

Информатички модели у анализи осећања засновани на језичким ресурсима докторска теза

студент

Миљана Младеновић

ментор

Душко Витас

Математички факултет
Београд, мај 2015.

Рачунарска лингвистика - Computational linguistics (CL)

интердисциплинарна област

- информатика
- вештачка интелигенција
- лингвистика

Примарни задаци рачунарске лингвистике:

- класификација докумената (*Document classification*),
- сегрегација (*Document clustering*),
- претраживање информација (*Information retrieving*),
- разрешавање вишезначности ентитета (*Word-sense disambiguation*),
- издвајање ентитета (*Text extraction*),
- машинско превођење (*Machine translation*),
- рачунарско препознавање говора (*Computer speech recognition*),
- рачунарско генерисање говора (*Computer speech synthesis*),
- генерисање природног језика (*Natural language generation*),
- корпусна лингвистика (Corpus linguistics),
- оптичко препознавање карактера (Optical character recognition - OCR)
- анализа осећања (Sentiment analysis)

Анализа осећања у тексту је специфична врста класификације текста.

Различити називи за област чији је предмет интересовања обрада емоција у тексту:

- **класификација осећања** (енг. *sentiment classification*),
- **истраживање мишљење** (енг. *opinion mining*),
- **анализа осећања** или **сентимент анализа** (енг. *sentiment analysis*),
- **екстракција осећања** (енг. *sentiment extraction*).

Три врсте сентимент класификације текстова:

- **идентификација субјективности** (енг. *opinion classification* или *subjectivity identification*) која се бави поделом докумената на оне који носе субјективне и оне који носе искључиво чињеничне информације.
- **класификација осећања** (енг. *sentiment classification*) или **идентификација поларитета сентимента** (енг. *polarity identification*)
- **одређивању снаге или интензитета емоционалног поларитета** (енг. *strength of orientation*)

У погледу нивоа на коме се анализа осећања врши, три методологије:

- **анализа осећања на нивоу документа** (сваки документ се посматра као јединствени објект сентимент анализе, а сва мишљења потичу од једног посматрача тј. извора емоција)
- **анализа осећања на нивоу реченице** (документ дели на реченице уз претпоставку да је свака реченица понаособ носилац емоција о једном објекту и све емоције у реченици потичу од једног носиоца емоција)
- **анализа осећања на нивоу атрибута** - пример фотоапарата који се посматра кроз појединачну оцелу свих његових значајних саставних делова.

Методи који се примењују у анализи емоција у тексту:

- **машинско учење** (machine learning)
 - **методе ненадгледаног учења** (unsupervised learning) - користе правила лексичко-семантичких мрежа као што су *WordNet*, *ConceptNet* или се користе знања о емоцијама похрањена у доменским онтологијама или онтологијама високог нивоа
 - **методе надгледаног учења** (supervised learning)
 - вероватносне класификационе методе машинског учења -методе засноване на Бајесовом правилу (енг. *Naive Bayes*), регресионе методе какава је метода максималне ентропије (енг. *Maximum Entropy*)
 - невероватносне класификационе методе каква је метода потпорних вектора (енг. *Support vector machines*).
- **системи засновани на правилима** (rule-based systems)
 - латентна семантичка анализа (Latent semantic analysis -LSA)
 - системи засновани на корпусима и лексиконима
 - правила изведена из узајамне корелације (Pointwise mutual information -PMI)
 - семантичке мреже: *WordNet*, *ConceptNet*, *WebNet*, *SentiWordNet*
 - закључивање уз помоћ доменских онтологија

Предмет истраживања: анализа емоција у тексту

Циљ истраживања: класификација осећања, на нивоу докумената, на српском језику

I фаза

Методе истраживања: вероватносна метода надгледаног машинског учења максималне ентропије.

У раду се предлажу, примењују и оцењују методе редукције векторског простора предиктора употребом лексичко-семантичке мреже *WordNet* и уводи се и имплементира специјализовани лексичко-семантички ресурс „речник емоција“.

II фаза

Циљ истраживања: побољшање резултата класификације осећања препознавањем фигуративног говора

Методе истраживања: онтолошко закључивање дескриптивном логиком

У другом делу тезе описује се израда и имплементација доменске онтологије реторских фигура српског језика, *WordNet* онтологија и њихова имплементација у процесу препознавања фигуративног говора у текстовима на српском језику.

Методолошке основе примењене у истраживању

Језички модели

Модели репрезентације текстуалног документа:

- логички модел (van Rijsbergen, 1986)
- модел векторског простора (Salton, 1971)
- вероватносни модел (Maron и Kuhns, 1960).

Вероватносни модели дају дистрибуцију вероватноће случајне променљиве по свим исходима.

Нпр. један од вероватносних модела језика (n-грам) сваком симболу ω_i у документу, у низу од m симбола (карактера или речи), додељују условну вероватноћу $P(\omega_i | \omega_1, \dots, \omega_{i-1})$ помоћу расподеле вероватноће претходних симбола у секвенци. Вероватноћа $P(\omega_1, \omega_2, \omega_3, \dots, \omega_n)$ n-грама (низа симбола $\omega_1, \omega_2, \omega_3, \dots, \omega_n$) апроксимира се производом условних вероватноћа

$$P(\omega_1, \omega_2, \omega_3, \dots, \omega_n) = \prod_{i=1}^n P(\omega_i | \omega_1, \dots, \omega_{i-1})$$

Вероватносни језички модели користе се у системима вероватносне класификације текста.

Регресија

Моделирање неког процеса помоћу кога се вредности једне или скупа више посматраних променљивих (предиктора процеса) пресликавају у неку излазну вредност.

Регресија - излазна вредност континуална (најчешће реалан број)

Класификација - излазна вредност из скупа дискретних вредности

Регресија и класификација могу узети у обзир и више од једног улазног предиктора (*multiple*)- вишедимензионална, а такође могу бити и са више излазних вредности или са више класа у случају класификације (*multinomial*) - вишезначна.

Линарни регресиони модел

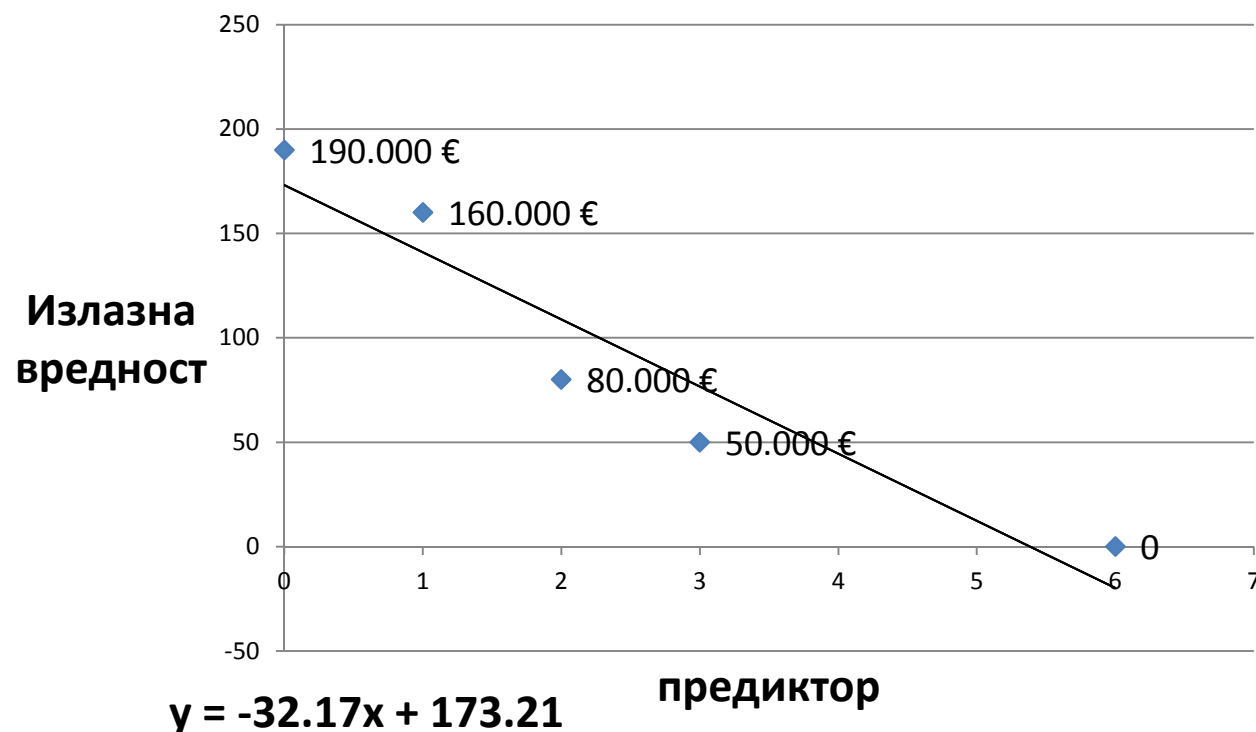
- Скуп онлајн огласа за продају некретнина
- Моделујемо систем за предикцију продајне цене неког стана
- Посматрамо једну променљиву (предиктор) коју ћемо назвати „број епитета којим се стан описује“.

