UDK: 811.163.2’322.2

УНАПРЕЂИВАЊЕ ДРУШТВЕНИХ ИНФОРМАТИВНИХ МЕДИЈА НА БУГАРСКОМ ПУТЕМ ОБРАДЕ ПРИРОДНИХ ЈЕЗИКА

Валентин Жиков, valentin.zhikov@ontotext.com Ontotext AD
Ивелина Николова, iva@lml.bas.bg Институт за информационо – комуникационе технологије (IICT), Бугарске академије наука и Ontotext AD
Лаура Толош, laura.tolosi@ontotext.com Ontotext AD
Јавор Иванов, yavor@xenium.bg Xenium Ltd.
Борислав Попов, borislav.popov@ontotext.com Ontotext AD
Георги Георгијев, georgiev@ontotext.com Ontotext AD

С енглеског превела Јелена Бајић

Апстракт
У овом раду, уводимо систем заснован на техникама обраде природних језика које имају за циљ унапређивање друштвених информативних медија на бугарском. Тим системом се решава задатак класификације докумената са више класа и више обележја. Примењујемо алгоритме на збирку чланака из медија објављених на сајту Svejo.net, популарном бугарском веб ресурсу који обухвата садржаје које креирају корисници. У нашим алгоритмима се користе методи класификације „један против свих“, широко распрострањени у рачунарској лингвистици. Описујемо алгоритме, својства која су употребљена и процењујемо утицај тих својстава на делотворност модела. Тиме показујемо да сазнања о кориснику и понашању корисника могу много да допринесу побољшању учинка. Осим тога, упркос чињеници да су целу нашу збирку направили корисници друштвених медија, квалитет резултата класификације је упоредив са квалитетом од раније познатих студија.
Бавимо се и задатком аутоматске екстракције кључних речи и израза из неструктурисаног текста и прилагођавамо га потребама Svejo.net-а ради индуковања „тема“. Теме се дефинишу као одломци текста у којима је сумирана суштина неког чланка. Оцењујемо делотворност неколиких генеричких метода екстракције кључних речи и израза примењених на корпус

се ослањају на широко прихваћене методе проналажења информација и машинског учења и функционишу независно од језика. Такође разматрамо дејство компоненте коју чини стемер на прецизност екстракције кључних израза. Задовољавајући учинак наших модела, упркос ограниченом лингвистичком знању уграђеном у њих, препоручује их да буду полазна основа за екстракцију кључних речи и израза у бугарском језику.

Кључне речи:
обрада природних језика, машинско учење, приступи који не зависе од језика, екстракција кључних речи, класификација текста

1. Увод

Потреба за техникама за обраду природних језика ради унапређења медија какав су друштвене мреже је неоспорна. Сајтови друштвених медија имају приступ огромној количини текстуалних података које је потребно ефикасно аутоматски обрађивати. У овом раду разматрамо два различита аспекта употребе обраде природних језика, са циљем да се унапреди сервис друштвене мреже Svejo.net на бугарском. Један од аспеката је аутоматска класификација текста збирки вести да би се убрзало додавање нових чланака овом сервису. Други аспект је аутоматска екстракција тема из чланака које се могу даље користи за груписање чланака.

Svejo.net је веома популаран сајт и један од првих бугарских друштвених медија који посећује преко 20% корисника бугарских информативних сајтова. Свакога дана корисници Svejo.net додају преко 1 500 вести и чланака, 3 000 коментара и гласају преко 15 000 пута. Садржајем Svejo.net-а у потпуности управљају корисници, јер у редакцији нема новинара, већ само неколико модератора чије су дужности што се тиче садржаја сајта веома ограничене. Svejo.net се, дакле, у потпуности ослања на друштвени елемент и омогућује корисницима да додају занимљиве линкове, при чему су у центру пажње текстови информативног карактера и мултимедијални садржаји (видео снимци, слике, итд.). Мада не постоје језичка ограничења, текстови до којих воде линкови са сајта Svejo.net су углавном на бугарском језику, а понеки прилози су на енглеском, француском, немачком, руском, итд.

Популарност друштвених медија је у многоме повезана са степеном интуитивности и лакоће коришћења њихове сумеђе. Приликом додавања текстуалног садржаја на сајту Svejo.net, корисници морају ручно да додају кратак опис, категоризују и идентификују теме о којима је реч у документу. Иако је процес делимично аутоматизован (на пример, кратки опис – извод из текста се предлаже кориснику), још увек се категоризовање и придруживање тема ради ручно. Један од корака да би се олакшало додавање нових чланака сајту Svejo.net је да се омогући аутоматско препознавања теме чланка и кључних термина и да се корисници ослободе спецификовања свих тих елемената, те да могу да додају чланак једним кликом. Овај проблем би могао делимично бити решен применом класичне технике обраде природних језика, аутоматске класификације текста. Класификација текста се већ годинама проучава и још

увек представља изазов када су у питању отворени домени. Примењујемо машинске класификаторе при решавању задатка класификације текста са више обележја, више класа и за више језика, који су осмишљени тако да одговарају конкретним потребама Svejo.net-а. Показујемо како се традиционалне технике машинског учења могу унапредити опцијама које су карактеристичне за поједине кориснике и одсликавају њихово понашање.

Теме представљају кратак сажетак и обухватају срж текстуалног документа. Корисне су за аутоматизовану и ефикасну категоризацију докумената, вођене упите, брзи преглед докумената, при чему се визуелно истичу важни изрази. Оне чине моћну основу за мерење сличности докумената (Gutwin et al. 1999, Jones 1998, Witten 2003). Популарни бугарски медијски ресурс, Svejo.net, користи теме за описивање докумената, за прегледање збирке докумената и као основу за груписање докумената.

