Софийски Университет „Св. Климент Охридски”

Факултет по Математика и Информатика

Катедра Софтуерни Технологии

Извличане и анализ на мнения и чувства от текст от онлайн източници

Дисертация
за присъждане на образователна и научна степен Доктор по научна специалност 01.01.12. „Информатика” в професионално направление „Изкуствен Интелект”

Автор: Борис Димитров Крайчев

Научен ръководител: доц. д-р Иван Койчев

14 януари 2014г.

Резюме

Настоящата дисертация разглежда актуалният проблем за откриване и автоматичен анализ на потребителски мнения в интернет. След като прави обзор на постиженията в областта до този момент се разглеждат методи за извличане на мнения от уеб, техният емоционален анализ, както и общ емоционален анализ на части на речта от българския език. Дисертацията описва и основните стъпки в приложението на изследваните методи в практиката чрез построяване на инструмент за анализ на онлайн репутация.

Основният принос на изследването се състои в разработването на алгоритъм за разпознаване на мнения в уеб страници чрез съпоставящи дървета (Wrapper trees), който има линейна изчислителна сложност. По информация на автора известните до момента алгоритми с квадратична изчислителна сложност и оптимизацията, макар и на цената на допълнителни ограничения е доказана теоретично и показана експериментално. Предлага се и метод за изработване на тематично специфични лексикони от емоционално значими думи. Методът използва синонимните връзки в WordNet и анализ на присъствието на термините, кандидати за речника, в тестово множество. Не на последно място приносът на настоящата дисертация се изразява и в изследване на емоционалното значение на прилагателните имена в българския език и разполагането им по своеобразни емоционални оси, построени по групи от ръчно избрани и сравнително често използвани думи. Изложените методи и алгоритми са внедрени в завършена система за анализ на мнения, която има международно признание.

Благодарности

Бих искал да изкажа специални благодарности на моя научен ръководител доц. д-р Иван Койчев първо за факта, че ме прие за свой докторант, за споделянето на ценни съвети относно извличането на информация и анализа на чувства от текст и за цялото отделено време и съвместна работа по настоящата дисертация, съвместните публикации и участия в научни проекти.

Бих искал също да изкажа благодарност на членовете на катедра Софтуерни технологии, които през изминалите четири години ми помогнаха със съвети и напътствия, а именно проф. д-р Силвия Илиева, проф. дтн Христо Кабакчиев, доц. д-р Александър Григоров, проф. д-р Боян Бончев, доц. д-р Петко Русков, доц. д-р Олга Георгиева, доц. д-р Александър Димов, доц. д-р Десислава Петрова-Антонова, доц. д-р Милен Петров, доц. д-р Елисавета Гурова, ас. Ирена Павлова и ас. Георги Къдрев.

Работата по настоящата дисертация се осъществи и благодарение на проекта „Укрепване и устойчиво развитие на научно-изследователския потенциал на Софийски университет в областта на електронното обучение" по договор № BG051PO001-3.3.04/52 (2009-2011) с ръководител доц. Румяна Пейчева-Форсайт.

Съдържание

Резюме 2
Благодарности 3
Списък на таблиците 8
Списък на фигурите 9
1. Увод 11
1.1. Мотивация 11
1.1.1. Мнения, споделени онлайн 11
1.1.2. Задачи на извличането и анализа на чувства от текст 13
1.2. Цели на дисертацията 14
1.3. Структура на дисертацията и използвана терминология 15
2. Обзор на анализа на чувства от текст 18
2.1. Анализ на чувства на ниво дума 18
2.1.1. Ръчно съставени лексикони 18
2.1.2. Лексикони с числова оценка (тегло) на думите. 19
2.1.3. Автоматично генериране на лексикони с WordNet. 20
2.1.4. Автоматично създаване на речници от емоционално заредени думи чрез машини за търсене в интернет 21
2.2. Анализ на положителност на ниво фраза и изречение 21
2.2.1. Анализ чрез лингвистични връзки 21
2.3. Анализ на положителност на ниво документ 22
2.3.1. Опростен подход (Bag-of-words) 23
2.3.2. Анализ на документи по стилометрични данни 23
2.3.3. Анализ на документи според личността на автора 24
2.4. Анализ на потребителски мнения 24
2.5. Анализ на чувства от текстове, засягащи повече от една области 26
2.6. Класификация на текстове чрез алгоритми за машинно самообучение без учител 28

2.7. Класификация на текстове чрез алгоритми за машинно самообучение с учител 32
3. Ефективен алгоритъм за извличане на потребителски мнения от уеб страници 37
3.1. Постановка на задачата 37
3.2. Методология 39
3.2.1. Извличане на данни 39
3.2.2. Съпоставяне с целевото дърво 44
3.2.3. Изчислителна сложност на алгоритъма 48
3.2.4. Ограничения на метода и решение 49
3.3. Експеримент 49
3.3.1. Дизайн 51
3.3.2. Провеждане на експеримента и резултати 52
3.4. Дискусия 56
3.5. Изводи 58
4. Класификация на потребителски мнения чрез семантични лексикони 59
4.1. Генериране на лексикон и анализ на положителност 60
4.1.1. Определяне на начални думи на лексикона и обогатяване чрез WordNet 61
4.1.2. Генериране на лексикон с Label Sequential Rules (LSR) 63
4.1.3. Класификация на мненията 66
4.2. Експеримент и анализ 67
4.2.1. Дизайн 67
4.2.2. Определяне на положителност чрез оценка на текста 68
4.2.3. Визуализация на атрибутите и резултати от оценката на положителност във WEKA 69
4.3. Дискусия 75
4.4. Обобщение 75
5. Автоматична класификация на български прилагателни имена по семантични оси 77

5.1. Предишни разработки по темата 77
5.2. Метод за групиране и оценка на думи по емоционални оси 79
5.3. Изчисление на емоционалната ориентация на думи от българския език по предварително дефинирани оси 82
5.3.1. Избор на машина за търсене 82
5.3.2. Избор на думи, определящи емоционални оси 83
5.3.3. Оценка на положителност 84
5.3.4. Оценка на разположение по емоционални оси 87
5.4. Дискусия 90
5.5. Обобщение 90
6. Внедряване и практическо приложение 92
6.1. Изисквания 92
6.1.1. Изисквания към обработката на естествен език 92
6.1.2. Технически изисквания към реализацията 93
6.1.3. Изисквания към потребителския интерфейс 93
6.1.4. Изисквания към технологичния процес 94
6.2. Реализация на проекта 94
6.3. Реакции и оценки на софтуерния пакет в публичното пространство 99
6.4. Обобщение 101
7. Заключение 102
7.1. Приноси на дисертационния труд 102
7.1.1. Научни приноси 102
7.1.2. Научно-приложен принос 102
7.2. Бъдеща работа 103
8. Публикации и цитирания 104
8.1. Публикации 104
8.2. Доклади 104
8.3. Цитирания 104

8.4. Цитирания на софтуерния проект 105
9. Библиография 106
10. Декларация за оригиналност 114

Списък на таблиците

Таблица 1 Съответствие между използваната терминология на български и английски език. 17
Таблица 2 Характеристики на документа, използвани при класификация на мнения..... 34
Таблица 3 Обзор на методите за анализ на чувства прилагащи машинно самообучение с учител. 36
Таблица 4. Класифициране на 100 ръчно подбрани положителни и отрицателни примери. 54
Таблица 5. Резултати от втората част на експеримента със 100 случайно подбрани сайта. 55
Таблица 6 Присвоени атрибути за оценката на потребителските мнения 66
Таблица 7 Резултати от прилагането на три различни алгоритми за машинно самообучение 74
Таблица 8 Оценка на популярните машини за търсене. 82
Таблица 9 Референтни прилагателни имена, формиращи емоционални оси 84
Таблица 10 Оценка на положителност на случайно избрани думи. 86
Таблица 11 Оценка на емоционално значими думи по семантични оси. 89

Списък на фигурите

Фигура 1 Примерно мнение, споделено в сайта viewlondon.co.uk (1) 12
Фигура 2 Структура на WordNet според (11) 20
Фигура 3 Оценка на положителност на изреченията по метода на Hu и Liu (3) 31
Фигура 4 Обзор на методите за анализ на чувства прилагащи машинно самообучение без учител. 32
Фигура 5 Съкращаване на структурата на уеб страница, представена в DOM (Document Object Model) 41
Фигура 6 Визуализация на входните данни преди съпоставянето на дървовидните структури 43
Фигура 7 Псевдо-код на алгоритъма Simple Tree Matcher (68) 44
Фигура 8 Псевдо-код на алгоритъма Center-Star (68) 45
Фигура 9 Псевдо-код за откриване на идентични под-дървета в програмен код по I. Baxter et al. (59) 45
Фигура 10 Прилагане на хеш-функция към обработеното съдържание на уеб документ 46
Фигура 11 Съпоставяне на целевото дърво с отпечатъка на уеб документ .. 47
Фигура 12 Псевдо-код на алгоритъма за съпоставяне на йерархични структури на документи 48
Фигура 13. Общ изглед на категориите за новини в Open Directory Project . 51
Фигура 14 Обхождане в ширина на уеб страниците от даден домейн или начален уеб адрес при фиксирана дълбочина 52
Фигура 15 Пример на положителен (успешно съпоставен) уеб документ, спрямо целевото съпоставящо дърво 53
Фигура 16 Основни етапи на извличане и филтриране на онлайн мнения ... 57
Фигура 17 Примери за етикетни последователни правила (LSR) 63
Фигура 18 Метод за конструиране на ефективни етикетни последователни правила (LSR) 65
Фигура 19 Визуализация на връзката между негативни термини и думи и признаци на противоречиво мнение 69
Фигура 20 Визуализация на връзката между термини, изразяващи положителна емоция и признаци за противоречиво мнение 70
Фигура 21 Визуализация на връзката между термините с положителни и отрицателни емоции. 71

Фигура 22 Основни модули на класификатор на мнения в реално време..... 95
Фигура 23 Демографски анализ, извършен върху публични изказвания за празника „Св. Валентин” 96
Фигура 24 Извличане на фрази чрез етикетни последователни правила (LSR) с оценка на обем и вложени емоции..... 97
Фигура 25 Количествен анализ на използваните чувства и географско разположение на източниците..... 98
Фигура 26 Автоматично класифицирани мнения..... 99
Фигура 27 Препоръка на Rankur.com в публичното пространство..... 100

1. Увод

Чувствата и емоциите изразяват психичното състояние на хората, в зависимост от влиянието на различни обстоятелства и субекти. Те могат да бъдат разделени на положителни (възторг, радост, щастие, обич, …) и отрицателни (тъга, яд, страх, скръб,…). Могат да бъдат и неутрални – апатия, безразличие. Не бихме могли да си представим човешкия живот без изразяването на чувства и емоции. Те са важна част от общуването, обуславят нашите действия и имат отражение върху психическото и физическото здраве на хората.

Автоматичният анализ на чувства от текст е обширно пространство, съдържащо множество предизвикателства. Въпреки, че компютърната лингвистика и обработката на естествен език (Natural Language Processing – NPL) имат богата история, извличането на мнения и чувства от текст е обект на активни изследвания едва от около десетина години. За този факт има няколко причини. Първо – темата има богат набор от комерсиални приложения, като осигуряване на обратна връзка от клиенти към производители/доставчици на продукти и услуги, а нуждите на бизнеса винаги провокират висока мотивация за изследвания и експерименти в дадена област. Второ, научните и техническите предизвикателства в областта са сериозни и решими само със съвременни методи като разпределени изчисления и обработка на големи масиви от данни. И трето – благодарение на развитието на социалните мрежи и приложенията за споделяне на впечатления от продукти и услуги – за първи път в човешката история имаме в цифров вид голямо количество потребителски мнения, споделени впечатления и изразени чувства и емоции. Именно благодарение на тези бази от споделени мнения стават възможни и съвременните изследвания и експерименти. Можем спокойно да кажем, че революцията на споделянето в интернет промени изцяло маркетинг индустрията, която от еднопосочен канал на електронни и печатни реклами се превръща в многопосочна комуникация между самите потребители.

1.1. Мотивация

1.1.1. Мнения, споделени онлайн

Нека разгледаме следния коментар за заведение за обществено хранене в Лондон от 10.01.2013 г.:

Reviewer London0890

10/01/2013

I had a brilliant night here recently. (2) It was a sat and the place is open until 2am. (3) The DJ played great songs for dancing, and the place had a good atmosphere. (4) Nice customers and welcoming staff.

Фигура 1 Примерно мнение, споделено в сайта viewlondon.co.uk (1)

От текста могат да се направят следните изводи:

Текстът съдържа няколко твърдения, описващи посещение на нощен клуб. Изречение (1) описва преживяването в клуба. Изречение (2) съдържа фактологическа информация за работното време в събота вечер, а (3) съдържа оценка за музикалния оператор на заведението и общата атмосфера. (4) съдържа оценка за посетителите и персонала.

Може да се заключи, че дадено мнение, което означаваме условно с *o*, има поне две компоненти: обект на мнението *g*, който в дадения пример представлява от *night, DJ, songs, atmosphere, customers, staff* и оценка *s*, която в примера по-горе се определя от прилагателните имена *brilliant, great, good, nice, welcoming*.

Други характеристики на показания текст са неговия автор *h = london0890* и общата оценка между 1 и 5 звезди. Източникът на мнение (авторът) е прието да се нарича *opinion source* или *opinion holder* (2).

Описаното преживяване е осъществено на конкретна дата, вероятно съвсем близка до показаната 10.01.2013г, което е датата на публикация на мнението.

С така извършеното наблюдение може да се даде следната **дефиниция за мнение**: Мнение е вектор със следните четири елемента

$$(g, s, h, t) \quad (1)$$

Където g е целевият обект, s е изразеното чувство за обекта, h е източникът на мнението и t е времето, в което е споделено мнението. Целта g на мнението

може да бъде конкретен продукт, име на компания, политик и др. както и характеристики на основният описван обект, които в този случай са персоналът на заведението, посетителите и общата атмосфера.

Целевият обект (entity) на мнението може да се означи с двойката

$$(T, W) \quad (2)$$

Където T е йерархия от под-обекти на дадения обект (части, елементи на продукт, и др.), а W е множеството от техните характеристики като брой, големина, и др. (3).

1.1.2. Задачи на извличането и анализа на чувства от текст

С така дефинираните понятия може да се определят основните задачи на извличането и анализа на мнения. Нека имаме множество от документи D. За това множество се дефинират следните задачи:

- Идентификация на обектите, които са отразени в мненията от D.
- Идентификация на източниците (авторите) на тези мнения.
- Извличане на времето на публикация на документа и стандартизиране в унифицирана часова зона.
- Класификация на положителност на дадено мнение.
- Генериране на числово представяне на мнението според дефиницията от (1).

Класификацията на мнения може да се задълбочи с въвеждането на няколко категории, относно целевите обекти на мненията. Например едно мнение може да се категоризира като *прямо (директно)*, когато то адресира пряко описваният обект. Примерът за мнение по-горе е именно от тази категория. Мнението може да бъде и *косвено* (*индиректно*), когато отношението се предава по непряк начин. Например следният израз „*След посещението на ресторанта (аз) останах с неприятно усещане*” изразява по не пряк път неодобрение за въпросният ресторант. Към класовете с мнения може да се добави и класа на сравнителните становища, когато в даден документ става дума за повече от един обект на оценка. Обикновено тези мнения правят паралел между характеристиките на два подобни обекта.

В настоящата дисертация се разглеждат основно директни мнения като примерът в началото за посещението на нощен клуб. Именно този вид мнения са преобладаващи в мрежата и затова са във фокуса на алгоритмите за автоматичен анализ.

1.2. Цели на дисертацията

Анализът на мнения от онлайн източници и автоматичното разпознаване на изразените чувства е обширна и динамично развиваща се тема. Настоящата дисертация ще се фокусира върху изследване на начините за разпознаване и извличане на публично достъпни мнения от уеб, начините за построяване на речник с емоционално заредени думи и фрази, като и оценката на самите мнения. Поради малката наличност на онлайн мнения на български език, дисертацията изследва мненията на английски език, а емоционално заредените български думи се анализират с помощта на готов индекс на популярна машина за търсене.