„Продаје се фантастичан стан, на атрактивној локацији, са изузетно функционалним распоредом“.

- Скуп за учење - низ огласа о **продатим** становима
- Направити корелацију између броја атрибута и вредности по којој је стан продат.

број епитета	цена продатог стана
6	0
3	50.000€
2	80.000€
1	160.000€
0	190.000€

Линарни регресиони модел



Променљива x је посматрана особина (предиктор) процеса. Коефицијенти a и b су тежине (енг. *weights*).

Линеарни модел је вектор тежина $\vec{w} = (173.21, -32.17)$

Линарни регресиони модел са већим бројем предиктора

- Претпоставка да на цену некретнине утичу и укупан број станова у понуди у датом делу града и висина каматне стопе стамбеног кредита.
- Линеарна регресиона функција
$$\text{cena} = w_0 + w_1 * \text{broj_epiteta} + w_2 * \text{stopa} + w_3 * \text{broj_stanova_u_ponudi}$$

У општем случају је излазна вредност зависна од већег броја улазних предиктора и њихових тежина:

$$y = w_0 + \sum_{i=1}^n w_i * f_i$$

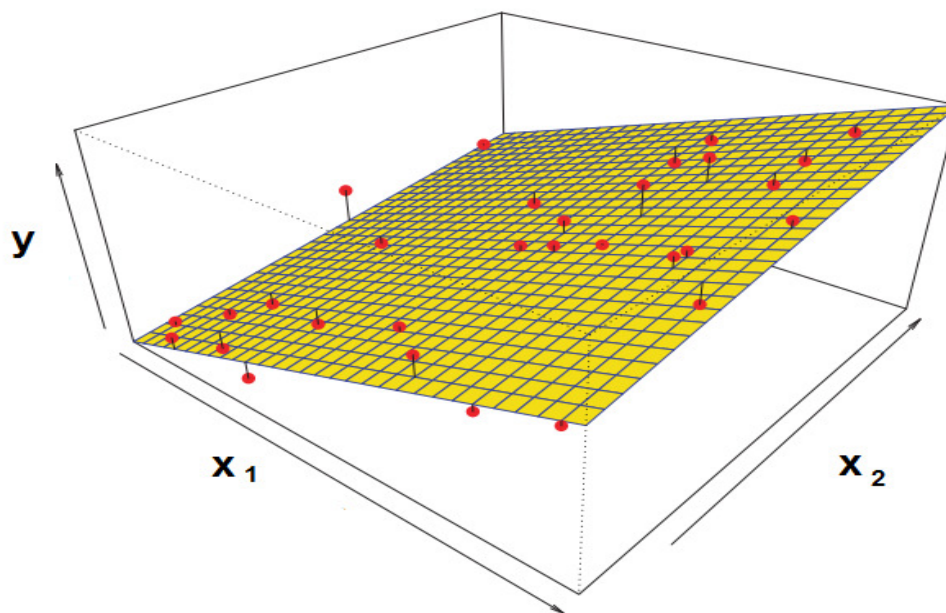
$$y = \sum_{i=0}^n w_i * f_i \quad \text{за } f_0=1$$

Ако са $\vec{w}$ означимо вектор тежина, а са $\vec{f}$ вектор предиктора, линеарна регресија се може представити и као скаларни производ два вектора.

$$y = \vec{w} \cdot \vec{f}$$

Линеарни регресиони модел са два предиктора где су посматране вредности y_{obs} скупа за учење дате црвеном бојом, а одговарајуће вредности процењене моделом y_{pred} леже на регресионој равни.

Линеарни регресиони модел са два предиктора x_1 и x_2 који минимизира суму квадрата разлика скупа за учење



Минимизација разлике између моделом предвиђене и стварне, односно посматране излазне вредности. Тражени вектор $\vec{w}$ налази се као функција минимума суме квадратних грешки:

$$\vec{w} = \min \sum_{j=0}^m (y_{pred}^{(j)} - y_{obs}^{(j)})^2$$

У општем случају, хиперраван у n -димензионалном простору одваја документе који припадају класи од докумената који јој не припадају.

Логистичка регресија

Класификација - посматрани документ или припада или не припада једној одређеној класи (нпр. документ може носити емоционалну информацију или не).

Често желимо да систем моделује степен припадности одређеној класи.

Вероватносни линеарни модел :

$$P(y = true|x) = \sum_{i=0}^n w_i * f_i = w \cdot f$$

Уместо вероватноће, посматрамо шансу (енг. *odds*) тј. однос вероватноће припадања класи и вероватноће неприпадања класи:

$$\frac{P(y = true|x)}{1 - P(y = true|x)} = \sum_{i=0}^n w_i * f_i$$

Логистичка регресија

$$\ln \left(\frac{P(y = \text{true}|x)}{1 - P(y = \text{true}|x)} \right) = w \cdot f$$

Логаритам шансе назива се логит функцијом:

$$\text{logit}(P(x)) = \ln \left(\frac{P(x)}{1 - P(x)} \right)$$

Модел регресије који користи линеарну функцију да процени логит функцију вероватноће да документ припада класи зове се логистичка или лог-линеарна регресија.

Вероватноћа да посматрани документ x припада класи y , на основу два вектора: вектора тежина $\vec{w}$ и вектора предиктора $\vec{f}$ посматраног документа x .

$$\frac{P(x)}{1 - P(x)} = e^{w \cdot f}$$

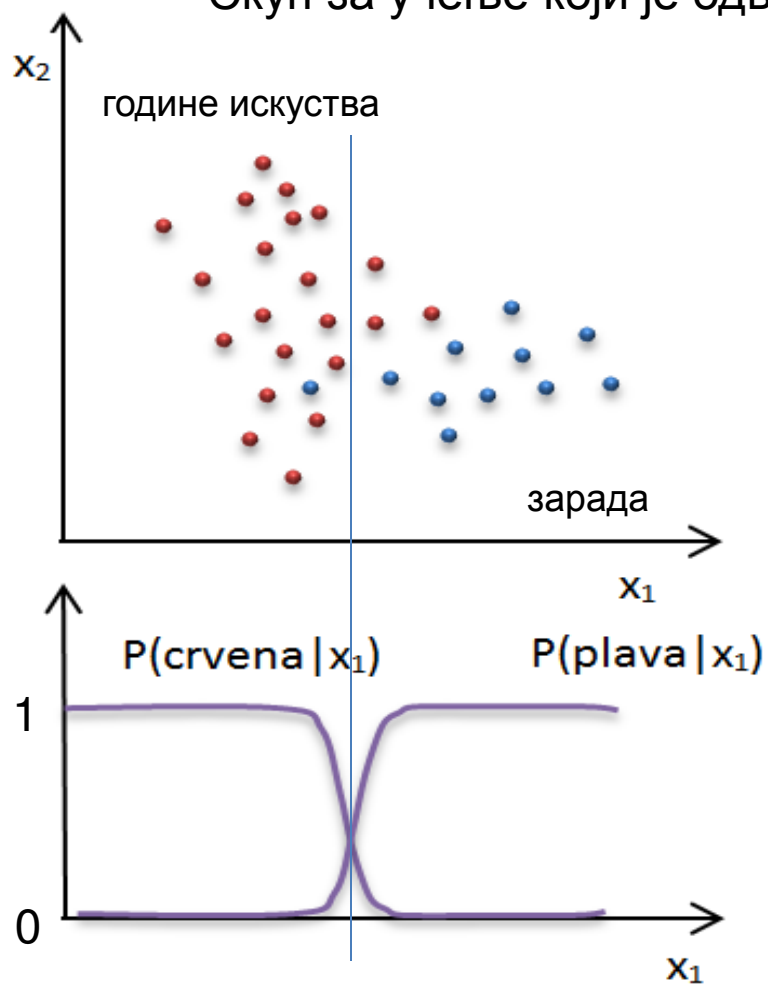
$$P(y = \text{true}|x) = \frac{e^{w \cdot f}}{1 + e^{w \cdot f}} \quad \text{сигмоидална функција}$$

$$P(y = \text{false}|x) = 1 - P(y = \text{true}|x) = \frac{1}{1 + e^{w \cdot f}}$$

Бинарна логистичка класификација

Сигмоидална или основна логистичка функција (дискриминативна).