У општем случају, када је потребно одабрати тему за збирке докумената (на пример, научних чланака), потенцијалне теме могу бити унапред одабране, кључне речи или кратки текст на који нису примењена ограничења. Теме на сајту Svejo.net се добијају од мета-етикете за кључне речи, када год она постоји у изворном садржају, или теме додељује тим за подршку овог сајта. Међутим, често се кључне речи из мета-етикете генеришу токенизацијом наслова чланка, чиме се не добија довољно прецизан опис. Истовремено, члановима тима за подршку, бављење свим прилозима који немају тему одузима сувише времена. Стога је аутоматска екстракција тема од изузетног интереса за Svejo.net. Овај задатак се у научној литератури такође назива екстракција кључних речи и израза.

Остатак овог чланка је структурисан на следећи начин: сродни радови су представљени у одељку 2, експериментални скупови података су описани у одељку 3, кратак преглед система је дат у одељку 4, у одељку 5 су представљени методи, док се у одељку 6 налазе резултати и анализа грешака, а закључак и упутства за будуће активности су дати у одељку 7.

2. Сродни радови

Разни алгоритми за надзирано учење, укључујући наивни бајесовски, машине са потпорним векторима, „boosting“, учење правила постижу релативно добре резултате у домену класификације текста (Lewis 1998, McCallum и Nigam 1998, Sahami 1996, Dumais et al. 1998, Joachims 1998, Schapire и Singer 2000, Cohen и Singer 1999, Slattery и Craven 1998, Yang 1999). Треба имати у виду чињеницу да од свих горе поменутих техника, ни за један метод не постоје докази да може значајно и конзистентно да надмаши друге методе када се примени на бројне домене и језике.

Модели максималне ентропије се често користе за класификацију текста (Nigam et al. 1999). Постоје радови у којима се описује избор својстава за такве моделе, као на пример (Mikheev 1998) (технички апстракти) за корпус RAPRA. У (Ratnaparkhi 1998), се пореде модели максималне ентропије и дрва одлучивања и показује се да модел максималне ентропије надмоћан у класификовању неких од класа у Ројтеровом скупу од 21 578 података.

Додељивање више од једног обележја документу, што се назива класификација текста са више класа и више обележја представља занимљив проблем и налази се у центру пажње овог рада. Задатку додајемо још један извор комплексности, односно, више језика. У (Luo и Nur Zincir-Heywood 2005) се пореде два алгоритма за машинско учење: kNN класификатори (засновани на алгоритму k-најближих суседа) и „латентно семантичко индексирање“. Аутори налазе да први систем има боље резултате када се примењује на до-

кументе са више обележја, док се други показао бољим за документе са једним обележјем. Закључују да учинак система зависи од примењеног скупа података, као и од циља примене.

Скорашње студије приступају решавању истог задатка применом више класификатора који функционишу по принципу „један против свих“ (Zelaia et al. 2011).

Можда је од избора метода класификације још важнији избор својстава. Студије су показале да су, за извршавање задатка класификације веб страница, својства екстрахована из полуструктурисаног HTML-а експресивнија него својства која се традиционално користе за класификацију чистог текста. У таква својства спадају породице HTML етикета, URL веб странице, HTML мета-етикете попут кључних речи, суседне странице, сидра, заглавља, итд. (Qi и Davison 2009).

Постоје два општа приступа извршавању задатка екстракције или додељивања кључних речи и кључних израза из неструктурисаног текста. Први приступ не подразумева надзирање и заснива се на претпоставци да се кључне речи често појављују у документу, али ређе у целој збирци докумената. Стога се користи популарна тежинска мера TF-IDF. Бројни радови показују да је метод TF-IDF веома ефикасан у одређеним доменима (Frank et al. 1999, Hulth 2003, Ha Cohen-Kerner et al. 2005). Да би се добили поуздани TF-IDF резултати, корпус мора бити релативно велики. У (Matsuo i Ishizuka 2004), предлаже се конкурентан метод који користи расподелу истовремених догађаја и стратегију груписања за екстраховање кључних речи која се не ослања на велики корпус. Други аутори користе додатне изворе знања са веба, а та идеја се примењује и у овом рукопису. У (Turney 2002, Inkpen i Desilets 2004), аутори процењују вредност тачкасте узајамне информације да би одабрали кључне речи. Методи засновани на графовима, слични Google-овом алгоритму PageRank (Brin и Page 1998) су такође предлагани. У (Wan et al. 2007), се примењује техника учења уз утврђивање за симултану екстракцију кључних речи и резимирање текста, а полази се од претпоставке да важне реченице обично садрже кључне речи. Извршавањем сродног задатка који се назива додељивање кључних речи, омогућује се додељивање кључних речи само из предефинисаног речника (Dumais et al. 1998). У овом раду не користимо предефинисани речник, јер желимо флексибилност и брзо прилагођавање новим темама које се брзо појављују на сајту Svejo.net.

Екстракција кључних речи се такође може дефинисати као надзирани задатак класификације који се може решити уз помоћ техника машинског учења (Frank et al. 1999, HaCohen-Kerner et al. 2005, Turney 2000, Turney 2002, Turney 2003). Алгоритам за учење класификује на основу скупа својстава, потенцијалне речи и изразе пронађене у документу у оне који су одговарајући - позитивне (кључне речи) и оне који нису – негативне (речи које нису кључне). Међу корисним својствима су TF-IDF и његове варијације, положај кључног израза у односу на почетак документа, врсте речи, корени, леме, релативне дужине израза итд. (Turney 2002).

