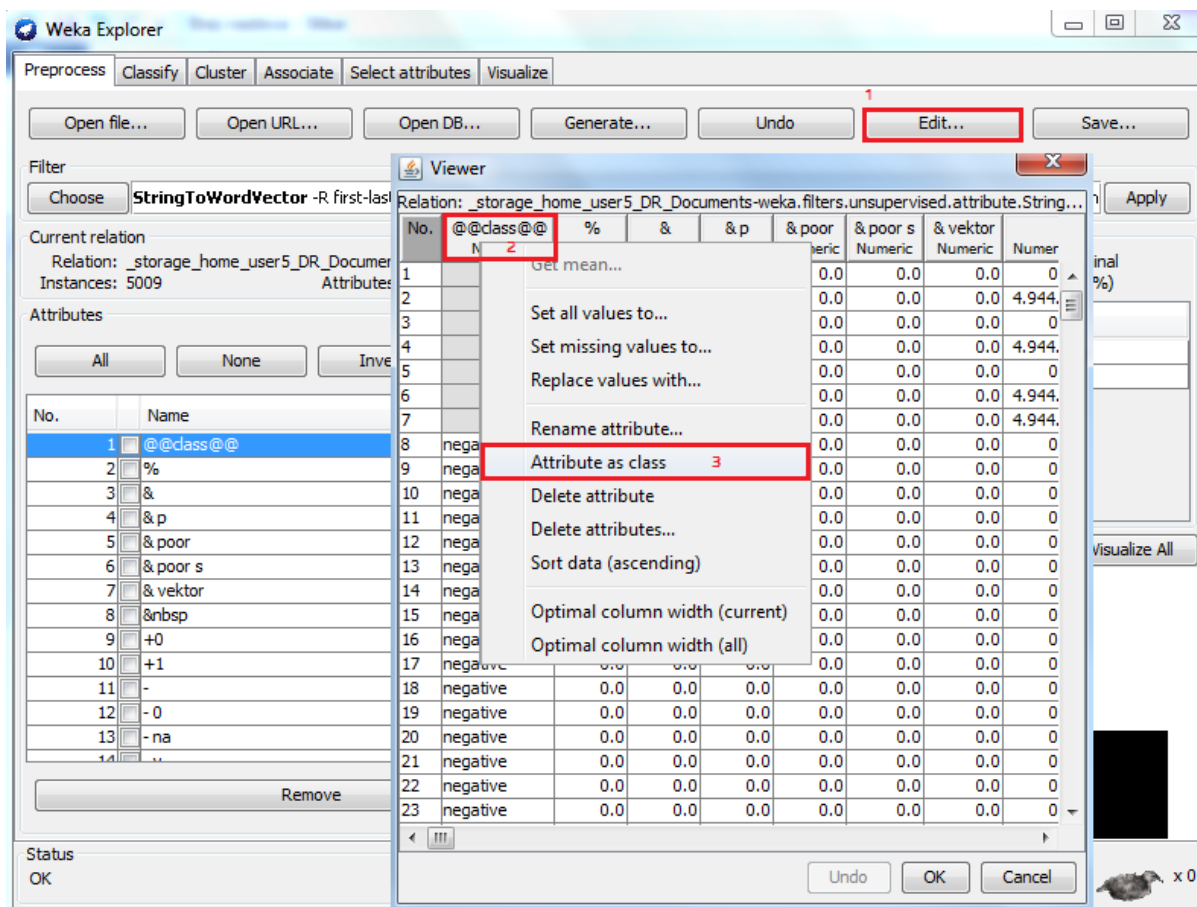
Filter-> filters -> unsupervised -> attribute -> StringToWordVector -> kliknete z levim gumbom na miški na filter da lahko urejate lastnosti filtra (IDFTTransform=True, TFFTransform=True, lowerCaseTokens=True,minTermFreq=2, tokenizer=NGramTokenizer (min=1,max=3), wordsToKeep=10000) (traja ~ 30s) -> Apply