Скуп за учење који је одвојив



Процес учења у логистичкој регресији

У линеарној регресији процес учења се састоји од проналажења оног вектора тежина $\vec{w}$ који минимизује суму квадратне грешке свих предиктора из скупа за учење.

У логистичкој регресији се до вектора тежина $\vec{w}$ долази оценом максималне веродостојности (MLE), што значи да **вектор $\vec{w}$ чине оне тежине које дају највећу вероватноћу (y_{pre}) да за сваки документ, из посматраног скупа за учење, класификација тог документа буде једнака оној коју тај документ заиста и има (y_{obs}).**

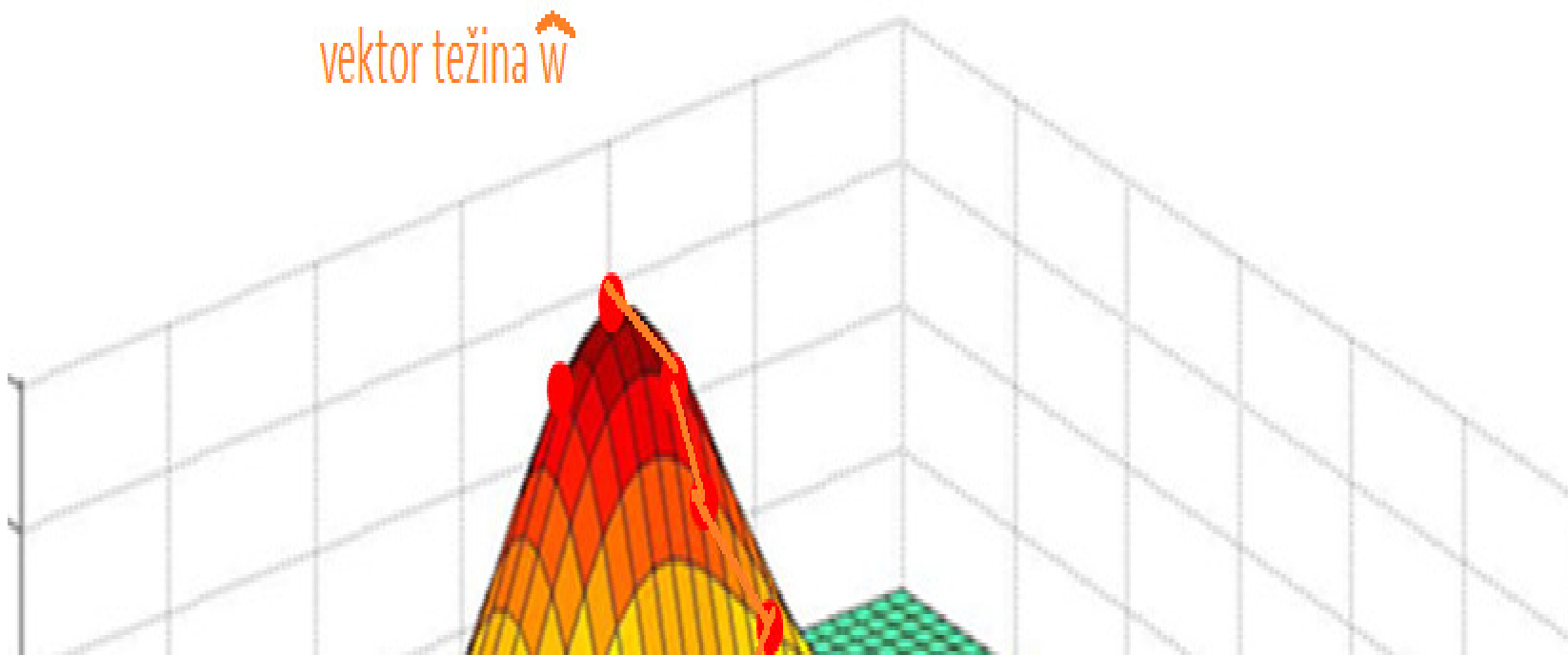
За све документе из скупа за учење функција MLE исказана је као:

$$l(w) = \prod_{j=1}^m \begin{cases} P(y^{(j)} = 1 | X^{(j)}) & \text{za } y^{(j)} = 1 \\ P(y^{(j)} = 0 | X^{(j)}) & \text{za } y^{(j)} = 0 \end{cases}$$

а тражени вектор тежина $\vec{w}$ добија се из услова:

$$\hat{w} = \underset{w}{argmax} P(y^{(j)} | x^{(j)})$$

Процес учења у логистичкој регресији – конвергенција Њутн-Рафсоновог метода



Метод максималне ентропије

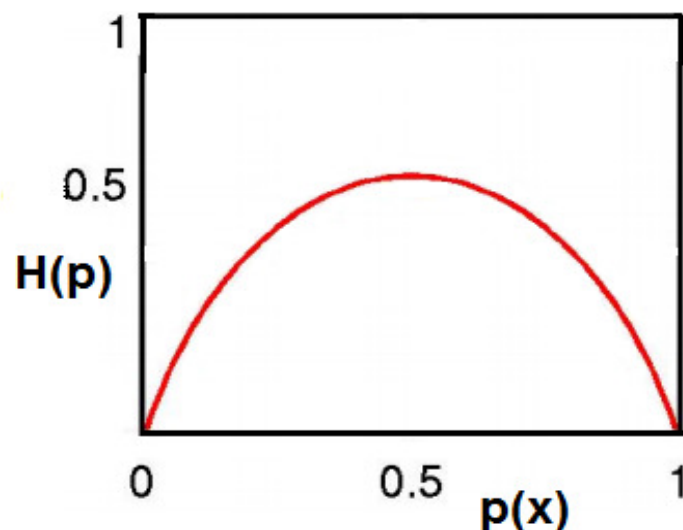
Бергер је 1996. доказао да је модел мултиномиалне логистичке регресије који постиже максималну веродостојност еквивалентан моделу $p^* \in \mathcal{C}$ са условљеном максималном ентропијом

$$p^* = \underset{p \in \mathcal{C}}{\operatorname{argmax}} H(p)$$

Ентропија је мера неуређености неког система, позната из термодинамике и теорије гасова. У теорији информација посебно је значајна функција бинарне ентропије $H(p)$ као мера неизвесности случајне променљиве која може узети једну од две вредности 0 и 1. Дефинише се као функција вероватноће неке случајне променљиве x :