3. Скупови података

Материјал за наш корпус, за задатак класификације је сакупљан на сајту Svejo.net неколико година и обухвата скоро 400 000 докумената на бугарском и другим језицима, укључујући енглески, француски, немачки и руски (колико год да су мало заступљени у односу на бугарски). Подаци су у XML формату и сваки документ садржи следеће елементе: *наслов*, *резиме*, *ид_корисника*, *тип_медија*, *тагови*, *категорије*, *направљен_у* и *ажуриран_у*. Последња два елемента садрже

датуме. *Резиме* сваког документа се екстрахује из онлајн чланка и садржи до 1 000 карактера са почетка текста. Све HTML етикете су уклоњене и остаје само текст. *Наслов* садржи наслов чланка, *тагови* су слободан текст и састоје се од кратких одломака текста који одражавају садржај чланка, а категорије додељују корисници са предефинисаних листа. Сваки документ може имати више *тагова* и *категорија*, а то могу бити: *друштво*, *технологија, наука, пословање, политика, спорт, уметност, здравље, забава, стил живота, куповина*. Преко 9% чланака је сврстано у више категорија. Заступљеност докумената по категоријама је дата у Табели 1. Најпопуларније категорије су *друштво* и *забава*. Одмах за њима следи *стил живота*, категорија додељена у случају 18% докумената. Око 10% докумената је сврстано у категорије *технологија, спорт* и *здравље*. Најмање популарна категорија на сајту Svejo.net је *куповина* у коју је сврстано само 930 чланака, што представља мање од четвртине процента целокупне збирке докумената.

Ради извршења задатка екстракције тема, користимо други скуп докумената, припремљен у ту посебну сврху. Иако наш систем може да обрађује корпусе на више језика, наша анализа се фокусира на текстове на бугарском, пошто је наша апликација намењена сајту Svejo.net. Скуп података који представља златни стандард садржи углавном документе и анализе информативног карактера, са нагласком на политичке теме. Да би се обезбедио добар квалитет анотација, одабрали смо само документе којима су кључне речи додали чланови тима за подршку сајта Svejo.net или аутори докумената. Коначни скуп података обухвата 1 798 чланака подељених у групе за обуку (70%), развој (10%) и тестирање (20%), извучених насумично из целе збирке.

Табела 1. Распоред чланака по категоријама

категорија	број чланака	% корпуса	просечан број речи
друштво	88425	22,53	38,54
забава	82839	21,11	30,44
стил живота	71151	18,13	41,54
технологија	42399	10,80	22,25
спорт	37092	9,45	36,87
здравље	36180	9,22	39,59
пословање	24759	6,31	38,11
политика	21692	5,53	44,17
уметност	17658	4,50	36,86
наука	12539	3,19	42,53
куповина	930	0,24	64,56

Осим скупа података који представља златни стандард, индексирали смо и обимну збирку чланака преузетих са Svejo.net-а, при том не узимајући у обзир тагове, да бисмо добили репрезентативнију статистику фреквенције речи.

Пре вршења експеримената, до извесне мере смо спровели претходну обраду (конверзија у мала слова, уклањање нумеричких токена, стеминг, итд.) кључних речи и израза из златног стандарда, да бисмо обезбедили компатибилност са нашим скупом кандидата.

4. Опис система

Циљ ове студије је пружање подршке систему који ради у реалном окружењу, чиме ће се олакшати дељење садржаја на сајту који по типу спада у друштвене медије. Развојни процес се састоји од неколико итерација обуке, тестирања и провере, што је скупо, с обзиром да се ради ручно и имајући у виду потребне рачунарске ресурсе. Итерације се врше у сарадњи са тимом за подршку сајта Svejo.net. Сваку итерацију ажурирања обавља тај тим путем специјализоване сумеће експониране кроз низ услуга на вебу.

У развојној фази су честа ажурирања

модела и скупова података. Развојни циклус укључује анализу грешака у класификацији, при чему се користе документи који раније нису виђени, ревизију скупова података из златног стандарда, прибављање додатних анотираних чланака и поновну обуку модела. Имајући све ово у виду, развили смо алгоритамско решење за пакетно учење које подржава итеративно ажурирање и омогућава да лаици са лакоћом користе систем. Саставни део система су веб методи за обележавање докумената који раније нису виђени, поновну обуку модела и проналажење информација о статусу система. Метод додељивања обележја прихвата да се чланци прилажу у XML формату и генерише одговор у том истом формату који се може машински обрађивати и који садржи предвиђања о категоријама. Збирке докумената којима су додељена обележја која служе за развој модела се унапред постављају у репозиторијум - фасциклу којој сервер приступа путем генеричког протокола за трансфер датотека. Збирке могу бити доступне као збирке докумената у XML формату које се налазе у подфасциклама репозиторијума или у архивским датотекама у .zip формату, у којима се налази пакет докумената произвољне величине. Стање и садржај репозиторијума у коме се налазе документи, број активних модела за додељивање обележја, као и број доступних дозвола за паралелни приступ систему се саопштавају када се позове специјализована метода која се односи на статус система.

Поновна обука система се може активирати позивом сервисног метода, тако што се прецизира путања која води до одређеног скупа података.

