АВТОМАТИЧНА КЛАСИФИКАЦИЯ НА БЪЛГАРСКИ ПРИЛАГАТЕЛНИ ИМЕНА ПО ЕМОЦИОНАЛНИ ОСИ

Борис Крайчев, Иван Койчев

Софийски Университет „Св. Климент Охридски”
{kraychev, koychev}@fmi.uni-sofia.bg

***Резюме**: Настоящото изследване представя метод за автоматична класификация на българските прилагателни имена, както за позитивност, така и за позициониране върху предварително избрани емоционални оси като любов – омраза, щедрост – алчност, добрина – злина и др. В основата на изследването стоят данните за честотата на срещане на думите в документи от индекса на популярната машина за търсене bing, която дава информация както за броя документи съдържащи дадена дума, така и за броя документи, в които дадени две думи се срещат в определена близост. Резултатите от изследването отразяват съвременното използване на българския език в глобалната мрежа.*

***Ключови думи**: класификация на думи; анализ на чувства; емоционални оси*

1. Въведение

Известните уеб адреси (URL) в интернет отдавна надхвърлят един трилион, а съдържанието, споделено онлайн предлага неизчерпаеми възможности за лексически анализ. Един ресурс за изследване на естествения език, използван в интернет представляват публично достъпните машини за търсене. Изхождайки от хипотезата, че граматическата близост на дадени две думи води и до вероятна семантична прилика между думите, настоящото изследване се фокусира върху прилагателните имена от българския език. Чрез статистическата информация, получена от популярна машина за търсене, изследването построява *n*-мерно пространство обусловено от думи, характеризиращи конкретни противоположни емоции. Предложеният по-долу метод предлага модел за автоматична класификация на българските прилагателни имена в изграденото емоционално пространство.

2. Предишни разработки по темата

Изучаването на емоционалните нюанси в даден текст е обект на лингвистични изследвания от няколко десетилетия, а напоследък, след като обемът на свободно споделените мнения в мрежата надхвърли многократно редакционните текстови от основните медии, автоматизираният анализ на мнения става все по-популярен. Известно е, че едно и също съдържание може да бъде представено с разнообразни емоционални нюанси. Например, едно и

също събитие може да бъде представено като грандиозен успех или провал само чрез избора на подходящи думи. Историята на лингвистичния анализ върху емоционалното изразяване на мнения не е никак кратка.

В психологически изследвания в началото на шестдесетте години на миналия век, [1] предполага, че думите могат да бъдат разположени по семантични оси, и разработва експерименти, които са били използвани да прогнозират позиционирането на думите по тези оси, като "голям-малък", "топло-студено" и т.н. Тези оси обикновено се наричат „лингвистична скала”, определена от [2] през 1983г. като множество от контрастни изрази, които могат да бъдат разположени по оси, в зависимост от силата на тяхното значение по съответната ос. Друга подобна разработка са семантичните области, [3] и [4], които съответстват на група от думи, обхващащи някои семантични измерения, като "цветове".

В допълнение към тези посоки в научните изследвания, насочени към условно позициониране по семантични оси, други изследователи като Stone и Lasswell са започнали изграждане на семантични лексикони, в който думите са етикетирани с тяхната емоционална стойност. Например, в Lasswell Value Dictionary [5], думата “възхищавам” (admire) се маркира с положителна стойност по оста “уважение” (respect). Този речник означава думи с двоични стойности по осем основни измерения като богатство, власт, справедливост, уважение, просвещение, вещина, привързаност и благополучие). Проектът General Inquirer[6] е активен и днес и информация за него може да се получи на http://www.wjh.harvard.edu/~inquirer/.

По-съвременни експерименти допълват работата на тези ръчно етикетирани лексикони. Hatzivassiloglou и McKeown [7] се опитват да намерят тагове като “positive” и “negative” автоматично чрез статистически анализ на корпус от текстове. Авторите вземат редица често срещащи се прилагателни, за които те считат, че притежават скаларна или полярна ориентация и след това използват статистически данни за да определят дали двете прилагателни се появяват заедно в корпус като модел „X и Y”. Думи, които се появяват едновременно в такива модели се считат като притежаващи една и съща полярност. Отделно, класът с по-голям брой думи се счита, че е съставен от отрицателни думи, тъй като има повече отрицателни, отколкото положителни думи в английския език. Те постигат 92% точност над набор от 236 прилагателни, които те класифицират като положителни или отрицателни.