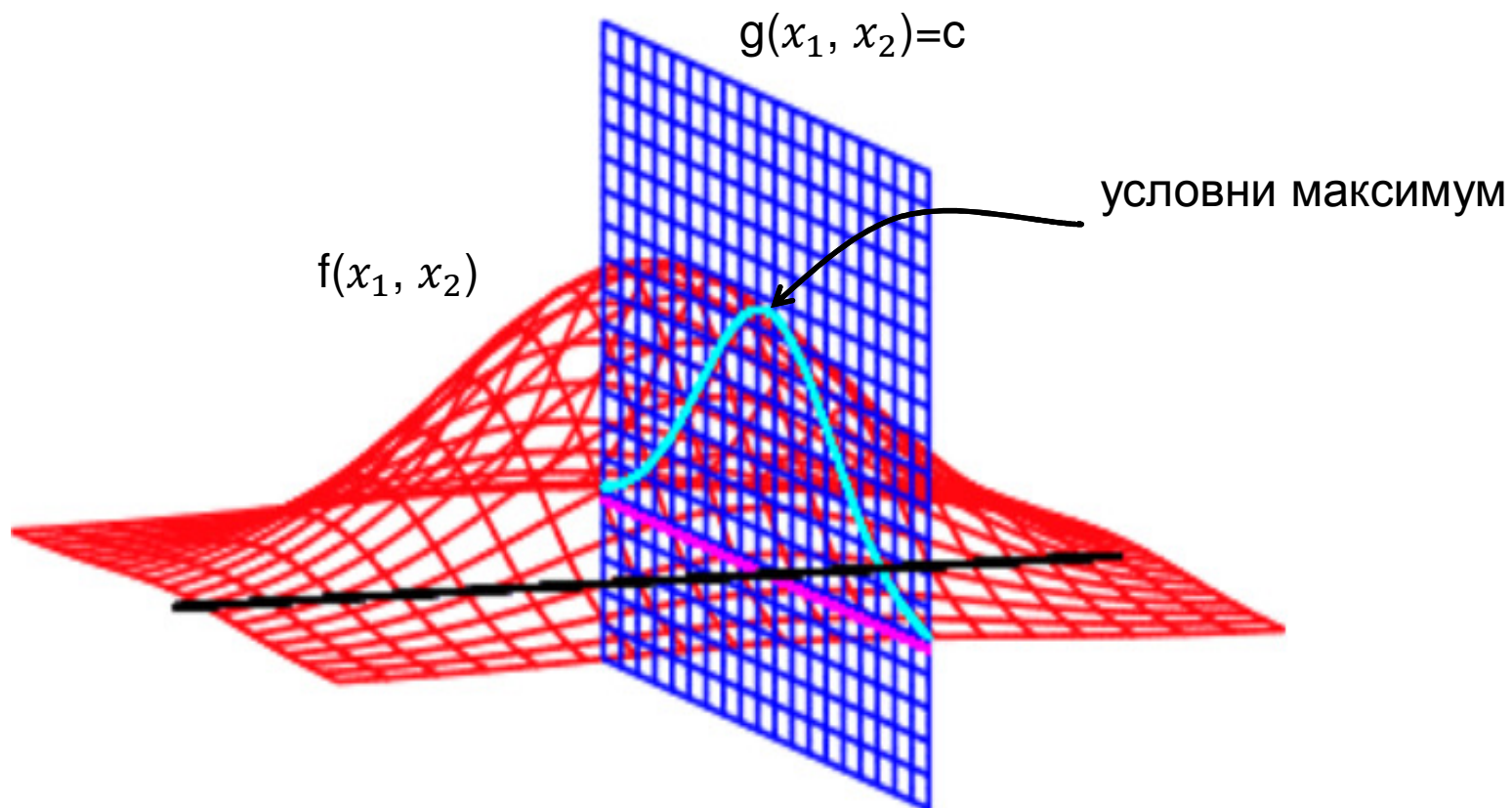
$$H(p) = -\sum_x p(x) \log_2(p(x)) .$$

Дистрибуција вероватноћа
дихотомне случајне променљиве x



**Процес учења у моделу максималне ентропије –
конвергенција методом Лагранжеових мултипликатора**

ЛМ је метода којом се налази екстремум функције више променљивих $f(x_1, x_2, \dots, x_n)$ под задатим ограничењима $g(x_1, x_2, \dots, x_n) = 0$



Индикатор функција максималне ентропије

У сентимент класификацији (а и осталим областима НЛП) ограничења система најчешће се дефинишу бинарним предикторима који описују систем.

Бинарни (дихотомни) (енг. *dummy*) предиктори - **индикатор функције** – моделују квалитативне промениљиве (боја очију, пол,...).

Предиктор да је i -та реч у документу ($word_i$) присвојни придев:

$$f_i(c, x) = \begin{cases} 1, & (is_lower_case(word_i[0])) \text{ and} \\ & (suffix(word_i) = \text{"ски"}) \text{ and } (c = PP) \\ 0, & otherwise \end{cases}$$

Могућност нарастања броја предиктора на више стотина или више хиљада у проблемима рачунарске лингвистике даје велику важност методама селекције предиктора.

Принцип максималне ентропије не бави се директно процесом селекције предиктора. Он једино даје најоптималнији модел од понуђених предиктора.

Селекција предиктора

Селекција предиктора (*feature selection*) је скуп техника којима се смањује векторски простор предиктора скупа за учење на ону меру која даје ефикасније процесирање (брже и једноставније) али без губитка информација.

Разликујемо 3 врсте приступа:

- **селекција подскупа предиктора** (*best subset selection*) - проналази подскуп p предиктора таквих да најбоље репрезентује скуп за учење;
- **регуларизација или ограничавање** (*shrinkage*) – примењује се над свим предикторима, али ефекат регуларизације делује највише на оне предикторе који имају највеће вредности. Корекција може бити двојака: смањењем вредности предиктора тако да она постаје блиска нули (тзв. *L2-norm*) или изједначавањем са нулом (тзв. *L1-norm*), чиме се овај метод може посматрати и као вид селекције подскупа;
- **смањење димензије векторског простора** (*dimension reduction*) – смањење димензије се постиже пројектовањем p -димензионалног векторског простора предиката у m -димензионални подпростор где је $m < p$.

Предложени метод смањења векторског простора предиктора

Пример:

Ekipe Elektrodistribucije uspele su da uspostave normalno snabdevanje električnom energijom u Zlatiborskom okrugu, izjavio je za Tanjug Dušan Šunjevarić iz Elektrodistribucije Užice. "U toku dana sve raspoložive ekipe Elektrodistribucije bile su na terenu i uspele su da otklone kvarove na elektro-energetskim objektima, prouzrokovane padom drveća, usled težine snega, na dalekovode i trafo stanice", kaže Šunjevarić. Trenutno, prema njegovim rečima, struju nema još šest domaćinstava na Pašinoj ravni, gde ekipa EPS-a radi na otklanjanju kvara na jednoj trafo stanici.

Претпоставка смањења векторског простора предиктора:

Текст се позитивно класификује и поларитетом сентимент речи и фраза: **успеле су, успоставе нормално, расположиве, отклоне кварове и отклањању квара.**

Појединачно посматрано, ове речи и фразе са својом фреквенцијом појављивања $f_2 = 1/77$, $f_3 = 1/77$, $f_{1,4,5} = 2/77$ могу се трансформисати у нови предиктор, па би, уместо 6 фраза са фреквенцијама појављивања 1 и 2, добили само концепт LexiconPos чија је фреквенција појављивања $6/77$, а векторски простор предиктора се смањује за 5.

Предложени метод смањења векторског простора предиктора

Пример:

*Ekipe Elektro distribucije **LexiconPos** su da **LexiconPos** snabdevanje električnom energijom u Zlatiborskom okrugu, izjavio je za Tanjug Dušan Šunjevarić iz Elektro distribucije Užice. "U toku dana sve **LexiconPos** ekipe Elektro distribucije bile su na terenu i **LexiconPos** su da **LexiconPos** na elektro-energetskim objektima, prouzrokovane padom drveća, usled težine snega, na dalekovode i trafo stanice", kaže Šunjevarić. Trenutno, prema njegovim rečima, struju nema još šest domaćinstava na Pašinoj ravni, gde ekipa EPS-a radi na **LexiconPos** na jednoj trafo stanici.*

Предложени метор смањења векторског простора предиктора

Концептуално пресликавање обезбеђују четири предиктора дефинисаних индикатор функцијама:

$$f1(c, x) = \begin{cases} 1, & x = \text{LexiconPos}, c = \text{Positive} \\ 0, & \textit{otherwise} \end{cases}$$

$$f2(c, x) = \begin{cases} 1, & x = \text{LexiconNeg}, c = \text{Negative} \\ 0, & \textit{otherwise} \end{cases}$$

$$f3(c, x) = \begin{cases} 1, & x = \text{WordNetPos}, c = \text{Positive} \\ 0, & \textit{otherwise} \end{cases}$$

$$f4(c, x) = \begin{cases} 1, & x = \text{WordNetNeg}, c = \text{Negative} \\ 0, & \textit{otherwise} \end{cases}$$

x је n -грам ($n \leq 3$) у скупу за учење, а c је ознака класе документа у коме се налази посматрани n -грам.