Поставят се следните цели:

- Изследване на алгоритмите за разпознаване на структури в HTML и подобряване на тяхната ефективност като точност и изчислителна сложност.
- Изследване на потребителски мнения в дадена област и търсене на ефективен алгоритъм за класифициране на мненията в тази област. Допълнителната под-задача тук е да се провери възможността за по-добри резултати при построяване на тематично специфичен речник на емоционално заредените думи и фрази.
- Построяване на автоматичен класификатор на думите от българския език, чийто резултат е числова оценка на емоционалния им заряд. Емоционалният заряд може да бъде оценяван в няколко аспекта. Базовата оценка е да се определи дали думата сама по себе си е част от положителна или отрицателна емоция, а допълнителната цел е да се групират думите и да се разположат по своеобразни емоционални оси като доверие, любов, страх и др.
- Построяване на софтуерно решение за автоматизиран анализ на публично споделени становища в цифрови и социални медии.

1.3. Структура на дисертацията и използвана терминология

Оставащите глави на дисертацията са структурирани по следния начин:

В глава 2 се прави обзор на изследванията и постиженията до момента, като се започва с най-ранните лингвистични изследвания и се стигне до съвременните методи за оценка на мнения.

В глава 3 се изследват начините за извличане на мнения от уеб документи. Предлага се алгоритъм за съпоставяне на йерархични структури и разпознаване на елементи от уеб страници.

В глава 4 се анализират мненията на потребители за ресторанти и се построява специфичен речник за анализ на положителността на мненията.

В глава 5 се изследва емоционалният заряд на думи в българския език и се построява автоматичен класификатор на думите по няколко емоционални оси.

Глава 6 показва комбинирането на изложените методи и алгоритми в софтуерен проект. Резултатът от проекта е инструмент за автоматичен анализ на онлайн репутация.

Дисертацията приключва с глава 7, която отчита резултатите и указва бъдещата работа по темата.

За да се улесни четенето на текста и сравнението с други материали на английски език по темата за извличане на чувства от текст, следва списък на условното съответствие между използваната българска и английска терминологии.

Термин на английски език	Съответствие на български език
Opinion mining	Извличане на мнения
Sentiment analysis	Анализ на чувства (от свободен текст)
Sentiment-aware words	Думи с емоционален заряд

Термин на английски език	Съответствие на български език
Lexicon	Лексикон, речник
Emotional axes	Емоционални оси
Wrapper tree	Съпоставящо дърво
Label Sequential Rules	Етикетни последователни правила
Pointwise Mutual Information	Взаимна информация
Web crawler	Уеб робот
Sentiment Score	Оценка на емоционален заряд
Feature extraction	Извличане на характеристики (атрибути, свойства)
Review spam detection	Откриване на фалшиви мнения
Review helpfulness	Полезност на мнение (по оценка на други потребители)
Opinion polarity	Полярност, положителност, позитивност на мнение
Sentiment shifters	Части на речта и композиции от части на речта, които уточняват предаваните чувства.
Structural Correspondence Learning	Обучение за структурно съответствие
Singular Value Decomposition	Разлагане на единични стойности
Similarity graph	Мрежи на сходство
Domain-independent words	Думи, чието емоционално значение е независимо от предметната област
Cross-domain Sentiment Classification	Класификация на чувства от текст върху повече от една предметни области
Word Lengthening	Удължаване на думите чрез повторение на знаци и символи: Coooooollllll!!!!
Probabilistic latent-semantic analysis	Вероятностен латентно-семантичен анализ

Термин на английски език	Съответствие на български език
Sentiment classification	Класификация на текстове според изразените чувства
Sentence level subjectivity classification	Класификация на субективността на ниво изречение
Feature extraction	Извличане на характеристики, споменати в мненията
Aspect extraction	Извличане и класификация на гледна точка, обобщено становище

Таблица 1 Съответствие между използваната терминология на български и английски език.

2. Обзор на анализа на чувства от текст

Bing Liu (4) разделя условно подходите за анализ на чувства от текст според анализираните съставни части. Началните методи за анализ са базирани на ниво дума – този анализ има сравнително добра изчислителна ефективност и разчита на откриване на думи, които носят емоционален заряд дори и без да са поставени в конкретен контекст.

Следващата група методи се базират на анализ върху цели фрази и изречения. Тези методи позволяват откриването на семантични конструкции като отрицание, откриване на конкретното значение на думите и използване на лингвистични конструкции за оценка на посланията.

Последният подход включва оценка на цялостен документ (смислово свързан текст) като например споделено мнение за филм, туристическа дестинация, събитие или произволен друг текст. В последните години се развиват статистически подходи за оценка цялостен текст. В следващите параграфи се разглеждат постиженията и историята на изследванията в тази област.

2.1. Анализ на чувства на ниво дума

Преобладаващото множество от думи в естественият език не носи конкретен емоционален заряд, но има думи, които притежават съвсем ясно значение и предават чувства независимо от контекста. Повечето автори работят върху групирането на тези думи в речници, тяхната автоматична класификация и принос за анализ на чувствата във фрази и документи. В следващите под секции се представят основните подходи за съставяне на речници от такива думи.

2.1.1. Ръчно съставени лексикони

През 1984г. Myrick и Erney (5) публикуват речникът на емоционалните коефициенти (Emotional Quotient Inventory – EQI), който се развива с годините. Към сегашния момент съдържа над 3000 думи и фрази, основно прилагателни и наречия, но думите не са подредени по емоционален заряд като позитивни, негативни или каквато и да е друга нотация.Това налага допълнителна работа върху речника преди той да бъде използван за анализа на чувства от текст.

Levin (6) през 1993г. анализира семантиката на глаголите в английския език, оценявайки 113 глагола като имащи отношение към положителност и отрицателност на изразеното мнение. 34 глагола са оценени като неутрални, т.е. не могат да бъдат оценени като позитивни или негативни.

Следва работата на Stone et al. (7) върху думи с емоционален заряд, които разработват речника General Inquirer, съдържащ 13000 думи и над 34000 значения, като думите присъстват в речника толкова пъти, колкото значения притежават. General Inquirer групира думите в различни семантични области (онтологии) - *tag* и *marker*. Най-значимите за нашето изследване думи от *tag* онтологията са етикетирани като позитивни и негативни думи. Например 1635 думи, притежаващи 1915 значения са етикетирани като положителни (позитивни), а 2005 думи с 2293 изброени значения са оценени като отрицателни (негативни). 17 думи са класирани като „противоречиви", т.е. неподлежащи на категорична класификация. Например думата *fun* е класирана като съществително с позитивно значение, но глаголът *fun* (*to make **fun** of*) е негативен.

В *marker* структурата на General Inquirer има 182 синтактични и семантични категории, от които най-важни за анализа на чувства от текст са Емоции (*anger, fury, happiness, distress, ...*) и оценяващи прилагателни като (*good, bad, nice, hard, awesome, easy, ...*). Думите от тези две категории принадлежат и на *tag* онтологията в категориите положителни и отрицателни думи (*Positiv, Negativ).*

Друга подобна разработка е Linguistic Inquiry and Word Count Dictionary на Pennebaker et al. (8), който съдържа над 2200 модели на думи. Моделите са разделени на позитивни и негативни, допълнително и по видове емоции като тъга, щастие и др.

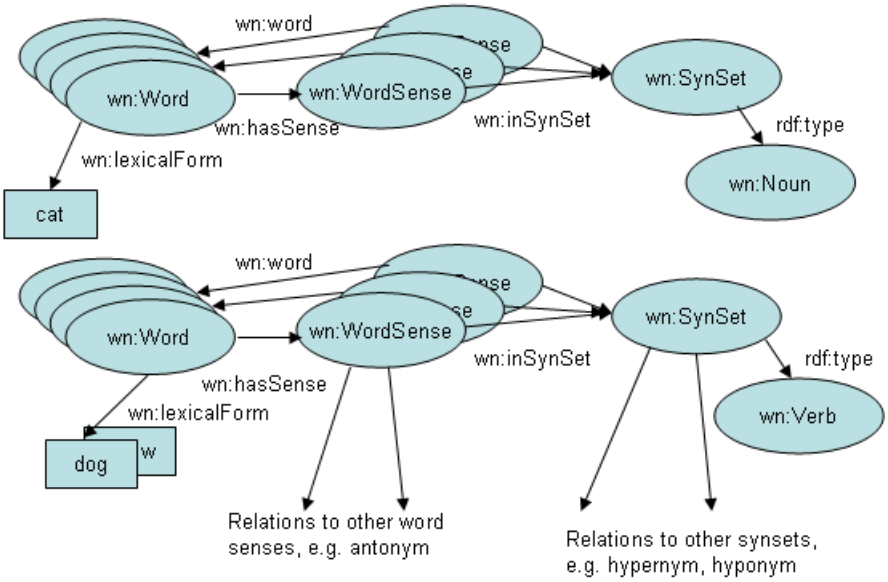
2.1.2. Лексикони с числова оценка (тегло) на думите.

Споменатите по-горе лексикони не притежават числова оценка на силата на емоцията, която изразяват. В тях характеристиките на дадена дума са ограничени до принадлежност на думата към дадена категория. Числовият израз на емоционалният заряд позволява да се сравнят думи с противоположно значение и да се използват методи за автоматична класификация на думи по емоционално значение.

Whissel (9), през 1989г. публикува Whissell's Dictionary of Affect – речник, който е създаден след оценка на корпус от един милион думи. Към момента думите в речника са над 8700 и могат да притежават еднакви корени (lemmata). Думите са оценени по три оси – оценка, сила и образност. Крайните оценки на дадена дума са създадени от 200 оценители от разнообразни демографски групи, като средното аритметично е запазено като окончателна оценка за думата.

2.1.3. Автоматично генериране на лексикони с WordNet.

Базата с думи и релации между тях WordNet (10) е създадена през 1995 година и съдържа над 150 000 думи и семантични връзки. Структурата може да се види на следната фигура:



Фигура 2 Структура на WordNet според (11)

WordNet съдържа части на речта като съществителни и прилагателни имена, глаголи и наречия, групиране в семантични единици – SynSet. Семантичните единици са организирани в мрежа и връзките между тях са реализирани с указатели. Връзките могат да бъдат едно от следните: синоними, хиперними (значение, обобщаващо текущото значение), хипоними (под-значение,

специфициращо смисълът на текущата дума). Значението на думите е класифицирано с категории, като например (*съществителни, чувства*) и др.

Различни автори използват WordNet за да изчислят семантична близост като например Budanitsky et al. през 2001 г.(12).

Друго приложение на WordNet е изчислението на емоционалната ориентация на думите. В тази посока работят авторите Kamps et al. (13), Andreewskaia et al. (14), както и Hassan и Радев (15). Като основна доработка е представена от Esuli et al. (16) с проекта SentiWordNet, в който семантичните единици (SynSet) са оценени с три числови оценки за обективност, положителност и отрицателност.

2.1.4. Автоматично създаване на речници от емоционално заредени думи чрез машини за търсене в интернет

Колкото и голям речник с емоционално заредени думи да се създаде, то той никога не е достатъчен. Turney et al. (17) през 2001г. описват метод за автоматична оценка на положителност, базиран на броя срещания на дадена дума в близост до референтни думи в машините за търсене в интернет като популярната по това време Altavista. Именно техният подход е в основата на изследването в глава 5 Автоматична класификация на български прилагателни имена по семантични оси.

2.2. Анализ на положителност на ниво фраза и изречение

Извличането на чувства от текст на ниво фраза и изречение се осъществява с намиране на лексикални конструкции, изразяващи емоционална оценка. Подходът позволява използване на семантичен анализ и граматична свързаност за да се оцени допълнителният контекст на думите и да се получи точна оценка за обекта на изразените емоции.

2.2.1. Анализ чрез лингвистични връзки

Базата с общи знания OMCS: (Open Mind Common Sense database) съдържа над 400 000 факти от „ общо” значение, организирани като понятия. Например: People can love; people can talk to each other; ...

Liu et al. (18) през 2003г. приемат, че емоционалното значение на думите зависи от поятие, например глагол, съществително или фраза с прилагателно име, участващи в текста. Авторите извеждат модели на изречения, изразяващи емоция и така могат да поставят оценка за положителност спрямо цяло изречение и да идентифицират конкретния обект на изразеното отношение.

Riloff и Wiebe през 2006г. използват статистически подход за да извлекат емоционално значими фрази. Техният подход използва модели за извличане (extraction patterns), които съдържат граматическа информация за конкретни думи и фрази. Такива модели са например: <subject> is happy; <subject> is satisfied; <noun>'s speech; endorsed <direct-object>

2.3. Анализ на положителност на ниво документ

Дефиниция на проблема: При даден документ d, изразяващ мнение за обект O, да се определи общата положителност на изразената оценка s. Като допълнителна задача може да се определи автора a и времето t на изразеното мнение. Резултатът от решението на задачата може да се представи като вектор с четири елемента:

$$d = (O, s, a, t) \quad (3)$$

В дефиницията на задачата се приема, че анализираният документ изразява мнение за един единствен обект. Ако даден документ съдържа мнения за различни обекти, то и поляритета на мненията може да бъде съвършено различен. Практически се губи смисълът да се даде обща оценка за изразените чувства, при условие, че те се отнасят за различни обекти.

Liu (19) приема, че задачата може да се раздели на още два типа. В случаите, когато целта е да се определи конкретна полярност на мнението като положително, отрицателно или неутрално, задачата е *класификационна*. Когато

целта е да се определи числова оценка на изразеното мнение, то задачата е от класа на *регресионния анализ*.

Така дефинираната задача е приложима за големи класове от онлайн текстове като продуктови мнения (product reviews), коментари по публикации, новини, обучителни статии и др. До голяма степен текстовете в Twitter, поради ограничението на дължината на текста, са именно такива документи.

Следва преглед на основните групи от методи за анализ на чувства от документи.

2.3.1. Опростен подход (Bag-of-words)

Този подход се заключава в откриване на даден брой емоционално значими думи, последвано от тяхното класифициране и сумиране, като крайният резултат дава предположението за емоционална ориентация на документа. Методът може да се схематизира по следния начин:

- Откриване на N емоционално значими думи;
- Класификация на откритите думи;
- Ако броят на думите с положителна ориентация надхвърля броя на думите с отрицателна, то документът се счита за позитивно настроен и обратно.

Отчетената точност на експерименти с такъв алгоритъм е до 74% (17), като по-добри резултати са постигани с допълнителна ръчна оценка на думи (20).

2.3.2. Анализ на документи по стилометрични данни

Описаните досега подходи са основно създадени за емоционалния анализ на текстовете. Други области като стилометрията, също се занимават с анализа на текстове. На пръв поглед дължината на думите, изреченията и други статистически данни за стила на автора могат да подскажат неговата психология. Pennebaker et al. (21) представят през 2003г. необяснимо на пръв поглед твърдение, че думите, използвани от хората, носят психологическа информация, която надхвърля рамките на тяхното буквално значение. Използваните думи могат да диагностицират социалния, психологическия и дори физическия статус на автора, дори и те да не предават директна информация за това. Chung et al. (22) през 2007г. показват, че неемоционалните думи също изпълняват психологически

функции. Все пак най-висока точност с този подход се постига при задачата за откриване на автор на текст. Интересно изследване предлагат Brody и Diakopoulos (23) през 2011г., като изследват удължаването на думите чрез повторение на знаци – често срещат прийом за предаване на настроение в микроблогинг платформи като Twitter и други социални медии.

По отношение на анализа на онлайн мнения, стилометрията може да бъде полезна при откриването на спам в сайтовете за продуктови мнения.

2.3.3. Анализ на документи според личността на автора

Mairesse et al. (24) през 2007г. предлагат метод за разпознаване на личността на автора, използвайки лингвистичен анализ на неговите думи в диалог и лично създаден текст. Ако се приеме, че основните личностни черти са: екстровертност, емоционална стабилност, кооперативност, съзнателност и любознателност, то изследването на авторите показва, че личността се изразява и в лингвистичните черти на текста. Авторите например установяват, че екстровертността на характера се изразява в честотата на използване на различните части на речта или по-точно може да бъде установена с формулата:

$$E = (freq(nouns) + freq(adj) + freq(prep) + freq(art) - freq(pronouns) - freq(verbs) - freq(adverbs) - freq(interj) + 100)/2 \quad (4)$$

Според авторите подобен подход може да бъде използван и в анализа на чувства от текст, като положителните чувства биха изисквали използването на различен модел части на речта от текстовете с негативни мнения.

2.4. Анализ на потребителски мнения

Една допълнителна причина за големият интерес към анализа на чувства от текст са споделените продуктови мнения в мрежата. Това са сайтове за филми като IMDB и Rotten Tomatoes, Epinions, електронни магазини като Amazon, туристически сайтове като Trip Advisor и сайтове, посветени на заведенията за

обществено хранене като Top Table (25). Анализът на споделените мнения може да се използва за предвиждане на приходите от билети за филми, продажбите на продукти и за подобрение на самите продукти чрез осъществяване и обобщаване на обратната връзка от потребителите.