Подобни разработки са ползвали Wiebe [8] и Hindle [9], които предлагат метод за обогатяване на начален набор от прилагателни (seed adjectives) и генериране на речник чрез класификация на допълнителни думи.

Turney и Littman [10] предлагат ефективен начин за преценка дали дадена дума е позитивна или негативна. Използвайки множество от предварително класифицирани думи, например множеството от [7], те тестват колко често

самата дума се появява в контекста на множество позитивни думи (good, nice, excellent, positive, fortunate, correct, superior) и в контекста на набор от съответните негативни антоними (bad, nasty, poor, negative, unfortunate, wrong, inferior). За целта те използват оператора NEAR на все още популярната през 2003г. машина за търсене Altavista. Идеята им е да класифицират като позитивни думите, които се срещат по-често до позитивни думи и като негативни думите, които се срещат по-често до думите от негативният набор. Използвайки този метод те постигат 98.2% точност (accuracy) за 334те най-често използвани думи от тестовото множество на Hatzivassiloglou и McKeown [7]. Изхождайки от метода на Turney и Littman, друга група автори, обединени от компанията Clairvoyance, през 2006г. публикуват метод за анализ и класификация на емоционално заредени думи върху семантични оси [11]. Те изследват конструкции, изразяващи емоция (emotive patterns) – комбинации от глаголи и подсилващи думи като например "seem almost …", "feel so …", "appear too ..." като от тях чрез ръчна класификация оценяват най-продуктивните модели. Така полученият речник от емоционално заредени думи, авторите класират в семантично срещуположни групи, определящи емоционални оси. На следващ етап, използвайки модификация на метода на Turney и Littman определят близостта на дадена емоционално заредена дума до полюсите на така определените оси.

Следващият метод е близък до описаният по-горе, като е адаптиран за наличните ресурси в големите машини за търсене за български език.

3. Метод за групиране и оценка на думи по емоционални оси

Целта на метода е да направи извод за семантична ориентация за дадена дума чрез оценка на силата на асоциирането си с набор от позитивни думи, минус силата на асоциирането си с набор от негативни думи. Нека $\mathbf{Pwords}$ е множество от избрани думи с позитивна ориентация, а $\mathbf{Nwords}$ е множество от избрани думи с негативна ориентация. С $\mathbf{A(word_1, word_2)}$ да означим функция, даваща ни асоциативната близост между две думи. По-големите стойности на тази функция означават по-голяма асоциативна близост между тестваните думи, а отрицателните стойности означават взаимно изключване между двете думи т.е. наличието на една от тях води до отсъствие на другата. Приема се, че няма значение редът на аргументите на функцията. Така може да се построи следващата функция:

$$SOA(word) = \sum_{p \in Pwords} A(p, word) - \sum_{n \in Nwords} A(n, word) \quad (1)$$

която дава семантичната ориентация според асоциативната близост на дадена дума с избраните референтни множества. Положителните стойности на този

израз указват позитивна семантична ориентация, докато отрицателните сочат негативна.

Turney (10) използва функцията за оценка на взаимната информация (Pointwise Mutual Information - PMI) от теорията на информацията, за да изчисли силата на семантичната асоциативност между думите, а именно:

$$PMI(x,y) = log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (2)$$

Тази функция е симетрична и отговаря на изискванията за измерване на асоциативността. В този случай с $p(x,y)$ се означава вероятността думите $word_1$ и $word_2$да се появят заедно в документ. Ако се приеме, че двете думи са статистически независими, то същата вероятност би следвало да е равна на произведението $p(x)p(y)$. Отношението между последните два израза представлява измерение на статистическата зависимост между две думи.

Израза от (2) приема положителни стойности, когато двете думи имат смислова свързаност и се срещат заедно по-често от стандартно разпределение и съответно отрицателни стойности, когато двете думи са смислово изключващи се и твърде рядко се срещат заедно.