Даља унапређења селекције векторског простора предиктора

Друга важна претпоставка је да у морфолошки богатим језицима **примена стеминга** (свођења лексеме на корен речи), која у енглеском језику представља једну од основних метода за смањење векторског простора предиктора, **може довести до губитка информација које се тичу сентимента**, а могуће је и генерисати неисправне предикторе. Уколико користимо стемер у процесу смањења векторског простора предиктора, то ће произвести два ефекта.

1. Добијања супротног поларитета речи које се генеришу додавањем различитих префикса на исти корен речи.

Речи: „мотивисан“ , „немотивисан“ , „демотивисан“ и „безмотиван“- исти корен „мотив“.
Речи „бедно“ и „безбедно“ имају исти корен „бед“.

2. Неадекватан сентимент поларитет за речи које су пароними.

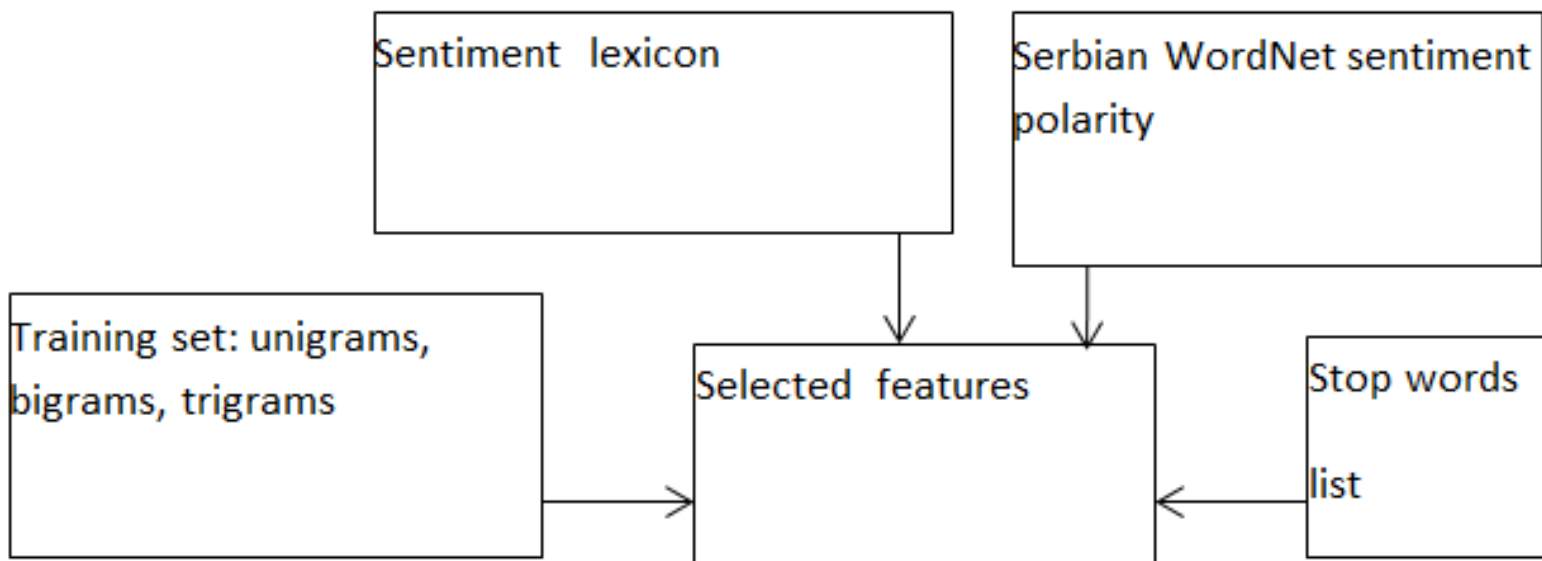
Пароними су речи изведени из истог корена који имају слично значење, али нису синоними. Примери:

„политичар“ и „политикант“ су пароними, имају исти корен „политик“.
„чистоћа“ и „чистота“ имају исти корен „чист“.

Селекције предиктора у САФОС-у

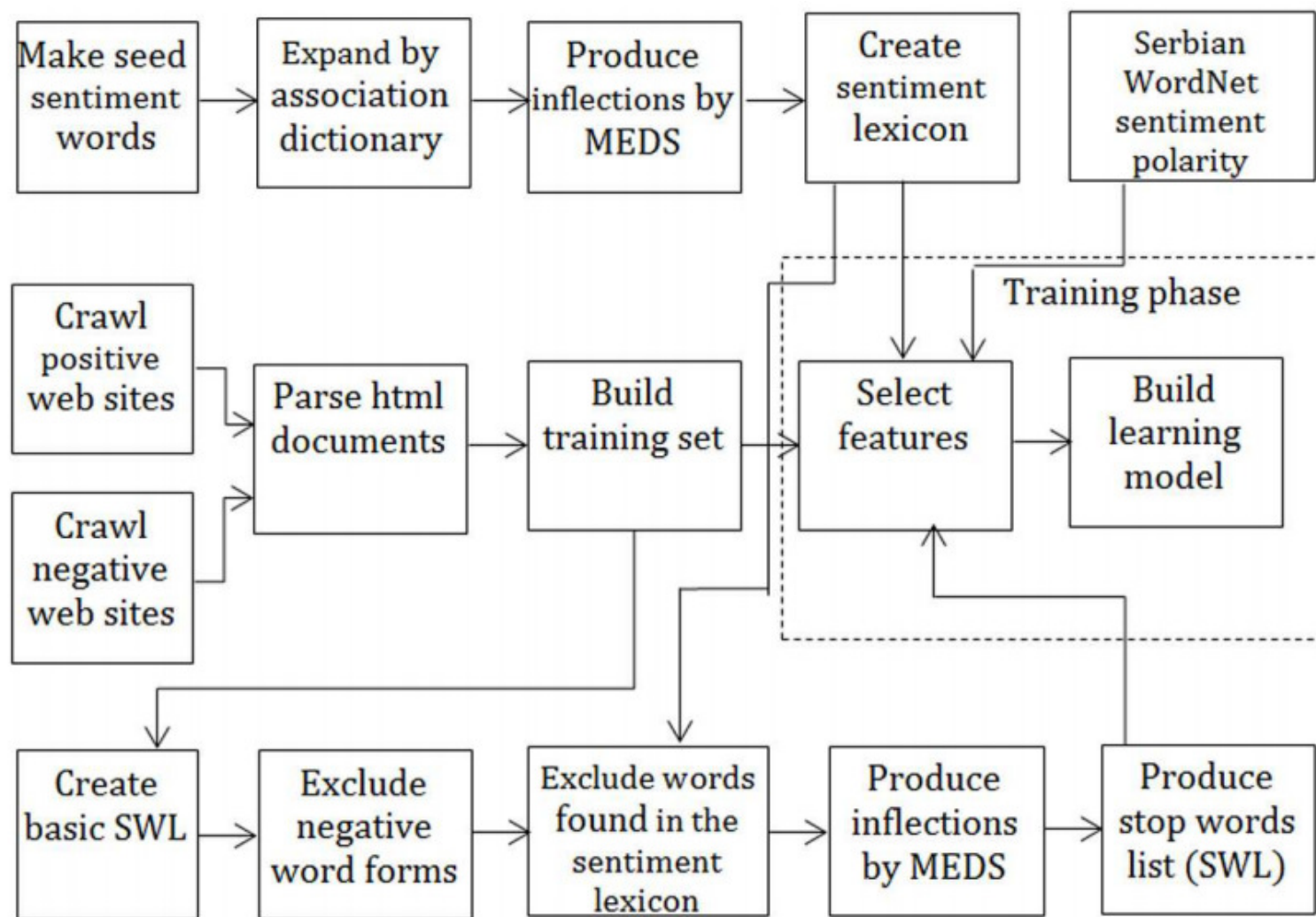
Процес генерисања и избор предиктора је важан део SAFOS система.