Могат да бъдат разграничени две основни посоки на научните изследвания: Първата се отнася до изграждането на лексикони от емоционално заредени думи, а втората група се състои от работата на цялостни системи за анализ на настроенията за документи и текстове.

През деветдесетте години на миналия век Hatzivassiloglou и McKeown(26) анализират корпус от текстове като статистически определят емоционалната ориентация на думите чрез близостта на срещанията им с референтни думи. Подходът им различава съюзите между самите думи.

Отзивите за филми са били предмет на изследване за Pang, Lee и Vathyanathan (27) и Yang Liu (28). Първата система постига точност от около 83% и показва, че техниките за машинно обучение се представят по-добре, отколкото подходът с обикновена сума на теглата на думите, имащи отношение към изразяване на чувства. Втората система използва линейна регресия за предвиждане на полярността на дадено мнение за филм, като интересно въведение в тази област е представено от C. Bishop (29) и съчетава информацията от приходите в бокс офиса от предишните дни с оценката на чувствата в мненията на публиката. С тази информация биват предвидени приходите от даден филм за текущия ден. Най-добрите резултати за този алгоритъм са точност от 88 %.

Някои от авторите като Pang et al. (27) се опитват да разделят текста на твърдения с фактологическа и емоционална ориентираност, докато други като Godbole et al. (30) считат, че всички споменати факти и становища допринасят за нагласите и полярността на един текст.

Друг подход за анализ на отзиви за продукти е анализът на настроения, базиран на характеристиките за даден продукт, предложени от B. Liu, Xy и Cheng (31), който извлича настроения според коментарите за атрибутите на даден обект. Използваната техника е наречена етикетни последователни правила (Label Sequential Rules - LSR). Общ преглед на методите за анализ на настроения е направена от Pang и Lee през 2008 година (32).

Сравнително скорошен подход е предложен от Hassan и Радев (15) през 2010 г., който определя емоционалната полярност на думи чрез прилагане вериги на Марков.

В настоящата дисертация се използва подход с числова оценка на емоционалното значение на думите, подобрен чрез етикетни последователни правила, за да се постигне по-задълбочена оценка на характеристиките на обекта. За крайното класифициране се използват познати алгоритми за машинно самообучение с учител.

2.5. Анализ на чувства от текстове, засягащи повече от една области

Анализът на чувства и класификацията на положителността на текстовете са силно зависими от терминологията и засегнатата област на изследваните текстове. Езиковите конструкции и дори смисълът на отделните думи носят съвсем различен емоционален заряд в зависимост от контекста и терминологията на областта (4). Един класификатор, обучен в дадена област, като например мнения за филми, често показва силно влошено представяне при оценка на становища от друга област като например потребителските мнения за продукти от електронни магазини.

Трансферът на класификатори към нови области е изследван от Aue et al. (33). Според авторите, за съжаление, съществуват много области, в които липсват етикетирани данни в големи количества. Такива данни са необходими за напълно контролирани подходи за обучение. В същото време, класификаторите на настроения трябва да бъдат адаптивни към нови области, за да бъдат полезни в практиката. Те се опитват да се справят с тези трудности и ограничения, като проучват четири различни подходи за персонализиране на система за класификация на настроения към нов целеви домейн, при липсата на големи обеми от етикетирани данни. Те основават експериментите си на данни от четири различни предметни области. Подходите, които използват са следните:

- Обучение върху смесица от етикетирани текстове от други области, в които има на разположение данни за обучение и тестване върху целевата област;
- Обучение на класификатор като в предишната точка, но ограничавайки разпознаваните характеристики до общи с целевия модел;
- Използване на различни класификатори от области с етикетирани данни и провеждане на тестове с целевата област;
- Комбиниране на малки количества етикетирани данни с големи количества не етикетирани данни от целевата област.

Резултатите показват, че последният подход е най-ефективен, и показва още веднъж, че класификацията на чувства от текст е силно зависима от контекстната област.

През 2007г. Blitzer et al. (34) предлагат метод, наречен „обучение за структурно съответствие" (Structural Correspondance Learning или SCL) за адаптиране на класификатори към нови предметни области. Вземайки етикетирани (ръчно класифицирани) текстове от началната област и неетикетирани текстове и от началната, и от целевата област техният класификатор избира множество от характеристики, които са общи за двата терминологични домейна и също са добри индикатори за текстовете от началната област. Изборът се води от формулата за взаимна информация (Pointwise Mutual Information – PMI). Комбинирайки избраните характеристики с останалите в корелационна матрица, авторите откриват нови характеристики, които са кандидати за ефективна класификация.

Подобен метод е използван и от Pan et al. (35) през 2010г. Авторите предлагат метод, наречен *спектърно съпоставяне на характеристики* (Spectral Feature Alignment - SFA). Те съпоставят специфичните за дадена област характеристики в клъстери като за целта използват думите, които не зависят от областта за установяване на връзката.

През 2011г. He et al. (36) изследват модели за установяване на свързани теми. Намерените теми от началната и целевата области се добавят като допълнителни класификационни характеристики. През същата година Gao et al. (37) прилагат вероятностно латентно-семантичен анализ, като също търсят съпоставяне на характеристики между началната и целевата области.

Съпоставянето се основава на съответствие на термини и изследване на тяхното взаимно присъствие.

Xia и Zong (38) през 2011г. откриват, че характеристиките, използвани за изразяване на становища в различни предметни области, се отличават по използваните части на речта, като други модели на изразяване остават тематично независими. Въз основа на тази констатация, те предлагат комбиниран модел, базиран на анализатор на части на речта, който подобрява представянето на класификатора чрез анализ на използваните изразни модели.

2.6. Класификация на текстове чрез алгоритми за машинно самообучение без учител

Едно от първите приложения на машинно самообучение без учител в анализа на чувства е от Turney и Littman през 2002г. (17). Според авторите семантичната ориентация варира както в две посоки (положителна или отрицателна), така и в своята степен (лека до силна). Тяхната автоматизирана система за измерване на семантична ориентация има приложение в класификацията и филтриране на текстове, проследяване на становища в онлайн дискусии, анализ на отговорите от проучване, както и автоматизирани системи за чат (chatbots). Те въвеждат метод за извод на семантичната ориентацията на една дума от своята статистическа асоциация с набор от положителни и отрицателни думи (парадигма). Две инстанции на този подход, се оценяват въз основа на две различни статистически мерки за асоциация: Формула за взаимна информация (PMI) и латентен семантичен анализ (LSA). Провеждат експеримент с 3596 части на речта (включително прилагателни, наречия, съществителни и глаголи), които са били ръчно етикетирани като положителни (1614 думи) и отрицателни (1982 думи). Методът постига точност от 82.8% на теста за цялото множество, но точността се повишава до над 95%, когато от класификацията са изключени думите с лека степен на емоционален заряд т.е. трудните за класификация думи.

През 2004г. Kim и Hovy (2) представят система, която по зададена тема, автоматично намира изразени мнения за тази тема и определя настроението на всяко становище. Тяхната система съдържа модул за определяне на положителност на дадена дума и втори, който определя положителност на ниво изречение. Системата им постига точност от около 70%, като за експеримента са

оценени 19748 английски прилагателни имена и 8011 глагола. Авторите използват подсилващи думи като *много*, *почти*, *изключително*, както и отрицанието в текста за да определят автоматично емоционалната ориентация на думите.

През същата година подобен подход е използван и от Hu и Liu (3), постигайки средна точност от 84%. Авторите на изследването се стремят да обобщят всички мнения на клиентите на даден продукт. Задачата на това обобщение е различна от традиционния преразказ/обобщаване на текст, защото включва оценки по различните характеристики на продуктите. Прави се изследване на положителност и отрицателност на изразените становища. Самият преразказ на мнения не се извършва чрез избор на конкретни представителни изречения, както това се прави в класическите алгоритми за обобщаване на текст, а алгоритмът работи в 3 стъпки: (1) Откриване на характеристики на продуктите, които са били коментирани от клиенти; (2) идентифициране на изречения, изразяващи мнение и за всяко твърдение в намерените изречения се определя дали е положително или отрицателно; (3) обобщаване на резултатите. Приложеният алгоритъм е представен на следващата фигура:

```
1. Procedure SentenceOrietation()
2. begin
3. for each opinion sentence si
4. begin
5. orientation = 0;
6. for each opinion word op in si
7. orientation += wordOrientation(op, si);
8. /*Positive = 1, Negative = -1, Neutral = 0*/
9. if (orientation > 0) si's orientation = Positive;
10. else if (orientation < 0) si's orientation = Negative;
11. else {
12. for each feature f in si
13. orientation +=
14. wordOrientation(f's effective opinion, si);
15. if (orientation > 0)
16. si's orientation = Positive;
17. else if (orientation < 0)
18. si's orientation = Negative;
19. else si's orientation = si-1's orientation;
20. }
21. endfor;
22. end
 1. Procedure wordOrientation(word, sentence)
2. begin
3. orientation = orientation of word in seed_list;
4. If (there is NEGATION_WORD appears closely
```

```
around word in sentence)
5. orientation = Opposite(orientation);
6. end
```

Фигура 3 Оценка на положителност на изреченията по метода на Hu и Liu (3)

През 2008г. Ding et al. (39) се фокусират върху изграждането на цялостен подход за лексикон, ориентиран към решаване на проблема чрез използване на външни признаци и езикови конвенции на изрази на естествен език. Този подход позволява на системата да се справи с думи, които са зависими от контекст и които причиняват сериозни затруднения за съществуващите до момента алгоритми. Той също се занимава с много специални думи, фрази и езикови конструкции, които имат въздействие върху мнението въз основа на дадени езикови модели. Методът им притежава функция за резолюция на струпване на множество противоречащи си думи в едно изречение. Средната точност постигната от авторите е 90%.

През 2011г. Taboada et al. (40) предлагат системата Semantic Orientation Calculator (SO-CAL), която е базирана на речник от думи, етикетирани с тяхната семантична ориентация (положителна или отрицателна) и включват разпознаване на подсилващи думи и отрицание. Предимството на метода е, че постига постоянна точност, независимо от областта на приложение. Авторите постигат точност от 74% до 90%.

Обобщение на разработките и техните резултати е представено в следвана таблица:

Авторски екип	Метод	Резултати (Accuracy)
Turney и Littman (17) – 2002г.	Measuring praise and criticism: Inference of semantic orientation from association	82.8%
Kim и Hovy (2) - 2004г.	Determining the sentiment of opinions	70%
Hu и Liu (3) – 2004г.	Mining and summarizing customer	84%

Авторски екип	Метод	Резултати (Accuracy)
	reviews	
Ding et al. (39) – 2008г.	A holistic lexicon-based approach to opinion mining	90%
Taboada et al. (40) – 2011г.	Lexicon-based methods for sentiment analysis	74% - 90%

Фигура 4 Обзор на методите за анализ на чувства прилагащи машинно самообучение без учител.

В глава 5 от настоящата дисертация се разглежда подход за класификация на думите от българския език, подобен на горе изброените. В него се прилага не само класификация на положителни и отрицателни думи, но и се прави оценка за семантична близост на думата по предварително определени емоционални оси.

2.7. Класификация на текстове чрез алгоритми за машинно самообучение с учител

Задачата за анализ на чувства от документ принадлежи на класа от задачи за класификация на текст. В повечето случаи авторите приемат, че дадено мнение може за изразява само положително или отрицателно мнение, то по-коректно е да се въведе и класът на неутралните мнения, когато авторът не изказва конкретно предпочитание. Пример за такива текстове са продуктовите мнения с 3 от 5 звезди. В тези случаи не може да се приеме, че авторът дава превес на предимствата или недостатъците на продукта.

Класификацията на текстове използва разпространени алгоритми като naïve Bayes и Support Vector Machines (SVM) използвани например от Joachims, 1999г. (41) и Shawe-Taylor et al. (42) през 2000г. През 2002г. Pang et al. (27) прилагат за първи път този подход за класификация на полярност на мнения.

Основните подходи за класификация на мнения са описани в следната таблица:

Термини и тяхната честота в документа	Най-често използваната характеристика на даден документ при класификация на текстове е честотата на използваните термини (униграми) и преценката спрямо даден корпус от подобни тестове. Популярен метод за оценка на използваните термини е TF-IDF (Term Frequency – Inverse Document Frequency), където имаме: $tf(t,d) = \frac{1}{2} + \frac{\frac{1}{2}freq(t,d)}{max\{freq(w,d):w \in d\}}$ $idf(t,D) = log\frac{\|D\|}{\|\{d \in D: t \in d\}\|}$ $TFIDF(t,d,D) = tf(t,d) \times idf(d,D)$ За даден термин *t*, документ *d* и корпус от документи *D*.
Анализ на части на речта	Анализът на частите на речта помага на извличането на чувства като отделя класове от думи, имащи пряко отношение към предадената положителност в мнението. Например това са прилагателните имена и глаголите. Друго приложение на анализа на части на речта са етикетните последователни правила (LSR), използвани в настоящата дисертация. Пример за такава работа е (43).
Речници от емоционално заредени думи	Емоционално заредените думи са думи, които сами по себе си изразяват положителност или отрицателност. Такива са например прилагателните имена *удобен*, *чудесен*, *страхотен*. Съответно примери за отрицателни думи в българския език са *отвратителен*, *неудобен* и др. Глаголи също могат да изразяват положителност като *харесвам*, *мразя* и др.
Конструкции на твърдения	Лексически конструкции, които изразяват или подсказват отношение към даден обект също са използвани за анализ на чувства. Идеята за тези конструкции е разработена още през миналия век през 1974г. от Montague (44) и през 1981г. от Dowty et al. (45). Правилата биват представени в Бекус-Наурова форма.
Уточняващи	Това са части на речта и композиции от части на речта,

конструкции	които уточняват предаваните чувства, модифицирайки значението на дадена емоционално заредена дума. Например „*никак не харесвам този хотел*”. Подобни конструкции трябва да се разглеждат с повишено внимание, защото привидното отрицание не винаги води до промяна на полярността на мнението: „*не само харесвам този хотел, но …*” Други конструкции, уточняващи предаваните чувства са думите *много*, *малко*, *изобщо*, *никак*, *винаги* и др.

Таблица 2 Характеристики на документа, използвани при класификация на мнения

Методите за класификация с машинно самообучение с учител могат да бъдат обобщени в следната таблица:

Авторски екип	Метод	Резултати (Accuracy)
Gamon M.(46) – 2004г.	Sentiment Classification on Customer Feedback Data (Microsoft research)	72-77%
Pang & Lee (47) – 2004г.	Sentiment analysis using subjectivity summarization based on minimal cuts	83% - 86%
Mullen et al. (48) – 2004г.	Sentiment analysis with Hybrid Support Vector Machines (PMI/Osgood and Lemmas)	86% - 89%
Kennedy et al. (49) – 2006г.	Sentiment Classification of Movie Reviews Using Contextual Valence Shifters	85% - 86%
Ng et al. (50)– 2006г.	Linguistic knowledge sources in the automatic identification and classification of reviews	86% - 90%
Abbasi et al.(51) – 2008г.	Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums.	87% - 90%
Li et al. (52) – 2009г.	A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge	75% - 79%

Авторски екип	Метод	Резултати (Accuracy)
Martineau et al.(53) - 2009	Delta TFIDF: An improved feature space for sentiment analysis	87% - 92%
Xia et al. (54) – 2010г.	Use of Word Relation Features for Sentiment Classification	83% - 85%
Bespalov et al. (55) – 2011г.	Supervised Latent N-gram Analysis (SLNA)	84% - 92%
Kamal et al.(56) – 2012 г.	Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources	85% - 89%

Таблица 3 Обзор на методите за анализ на чувства прилагащи машинно самообучение с учител.

В настоящата дисертация класификация на мнения чрез машинно самообучение с учител се разглеждат в глава 4, като постигнатата максимална точност е от 88.5%.

3. Ефективен алгоритъм за извличане на потребителски мнения от уеб страници

Интернет осигурява непрекъснати източници на информация с близка семантична структура като новинарски емисии, публикации в блогове, форуми, продуктови потребителски мнения и коментари на потребители на различни теми. Тези източници са от съществено значение за целта на извличането на мнения от мрежата. В тази глава се предлага ефективен алгоритъм за извличане на структурирана информация от уеб страници. Алгоритмът се основава на комбинация от анализ на структурирани данни и обработка на естествен език на съдържанието на текста. Той съпоставя HTML страници, съдържащи новини, ревюта или потребителски коментари в потребителски проектирани структури като RSS поток. Тази информация обикновено включва текстове с мнения и фактическата информация, като дата на публикуване, цената на продукт, името на автора и неговата оценка. Потоците от подобни данни се генерират в реално време и времето за тяхната обработка е силно ограничено. Изчислителната сложност на обработките трябва да бъде линейна или близо до линейна. Изчислителната сложност на предложения алгоритъм е именно линейна, с което може да се твърди, че се постига необходимата скорост на обработка на информацията. За сравнение, подобни по-рано публикувани подходи имат сложност не по-малък от $O(n^2)$. Следва описание на алгоритъма и метода за неговата реализация. Проведените експерименти с данни от реалния свят постигат точност на съпоставяне от 84 % до 92%, което е сравнимо с последните най-добри резултати в тази област.