5. Метод

5.1 Аутоматска класификација текста

Пошто је скуп својстава која се користе за обуку најважнији фактор који утиче на учинак модела, један од важних доприноса овог рада је обликовање својстава. Дефинисали смо известан број својстава, нека која зависе од текстуалног контекста („врећа речи“) и нека друга која се ослањају на мета-податке добијене уз документ: *тип медија, идентификатор корисника* и *тагове* које прилажу корисници. Такође смо експериментисали са везама између *тага* и *идентификатора корисника* и над таговима израчунали *n*-граме карактера. Оцењивали смо какав је допринос сваког типа својстава перформанси система. Својства су независна од језика и не користимо никакво лингвистичко знање ни ресурсе за бугарски (главни језик по заступљености у нашем скупу података). Осмислили смо наш алгоритам као чврсту класификацију. На пример, документ је класификован тако што је примењен приступ „један против свих“. Обучавамо бинарни класификатор за сваку категорију и сакупљамо све позитивне класификације да би било могуће додељивање више обележја једном документу. Класификација се обавља уз помоћ Edlin-a[1] са DSL-ом и софтверског слоја за обликовање својстава (Ganchev и Georgiev 2009). Систем се појављује на Svejo.net-у у форми Ontotext-ових услуга KIM Enterprise[2].

Методи класификације који се користе су наивни: бајесовски, максимална ентропија, перцептронски (Crammer et al. 2006) и MIRA (Rosenblatt 1958). За обуку користимо око 70% целокупне збирке, 15% је резервисано за развојни скуп, док је преосталих 15% намењено за оцену резултата примене класификатора. За класификаторе који користе наивни бајесовски метод, оптимизујемо хипер-параметар који

1 http://www.edlin.sourceforge.net
2 http://www.ontotext.com/kim

контролише степен заглађивања (користимо лапласовско заглађивање) у развојном скупу. Пошто је наш циљ да направимо систем који функционише у стварном свету, постоје својства која омогућавају обуку крајњих корисника у реалном времену. Параметар заглађивања је за овај систем постављен унапред, стратификовани сегменти за обуку и тестирање су динамично изграђени за обучавање сваког класификатора, а однос између података који се користе за обуку и тестирање је 9:1. Обука и евалуација се обављају применом аутоматске рутине која врши екстракцију свих класа заступљених у датој збирци докумената, припрема стратификоване и насумично изабране сегменте за обучавање и тестирање за сваку класу, анализира и складишти резултате и меморише генерисане моделе.

5.2 Екстракција/додела тема

Кључне речи и изразе бирамо из скупа кандидата који се састоји од *n*-грама предефинисаних величина (у нашим експериментима су коришћени униграми и биграми). Током фазе која претходи обради, врши се токенизација докумената, одбацују се речи које нису од значаја за обраду („стоп-речи“), токени се конвертују тако да се пишу малим словима, а они који садрже нежељене карактере (на пример, бројеве и знакове интерпункције) се елиминишу.

Упоређујемо два приступа екстракцији кључних речи у бугарском – ненадзиран и надзиран приступ.

Ненадзирани приступ се базира на хеуристици TF-IDF типа. TF-IDF речи која је кандидат се израчунава уз помоћ традиционалне формуле. TF-IDF оцена кандидата за кључне изразе се израчунава на два начина: (*i*) применом традиционалне формуле, при чему кључни израз представља један токен (тај метод се овде назива **микс**) и (*ii*) на основу TF-IDF резултата његових саставних делова (метод **средње вредности**). Прецизније речено, ако је израз састављен од две речи, израчунавамо TF-IDF двеју речи које чине саставне делове израза и целог биграма, а затим рачунамо средњу вредност та три броја, да бисмо добили јединствени резултат за тај израз. Осим тога, када се рачуна резултат читавог израза дозвољено је филтрирање конституената чији индивидуални резултат је испод постављеног прага.

Кандидати су рангирани према својим TF-IDF оценама или просечној вредности тих оцена. Да би се одредио праг према коме се одређују најбољи кандидати, израчунава се најближи цели број већи од средње вредности оцене израза у нашем скупу података на основу којих се врши евалуација. За **микс** алгоритам је неопходно нормализовати TF-IDF вредности добијене у неком документу, односно, свести их на вредности између 0 и 1.

Други приступ је надзирана класификација. Тај метод предвиђа кључне речи из скупа кандидата, на основу групе ручно обележених примера који се користе у сврху обуке. Инспирисан је алгоритмом KEA (Witten 1999), који користи два основна својства: TF-IDF оцену сваког кандидата (ознака је **TFIDF**) и позициони помак (ознака је **pos**) који представља број токена који претходе првом појављивању израза-кандидата у тексту. Као и у изворном методу, дискретизовали смо ова својства применивши надзирани метод (Fayyad и Irani 1993), и упоредили резултате са онима добијеним применом ненадзиране стратегије дискретизације која их групише у одлагалишта једнаке величине. Групи својстава предложеној у изворном чланку, (Witten et al. 1999) додали смо и друга, што је резултирало бољим учинком. Конкретно, додали смо дужину кандидата у токенима (ознака је **len**) и буловски атрибут који показује да ли је токен уврштен у наслов или није. На крају, узели смо у обзир различите до сада описане

везе између обележја.

За класификацију смо користили два алгоритма примењена у Edlin-u (Ganchev и Georgiev 2009) – полиномни наивни бајесовски (**MNB**), перцептронски (**PER**) (Rosenblatt 58) и MIRA (**MIRA**) (Crammer et al. 2006). Као и у (Witten et al. 1999), из групе кандидата избацујемо оне кључне речи и изразе који се појављују у току анализе текста документа само једном.