- основни скуп предиктора(*baseline*) - н-грами речи (unigrams, bigrams, trigrams)
- селекција подскупа предиктора - елиминацијом стоп речи
- повећање векторског простора предиктора – генерисањем флективних облика сентимент предиктора
- смањење векторског простора предиктора - концептуалним пресликавањем сентимент предиктора

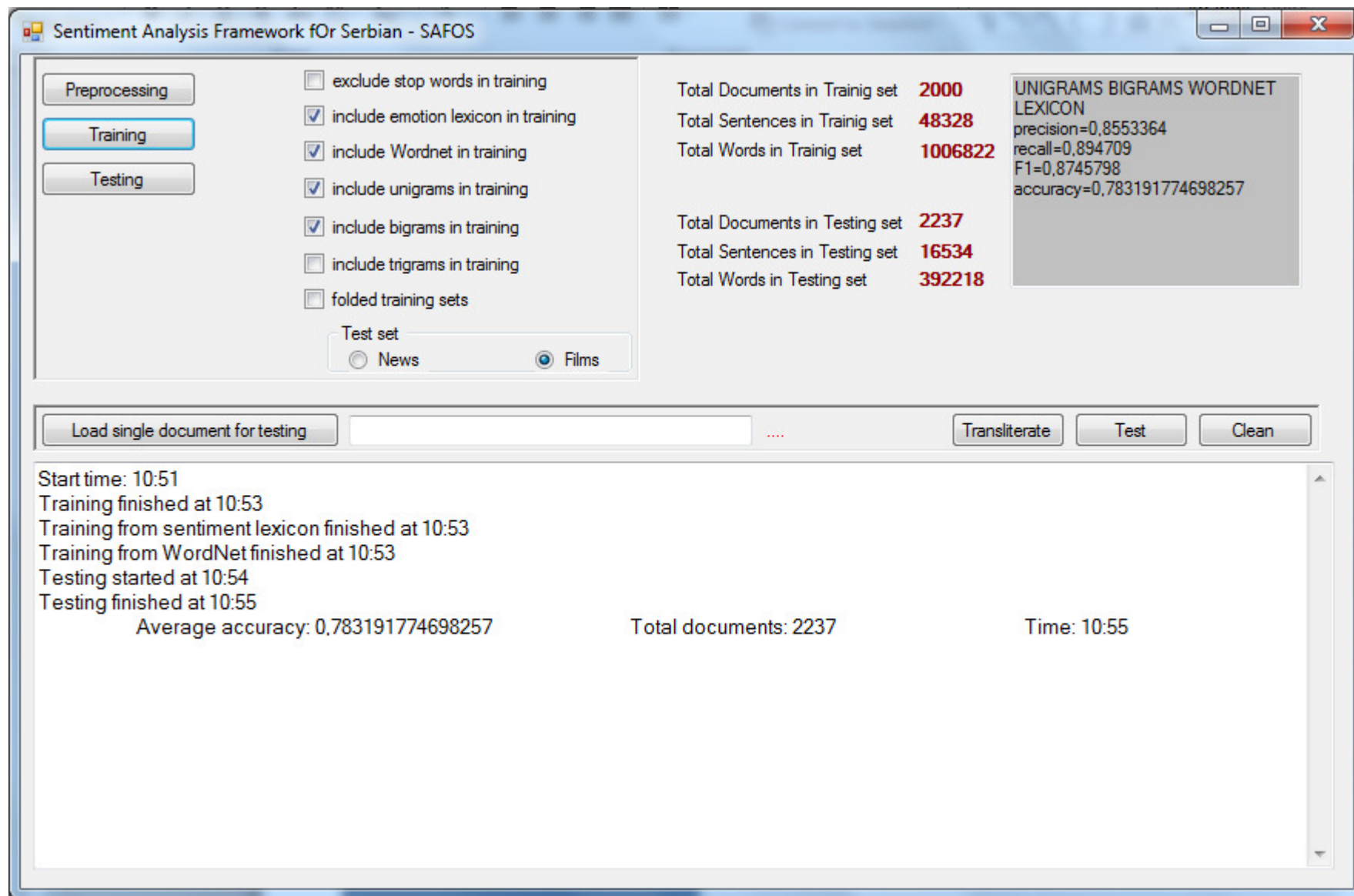


САФОС структура - активности укључене у препроцесирање и процес обуке SAFOS система

Препроцесирање укључује: стварање сентимент лексикона, генерисање листе стоп-речи, припрему скупа докумената на српском језику који се користе као скуп за учење и скупови за тестирање и генерисање предиктора који се користе у фази учења. У фази обуке генеришемо модел максималне ентропије на основу одабраног скупа предиктора.



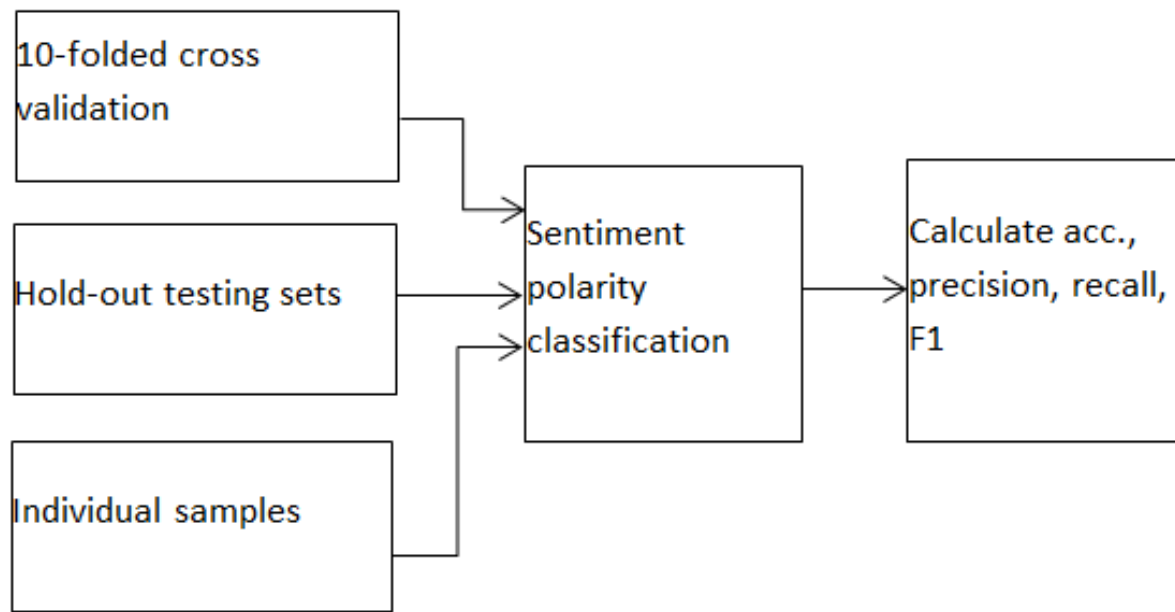
САФОС кориснички интерфејс



САФОС – Тестирање система

У фази тестирања, користили смо **три различита експеримента.**:

- **10-струку унакрсну валидацију скупа за обуку**
(<http://www.kurir-info.rs/crna-hronika>, <http://www.dobrevesti.rs>;
<http://www.svrljig.info/vesti/dobre-vesti>;
<http://www.ilovezrenjanin.com/category/drustvena-odgovornost-2/>)
- **скуп за тестирање** текстова са новинских портала ТЕСТ1
(<http://pistaljka.rs/> <http://www.vesti.rs/Dobre-vesti/>)
- **скуп текстова** који представљају оцене филмова ТЕСТ2
(www.2kokice.com)



САФОС – Евалуација система 10-струком унакрсном валидацијом

1. **10-струки унакрсни валидациони скуп** (1000 позитивних, 1000 негативних, 48.328 реченица, 1.006.822 речи)
 - скуп за учење - 1800 докумената и тестирање - 200 докумената.
 - Тест утицаја изабраног скупа предиктора на процес сентимент класификације помоћу четири параметра: **прецизности, одзива, F1 мере и тачности**.
 - Листа стоп-речи - генерички листа добијена из Референтног корпуса (<http://korpus.matf.bg.ac.rs/>).
 - Оцена статистичке значајности помоћу t-теста за зависне узорке