3.1. Постановка на задачата

Уеб е най-големия източник на обществено достъпни лични мнения, коментари и отзиви. Броят на нови страници публикувани всеки ден в интернет нараства експоненциално, както и количеството на наличната информация. С разрастването на Web 2.0 и мобилните устройства интернет потребителите все по-лесно споделят своето мнение за продукти, коментират новини, статии, книги и всякакъв вид продукти. Става практически невъзможно своевременно да се анализира и обработи ръчно важната информация като например наблюдението за нови тенденции. Целта на това изследване е да предостави решение за

автоматизирано извличане на метаданни от уеб източници като автор, дата на публикуване, основното съдържание, поток с коментари и други. Допълнително изискване, наложено от количеството налична информация, е използваният алгоритъм да бъде с минимална изчислителна сложност.

Програмите, които изпълняват дейността по извличане на информация от нестандартни структури като HTML документи се наричат „съпоставящи дървета" или Wrapper Trees (57). Всяка уеб страница може да бъде представена като йерархична структура от елементи, които често биват повече от 5000 на документ. Целта е да се опрости и уеднакви тази структура с много по-малка, наподобяваща RSS поток и съдържаща желаните метаданни. Такава структура съдържа до 5 елемента на извлечено потребителско мнение или средно от 20 до 100 елемента. Основният подход е да се изключат неподходящи елементи чрез аналитична оценка на текстово съдържание. Генераторът на съпоставящи дървета е снабден с инструменти за обработка на естествен език като анализатор на части на речта (POS), речник на емоционално заредени думи, синонимен речник.

Доколкото е известно на автора, съществуващите алгоритми за съпоставяне на дърветата, имат изчислителна сложност от поне $O(n^2)$ или по-висока: Zhai и Liu (57), Ferrara и Baumgartner (58). Проблемите на съпоставяне на дървовидни структури са изследвани в теорията на компилаторите, където задачата е да се открие и оптимизира дублиран програмен код, като за целта е нужен алгоритъм с линейна изчислителна сложност. Тази област е изследвана от Baxter et Al през 1998г (59).

Настоящата дисертация прилага подобен подход в построяването на съпоставящо дърво и определя хеш-функция, която сравнява линейно елементите от уеб страницата, които са кандидати за корен на дървото със съпоставящото дърво.

Проблемът с построяване на съпоставящо дърво е проучен в системи като Stalker (60) от Musle et al. През 1999г., които предлагат символен метод за извличане на информация. За съжаление, той строго зависи от синтаксиса на конкретната уеб страница и се нуждае от актуализация, след всяка промяна на HTML дизайна. Yang (61) предлага през 1991г алгоритъм за измерване на разстоянието между две дървета. Той има изчислителна сложност от порядъка на произведението от размерите на двете дървета. Методът е зависим от топологията на сравняваните дървовидни структури. Zhai et al. (62) предлагат алгоритъм за

извличане на уеб данни, основани на частично съпоставяне. Тяхното решение има квадратична сложност и може да се подобри при определени ограничения. Freitag (63) и по-късно други автори като (64) работят по алгоритми, базирани на синтаксис, които също са обвързани с точната структура на уеб страниците. Последните проучвания изследват автоматична адаптация на съпоставящото дърво (58) и използване на логически мрежи на Марков(65).

Преглед на последните приложения на техники за съпоставяне на дървовидни структури може да се намери в (66) и (67).

3.2. Методология

Основната идея за построяване на съпоставящото дърво може да се обясни по следния начин: Преди всичко се зарежда съдържанието на уеб страницата и се подрежда в дървовидна йерархична структура според съставните ѝ HTML елементи. Структурата на една „ стандартна” статия от електронна медия съдържа над 5000 елемента. На втори етап се извършва разпознаване на типовете данни, които се съдържат във всеки възел на получената структура, започвайки от листата. Типовете данни могат да бъдат: лични имена, дата и час, текстово съдържание, както и емоционална оценка на текстовото съдържание – проверява се наличието на емоционално заредени думи. В последния трети етап се прилага алгоритъмът за съпоставяне на дървета, който сравнява възлите на дървото на уеб страницата с целевата структура чрез хеш-функция.

Ограниченията на алгоритъма са свързани с валидността на получения HTML код и структурата на получените уеб страници. Този проблем се обсъжда по-подробно в следващите точки.

3.2.1. Извличане на данни

Анализираните уеб страници се получават с уеб робот (web crawler), след това се анализират с HTML parser, който създава йерархичната структура, представяща съответния документ (Document Object Model или DOM). В третия етап се извършва присвояване на типовете на съдържанието на всеки възел. Започвайки от листата се присвоява един от следните типове:

- *Твърдение за дата на публикация*

Кратък текст, съдържащ дата и/или час, фраза за относително време, като „2 days ago” например: *Добавено от { Автор} на 20/11/2011*. Във фигурите и диаграмите този тип е означен с O. Допълнителен разпознавателен признак за подобни елементи са атрибутите на HTML таговете, като class, id, title и атрибути със специфични имена, които съдържат ключови думи като time, published и др.

- *Вероятен Коментар/Review*

Възел, който съдържа текст с емоционално заредени думи. Във фигурите и диаграмите този тип е означен с C.

- *Вероятен основен текст*

Най-дългият текст на страницата, не е връзка и съдържа морфологично пълни изречения. Във фигурите и диаграмите този тип е означен с T. Тъй като самият текст има форматираща структура, от оценяващия алгоритъм се изисква да слива съседните възли с такъв тип.

- *Име*

Засечен от регулярен израз и списък на известните лични имена. Във фигурите и диаграмите този тип е означен с N. HTML възлите, съдържащи подобни текстове, често съдържат ключови думи в атрибутите и техните имена. Такива думи са author, publisher, reviewer и др.

- *Страничен тип*

Възли, за които не е определен конкретен тип данни. Във фигурите и диаграмите този тип е означен с E.

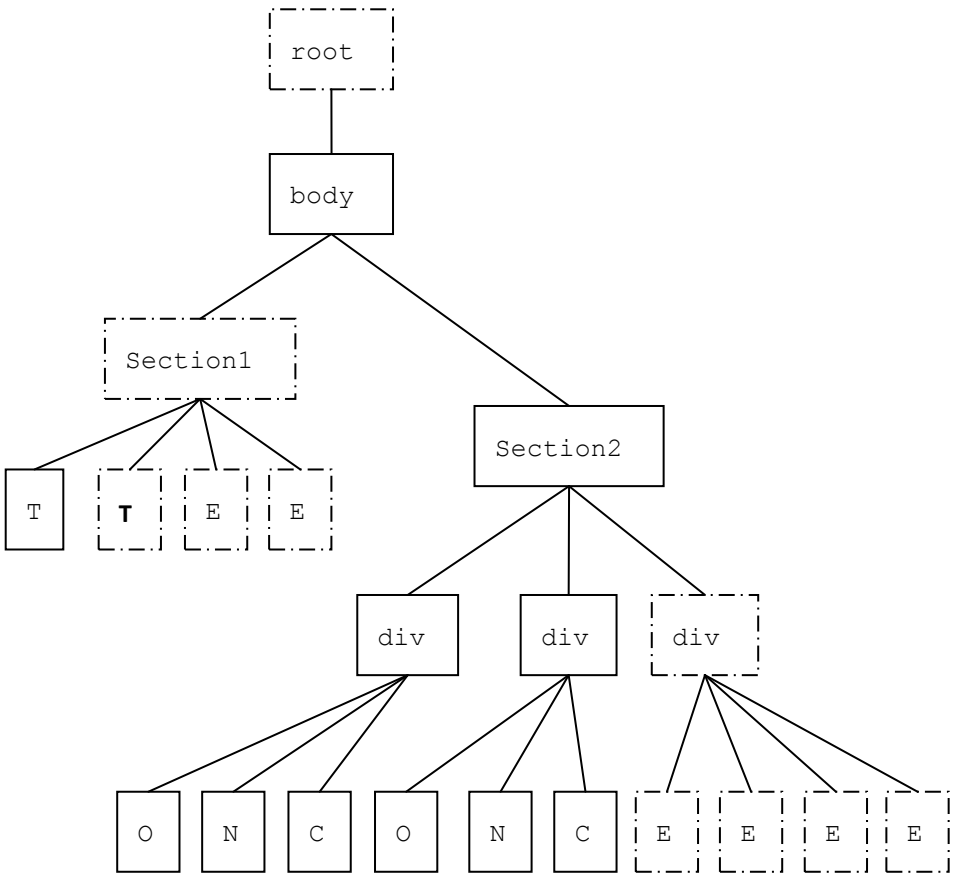
Класификацията на текстовото съдържание е класическа задача за текстова класификация и включва анализ на частите на речта, анализ на емоционалния заряд на използваните думи и фрази, извличане на текст от DOM обект.

След като на всички листа е присвоен тип, се изтриват следните елементи:

- всички листа от тип *Страничен*,
- всички възли, които не се водят до някое от оставащите листа, с присвоен тип.

- всички възли, които имат единствен наследник.

Процесът на съкращаване на структурата на документа е изобразен на следната фигура, която използва силно опростена структура с цел да не затруднява визуализацията. Възлите, подлежащи на изтриване са означени с прекъсната линия:



Фигура 5 Съкращаване на структурата на уеб страница, представена в DOM (Document Object Model)

За целите на съпоставянето с целевото дърво, на следваща стъпка, за всеки възел се задават следните стойности:

- Сортиран списък на типовете листа в под-дървото, като последователните, идентични под-дървета се обобщават с (...)+. За илюстрация, възелът *Section2* от горната фигура би бил записан като *(CNO)+*

Типовете листа в под-дървото са най-важната характеристика за семантичната стойност на описваното под-дърво. Най-лекият пример за съпоставяща хеш-функция е описваната текстова стойност.

- Брой на възли-наследници (включително листа)

Информацията за възлите-наследници е определяща за структурата на дървото. В бъдещата хеш-функция може тази стойност да се редуцира само до нейният знак: положителен, отрицателен или нула.

- Брой на листа в под-дървото

Подобно на предходната стойност, броят листа, техният знак или друга производна стойност за указатели за структурата и съдържанието на дървото.

- Общ брой на думите във възлите-наследници
- Числова оценка (score) на емоционално заредените думи.

Междинните резултати от текстовата класификация на съдържанието на възлите също могат да се използват като входни данни за съпоставящата функция. С тяхна помощ могат да бъдат изключени грешно идентифицирани текстове и могат да се използват за филтър, който да повиши производителността на алгоритъма. Стойността на числовата оценка на емоционално заредените думи се изчислява с техника bag-of-words, използвайки лексикон на емоционално заредени думи с тегла, без оценка на подредбата и фразите (30). Стойността се използва, за да открие потребителски мнения и коментари, които съдържат именно емоционална оценка на останалото съдържание.

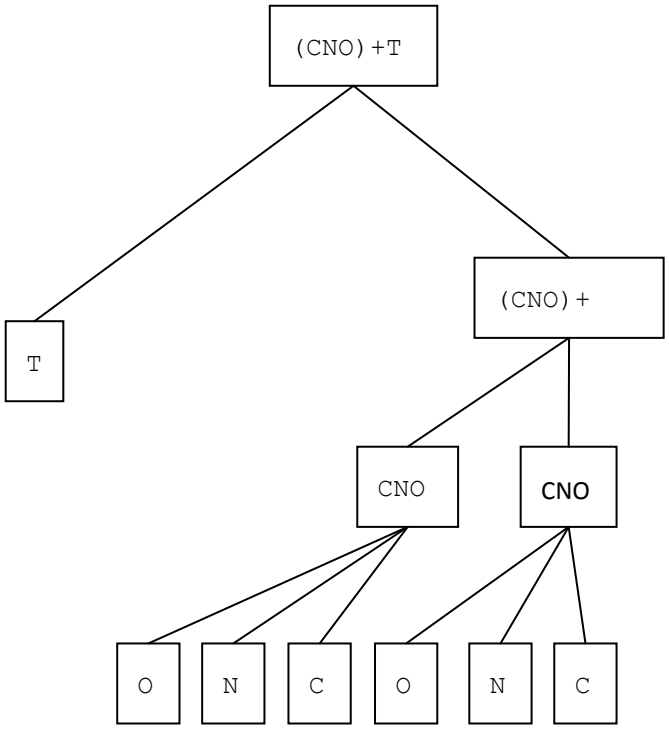
- Оценка за фактологически-ориентирани думи

Стойността е разликата между общия брой думи и броя на емоционално заредените думи.

- Брой на подобни съседни възли

  Броят на съседните възли с идентично съдържание като листа указва повторяемост на съдържанието. Подобна повторяемост подсказва наличие на потребителски мнения и коментари. Повторяемо съдържание, съдържащо информация за време и лични имена е силен признак за поток от потребителско съдържание.

Накратко, за всеки възел се присвоява информация за под-дървото, което представлява. След това стойностите се използват като вход за съпоставяне на йерархичните структури чрез хеш-функция. Резултатите от прилагането на оценката за текущия пример могат да се видят на следващата графика. Числовите стойности са изключени от представянето за опростяване на визуализацията:



Фигура 6 Визуализация на входните данни преди съпоставянето на дървовидните структури

Техника, която може да се приложи, за да се постигнат по-чисти резултати е да се изтегля съдържанието на анализирания документ през определен интервал от време и да се изключат вариращите блокове, които съдържат реклами или нерелевантен текст като например най-нови публикации, засягащи други продукти или оценявани обекти..

3.2.2. Съпоставяне с целевото дърво

Основната цел е да се извлече автоматично Rich Site Summary (RSS) поток от онлайн становища и коментари от уеб страница. Желаният резултат е йерархична дървовидна структура, която трябва да бъде съпоставена с елементите на анализираната уеб страница по метода, описан в предходната секция.