За да се изчисли вероятността от срещане на дадена дума може да се използва някоя от големите машини за търсене, която предлага оператор за близост и достъп през програмен интерфейс.

Вероятността от срещане на дадена дума може да се опише като:

$$p(word_1) = \frac{hits(word_1)}{N} \quad (3)$$

където N е общият брой индексирани документи, а функцията hits дава броя на документите от индекса, съдържащи термина, аргумент на функцията. Следователно асоциативната близост между две думи се оценява с израза:

$$PMI(word_1,word_2) = \log\left(\frac{\frac{1}{N}hits(word_1\ NEAR\ word_2)}{\frac{1}{N}hits(word_1)\frac{1}{N}hits(word_2)}\right) = log\left(N\frac{hits(word_1\ NEAR\ word_2)}{hits(word_1)hits(word_2)}\right) \quad (4)$$

Последният израз не зависи от броя на индексираните документи и комбинирайки го с (1) се получава обща формула за изчисление на семантична ориентация:

$$SOPMI(word) = \sum_{p \in Pwords} PMI(p,word) - \sum_{n \in Nwords} PMI(n, \quad (5)$$

или

$$SOPMI(word) = log\left(\frac{\prod_{p \in Pwords} hits(word\ NEAR\ p).\prod_{n \in Nwords} hits(n)}{\prod_{p \in Pwords} hits(p).\prod_{n \in Nwords} hits(word\ NEAR\ n)}\right) \quad (6)$$

Така всяко изчисление на семантична ориентация изисква равен брой заявки към машината за търсене на броя на референтните думи от множествата Pwords и Nwords.

Последният резултат (6) може да бъде използван както за изчисление на позитивна и негативна семантична ориентация, така и за оценка на думи спрямо поляризирани множества от думи с емоционална стойност, които определят своеобразна емоционална ос. Такива примери са например *обичан-мразен*, *вълнуващ-скучен*, *добър-лош*, *щастлив-нещастен* и др.

4. Изчисление на емоционалната ориентация на думи от българския език по предварително дефинирани оси

4.1. Избор на машина за търсене

За да бъде изследването максимално пълно има нужда от голям текстов индекс на възможно повече индексирани български документи. Поради обема на изследването има нужда и от програмен интерфейс, по който да бъдат зададени около десет хиляди заявки за търсене. За оценката на близост на думи е нужно наличието и на оператор за оценка на близост на думи. Изискванията към машините за търсене могат да се систематизират в следващата таблица:

Таблица 1 Оценка на популярните машини за търсене.

Име	Наличие на глобален индекс	Филтриране по език	Филтриране по близост на думи	Програмен интерфейс (API)	Информация за броя резултати (hits count)
Google.com	Да	Да	AROUND(n)	До 2009г.	Приблизителен
Yahoo.com	Да	Да	NEAR - Неофициална поддръжка	Да	Не
Bing.com	Да	Да	NEAR:n	Да	Да

Мярката „брой резултати" е от най-голяма важност и това се потвърждава от факта, че Google.com оценява резултатите на думата „честен" като 437 000, а при заявка „смел AROUND(10) честен" резултатите са оценени на 504 000. Липсата на програмен интерфейс също прави невъзможен изборът на тази машина за търсене въпреки, че по субективна оценка индекс □ е най-пълен.

Yahoo BOSS е добра алтернатива за търсене, но при тази услуга липсва информация за броя на намерените резултати. Третият избор - Bing.com предлага приятна изненада с точна оценка на броя върнати резултати, сравнително пълен индекс за български документи (по наша субективна оценка), филтриране по език, оператор за близост на думи и програмен интерфейс с информация за броя на резултатите.