TF-IDF оцене и неки од корака претходне обраде се спроводе кроз оквир Lucene[3]. Сви алгоритми за машинско учење и експерименти са надзираном дискретизацијом су спроведени у Edlin-у. Такође се користи и стемер Bulstem[4] описан у (Nakov 1998). Систем се појављује на Svejo.net-у у форми Ontotext-ових услуга KIM Enterprise.

6. Резултати и дискусија

6.1 Аутоматска класификација текста

Увидели смо да је наивни бајесовски метод изразито надмоћан у поређењу са осталим методама класификације. У наставку ће бити приказани само експерименти у којима се користи наивни бајесовски класификатор.

У Табли 2. је дат резиме учинка класификатора за сваку циљну категорију. Наведени резултати се односе на прецизност, одзив и макро F1 меру и представљају средњу вредност добијену у свакој категорији у 10 независних експеримената.

Основу поређења чини „вреће речи“ из наслова и самих текстова и макро F1 оцена нижа од 60% у свим категоријама. Разлог за лош учинак је вероватно то што сваки документ садржи ограничену количину текста. Укључивање тагова доводи до значајног побољшања учинка, а оцена F1 се пење до 67%. Својство *тип_медија* не побољшава оцену F1, ни када се користи самостално, нити у комбинацији са другим атрибутима мета-података. У другим моделима *ид_корисника* је најинформативније од три својства мета-података, нарочито у комбинацији са таговима докумената. То објашњавамо или тенденцијом неких корисника да додељују одређене тагове, или интересовањем сваког корисника за одређену категорију чланака. У нашим експериментима, највећу прецизност постиже систем који користи скуп својстава који се састоји од „вреће речи“ из текстуалног садржаја, свих својстава мета-података и комбинације *ид_корисника* и *тагова*. Систем у просеку постиже за 5% бољи резултат од основног модела. Проширење скупа обележја додавањем триграма карактера из тагова додељених одређеном документу доводи до смањења просечне оцене. *n*-грами из *тагова* се укључују да би се решили проблеми, попут појаве *тагова* у множини и једнини, са чланом или без члана, пошто су *тагови* слободан текст који исписују корисници приликом додавања новог ресурса на сајт Svejo.net. Иако је примећен известан напредак када се користе *n*-грами, а нарочито триграми из тагова, посебно у поређењу са случајевима када су тагови укључени као речи, не чини се да то својство доводи до побољшања када су присутни други, информативнији, атрибути, као што је *ид_корисника*.

Резултат је, очекивано, најнижи у категорији *куповина*, међутим, када се примењује овај модел, сведоци смо комбинованог побољшања прецизности и одзива од скоро 12%, у поређењу са моделом који је други по учинку, док укупни резултат достиже готово 42% при овим поставкама. Модел за категорију *спорт* даје најбољу оцену F1 од свих модела из скупа (92%).

Анализа грешака

Ограничена количина текста у документима који се класификују представља главни извор грешака, и оне се јављају у свим категоријама, без обзира на њихову бројност у корпусу.

3 http://lucene.apache.org/core/

4 http://lml.bas.bg/~nakov/bulstem/index.html

Табела 2. Резултати класификације текста

	Прецизност	Одзив	F-мера
Просечне вредности за основу групу (врећа речи)	0,53	0,61	0,57
+ тагови			
Куповина	0,25	0,35	0,29
Спорт	0,90	0,87	0,89
Уметност	0,61	0,65	0,63
Пословање	0,52	0,68	0,59
Политика	0,50	0,75	0,60
Друштво	0,69	0,81	0,75
Забава	0,81	0,80	0,80
Наука	0,54	0,47	0,50
Стил живота	0,70	0,77	0,73
Технологија	0,80	0,85	0,82
Здравље	0,78	0,84	0,81
Остало	0,98	0,96	0,97
Укупна средња вредност	**0,64**	**0,71**	**0,67**
+ тип_медија			
Куповина	0,23	0,32	0,27
Спорт	0,90	0,87	0,89
Уметност	0,61	0,66	0,64
Пословање	0,51	0,68	0,58
Политика	0,50	0,74	0,60
Друштво	0,69	0,81	0,75
Забава	0,80	0,80	0,80
Наука	0,53	0,48	0,51
Стил живота	0,69	0,77	0,73
Технологија	0,80	0,85	0,82
Наука	0,78	0,83	0,81
Остало	0,98	0,96	0,97
Укупна средња вредност	**0,64**	**0,71**	**0,67**
+ид_корисника +ид_корисника и тагови			
Куповина	0,37	0,45	0,41
Спорт	0,93	0,91	0,92
Уметност	0,72	0,74	0,73
Пословање	0,63	0,75	0,69
Политика	0,60	0,81	0,69
Друштво	0,77	0,87	0,82
Забава	0,85	0,86	0,86
Наука	0,66	0,57	0,61
Стил живота	0,79	0,83	0,81
Технологија	0,85	0,88	0,87
Здравље	0,83	0,87	0,85
Остало	0,98	0,97	0,98
Укупна средња вредност	**0,73**	**0,78**	**0,75**
+ триграми из тагова			
Куповина	0,21	0,47	0,29
Спорт	0,91	0,88	0,90
Уметност	0,60	0,73	0,66
Пословање	0,53	0,71	0,61
Политика	0,49	0,80	0,61
Друштво	0,69	0,85	0,76
Забава	0,83	0,82	0,82
Наука	0,52	0,56	0,54
Стил живота	0,72	0,80	0,76
Технологија	0,81	0,85	0,83
Здравље	0,78	0,86	0,82
Остало	0,98	0,96	0,97
Укупна средња вредност	**0,64**	**0,76**	**0,69**

Такође, примећујемо да је категорије тешко разликовати само по њиховим лексичким својствима. Многе категорије, укључујући и оне највеће (*друштво* и *забава*) имају прилично уопштене лексиконе и мало специфичних термина. То доводи до лошег учинка, ако се за предвиђање користи „врећа речи“. Терминологија је експлицитнија у само две категорије: *спорт* и *технологија*.