Feature set	precision	recall	F1	accuracy
Cf ₁ (U)	0.925	0.944	0.934	0.950
Cf ₂ (U+mapping)	0.937	0.950	0.943	0.956
Cf ₁ (U+SWL)	0.922	0.943	0.932	0.948
Cf ₂ (U+SWL+mapping)	0.934	0.946	0.940	0.954
Cf ₁ (U+B)	0.889	0.914	0.910	0.909
Cf ₂ (U+B+mapping)	0.904	0.919	0.912	0.914
Cf ₁ (U+B+SWL)	0.919	0.943	0.931	0.946
Cf ₂ (U+B+SWL+mapping)	0.929	0.951	0.940	0.950
Cf ₁ (U+B+T)	0.779	0.825	0.800	0.787
Cf ₂ (U+B+ T+mapping)	0.783	0.820	0.800	0.788
Cf ₁ (U+B+T+SWL)	0.899	0.921	0.910	0.920
Cf ₂ (U+B+T+SWL+mapping)	0.908	0.924	0.916	0.923

САФОС – Евалуација система тест скупом новинских чланака

2. Скуп за тестирање „онлајн вести“ (“news test set”) са 779 докумената (408 позитивних и 371 негативних), 16.304 реченица, 438.025 речи

- Скуп за учење од 2000 докумената и тестирање.
- Оцена статистичке значајности помоћу z-теста

Feature sets	precision	recall	F1	accuracy
Cf ₁ (U)	0.708	0.821	0.760	0.729
Cf ₂ (U+mapping)	0.727	0.824	0.772	0.746
Cf ₁ (U+SWL)	0.710	0.824	0.763	0.732
Cf ₂ (U+SWL+mapping)	0.730	0.821	0.773	0.747
Cf ₁ (U+B)	0.727	0.880	0.796	0.764
Cf ₂ (U+ B+mapping)	0.752	0.887	0.814	0.788
Cf ₁ (U+B+SWL)	0.736	0.855	0.791	0.764
Cf ₂ (U+B+SWL+mapping)	0.767	0.865	0.813	0.792
Cf ₁ (U+B+T)	0.720	0.814	0.764	0.737
Cf ₂ (U+B+T+mapping)	0.745	0.824	0.782	0.760
Cf ₁ (U+B+T+SWL)	0.736	0.841	0.785	0.759
Cf ₂ (U+B+T+SWL+mapping)	0.768	0.850	0.807	0.787

САФОС – Евалуација система тест скупом оцена филмова

3. Скуп за тестирање „оцене филмова“ (“films test set”) са 2237 докумената (1890 позитивних, 347 негативних), 16.534 реченица, 392.218 речи

- Скуп за учење од 2000 докумената
- Оцена статистичке значајности помоћу z-теста

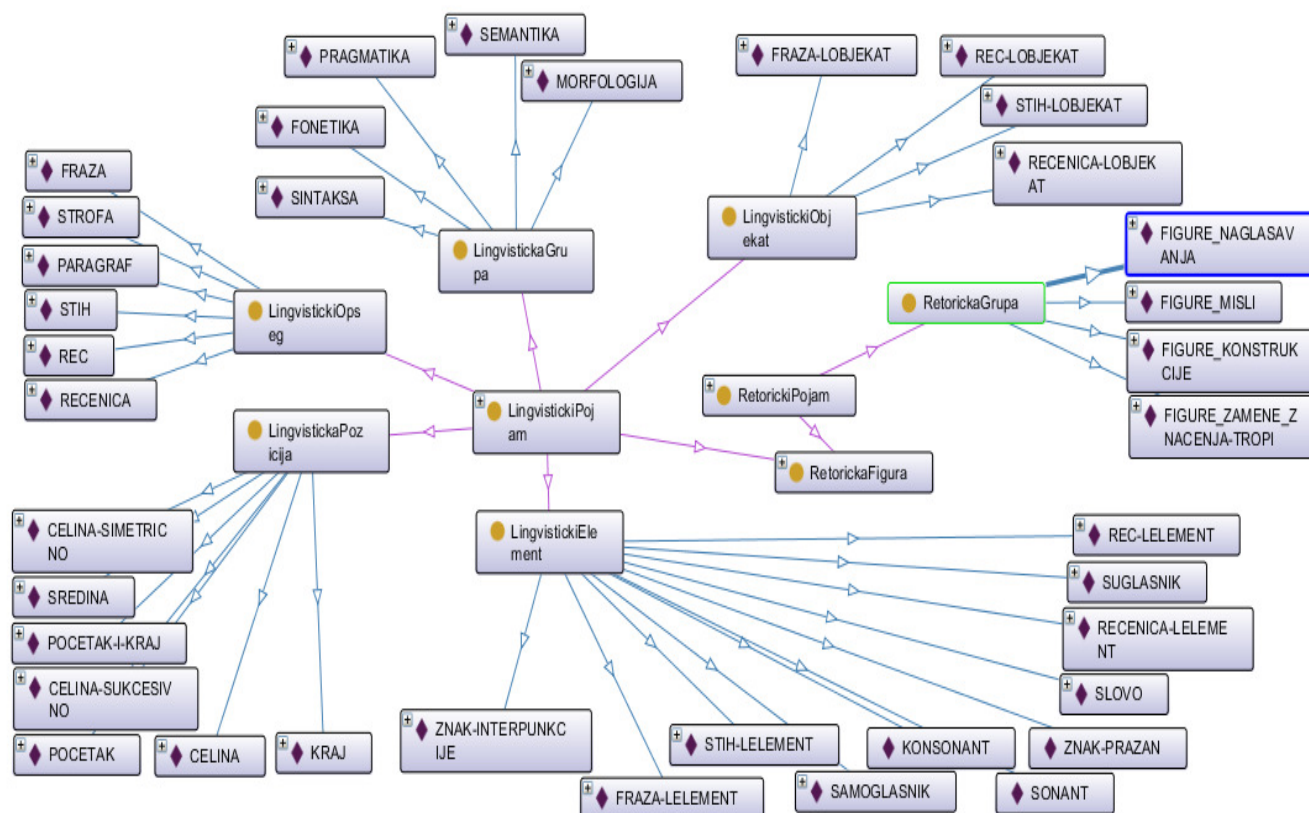
Feature sets	precision	recall	F1	accuracy
Cf ₁ (U)	0.831	0.736	0.781	0.651
Cf ₂ (U+mapping)	0.877	0.800	0.837	0.736
Cf ₁ (U+SWL)	0.828	0.728	0.775	0.642
Cf ₂ (U+SWL+mapping)	0.877	0.802	0.838	0.738
Cf ₁ (U+B)	0.839	0.882	0.860	0.758
Cf ₂ (U+B+mapping)	0.855	0.895	0.874	0.783
Cf ₁ (U+B+SWL)	0.839	0.818	0.828	0.714
Cf ₂ (U+B+SWL+mapping)	0.869	0.851	0.860	0.766
Cf ₁ (U+B+T)	0.839	0.879	0.859	0.755
Cf ₂ (U+B+T+mapping)	0.853	0.889	0.871	0.777
Cf ₁ (U+B+T+SWL)	0.836	0.848	0.842	0.731
Cf ₂ (U+B+T+SWL+mapping)	0.858	0.875	0.867	0.772

Неки упоредни резултати сентимент анализе МЕ методом

Аутор	Тест скуп	Резултати сентимент анализе МЕ методом	Језик
Pang and Lee 2002.	movie	Average three-fold cross-validation: Unigrams Acc=0,804 Uni+Big Acc=0,808 Bigrams Acc=0,774	енглески
Mehra et al. 2002.	movie	Acc=0,841	енглески
Go et al. 2009.	Twitter Data	Unigrams Acc=0,805 Bigrams Acc=0,791 Uni+Big Acc=0,830	енглески
Blinov et al. 2013.	movie	Acc=0,806	руски
Batista and Ribeiro 2013.	Twitter Data 7200 tweets-train. data 60800 tweets – testing data	Development data Unigrams Acc=0,552 Uni+Big Acc=0,538 Sent. Lexicon Acc=0,548 Test data Acc=0,634 Acc=0,622 Acc=0,632	шпански
Brychcin& Habernal 2013. RANLP	Czech CSFD movie dataset 90000 reviews	Average 10-fold cross-validation: ME baseline Acc=0,775 ME + sspace Acc=0,787 ME + Dir Acc=0,805 ME + sspace + Dir Acc=0,815	чешки