Конвенционалните алгоритми като Simple Tree Matcher и Adapted Center-Star (57) имат квадратична или по-висока изчислителна сложност. За справка следва псевдо-код на въпросните алгоритми:

```
Algorithm: STM(A, B)
1. if the roots of the two trees A and B contain distinct symbols then
2. return (0)
3. else k ←the number of first-level sub-trees of A;
4. n ← the number of first-level sub-trees of B;
5. Initialization: m[i, 0] ← 0 for i = 0, …, k;
m[0, j] ← 0 for j = 0, …, n;
6. for i = 1 to k do
7. for j = 1 to n do
8. m[i, j] ← max(m[i, j-1], m[i-1, j], m[i-1, j-1]+W[i, j]),
where W[i, j] ← STM(A i , B j )
9. end-for
10. end-for
11. return (m[k, n]+1)
12. end-if
```

Фигура 7 Псевдо-код на алгоритъма Simple Tree Matcher (68)

Алгоритъмът Center-Star (69) е използван за съпоставяне на текстови низове (strings) и откриване на повтарящи се последователности от знаци. Той може да бъде приложен и за дървовидни структури, както това е направено в (68):

```
CenterStar(S)
1. choose the center star s c using Equation (3);
2. initialize the multiple sequence alignment M that contains only s
c;
4. for each s in S-{s c } do
5. let c* be the aligned version of s c in M;
6. let s□ and c*□ be the optimally aligned strings of s and c*;
```

```
7. add aligned strings s□ and c*□ into the multiple alignment M;
8. add spaces to each string in M, except, s□ and c*□, at locations
where new
spaces are added to c*
9. endfor
10. return multiple string alignment M
```

Фигура 8 Псевдо-код на алгоритъма Center-Star (68)

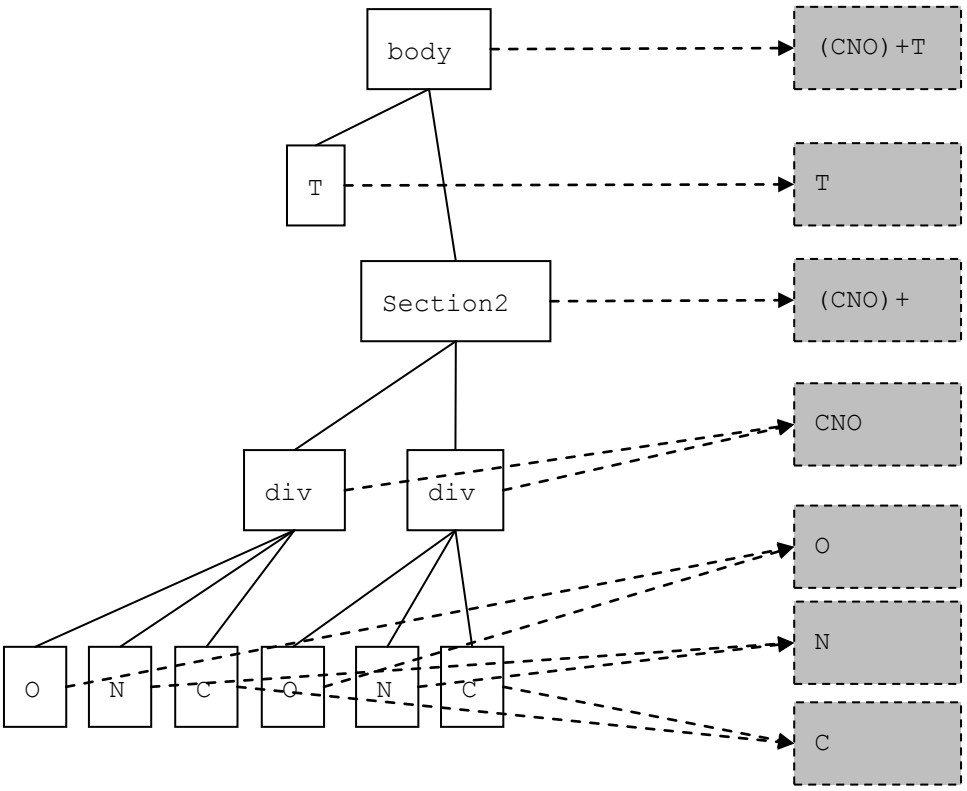
Целта е да се намали значително изчислителната сложност чрез прилагане на хеш-функция върху оценката на под-структурата на всеки възел. Методът се прилага в дизайна на компилатори и е предложен от Ira D. Baxter още през 1998г. (59) и е уместен за приложение за целите на извличане на информация от структуриран текст. Този алгоритъм трябва да открива идентични под-дървета програмен код в големи масиви от структурирани програмни текстове. Псевдо-кодът на алгоритъма е показан в следващата фигура.

```
1. Build the list structures describing sequences
2. For k = MinimumSequenceLengthThreshold
to MaximumSequenceLength
3. Place all subsequences of length k
into buckets according to subsequence hash
4. For each subsequence i and j in same bucket
5. If CompareSequences (i,j,k) > SimilarityThreshold
Then {
RemoveSequenceSubclonesOf(clones,i,j,k)
AddSequenceClonePair(Clones,i,j,k)
}
```

Фигура 9 Псевдо-код за откриване на идентични под-дървета в програмен код по I. Baxter et al. (59)

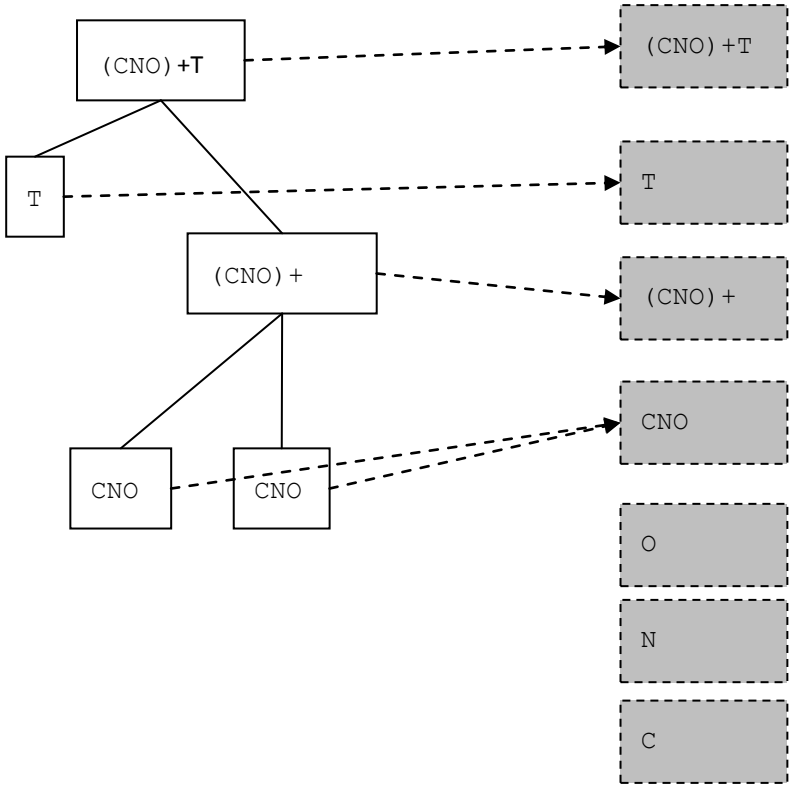
Имайки предвид факта, че функцията разпределя дърво с N на брой възли в B на брой множества (buckets), сложността на алгоритъма би била $O(\frac{N}{B})$. В зависимост от дизайна на хеш-функцията може да се пренебрегне подредбата на листата или вътрешната структура на дървото и да се сравнят дърветата само по съдържащите информация листа. Например резултатът на функцията може да бъде сортираният списък на типовете данни в листата на под-структурата на всеки възел. "CNO" може да бъде хеш-стойност за всички предци на точно едно листо от тип „Име”, едно листо от тип „коментар” и едно листо от тип „Дата”, игнорирайки точния им ред в под-дървото.

Използвайки примера от предходната точка може да се илюстрира съпоставянето на редуцираното дърво по следния начин:



Фигура 10 Прилагане на хеш-функция към обработеното съдържание на уеб документ

Съответното приложение на хеш-функцията към целевото дърво и получаването на съответствие е илюстрирано на следващата фигура:



Фигура 11 Съпоставяне на целевото дърво с отпечатъка на уеб документ

Съпоставянето на целевото дърво с входния уеб документ се счита за успешно, тогава и само тогава, когато има еднозначно съответствие между възлите на целевата структура и възлите на обработения документ.

Цялостният алгоритъм може да бъде представен със следният псевдо-код:

```
1. Foreach node ϵ DocumentTree, direction = leafs to root {
      AssignTypeToNode(node);//types in E,N,T,C,O
   }
2. Foreach node ϵ DocumentTree, direction = leafs to root {
      if (node.type==E) node.delete();
      else if (node.numberOfChildren()<=1) node.delete();
      else   node.assignAttributes();//#descendants,   emotional   score
   }
```

```
3. Foreach node ϵ DocumentTree, direction = leafs to root {
      applyHash(node,hashMapDocument);
   }
4. Foreach node ϵ MatchingTree, direction = leafs to root {
      applyHash(node,hashMapMatching);
   }
5. Foreach result ϵ hashMapMatching {
      if (hashMapDocument.get(result)!=1) return false;
   }
   return true;//correspondence between trees found.
```

Фигура 12 Псевдо-код на алгоритъма за съпоставяне на йерархични структури на документи

3.2.3. Изчислителна сложност на алгоритъма

Назначението на типове данни за възлите на листата има линейна сложност, защото се извършва поотделно за всички листа и в последователен ред. Сложността на самия анализ на текст може да бъде нелинейна, но това е извън предмета на текущото изследване и също така може да се приеме, че дължината на текстовете е ограничена от дадена константа C.

Фазата на редуциране на дървото, когато последното се почиства от листата с ненужни типове данни и техните възли-предшественици, които не водят до значими листа, има също линейна сложност и се прави последователно от листата към корена на структурата.

Накрая, сравнението с хеш-функция на полученото опростено и оценено дърво с дървото-обвивка е също от порядъка на N т.е. има линейна сложност, защото изчислението на стойностите на хеш-функцията зависи само от присвоената информация към всеки възел. Сложността на фазите на алгоритъма се изчислява, както следва:

• Оценка да типа на данните в листата – $O(N)$

• Редукция на входящата структура – $O(N)$

• Съпоставяне на анализираната структура с целевата структура – $O(N/B)$, където B е броят на различните стойности (отпечатъци) на хеш-функцията. Тъй като B е от порядъка на N следва, че общата изчислителна сложност на алгоритъма е линейна.

Описаният процес не изследва анализа на самата HTML страница до структуриран йерархично документ (HTML parsing), тъй като този етап не е предмет на изследването и има много налични софтуерни решения за този проблем – както комерсиални, така и със свободен достъп.

3.2.4. Ограничения на метода и решение

Идеалният вход за описания съпоставящ алгоритъм са валидни уеб страници, следващи стандартен XHTML или HTML език, с чиста дървовидна йерархия на елементите. Определено има много изключения в мрежата, но софтуерните продукти за анализ на HTML документи се справят успешно с „поправката" на подобни страници и предоставят добра отправна точка за настоящия метод.

Следващият набор от ограничения е свързан със съпоставящата хеш-функция. Откриването и съчетаване на повтарящи се под-структури трябва да бъде решено чрез добавяне на допълнителни знаци в резултата на функцията, например добавянето на скоби, както бе показано в примерите по-горе е едно възможно решение.

За да бъде ефективен методът, той би трябвало да съпоставя повече от един модел дърво на дадена уеб страница. Това се налага от факта, че различните сайтове за новини с потребителски коментари, продуктови или туристически мнения имат различна структура, но все пак следват общи модели на представяне. Без да се губи линейна изчислителна сложност, могат да се добавят още възможни съпоставящи дървета към метода. Това ще увеличи сложността до $O(MxN)$, където M е броят на съвпадащите дървета (Wrapper Trees), който е по-малък от определена константа N и може да бъде игнориран.

3.3. Експеримент

Целта на експеримента е да се демонстрира ефективността на алгоритъма за извличане на потребителски мнения с данни от реалния свят.

За тестова база е избран проекта Open Directory (http://www.dmoz.org/), защото той все още представлява програмно достъпен масив от класифицирани уеб адреси по категории. Съдържанието на целият масив от уеб адреси с тяхното описание и класификация по категории, може да бъде изтеглено във формат RDF (Resource Description Framework) и обработвано автоматично. Списъкът с уеб сайтове често е използван като отправна точка за мащабно уеб индексиране. Уеб сайтовете включени в проекта също са класифицирани по езиков и географски признак. По време на експеримента директорията съдържаше около 4.9 милиона уеб сайта.

Open Directory Project предоставя интернет адреси, които съдържат, като под-страници, потенциални кандидати за тест със съпоставящи дървета. Целта на настоящият експеримент е да бъде подготвен набор от съпоставящи дървета (Wrapper Trees), който да бъде приложен към под-страниците на даден списък от уеб директорията. За успешно съпоставяне се счита всяко еднозначно съпоставяне на целево дърво с уеб страница от уеб домейн, или под-домейн, включен в проекта. Това означава, че са необходими следните условия:

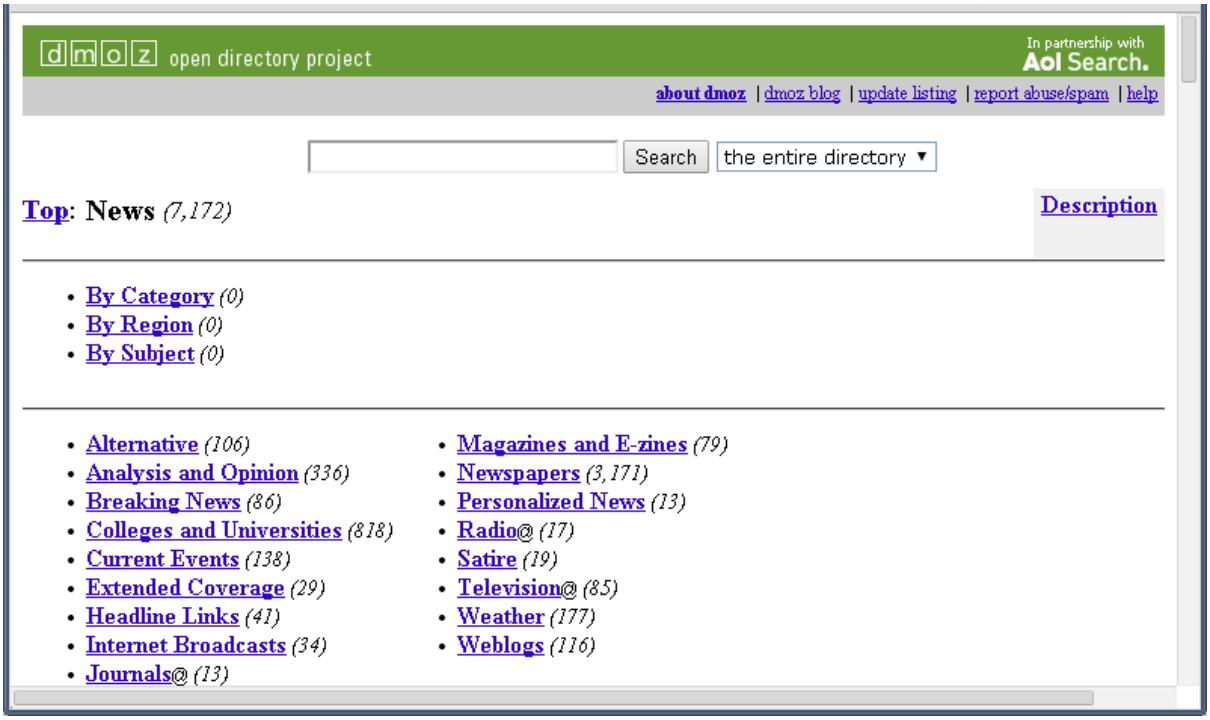
(1) да се избере категория от Open Directory Project, която съдържа подходящи уеб сайтове,

(2) сайтовете от въпросната категория на се обходят до определена дълбочина и

(3) получените уеб страници да се обработят допълнително по метода, описан в настоящата глава и

(4) да се извърши съпоставяне на целевите дървета с получените уеб документи.

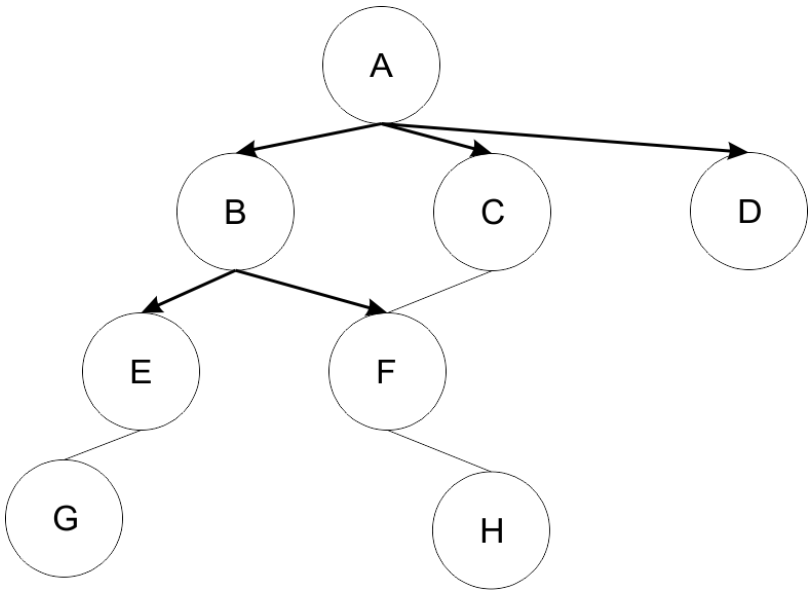
Потребителският интерфейс на Open Directory Project дава информация за наличните категории и тяхното съдържание:



Фигура 13. Общ изглед на категориите за новини в Open Directory Project

3.3.1. Дизайн

Методът се прилага в два тестови набора от уеб сайтове. Първата оценка на резултатите се базира на 100 ръчно подбрани уеб страници на английски език, разделени поравно на отрицателен (несъпоставим) и положителен (съпоставим) клас. Вторият набор от тестове е проектиран да оцени по-голям брой сайтове, взети от определени секции на уеб директорията. При този вид тест, проверяваните уеб сайтове се обхождат до определена дълбочина и се съпоставят всички намерени страници с целевите структури – съпоставящи дървета. Редът на обхождане на страниците е илюстриран на следната фигура:



Фигура 14 Обхождане в ширина на уеб страниците от даден домейн или начален уеб адрес при фиксирана дълбочина

За целта на индексирането на големия брой страници е използвана платформата за разпределени изчисления Apache Hadoop (http://hadoop.apache.org/) и уеб робота Nutch (http://nutch.apache.org/).