Документ из категорије *пословање* у коме се говори о бугарском министру финансија је по нашем систему, сврстан у категорију *политика*. Разлог томе може бити чињеница да се бугарски министар финансија три пута помиње у тексту који се прослеђује систему, а чија је дужина ограничена на 1 000 карактера. Следи још један сличан пример: прича о студенту који је направио филмски „trailer“ за свој омиљени фудбалски клуб је класификована као *спорт*, а у ствари спада у *стил живота*. У другом документу који припада класи *стил живота*, говори се о популарном бугарском фудбалском тиму, чији чланови су се опуштали у једном ресторану у Бургасу после освајања бугарског фудбалског супер купа, а класификован је као *спорт*. Други извор грешака је скуп докумената који садрже рекламе, смештен у класу *куповина*, иако спада у групу *здравље*. По нашем мишљењу,

горенаведене грешке у нашем систему су, у ствари, прихватљиви предлози за категоризацију. Оне су последица преклапања међу темама у различитим категоријама и верујемо да би оптимално придруживање тагова дозвољавало и категорије које се предвиђају и оне које припадају златном стандарду.

Јасну грешку, која се не може објаснити преклапањем термина, представља, на пример, погрешно стављање чланка о ступању Лихтенштајна у шенгенски простор у класу *куповина*. Верујемо да до тога долази због претераног ослањања алгоритма на нека мета-својства као што су *ид_корисника* и други тагови. То се дешава углавном у категорији *куповина* која је мање од других заступљена у корпусу.

Још један извор грешака су документи написани на страним језицима којих у корпусу има мање него других врста докумената.

6.2 Екстракција/додела тема

Резултати експеримената из домена екстракције/доделе тема су представљени у Табелама 3 и 4. Евалуациона метрика обухвата прецизност, одзив и F1 меру који се израчунавају за циљне класе (праве кључне речи и изразе). Кандидат је стварно одговарајући само ако у потпуности одговара одредници пронађеној у скупу који сачињава златни стандард. Када је реч о ненадзираном приступу, саопштавамо само резултате добијене уз оптималне поставке параметра који ограничава број добијених резултата.

Стеминг не утиче значајно на учинак ненадзираних метода (Табела 3), вероватно због бројних ограничења која се односе на скуп кандидата (игнорисање кандидата са ниском фреквенцијом, изузимање дужих кандидата који садрже небитне речи и др.). **Микс** метод је надмоћан у односу на метод **средња вредност** и са стемингом (ред 8, F1=13.79%) и без стеминга (ред 1, F1=13.78%). Међу варјантама **средње вредност**и, најбољи резултат се добија када се одстрани 75% најмање важних токена (ред 7, F1=13.63%).

Модели који користе основна својства (Witten et al. 1999), односно **TFIDF+pos**, по учинку нису надмашили ненадзиране методе, без обзира да ли је употребљаван стеминг или не. Ако се дода својство **len** надзирани приступ има значајно бољи учинак од ненадзиране основе (упоредити Табелу 3, ред 2 и Табелу 4, ред 2). Примећено је да се учинак изузетно побољшава ако се у модел уведе спајање својстава, а најбољи резултати које смо добили укључују комбинације **TFIDF&pos** и **pos&len**.

У нашим експериментима смо дискретизовали својства **TFIDF** и **pos** која имају континуалне вредности, као што је објашњено у одељку у коме се говори о методима. Надзирана дискретизација није по учинку надмашила ненадзирану дискретизацију када се користе само основна својства, док су модели који укључују комбинације својстава дали само мало боље резултате уз ненадзирану дискретизацију (повећање F1 од >1%).

Табела 3. Експерименти спроведени уз примену ненадзираног метода

	Метод	n	F	L	F1	P %	R %
1	микс	1, 2	-	7	11.95	8.28	21.44
2	микс	1, 2	-	5	13.78	10.61	19.66
3	микс	1	0%	5	12.76	09.82	18.21
4	с. вредност	1, 2	0%	5	11.52	8.87	16.44
5	с. вредност	1, 2	25%	7	11.54	8.00	20.71
6	с. вредност	1, 2	50%	4	12.76	9.82	18.21
7	с. вредност	1, 2	75%	5	13.63	10.49	19.46
8	микс	1, 2	-	5	13.79	10.61	19.67

У горњем и доњем делу табеле, који су раздвојени двоструком линијом, дати су резултати са речима, односно, основама

Легенда: L (граница), F (филтер), n (*n*-грами)

Табела 4. Експерименти спроведени уз примену надзираног метода

	Својства	Algo	F1 %	P %	R %
1	TFIDF +pos	MNB	9.32	47.62	5.16
2	TFIDF +pos	MNB	10.31	32.31	6.14
3	+len	MNB	25.60	26.90	24.42
4	+ TFIDF &pos	MNB	29.73	22.81	42.70
5	+ len &pos	MNB	**30.02**	22.15	46.58
6	TFIDF +pos	PER	11.58	36.68	6.88
7	TFIDF +pos	MIRA	10.37	27.35	6.36
8	TFIDF +pos	MNB	10.26	36.65	5.97
9	+len	MNB	**27.35**	19.40	46.33
10	TFIDF +pos	PER	11.58	36.68	6.88
11	TFIDF +pos	MIRA	15.86	29.07	10.90
12	Све спојено	MIRA	20.69	44.92	13.44

У првом делу табеле су дати резултати добијени са речима, а други са основама (та два дела су раздвојена двоструком линијом). „+“ означава додавање својства из претходног реда.