Унапређење скупа предиктора

- Онтолошка детекција реторских фигура
- Развој система онтолошког закључивања
- Детекција предиктора фигуративног говора



Унапређење скупа предиктора

(a) SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ont: <http://www.mmijana.com/ontologies/ont.owl#>
SELECT DISTINCT ?figura
WHERE {
    ?figura ont:jeNadOpsegom ?opseg.
    ?opseg ont:naziv ?nazivOpsega.
    FILTER( ?nazivOpsega = "REC" ) }
```

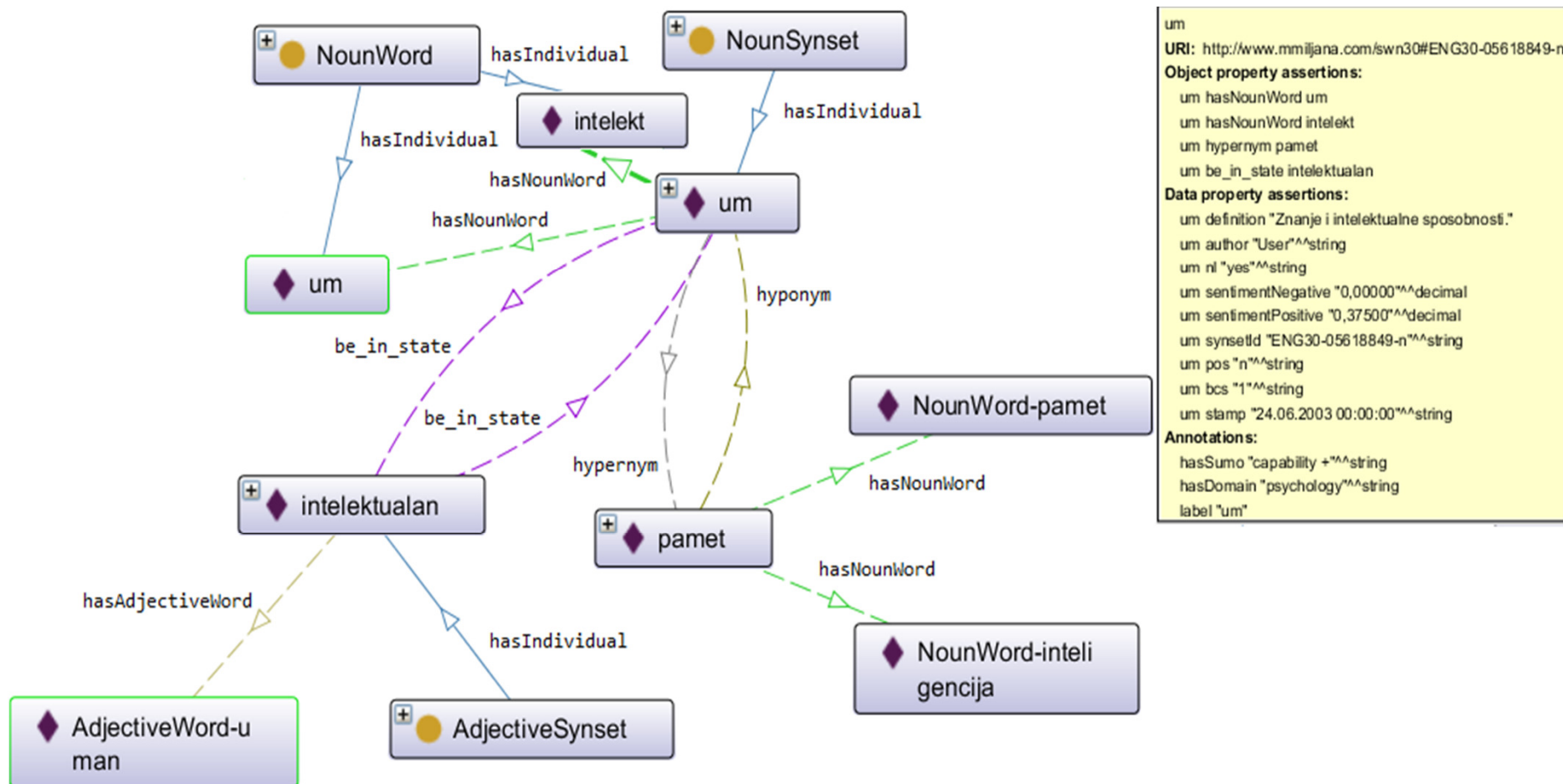
figura
SIMBOL
PROTEZA
PERIFRAZA
AFEREZA
ANAGRAM
EPENTEZA
DIJAREZA
METATEZA
APOKOPA

(b) SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ont: <http://www.mmijana.com/ontologies/ont.owl#>
SELECT DISTINCT ?figura
WHERE {
    ?figura ont:selzostavlja ?element.
    ?element ont:naziv ?naziv.
    FILTER( ?naziv = "SLOVO" ) }
```

figura
ELIZIJA
APOKOPA
SINKOPA
AFEREZA

Унапређење скупа предиктора – онтологија српског *WordNet-a*



Унапређење скупа предиктора – препознавање фигура којима се мења сентимент поларитет речи и фраза

иронија, сарказам, оксиморон

“**Brz** je kao **kornjača**. “

“**Hrabar** je kao **zec**! “

$\{?n : \text{characteristic } ?p1. \ ?p1 : \text{near_antonym } ?p2\} \Rightarrow \{?n : \text{irony } ?p2\}$

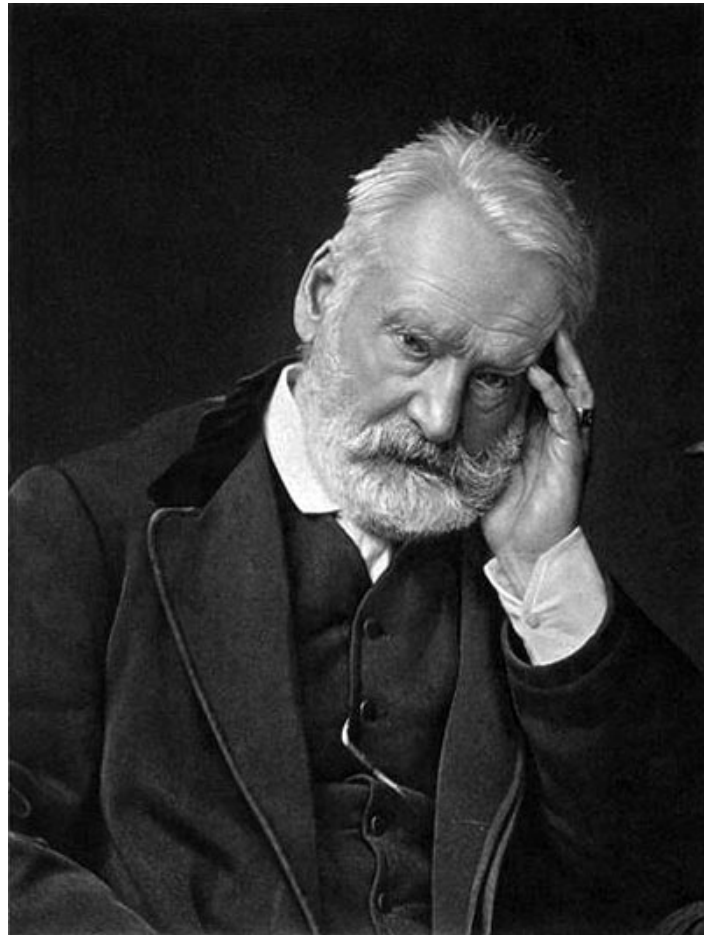
SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swm30: <http://www.mmljana.com/swm30#>
SELECT ?noun ?adjective1 ?adjective2
  WHERE { ?adjective1 swm30:characteristic ?noun.
          ?adjective2 swm30:near_antonym ?adjective1. }
ORDER BY ?noun
```

noun	adjective1	adjective2
pogled	hladan	vruć
pogled	hladan	topao
kornjača	spor	brz
puž	spor	brz
zec	plašljiv	hrabar
zec	brz	spor
strela	brz	spor
slika	lep	ružan
ideja	brz	spor
vetar	brz	spor
vatra	vruć	hladan
kamen	hladan	vruć
ugljenohidrat	sladak	gorak
led	hladan	vruć

1862.

?



!