4.2 Избор на думи, определящи емоционални оси

Поради ограничения брой индексирани документи в Bing.com – за сравнение, думата „честен” е индексирана около 100 000 пъти в bing.com и около 500 000 пъти в google.com – трябва да се изберат относително разпространени референтни думи, които да определят емоционалните оси, по които да се класират останалите думи. За целта се избират няколко основни емоционални противоположности, като:

- **Любов – Омраза**
- **Щастие – Нещастие**
- **Очакван – Изненадващ**
- **Полезен – Вреден**
- **Щедрост – Алчност**
- **Доброта – Злина**
- **Красота – Грозота**
- **Ум – Глупост**
- **Яснота – Оборканост**
- **Вълнуващ – Скучен**
- **Смелост – Страх**
- **Трудолюбие – Мързел**
- **Уверен – Подтиснат**
- **Богатство – Бедност**

Като втори етап се избират прилагателни, съответстващи на референтните области, например алчен – щедър, ясен – оборкан, умен – глупав и др. Чрез синонимен речник се разширява множеството от предложения за референтни прилагателни, достигайки до около 250 предложения и чрез програмния интерфейс на bing.com се записва разпространеността на думите в индексираните документи. След сортиране на думите по разпространеност, се избират най-разпространените прилагателни за референтни. Така се получават окончателните емоционални оси:

Таблица 2 Референтни прилагателни имена, формиращи емоционални оси.

Положителна	Отрицателни
Разбран	Оборкан
Обичан	Мразен
Хубав	Грозен
Умен	Глупав
Очакван	Изненадващ
Вълнуващ	Скучен
Смел	Страхлив

Активен	Мързелив
Щедър	Алчен
Верен	Подъл
Уверен	Подтиснат
Богат	Беден
Добър	Лош
Полезен	Вреден
Щастлив	Нещастен

Оценка на положителност

Колоните от референтни прилагателни имена от представляват, сами по себе си, набор от референтни позитивни и негативни думи. Следователно, използвайки формулата (6), може да се оцени дадено прилагателно име за положителност. Използвайки програмният интерфейс на bing.com се оценява броят взаимни срещания на две думи. Това става със заявки от типа:

$$(word\ NEAR{:}\ 10\ Rword)\ language{:}\ bg \qquad (7)$$

където $word$ е тестваната дума, а $Rword$ се замества с референтните думи от позитивното и негативното множество. Броят заявки към програмния интерфейс на машината за търсене е равен на произведението на броя на елементите от референтните множества, броят на тестваните думи и броя на граматичните форми на прилагателните имена, като се поставя изискване за съгласуваност по род и число на двете тествани думи. Последното изискване има за цел да изключи наличието на близост между тестваните думи, когато те адресират различни обекти и евентуално изразяват емоции без пряка връзка в употребата на термините.

След прилагане на теста върху 70 случайно избрани прилагателни имена се получават следните резултати:

Таблица 3. Оценка на положителност на случайно избрани думи.

Дума	Оценка
нов	49.12
жив	48.03
страхотен	38.92
честен	38.61
млад	38.15
истински	37.76
мощен	36.76
Бърз	35.12
малък	33.23

делови	28.98
Прав	25.89

…

Стар	1.41
разумен	-0.06
грешен	-0.61
ужасен	-1.15

…

напрегнат	-2.63
противен	-3.21

незаконен	-3.21
неприятен	-9.28
Мазен	-10.28
неправилен	-14.95
мръсен	-14.95
Опасен	-18.74
ограничен	-21.61
Долен	-25.84

От резултатите представени в таблица 3 е видно, че избраните референтни думи отчетливо разграничават позитивни и негативни думи и алгоритъмът притежава висока прецизност за думите с голяма абсолютна стойност на оценката. Добавянето на допълнителни референтни думи и емоционални оси ще подобри класификационните способности на алгоритъма. Съществено значение за успешната класификация е богатството на текстовия индекс с български документи.

Таблица 4. Оценка на емоционално значими думи по семантични оси.