У Табели 4, у редовима 1 и 2 је приказан учинак модела када се за обуку користи скуп који сачињава златни стандард, односно, целокупна збирка чланака (видети одељак о подацима) за потребе евалуације статистике о учесталости. Пошто је учинак са целокупном збирком најбољи, користимо цели корпус да бисмо добили статистику учесталости у следећим експериментима (модели 3-12).

Као и у случају ненадзираног модела, коришћење стемера не утиче много на учинак модела, осим у случају класификатора **MIRA**, где примећујемо побољшање од 5% у F1 мери (Табела 4, редови 7 и 11). Додавање више спојева својстава, скупу већ постојећих својстава, додатно повећава учинак (Табела 4, ред 12).

Како се повећава број својстава, прецизност тих алгоритама мерена у односу на позитивну класу се побољшава и на крају достиже ниво од преко 44%, али тај раст прати и изразит пад у одзиву и F1 мери.

Када се пореди учинак три класификатора, примећујемо да и **PER** и **MIRA** имају бољи учинак од **MNB-a** примењени на основни скуп обележја. И **MIRA** и **ПЕР** имају бољи учинак од ненадзиране основе поређења, што се тиче F1 мере и најбољи резултат прелази 20%. Најбољи резултат (F1=30.02%) смо постигли применом алгоритма **MNB**, уз коришћење својстава **TFIDF**, **pos**, **len** и спојева **TFIDF&pos** и **pos&len**.

Анализа грешака

Анализа грешака се врши у развојном скупу. До многих грешака долази зато што се у документима чешће од имена и презимена јављају само презимена бугарских политичара и зато им наш ненадзирани алгоритам даје предност. Са друге стране, златни стандард препоручује и имена и презимена. На пример, златни стандард предлаже кључне речи *Сергеи Станишев* и *Траичо Траиков* за неке документе. Наш алгоритам даје резултат *Траичо Траиков*, који је стварно позитиван, али даје и *Траичо*, *Станишев* и *Траиков*, што су према нашем систему евалуације, лажно позитивни резултати. Приступ решавању овог проблема подразумева додавање **len** нашем алгоритму за надзирано учење.

Други проблем је склоност аутора да додају имена политичара и политичких организација, чак и када се они не помињу изричито у чланку. На пример, извештај о новом *министру здравља* је тагован његовим личним именом (златни стандард), док ми дајемо резултате *министар здравља, нови, здравље, министар* итд. Svejo.net процењује да тачност може

бити повећана за највише 30%, ако укључимо и анализу врста речи и уклонимо глаголе из нашег избора.

Када је наш модел направљен без стеминга, многи стварно позитивни резултати не бивају узети у обзир током евалуације, зато што кључне речи које као резултат враћа алгоритам, садрже, на пример, чланове, док их кључне речи, из златног стандарда, не садрже. На пример, за кључну реч *medal* из златног стандарда, алгоритам даје резултат *the medal*. Наши алгоритми такође дају и облике множине и једнине кључне речи, које златни стандард не даје. То указује да би укључивање лема (основних облика речи) допринело побољшању учинка наших модела.

7. Закључак

Представили смо преглед система заснованог на техникама обраде природних језика који олакшава дељење садржаја друштвеним сајтовима на мрежи.

Први приложени задатак је класификација са више класа више обележја и за више језика. Примењени модули омогућавају аутоматизацију веома напорног задатка који се обавља ручно. Такође стварају способност да се побољша тачност моделирања, при чему није потребно имати увид у детаље функционисања система. Описали смо примењене алгоритме и својства, оценили утицај својстава која учествују у процесу моделирања и показали да сазнања о корисницима могу у великој мери да побољшају учинак.

Такође смо се позабавили задатком аутоматске екстракције кључних речи и израза и његовом применом на популарни бугарски ресурс на мрежи, Svejo.net. Колико нам је познато, то је прва студија о екстракцији кључних речи и израза у бугарском, језику на коме има мало ресурса и који је недовољно проучаван. Представили смо два једноставна приступа који се не ослањају на скупе алате за лингвистичку анализу. Испитали смо различите потенцијалне стратегије селекције и оценили дејство неколико типова својстава и њихових комбинација на наше моделе. Осим тога, оценили смо колико се добија на тачности увођењем компоненте стемера.

У будућности планирамо да повећамо укључивање језичког знања. Многе студије указују на могућност великог побољшања учинка додавањем тагова врста речи. Један скорашњи рад (Georgiev et al. 2012) указује да додавање морфолошких својстава може да унапреди надзирану класификацију. Анализа грешака, пак, наводи на закључак да ортографска својства (на пример, „реч се састоји искључиво од великих слова“) могу да издвоје имена и називе организација у реченици, пошто се ти изрази јављају уз упадљиву промену малих у велика слова. Још један користан правац развоја може да буде аутоматско увођење комбинација својстава (McCallum 2003).

Литература

Brin, Sergey and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30:107-117, ISSN 0169-7552, DOI: 10.1016/S0169-7552(98)00110-X.