3.3.2. Провеждане на експеримента и резултати

Първият етап на експеримента е класификация на 100 ръчно подбрани страници, в които се търси директна съпоставка с целево дърво (Wrapper Tree). Целевото дърво е именно дървото от *Фигура 11 Съпоставяне на целевото дърво с отпечатъка на уеб документ*. За целта на експеримента са избрани равен брой положителни и отрицателни страници. Понеже съпоставянето се извършва на ниво страница, то в един и същи сайт и модел на управление на съдържанието могат да се открият както положителни, така и отрицателни примери. На следващата фигура е илюстриран положителен пример на уеб документ, съдържащ основен текст и потребителски коментари.

Фигура 15 Пример на положителен (успешно съпоставен) уеб документ, спрямо целевото съпоставящо дърво

Крайният резултат може да се намери в следната таблица, като постигнатата точност е от 91%. Разбира се, става дума за сравнително еднотипни документи, някои от които с идентични елементи от структурата.

	Actual positive	Actual negative	Total
Classified Positive	43	2	45
Classified Negative	7	48	55
Total	50	50	100

Таблица 4. Класифициране на 100 ръчно подбрани положителни и отрицателни примери.

Вторият набор от експерименти, беше насочен към изследване на реални данни. Избира се категория уебсайтове, за които се предполага, че отговарят на модела на съпоставящото дърво. В този случай също става дума за страници с модел основен текст + коментари на читателите. Такава категория е секцията *Новини* от Open Directory (www.dmoz.org/News), която в изданието, достъпно за сваляне и автоматична обработка съдържа 7947 препратки към времето на провеждане на експеримента. Сайтовете в този раздел се очаква да представят новини, статии с техния автор и дата на публикация, евентуално, съдържащи повтарящи се коментари от потребители.

За положително класифициран от алгоритъма уебсайт се счита този, за когото е налична поне една положително класифицирана под-страница, т.е. в която има успешно съпоставяне с целевото дърво (Wrapper Tree). Количеството уеб страници може да се предвиди по следния начин. Ако се приеме, че средният брой нови под-страници, посочени от дадена страница P е l, то на всяка нова итерация на уеб робота при обхождането на даден сайт, наличните документи ще се увеличават l пъти. Тоест, при провеждане на n итерации, общият брой страници ще бъде:

$$PageCount = l^n \quad (5)$$

Където l е средният брой нови под-страници, а n е броят на итерациите на обхождане на уеб робота. Тъй като сайтовете с новини са с висока степен на взаимно свързване между различните страници, то по предварителна оценка, подкрепена от данните от документите от първата част на експеримента, един уеб документ с публична новина съдържа до 50 връзки към други части и страници на сайта. Това са секции, категории, подобни новини, профили на читатели и др. Затова изборът на дълбочината на обхождане е определящ за практическото осъществяване на самия експеримент. Поради ограничените хардуерни възможности и за да се запази практическата обозримост на данните, за стойност на n бе избрано числото 3, т.е. началните уеб адреси бяха обходени в дълбочина до 3 нива. Към момента на индексирането на големия брой страници е използвана платформата за разпределени изчисления Apache Hadoop (http://hadoop.apache.org/) и уеб робота Nutch (http://nutch.apache.org/), базиран на текстовия индекс Lucene (http://lucene.apache.org/core/). Практическото приложение на този експеримент е построяването на база от модели за извличане на потребителски коментари от статии на английски език.

След извършване на експеримента, а именно – обхождане на първоначалното множество от 7947 сайта на електронни медии до 3-то ниво вътрешни препратки, настоящият алгоритъм сканира над 19 милиона страници и откри положителни съпоставяния в 6853 от сайтовете или 86.23%. Ръчна проверка на 100 случайно подбрани уебсайтове сред тест множеството показа резултатите в следната таблица:

	Actual positive	Actual negative	Total
Classified Positive	81	5	86
Classified Negative	11	3	14
Total	92	8	100

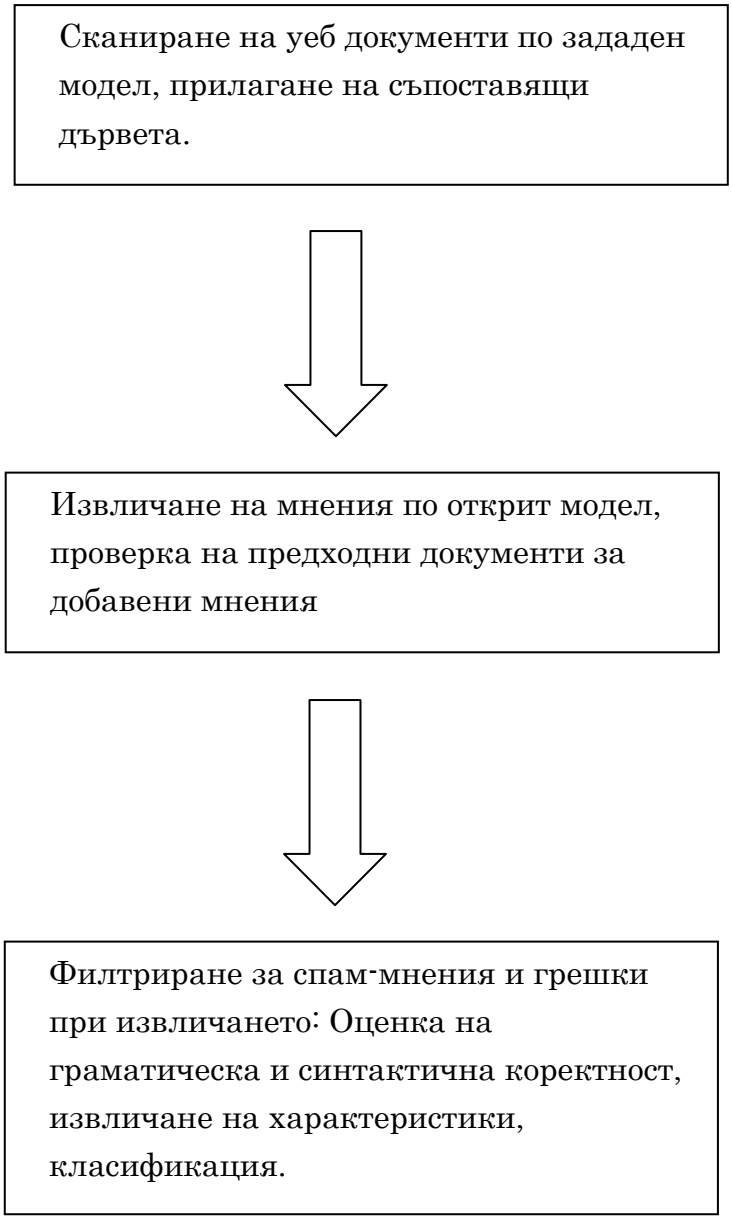
Таблица 5. Резултати от втората част на експеримента със 100 случайно подбрани сайта.

3.4. Дискусия

Предложеният алгоритъм показва точност 91% при ръчно избрания тестов набор, докато точността пада на 84% при произволно конструирания тестов набор, като резултатите са сравними с точността постигната от Peng et al. през 2004г. (70). Спадът от 91 на 84% в резултатите между двете изследвания може да се дължи на най-малко две причини: Първата е, че преди всичко ръчно изработеното тестово множество, волно или неволно, дава предимство на определен стил на уеб страници, които са по-лесни за автоматичен анализ. Втората причина, която може да обясни разликата е фактът, че не всички от интернет страниците, присъстващи в проекта Open Directory следват съвременните тенденции и някои от тях са запазили своят архаичен (за разбиранията на мрежата) HTML дизайн, който е остарял и не е подходящ за съвременен анализ на структурата. Трябва да се отбележи, че тази отворена директория губи своята популярност и бе изоставена от компанията Google, които прилагат автоматизиран семантичен анализ пред ръчното класиране на dmoz.org. Данните в този проект постепенно остаряват и субективно наблюдение показва, че въпреки големия брой сайтове, обявени за предоставящи "Breaking News", не малък брой от тях не предлагат социални функции като коментиране и споделяне на мнения от читатели. Това създава по-големи трудности в процеса на съпоставяне и разпознаване на целевата структура.

Основен проблем в извличането на мнения чрез прилагане на посочения метод представляват потоците, генерирани от страниците, които са класифицирани като положителни, но всъщност са отрицателни (true negative, classified positive). Тези потоци от становища всъщност представляват текстове с отрицателен принос (спам становища) за крайния резултат. Един от начините за справяне с проблема е да се разработи филтър за оценка на приноса на дадено становище, чрез статистическа оценка на използваната терминология. Yang Liu (28) предлага именно такъв подход през 2010г. Jindal et al. (71) прилагат анализ чрез линейна регресия и извличане на дискутирани характеристики. Допълнително филтриране, което е щадящо изчислителните ресурси, е обикновената граматическа проверка за правопис и части на речта. Становищата с граматически грешки и проблеми в изказа показват ниска стойност поради техническа грешка в извличането, ниска съпричастност на автора към темата, или ниска ангажираност на читателите на мнението. Етапите на обработка на уеб

документи в процеса на извличане на мнения може да се обобщи в следната диаграма:



Фигура 16 Основни етапи на извличане и филтриране на онлайн мнения

Точността на предложения класификатор е силно зависеща от методите за обработка на естествен език като анализ на чувства и разпознаване на възлите за дата и имена (потребителски имена, име на автор, ключови думи в атрибутите на възлите). Качеството на този предварителен анализ е от решаващо значение за крайните резултати и следователно има влияние върху оценката на алгоритъма за съпоставящи дървета.

Друг ключов момент в експеримента е проектирането на съпоставящата хеш-функция. Тя определя чувствителността на съпоставянето на дървото и неговата прецизност. Подобни проблеми са били проучени от (59) за откриване на идентични под-дървета в йерархични структури и са в основата на настоящото предложение за изчислително ефективен метод за извличане на онлайн мнения. Експериментът установи, че, за целите на извличането на онлайн становища, няма значение хоризонталната подредба на наследниците в йерархично структурирания документ и че основна роля за ефективността на алгоритъма играе успешното присвояване на типовете данни на листата на структурата.

3.5. Изводи

Предложеният метод за извличане на информация се основава на хеш-функция, дизайн, който позволява сравняване и съпоставяне на дървовидни структури с линейна изчислителна сложност. Точността на метода е демонстрирана с набор от реални уеб страници. Ключов момент в изследването е прилагането на методи за анализ на чувства и текст на естествен език, които подготвят съдържанието на HTML документите и обобщават текстовото съдържание в типове от високо ниво. Експериментът е фокусиран върху извличането на статии и коментари към тях от уеб страници. Това е, според лично убеждение, една от най-широко разпространените форми на онлайн споделяне на мнение по актуални въпроси. По-нататъшната работа може да включва извличане на допълнителна информация, като указатели на одобрение от останалите читатели и подобряване на началния текстов анализ.

4. Класификация на потребителски мнения чрез семантични лексикони

Това изследване представя метод за класификация на положителност на онлайн коментари за ресторанти в Обединеното кралство. Коментарите са събрани чрез уеб робот (web crawler) и анализирани на няколко стъпки с разпознаване на части на речта, конструктор на лексикони на емоционално заредени думи, алгоритъм за числова оценка на положителност и експерименти с различни алгоритми за машинно самообучение, класифициращи събраните мнения. Методът е тестван върху 500 000 онлайн ревюта на ресторанти и хотели, представляващи сравнително прости и кратки текстове, които също са самооценени от техните автори. Използваният речник от авторите също е ограничен до набор от теми като обслужване, качество на храна, заобикаляща среда като паркинг и др. теми, свързани с общественото хранене. Резултатите от проведения експеримент показват, че този метод постига точност до 88%, което е сравнимо с най-добрите резултати, отразени от други подходи, изредени в обзорната част на настоящата теза.

Задачата за анализ на мнения в големи обеми от данни предизвиква увеличаване на изследователския интерес през последните години. С възхода на социалните мрежи и различни видове уеб медии като форуми, блогове и видео споделяне, се повиши и нуждата от методи и инструменти, които са в състояние да обработват информационния поток в реално време и автоматично да анализират мнения и настроения от онлайн текстове. Този анализ има различни приложения в бизнеса и социалните мрежи.

Настоящото изследване представя метод, който изгражда семантични лексикони за онлайн класификация на положителност. Той включва изграждане на речник от емоционално заредени думи, морфологичен подход за извличане на части на речта, етикетни последователни правила (Label Sequential Rules – LSR), идентификация на положителност на мнение чрез точкуване (scoring). В експерименталната част методът се обучава и прилага върху реални и скорошни потребителски отзиви за ресторанти от Обединеното кралство.

Областта на ресторантските ревюта предполага използването на анализ, ориентиран към конкретната тема и характеристиките на обектите, които се обсъждат в текстовете. Относно заведенията за обществено хранене клиентите

обикновено обсъждат няколко аспекта като храна, обслужване, местоположение, цена и обща атмосфера. Нашата цел е да се оцени автоматично положителността в мненията. Изграждат се два независими лексикона: първият се състои от емоционално заредени единични части на речта (униграми), а вторият представлява двойки на прилагателни и съществителни имена, извлечени от прегледите. Вторият лексикон всъщност представлява набор от характеристики, заедно с техните оценки, извлечени от онлайн прегледите.

За оценка на резултатите на алгоритъма служи самооценката на онлайн мненията, дадена от самите автори. Връщайки се към примера от Фигура 1 Примерно мнение, споделено в сайта viewlondon.co.uk (стр.12) се вижда, че авторът на мнението дава освен текстова и числова оценка на своето преживяване изразяваща се в избора на определен брой (от 1 до 5 или от 0 до 5) „звезди", даващи бърза представа за положителността на мнението. Тази оценка, в източниците на мнения е именно от 1 до 5 звезди и условно се приема, че мненията с 1 или 2 звезди са негативни, мненията с 4 и 5 звезди са позитивни, а мненията с 3 звезди са неутрални. След провеждане на експеримента, така построеният класификатор постигна точност от 88%, която може да се счита за много добър резултат, имайки предвид, че суровите онлайн данни съдържат спам-мнения и човешки грешки в собствената самооценка (броят "звездички", определен от автора на текста).

4.1. Генериране на лексикон и анализ на положителност

Настоящото изследване прилага обединение на два метода за построяване на класификатор на мнения, които, доколкото е известно на автора, не са били изследвани досега. Първият от тях е разширяването на речника чрез WordNet (72) като се прилага филтър чрез уточняване на числовата оценка за емоционалният заряд на думите. Филтърът е базиран на генерирана хистограма на срещанията на дадения термин в текстовете от обучаващото множество. Вторият метод е откриването на модели на твърдения (propositional patterns), определени като етикетни последователни правила (LSR, (68), (4)), използвайки относително голям тестов набор от онлайн мнения (250 000).

Основните стъпки за реализация са следните:

1. Изграждане на лексикон на думите с емоционален заряд.

Построяват се два речника с думи, имащи отношение към положителност/отрицателност на изразеното настроение. Това са:

(а) Речник на емоционално заредени прилагателни имена и глаголи - ръчно построен списък с начални (seed) думи, разширени чрез връзките за синоними в WordNet до окончателен списък на кандидати за емоционално значими думи.

(б) Речник на двойки прилагателно-съществително, изразяващи позитивно или негативно отношение

Вторият речник се получава с помощта на техники за извличане на характеристики с етикетни последователни правила (LSR), въведени от Liu et al. (73). LSR откриват модели от последователни части на речта. Те са много ефективни за извличане на оценка, базирана на специфични характеристики, споменати в потребителското мнение. Например двойките (*cool, soup*) и (*cool, ambience*) са с коренно противоположна емоционална ориентация, когато става дума за места за обществено хранене, докато, ако се оценява думата *cool* като униграма няма как да се вземе предвид конкретната имплементация на термина и да се извърши разпознаване на вложеното настроение.

2. Алгоритъм за оценка на чувства чрез точкуване (scoring).

Използвайки числова оценка на емоционалната ориентация на думите, се построява набор от атрибути за всеки текст. Целта е да се изгради числено изображение на текста, представящо неговия емоционален заряд, което да послужи за вход на алгоритмите за машинно самообучение от следващия етап.

3. Класификация на мнения чрез машинно самообучение без учител.

Полученото множество от атрибути от стъпка 2 се прилага като вход на различни методи за машинно самообучение, което позволява да се избере най-ефективният метод за конкретната среда. Детайлни резултати биват представени в следващите параграфи.