Думи	bing count	хубав-грозен	умен-глупав	активен-мързелив	верен-подъл	богат-беден	добър-лош	полезен-вреден	щастлив-нещастен
Нов	2 100 000	1.61	7.56	9.39	7.38	8.76	2.34	1.01	11.06
Жив	224 000	8.38	6.57	6.41	6.39	-0.21	9.06	7.44	3.99
Страхотен	224 000	9.97	6.57	-0.25	6.39	-0.21	2.53	7.44	6.49
Честен	50 200	6.39	9.36	-0.25	6.39	6.45	10.65	-0.21	-0.17
Млад	283 000	8.7	1.49	7.98	-0.27	2.24	4.07	7.44	6.49
Истински	607 000	8.97	-0.09	7.4	7.38	-0.21	3.52	6.45	4.34
Мощен	139 000	6.39	6.57	-0.25	6.39	8.76	9.28	-0.21	-0.17
Бърз	440 000	7.97	8.15	0.74	-0.27	-0.21	10.48	8.44	-0.17
Малък	693 000	8.97	-1.08	7.98	6.39	-7.86	10.56	8.44	-0.17
Делови	78 900	8.38	-0.09	-0.25	6.39	6.45	8.48	-0.21	-0.17
Прав	217 000	7.97	-0.09	8.72	6.39	-6.87	0.51	-0.21	9.48
Успешен	151 000	9.19	-0.09	-0.25	-0.27	6.45	2.28	-0.21	8.48
…									
Стар	404 000	0.47	-8.32	7.4	-0.27	0.37	2.14	6.45	-6.83
Разумен	39 700	-6.92	-0.09	-0.25	-0.27	-0.21	8.06	-0.21	-0.17
Грешен	94 500	-0.26	-0.09	-0.25	0.09	-0.17	0.03	-0.27	0.31
Ужасен	99 600	-0.26	-0.09	-0.25	-0.27	-0.21	0.32	-0.21	-0.17
…									
Противен	136 000	-8.5	-0.09	-0.25	-0.27	-0.21	6.49	-0.21	-0.17
Незаконен	50 200	-0.26	-0.09	-0.25	-0.27	-8.45	6.49	-0.21	-0.17
Неприятен	43 600	-0.26	-0.09	-7.9	-0.27	-0.21	6.49	-6.87	-0.17
Мръсен	42 500	-6.92	-0.09	-0.25	-0.27	-0.21	-0.17	-0.21	-6.83
Опасен	111 000	-0.26	-0.09	-0.25	-0.27	-0.21	-7.82	-9.66	-0.17
Ограничен	87 500	-0.26	-7.74	-6.91	-0.27	-6.87	7.48	-6.87	-0.17
Долен	102 000	-6.92	-0.09	-0.25	-6.93	-0.21	-11.06	-0.21	-0.17

Оценка на разположение по емоционални оси

Друга възможност за оценка е прилагането на подхода на Turney и Littman (10) не само върху групи от позитивно и негативно ориентирани думи, а върху семантично противоположни термини, определящи емоционални оси. Прилагайки изчислението (6) върху думите от Таблица 2, се получава оценка за семантична близост на дадена дума с върховете на избраните оси. Например за думата *жизнен*, срещаща се 54800 пъти в индекса, се получава оценка за близост с *хубав – 6.39*, *добър – 7.48*, *полезен – 6.45* и неутрално отношение към осите *умен-глупав: -0.09*, *верен-подъл: -0.27*, *щастлив-нещастен: -0.17*. Сред негативните думи, интересен извод се постига за думата мръсен – *грозен: -6.92*, *нещастен: -6.83*. Повече информация може да се получи от таблица 4.

Дискусия

Предложеният метод за класификация на думи по емоционални оси представлява машинно самообучение без учител. Основният източник на знания за алгоритъма са индексираните документи в корпуса на машината за търсене. Можем само да съжаляваме, че по технически причини (липса на подходящ програмен интерфейс) бе избрана машина с приблизително десет пъти по-беден корпус от български документи спрямо най-обхватната машина за търсене на компанията Google.

С нарастване броя на онлайн публикациите подобни изследвания биха се превърнали в коректен индикатор за еволюцията на българския език и за емоциите, които вълнуват онлайн потребителите. Разбира се, методът може да бъде приложен към всяко цифрово съдържание. Например при наличие на голям и разнообразен корпус от художествена литература на български език, класификаторът може да бъде се приложи за разпознаване на жанрове и автоматична класификация на литературни произведения.