Cohen, William W. and Yoram Singer. 1999. Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems* 17:141-173 ISSN 1046-8188, DOI: 10.1145/306686.306688.

Crammer, Koby, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz and Yoram Singer. 2006. Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research* 7:551-585.

Dumais, Susan, John Platt, David Heckerman and Mehran Sahami. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge*

Management, 148-155, ISBN 1-58113-061-9, Bethesda, Maryland, United States, DOI: 10.1145/288627.288651.

Fayyad, Usama M. and Keki B. Irani. 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, 1022-1029.

Frank, Eibe, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, et al. 1999. Domain-specific keyphrase extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 668-673.

Ganchev, Kuzman and Georgi Georgiev. 2009. Edlin: an easy to read linear learning framework. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 94-98.

Georgiev, Georgi, Kiril Simov, Petya Osenova, Valentin Zhikov and Nakov, Preslav. 2012. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, France*, 492-502.

Gutwin, Carl, Gordon Paynter, Ian Witten, Craig Nevill-Manning and Eibe Frank. 1999. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems* 27:81-104.

HaCohen-Kerner, Yaakov, Zuriel Gross and Masa, Asaf. 2005. Automatic extraction and learning of keyphrases from scientific articles. In *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico*, 657-669, ISBN 3-540-24523-5, ISSN 0302-9743, DOI: 10.1007/978-3-540-30586-6_74.

Hulth, Anette. 2003. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 216-223, DOI: 10.3115/1119355.1119383.

Inkpen, Diana and Alain Desilets. 2004. Extracting Semantically-Coherent Keyphrases from Speech, *Canadian Acoustics* 32:130-131.

Joachims, Thorsten. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning*, 137-142, ISBN:3-540-64417-2.

Jones, Steve. 1998. Link as you type: using key phrases for automated dynamic link generation (Working paper 98/16). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

Lewis, David D. 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, 4-15.

Luo, Xiao and A. Nur Zincir-heywood. 2005. Evaluation of Two Systems on Multi-class Multi-label Document Classification. In *Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems*, 161-169, DOI:10.1007/11425274_17.

Matsuo, Y. and M. Ishizuka. 2004. Keyword Extraction From A Single Document Using Word Co-Occurrence Statistical Information, *International Journal on Artificial Intelligence Tools* 13:157-170

McCallum, Andrew and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In *AAAI Workshop on Learning for Text Categorization. Technical report WS-98-05*.

McCallum, Andrew. 2003. Efficiently Inducing Features of Conditional Random Fields. In *Nineteenth Conference on Uncertainty in Artificial Intelligence*, 403-410, ISBN:0-127-05664-5.

Mikheev, Andrei. 1998. Feature lattices for maximum entropy modeling. In *Proceedings of the 17th International Conference on Computational linguistics, Vol.2, Montreal, Quebec, Can-*

ada, 848-854, DOI: 10.3115/980432.980709.

Nakov, Preslav. 1998. Design and Evaluation of Inflectional Stemmer for Bulgarian. In *Proceedings of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics)*.

Nigam, Kamal, John Lafferty and Andrew Mccallum. 1999. Using Maximum Entropy for Text Classification. In *IJCAI Workshop on Machine Learning for Information Filtering*, 61-67.

Qi, Xiaoguang and Brian D.Davison. 2009. Web page classification: Features and algorithms. *ACM Computing Surveys* 41(2)1-31, ISSN 0360-0300, DOI: 10.1145/1459352.1459357.

Ratnaparkhi, Adwait. 1998. Maximum Entropy Models for Natural Language Ambiguity Resolution, PhD dissertation. University of Pennsylvania Philadelphia, PA, USA, ISBN:0-591-94112-0.

Rosenblatt, Frank. 1958. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain, *Psychological Review* 65:386-408.

Sahami, Mehran. 1996. Learning Limited Dependence Bayesian Classifiers. In *KDD-96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 335-338.

Schapire, Robert E. and Yoram Singer. 2000. BoosTexter: A Boosting-based System for Text Categorization, *Machine Learning* 39:135-168.

Slattery, Sean and Mark Craven. 1998. Combining Statistical and Relational Methods for Learning in Hypertext Domains, In *Proceedings of the 8th International Conference on Inductive Logic Programming*, 38-52.

Turney, Peter D. 2000. Learning Algorithms for Keyphrase Extraction, *Information Retrieval* 2:303-336.

Turney, Peter D. 2002. Mining the web for lexical knowledge to improve key phrase extraction: Learning from labeled and unlabeled data. In *ERB-1096NRC#44947, National Research Council, Institute for Information Technology*.

Turney, Peter D. 2003. Coherent keyphrase extraction via web mining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico*, 434-439.

Wan, Xiaojun, Jianwu Yang and Jianguo Xiao. 2007. Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, June,* 552–559.

Witten, Ian H., Gordon W. Paynter, Eibe Frank, and Carl Gutwin and Craig G. 1999. Nevill-Manning, KEA: practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries, Berkeley, California, United States*, 254-255, ISBN 1-58113-145-3, DOI: 10.1145/313238.313437.

Witten, Ian H.. 2003. Browsing around a digital library. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, Baltimore, Maryland*, 90-99, ISBN 0-89871-538-5.

Yang, Yiming. 1999. An evaluation of statistical approaches to text categorization, *Journal of Information Retrieval* 1:67-88.

Zelaia A. and I. Alegria and O. Arregi and B. Sierra. 2011. A multiclass/multilabel document categorization system: Combining multiple classifiers in a reduced dimension, *Applied Soft Computing* 11: 4981-4990.