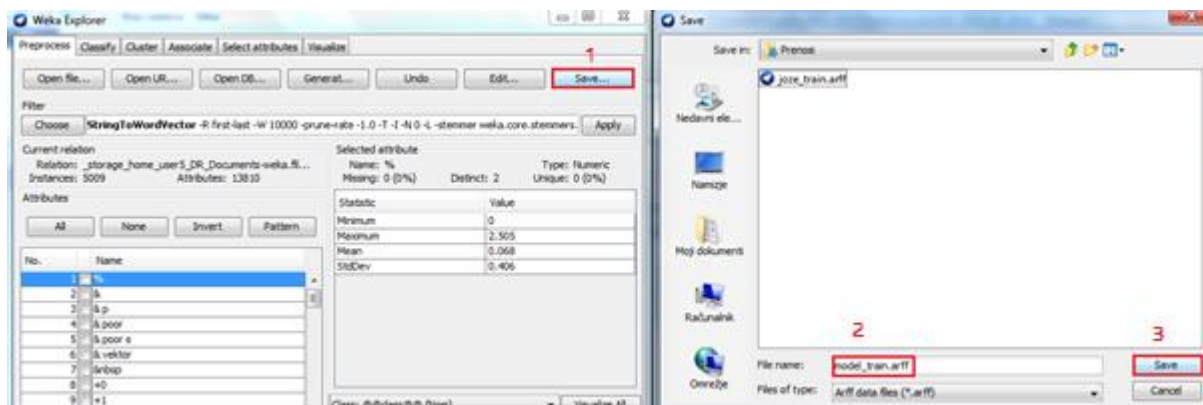


Slika 16: Izbira ustreznega filtra



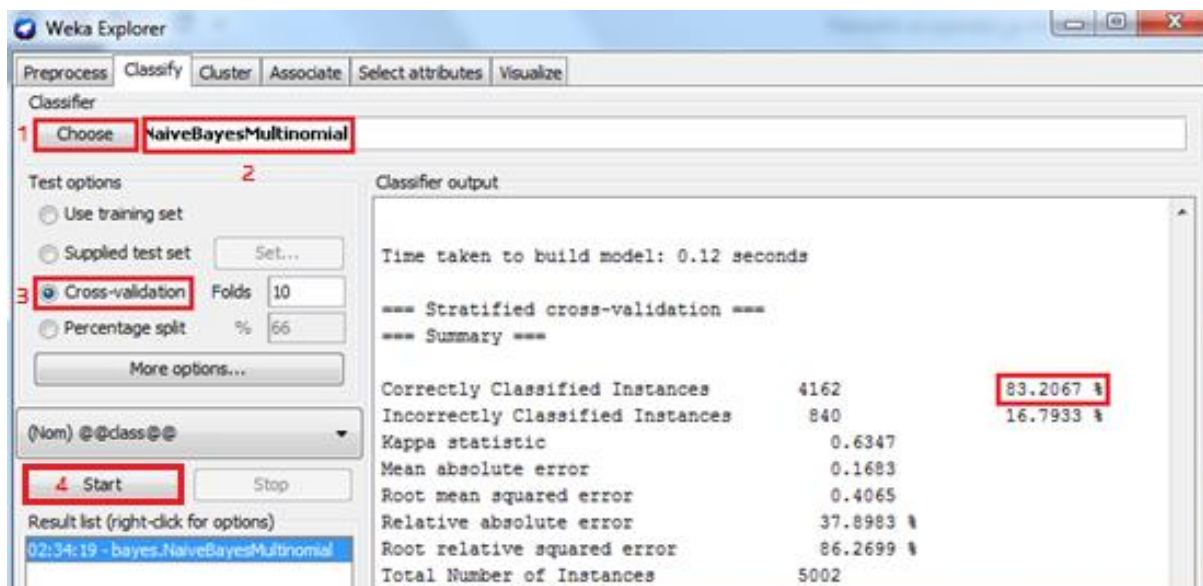


**Slika 17: Izbira ustreznega atributa kot ciljnega razreda v postopku določitve sentimenta novih spletnih besedil**



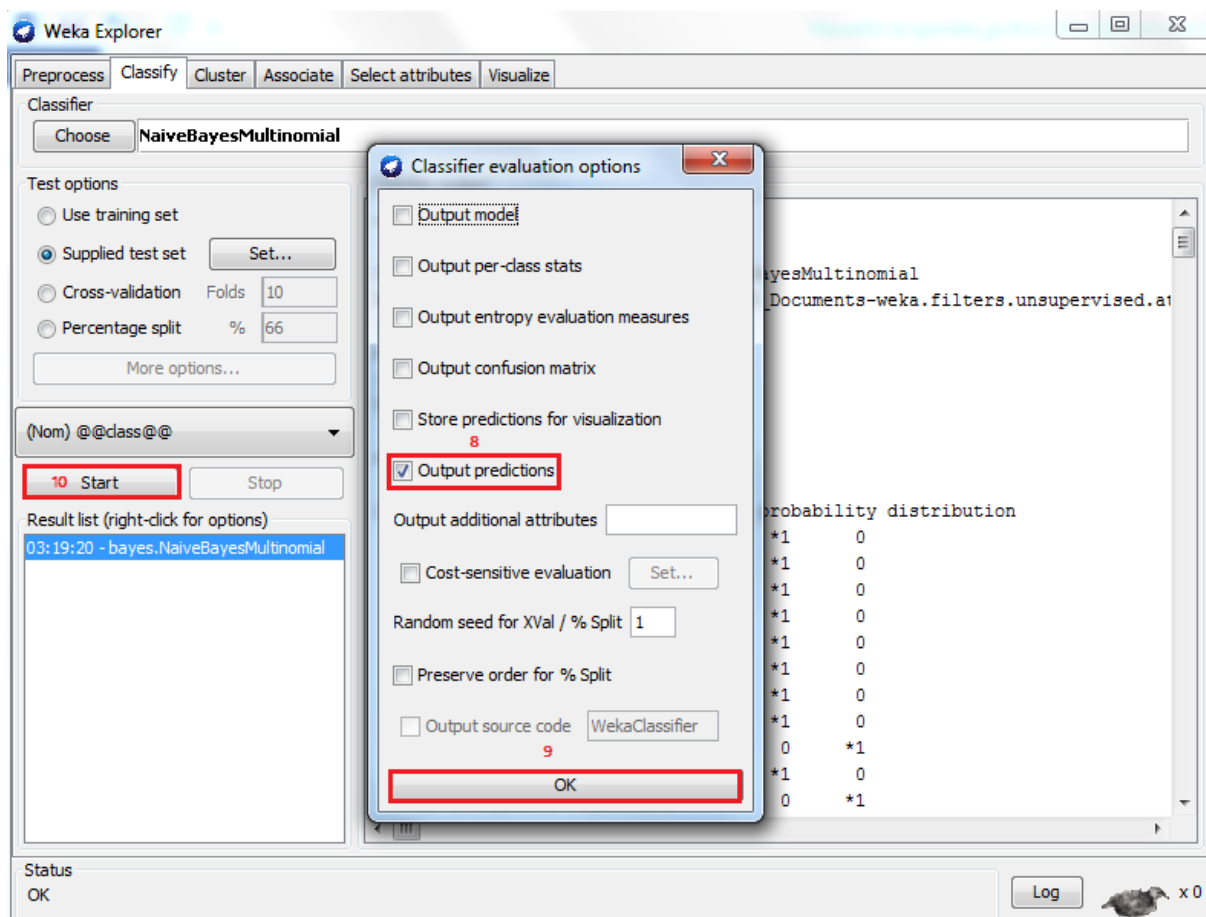
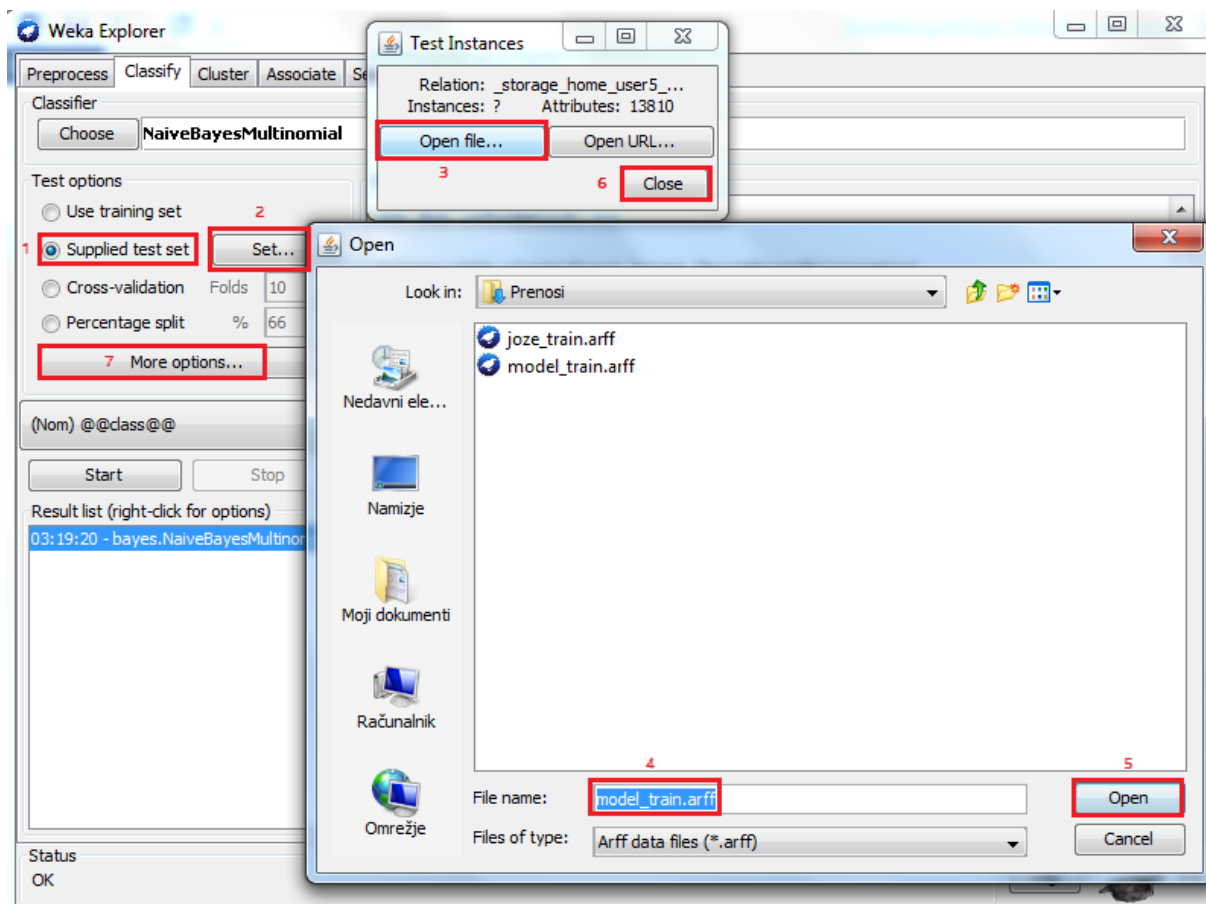
**Slika 18: Shranjevanje modela (model\_train.arff)**

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 | Select attributes | Visualize

Classifier: **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  
03:25:34 - bayes.NaiveBayesMultinomial

Classifier output:

```
=== Run information ===  
  
Scheme:weka.classifiers.bayes.NaiveBayesMultinomial  
Relation: _storage_home_user5_DR_Documents-weka.filters.unsupervised.a  
Instances: 5009  
Attributes: 13810  
[list of attributes omitted]  
Test mode:user supplied test set: size unknown (reading incrementally)  
  
=== Predictions on test split ===  
  
inst#,    actual, predicted, error, probability distribution  
1         ? 1:negative + *0.667 0.333  
2         ? 2:positive + 0 *1  
3         ? 2:positive + 0 *1  
4         ? 2:positive + 0 *1  
5         ? 2:positive + 0 *1  
6         ? 2:positive + 0 *1  
7         ? 2:positive + 0 *1  
8 1:negative 1:negative *1 0  
9 1:negative 1:negative *1 0  
10 1:negative 1:negative *0.998 0.002  
11 1:negative 1:negative *1 0
```

Status: OK Log x 0

**Slika 20: Določitev sentimenta novih spletnih besedil**