Резултатите в матрицата от **Error! Reference source not found.** загатват и един интересен факт: Вероятно актуалните теми в бита на българите увеличават броя на публикациите в интернет по конкретната тема и това подсилва връзките между думите, използвани в дадената предметна област. Този извод е породен от резултати като близостта на *опасен* с *вреден*, която вероятно е предизвикана от публикации на тема здравословно хранене. Подобна мотивация има вероятно и за близостта на думата *ограничен* с думата *добър*, защото *ограничаването* на дадени вещества в диетите за хранене е насърчавано като *добро* за човешкото здраве. Разбира се това са само предположения на автора и доказателство или отхвърляне може да се получи от корпус, позволяващ заявки с изисквания към частите на речта и филтриране на резултатите с етикетни последователни правила (LSR).

Друго интересно приложение на представения класификатор би била оценката на споделените емоции от интернет потребители във връзка с конкретни събития. Това би представлявало неформален индикатор на обобщеното мнение на онлайн аудиторията.

Обобщение

Настоящото изследване по информацията, достъпна на авторите, е първото онлайн изследване на българския език, което адресира настроенията, споделени в глобалната мрежа и класифицира българските прилагателни имена по емоционални оси. Несъмнено с увеличаване на броя на публикувани и индексирани документи в глобалната мрежа ще има възможност за допълнителни изследвания както на българския език, така и на методите за автоматична класификация на текстове. Бъдещата работа по темата включва усъвършенстване на изискването за близост на думите с включване на езикови

конструкции (Label Sequential Rules) и подобряване на избора на референтни думи чрез предварителен анализ на статистическите данни за честотата на срещане на всеки термин. Макар и изследването да не съдържа формална оценка за точност на алгоритъма, по субективна оценка на автора класификацията е оценена като успешна. Задачата за класификация на чувства и настроения от свободен текст не винаги има категоричен отговор дори и при ръчна обработка от лингвисти, а автоматичната класификация на текстове може да бъде използвана за индикатор на тенденции и обработка на големи обеми от информация.

Литература

1. J. Deese, The Associative Structure of some Common English Adjectives, Journal of Verbal Learning and Verbal Behavior 3(5), pp. 347-357, 1964.
2. L. S. C., Pragmatics., Cambridge University Press, 1983.
3. Berlin B. and Kay P., Basic color terms: their universality, Berkeley: Oxford: University of California, 1969.
4. A. Lehrer, Semantic Fields and Lexical Structure, London: North Holland, 1974.
5. Lasswell, H.D. and Namenwirth, J.Z., The Lasswell Value Dictionary, 1969: Yale University Press, New Haven.
6. Stone, P. J., Dunphy, D. C.,. Smith, M. S and Ogilvie, D., The General Inquirer: A Computer Approach to Content Analysis, Cambridge, MA: MIT Press, 1966.
7. Hatzivassiloglou, V., and McKeown, K. R., Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning, Proceedings of 31st Annual Meeting of the Association for Computational Linguistics, pp. 172—182, 1993.
8. J. Wiebe, Learning subjective adjectives from corpora, Proceedings of AAAI/IAAI 2000, pages 735—740, 2000.
9. D. Hindle, Noun classification from predicate argument structures, Proceedings of the 28th Annual Meeting of the ACL, pages 268-275. ACL., 1990.
10. Turney, P.D., and Littman, M.L., Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems (TOIS), 21 (4), 315-346, 2003.
11. Grefenstette G., Qu Y., Evans D., Shanahan J., Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes, Computing Attitude and Affect in Text, Springer, 2006.

AUTOMATIC CLASSIFICATION OF BULGARIAN ADJECTIBES ON EMOTIONAL SEMANTIC AXES

Summary: *This study presents a method for automatic classification of Bulgarian adjectives as both positivity and positioning on pre-selected emotional axes like love - hate, generosity - greed, goodness - evil and others. In the basis of the study stands the frequency of occurrence of words in documents from the index of the popular search engine Bing, which provides information as to the number of documents containing a word and the number of documents where two words are found in a nearby. The survey results reflect the use of modern Bulgarian language on the Internet.*