4.1.1. Определяне на начални думи на лексикона и обогатяване чрез WordNet

След пълен (автоматичен) анализ на частите на речта на текстовете от обучаващото множество се подрежда списъкът на най-използваните прилагателни

имена и глаголи. От тях ръчно се избират най-употребявани и най-характерни думи, които да послужат за начален списък в лексикона. Бъдещото разширение на речника се базира именно на тези думи.

Както вече бе споменато, разширението на речника става с помощта на синонимните връзки в WordNet. Добре известно е, че WordNet предлага много голям набор от синоними и има пътеки, които свързват дори *good* и *bad* като косвени синоними с 4 прехода. За целта на изследването се ограничава разрастването на два прехода и се прилага процент за намаляване на теглото на доверие на думите, намерени по този метод.

Godbole et al. (30) предлагат метод за изчисление на емоционалното тегло (sentiment score) на синонимите. Те използват следната формула за изчисление:

$$w = \frac{1}{c^d} \quad (6)$$

където c е константа, по-голяма от 1 и d е разстоянието между новата дума и сумата от началният списък. За целта на настоящото изследване избраната стойност е 1.25. Трябва да се обърне внимание върху факта, че при интегриране на речника WordNet в релационна база от данни, този подход предлага автоматична и изчислително ефективна възможност за получаване на термини, кандидати за включване в речника без да се налага участието на учител/помощник.

Разширението е планирано на два етапа – първият етап е да се открият синонимите от първо и второ ниво, а след това, като втори етап, се прилага филтър върху получените думи за да се премахнат тези, които имат противоречива оценка за положителност и не биха били ефективни за разпознаването на целевите класове от мнения. Тези думи не биха били характерни за нито един клас. Това може да стане чрез изграждане на хистограма за всяка дума над мненията от обучаващото множество. След това се изключват кандидатите, които имат различна хистограма от техните начални предшественици и накрая се изчислява крайното тегло на емоционална ориентация, както следва:

За даден термин t се означава с p и n срещанията, съответно в позитивни и негативни текстове. С P, N и U се означава общият брой позитивни, негативни и неутрални текстове. Тогава теглото на полярност се изчислява от равенството:

$$PositivityWeight = \frac{p - n}{P + N + U} w \quad (7)$$

Думи, които не са споменати в обучаващото множество се добавят в речника с теглото, изчислено по формула (6). Поради големината на обучаващото множество се предполага, че такива думи не биха оказали съществено влияние върху резултата, но се запазва тяхното присъствие, тъй като тези термини представляват знанието, кодирано в речника WordNet (72).

4.1.2. Генериране на лексикон с Label Sequential Rules (LSR)

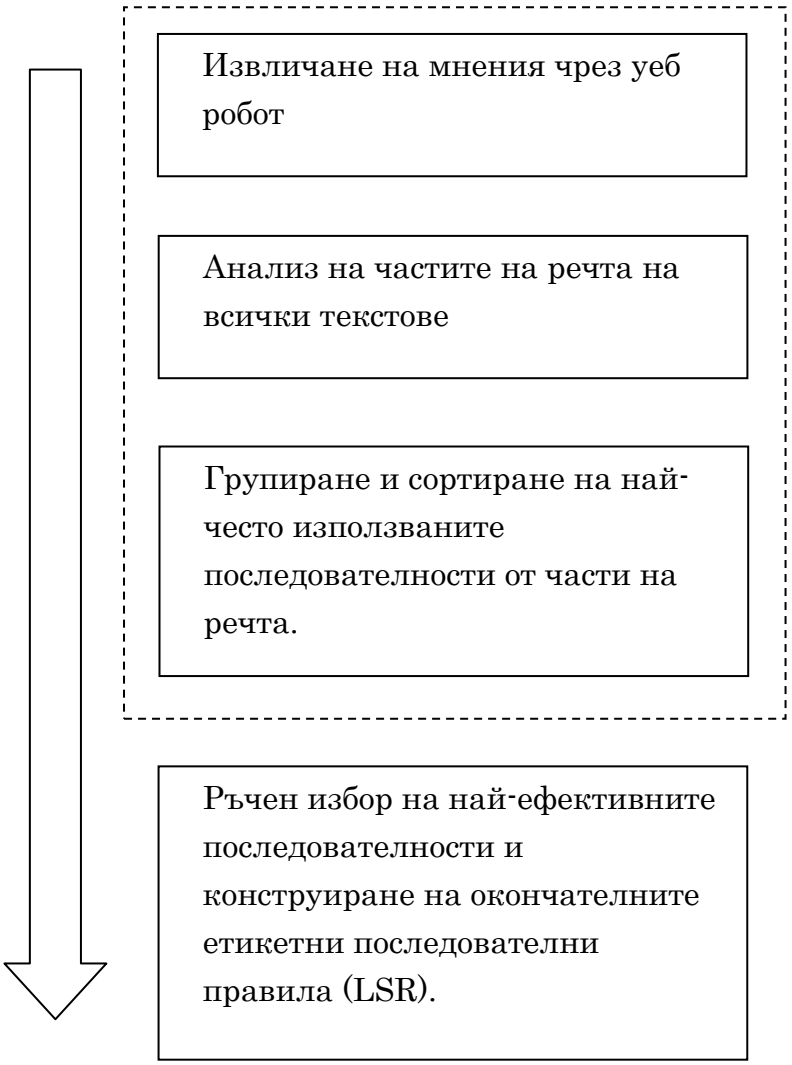
Етикетните последователни правила (LSR), осигуряват метод за извличане и откриване на общи модели на изразяване. Целевата област на кратки онлайн мнения предполага, че хората ще използват подобни експресивни модели. LSR представляват картографски последователности от части на речта и са генерирани в следния вид:

```
{$feature, noun}{(be),verb}{$quality, adjective}
[{and, conjunction}{$quality, adjective}] => 90%
{$actor, pronoun}{*,verb}
[{*,determiner}]{$feature, noun} => 90%
```

Фигура 17 Примери за етикетни последователни правила (LSR)

където квадратните скоби показват, че заградената част на речта не е задължителна и всяко правило има тегло на доверие, което пряко указва теглото в извода на двойки – кандидати за лексикона. Правилата позволяват използването на съюзи като "и" и "но", които допускат от дадено правило да бъдат изведени повече от една комбинации. Важно е да се отбележи, че методът LSR дава възможност за разделяне на оценката по характеристики (feature extraction) и чрез връзките в WordNet за хиперними и хипоними може допълнително да се обобщи групата от мнения по различните характеристики на оценения обект като качество на храната, обслужването и др. Други автори като Yang Liu (28) и Kim и Hovy (2) използват подобен подход за генериране на началния лексикон от униграми, докато при предлагания метод етикетните последователни правила се прилагат за извличане на оценки на характеристики на оценяваните обекти.

Изборът на LSR модели е важна част от алгоритъма на обучение. Един от подходите за намиране на ефективни модели е да се направи пълен анализ на частите на речта във всички текстове и после да се групират по двойки, тройки, четворки и, доколкото се прецени за нужно, n-орки от еднакви последователности от части на речта. След автоматично групиране и преброяване на n-грамите от части на речта, се извършва сортиране и ръчно отсяване на предполагаемите най-ефективни последователности. При процеса на ръчен избор се извършва и обединение на последователности (n-грами) чрез оператора за незадължителни елементи. Последователността на действията е илюстрирана на следната диаграма:



Фигура 18 Метод за конструиране на ефективни етикетни последователни правила (LSR)

Самият метод за построяване на етикетни последователни правила е извън основните цели на това изследване, но предходната илюстрация показва методологията, използвана в конкретния случай. Методът предоставя допълнителна възможност за анализ на характеристиките на даден обект, т. е. това е начин за извличане на дискутираните характеристики в предметната област (feature extraction).

4.1.3. Класификация на мненията

Използвайки резултатите от предните два етапа, а именно наличие на речник с емоционално заредени думи, оценки на характеристики и числово представяне на емоционалната им сила (тегло), се предприема последователно обхождане на всички текстове, като за всяко потребителско мнение се изчисляват следните първични и комбинирани числови атрибути:

Атрибут	Описание	Стойност
PosW	Сума на теглата на думите и фразите с позитивен заряд	Sentiment Scoring
NegW	Сума на теглата на думите и фразите с негативен заряд	Sentiment Scoring
Contr	Брой на думите с противоречив емоционален заряд като „но", „обаче" и др.	Sentiment Scoring
score1	Функция на PosW и NegW	{posw}+{negw}
score2	Функция на PosW и NegW	{posw}+2*{negw}
score3	Функция на PosW и NegW	2*{posw}+{negw}
score4	Функция на PosW, NegW и `contr`	{posw}+{negw}-{contr}
Class	Самооценка на положителността на мнението, според присвоените указателни звездички от автора (Фигура 1 Примерно мнение, споделено в сайта viewlondon.co.uk)	Positive, Negative

Таблица 6 Присвоени атрибути за оценката на потребителските мнения

Крайният резултат след присвояването на посочените атрибути е стандартна задача за машинното самообучение и позволява извършване на експерименти със средата за машинно самообучение WEKA (74).

Средата за машинно самообучение WEKA (Waikato Environment for Knowledge Analysis) предлага възможност за интегриране на числовите резултати от текущия етап, извършване на статистически анализ на атрибутите и експериментиране с различни постановки и методи за машинно самообучение. Предимствата на WEKA включват:

- безплатна наличност под GNU General Public License;
- преносимост, тъй като средата е разработена изцяло на езика Java и по този начин работи на почти всяка съвременна изчислителна платформа;
- изчерпателна колекция на инструменти за предварителна обработка на данни и техники за моделиране;
- лекота на използване се дължи на удобният графичен потребителски интерфейс.

Част от следващия експеримент е намирането на най-ефективния алгоритъм за машинно самообучение, приложим към настоящата приложна област – потребителски мнения за заведения за обществено хранене.

4.2. Експеримент и анализ

4.2.1. Дизайн

Следвайки описанието на метода, експериментът включва следните стъпки:

1. Уеб робот (web crawler) събира нужните потребителски мнения, заедно с тяхната самооценка от автора – брой звездички от 1 до 5, като 5 е максималната положителна оценка.

2. Извършва се пълен автоматичен анализ на частите на речта чрез MorphAdorner (75).

3. Сортират се частите на речта от тестовото множество и се определят началният набор от думи в речника от униграми, като се избират най-

характерните думи за начални (например по метода на взаимната информация (17)).

4. Генерира се първия лексикон чрез добавяне на синоними от WordNet и филтриране на получените думи чрез хистограма в тестовото множество.

5. Сортира се n-грами от последователни части на речта за n ϵ [2,5]; Конструиране на етикетни последователни правила (LSR) и прилагане на LSR. Генериране на втория лексикон от двойки прилагателни и съществителни имена.

6. Прави се оценка на текстовете с изчисление на теглата на участващите емоционално заредени думи от двата лексикона. Цифрово представяне на текстовете чрез изчисление на всички заложени атрибути.

7. Зареждат се резултатите в средата WEKA. Експериментиране с алгоритми за машинно самообучение над пространството от обекти и атрибути.

Целта на експеримента е да се извлекат реални данни от мрежата, след това да се анализират частите на речта и да се извадят началните думи и модели за производство на двата лексикона.

Крайният анализ на положителността в текстовете се състои от изчисляване на числови атрибути като сума от претеглените позитивни и негативни елементи, брой думи, изразяващи противоречиво мнение (думи като „„но”, „обаче”, „въпреки” са сигнал за противоречива оценка) и математически изрази, използващи предварително изчислените параметри и тегла. Изразите са всъщност входните елементи за по-нататъшната оценка от алгоритмите за машинно самообучение. Като последен етап се прилага класификацията на текстовете в средата за машинно обучение и „бенчмаркинг” WEKA (74).

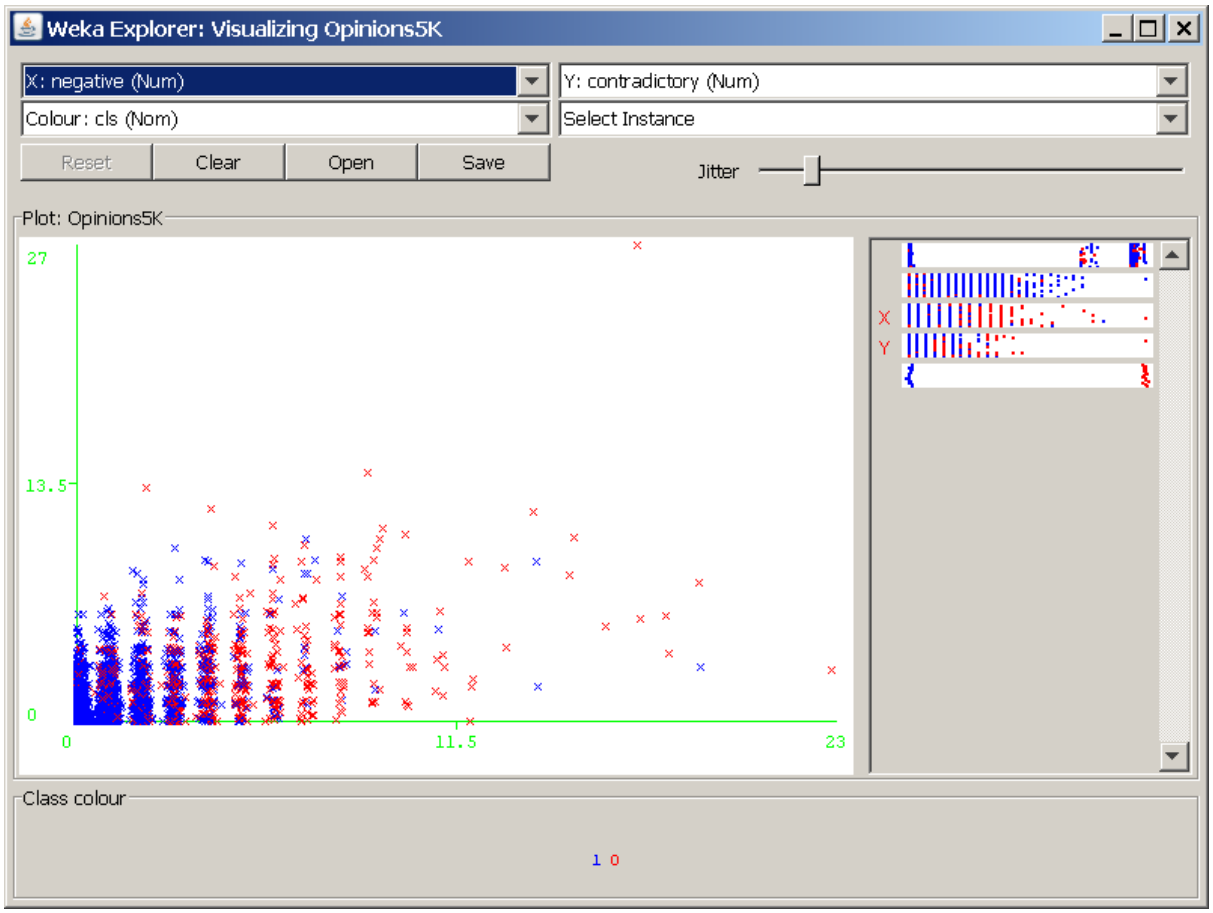
4.2.2. Определяне на положителност чрез оценка на текста

Сумата от теглата на позитивни и негативни елементи в текста образуват първите две характеристики за класификация. Тези суми се получават чрез алгоритъма за точкуване, който идентифицира емоционално заредените думи и фрази от двата лексикона. Тя се грижи за основни форми на отрицание и резолюция на местоимения. Например, ако думата е предшествана от отрицание като "is not …", знака на положителността на елемента се сменя. Например "is not good" отива към сумата на негативните думи, вместо тази на позитивните, като за

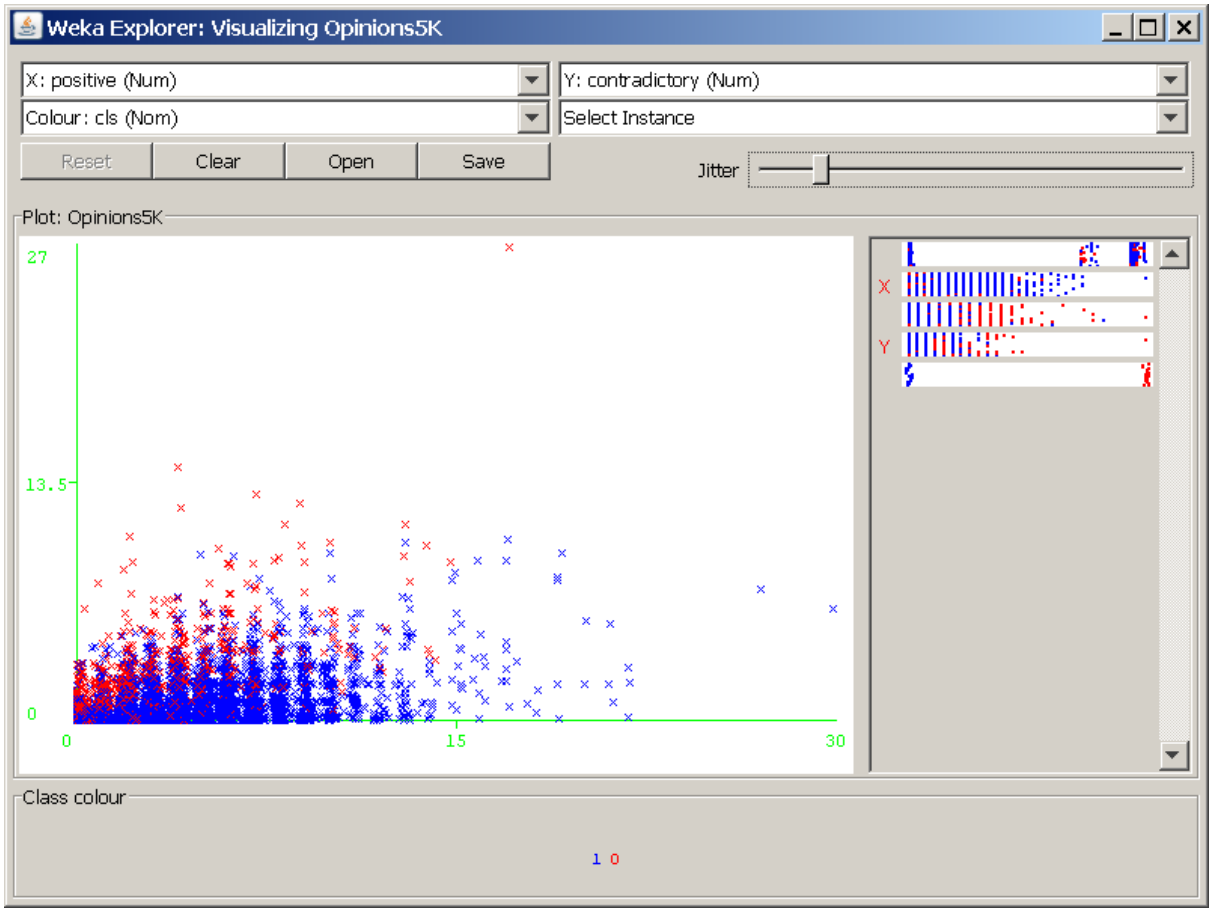
тегло се използва теглото от речника, но с обратен знак. Използваните атрибути са описани в Таблица 6 Присвоени атрибути за оценката на потребителските мнения.

4.2.3. Визуализация на атрибутите и резултати от оценката на положителност във WEKA

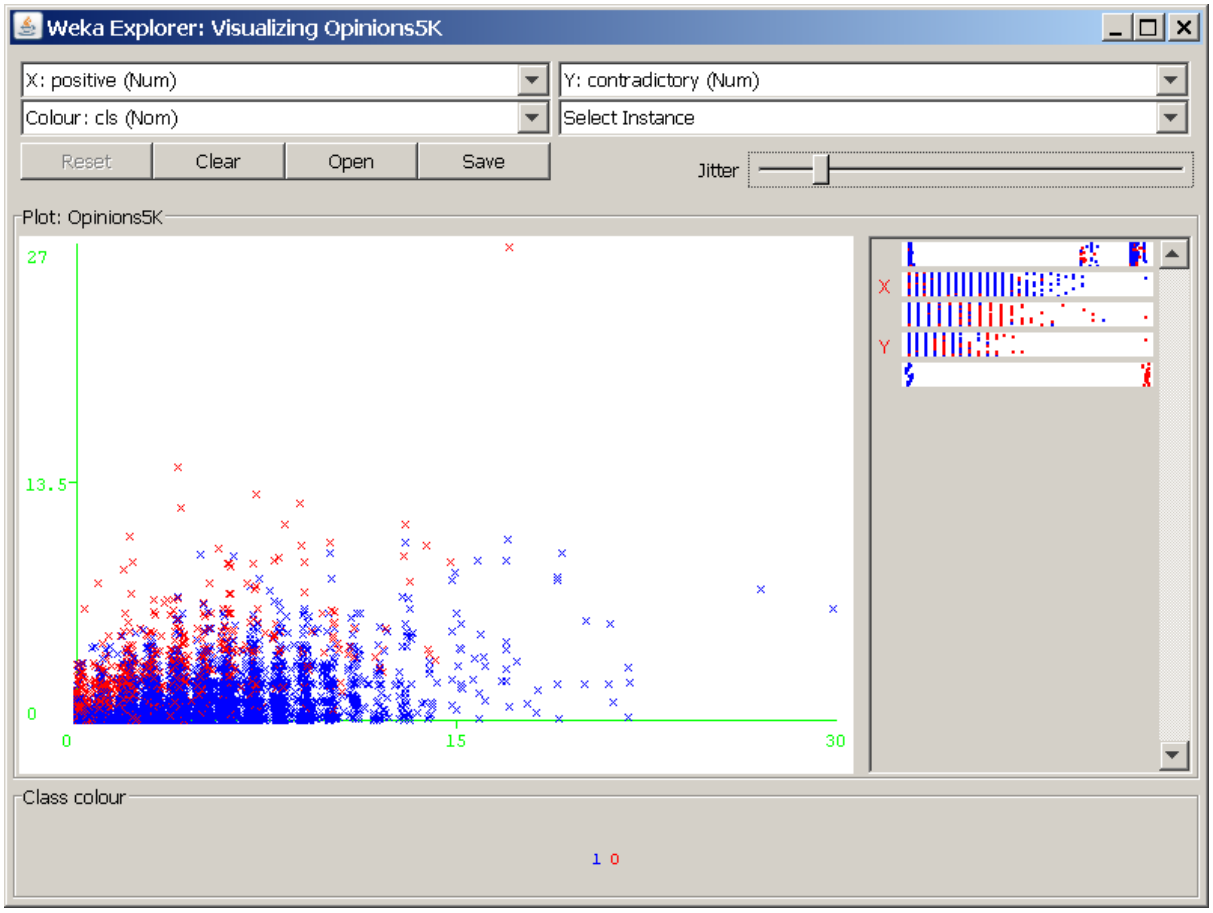
Средата за машинно самообучение WEKA предлага инструмент за визуализация на атрибутите, който помага за откриване на геометрични зависимости между техните стойности. Чрез визуализация на основните атрибути може да се заключи, че отрицателните мнения имат по-висока сума на теглата на термини и фрази с негативни мнения и по-висок брой на думи, изразяващи противоречие. Това е видно от следващите графики:



Фигура 19 Визуализация на връзката между негативни термини и думи и признаци на противоречиво мнение



Фигура 20 Визуализация на връзката между термини, изразяващи положителна емоция и признаци за противоречиво мнение



Фигура 21 Визуализация на връзката между термините с положителни и отрицателни емоции.

В предходната секция беше прието, числовите атрибути да са разширени с допълнителни, образувани от първоначалните три, за да може да се експериментира с повече алгоритми за машинно обучение и да се проведе разнообразен експеримент. Това са линейни функции, чиито резултат зависи само от стойностите на началните три атрибута, а именно: сума на позитивните тегла, сума на негативните тегла и брой на думите, изразяващи противоречие. Също така се експериментира с удвояване на стойността на негативните или позитивните елементи за да се даде шанс на хипотезата, че потребителите спестяват от силата на някой от двата вида термини, но тяхното наличие предполага по-силни емоции, изразени в крайната оценка с маркиране на звездички.

Класификацията през три вида алгоритми за машинно самообучение дава резултатите, показани в Таблица 7 Резултати от прилагането на три различни алгоритми за машинно самообучение. Точността от 87-88% удовлетворява очакванията, защото самите данни съдържат грешки при самооценката на мненията, които могат да се избегнат, освен ако не се инвестира огромен човешки ресурс за проверка на съответствието между самооценката и текста на мненията.

Algorithm	Accuracy	Precision
NaiveBayes	87%	87%

```
=== Summary ===

Correctly Classified Instances        432614               86.52   %
Incorrectly Classified Instances       67386               13.48   %
Kappa statistic                          0.4223
Mean absolute error                      0.1657
Root mean squared error                  0.3284
Relative absolute error                 58.7486 %
Root relative squared error             87.4686 %
Total Number of Instances              500000
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.963     0.614      0.885     0.963      0.922      0.845    1
                 0.386     0.037      0.682     0.386      0.493      0.845    0
Weighted Avg.    0.865     0.516      0.85      0.865      0.849      0.845
```

Algorithm	Accuracy	Precision
VotedPerceptron	83%	69%

```
=== Summary ===

Correctly Classified Instances        415108               83.02   %
Incorrectly Classified Instances       84892               16.98   %
Kappa statistic                          0
Mean absolute error                      0.1698
Root mean squared error                  0.4121
Relative absolute error                 60.1892 %
Root relative squared error            109.751  %
Total Number of Instances              500000
=== Detailed Accuracy By Class ===
```

Algorithm	Accuracy	Precision

```
               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         1          0.83      1         0.907      0.5      1
                 0         0          0         0         0          0.5      0
Weighted Avg.    0.83      0.83       0.689     0.83      0.753      0.5
```

ADTree	88%	87%

```
=== Summary ===

Correctly Classified Instances         442723               88.54   %
Incorrectly Classified Instances        57277               11.46   %
Kappa statistic                          0.5428
Mean absolute error                      0.2543
Root mean squared error                  0.3153
Relative absolute error                 90.1448 %
Root relative squared error             83.9655 %
Total Number of Instances               500000
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.96      0.477      0.908     0.96      0.933      0.881    1
                 0.523     0.04       0.725     0.523     0.608      0.881    0
Weighted Avg.    0.885     0.403      0.877     0.885     0.878      0.881
```

Таблица 7 Резултати от прилагането на три различни алгоритми за машинно самообучение

4.3. Дискусия

Задачата за класификация на чувства от текст варира за различните области на приложение. В настоящото изследване се показа, че алгоритмите за анализ чувства се представят по-добре, когато са ограничени в текстове от еднородна тематика т.е. в дадена приложна област. Интересни резултати могат да бъдат получени чрез изследване на изразените чувства общо за всички сканирани мнения за британските ресторанти по характеристики като храна, персонал, атмосфера и т.н. Това би разкрило евентуално географски и културни зависимости.

По лично мнение на автора, може да се отбележи, че общественото хранене е силно конкурентна област и клиентите са изключително взискателни. Характеристиката, която дразни повечето от клиентите е недостатъчната учтивост на персонала. Непосредствено до нея стои качеството на храната и цената идва като третата най-притеснителна тема.

Ако се обобщи общото настроение на клиентите за всички оценявани ресторанти можем спокойно да се заключи, че потребителите са категорично позитивно настроени и посещенията на местата за обществено хранене предоставят една положителна емоция. Това се вижда и от факта, че по-голямата част от изразените мнения са положителни.

4.4. Обобщение

В настоящото изследване се изгради метод за класификация на онлайн мнения, като проведеният експеримент показва успешното му приложение върху големи количества данни – оценки на ресторанти от Обединеното кралство. Подходът изгражда лексикон на емоционално заредени думи и фрази, специфични за областта на приложение, използвайки сравнително голямо тестово множество. След това се оценява полярността на чувствата в текстовете чрез прилагане на методите за теглови оценки на текст и резултатите се записват като атрибути на входни елементи за алгоритмите за машинно самообучение. Крайното класиране се извършва с помощта средата за експерименти с алгоритми за машинно самообучение WEKA.

Резултатите показват ясен път да следване – темата, свързана с автоматичният анализ на потребителски мнения може да се разглежда като ефективен и стабилен инструмент за мониторинг. Бъдещи изследвания могат да включват демографски и географски данни, за да се покажат предпочитанията на потребителите и да се осигури по-задълбочен анализ.

Бъдеща работа в тази област може да включва подобряване на алгоритъма за теглова оценка на емоционално заредените думи – по-добра резолюция на местоимения, подобряване на откриването на условни твърдения. Генерирането на лексикона може да се подобри в областта на извличане прилагането етикетните последователни правила (LSR) като се добавят повече модели. Също така, алгоритъма за изграждане на лексикони може да се приложи в различни тематични области като анализ на новини, филмови ревюта и др. за да се сравни работата му.

5. Автоматична класификация на български прилагателни имена по семантични оси

5.1. 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, (76) постулира, 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 идеи обикновено се указват като „лингвистична скала”, определена от (77) през далечната 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, (78) и (79), 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 (80), 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(7) 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 (26) 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 (81) и Hindle (82), които предлагат метод за обогатяване на начален набор от прилагателни (seed adjectives) и генериране на речник чрез класификация на допълнителни думи.

Turney и Littman (17) предлагат ефективен начин за преценка дали дадена дума е позитивна или негативна. Използвайки множество от предварително класифицирани думи, например множеството от (26), 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 операторът 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 (26).

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 (83). 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 в голяма степен близък до описаният по-горе, като е адаптиран за наличните ресурси в големите машини за търсене за български език.

5.2. 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

$$\text{Pwords} \tag{8}$$

е множество от избрани думи с позитивна ориентация, а

$$\text{Nwords} \tag{9}$$

е множество от избрани думи с негативна ориентация. Също с

$$\text{A}(\text{word}_1, \text{word}_2) \tag{10}$$

може да се означи 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:

$$SOA(word) = \sum_{p \in Pwords} A(p, word) - \sum_{n \in Nwords} A(n, word) \quad (11)$$

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 (17) 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:

$$PMI(x, y) = log\left(\frac{p(x, y)}{p(x)p(y)}\right) \quad (12)$$

Тази функция е симетрична и отговаря на изискванията за измерване на асоциативността от (10). 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.

Ако се логаритмува това отношение, се получава израза от (12), който 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

$$p(word_1) = \frac{hits(word_1)}{N} \quad (13)$$

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:

$$PMI(word_1, word_2) = \log\left(\frac{\frac{1}{N} hits(word_1\ NEAR\ word_2)}{\frac{1}{N} hits(word_1)\frac{1}{N} hits(word_2)}\right) \quad (14)$$
$$= log\left(\mathrm{N}\frac{hits(word_1\ NEAR\ word_2)}{hits(word_1)hits(word_2)}\right)$$

Последният израз не зависи от броя на индексираните документи и комбинирайки го с (11) се получава обща формула за изчисление на семантична ориентация:

$$SOPMI(word) = \sum_{p\ \in\ Pwords} PMI(p, word) - \sum_{n\ \in\ Nwords} PMI(n, word) \quad (15)$$

или

$$SOPMI(word) \quad (16)$$
$$= log\left(\frac{\prod_{p\ \in Pwords} hits(word\ NEAR\ p)\,.\prod_{n\ \in Nwords} hits(n)}{\prod_{p\ \in Pwords} hits(p)\,.\prod_{n\ \in Nwords} hits(word\ NEAR\ n)}\right)$$

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 (16) 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.

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

5.3.1. 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:

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

Таблица 8 Оценка на популярните машини за търсене.

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 нейният 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 субективна оценка на автора), филтриране по език, оператор за близост на думи и програмен интерфейс с информация за броя на резултатите.

5.3.2. 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 – Обърканост
- Вълнуващ – Скучен
- Смелост – Страх
- Трудолюбие – Мързел
- Уверен – Подтиснат
- Богатство – Бедност

Като втори етап се избират прилагателни, съответстващи на референтните области, например алчен – щедър, ясен – объркан, 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:

Позитивна референция	Негативна референция
Разбран	Объркан
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

Таблица 9 Референтни прилагателни имена, формиращи емоционални оси

5.3.3. Оценка на положителност

Колоните от референтни прилагателни имена от Таблица 9 Референтни прилагателни имена, формиращи емоционални оси представляват, сами по себе си, набор от референтни позитивни и негативни думи. Следователно, използвайки формулата (16), 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:

$$(word\ NEAR{:}\ 10\ Rword)\ language{:}\ bg \qquad (17)$$

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:

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

Таблица 10 Оценка на положителност на случайно избрани думи.

Видно от резултатите избраните референтни думи отчетливо разграничават позитивни и негативни думи и алгоритъмът притежава висока прецизност за думите с голяма абсолютна стойност на оценката. Вероятно 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.

5.3.4. Оценка на разположение по емоционални оси

Друга възможност за оценка е прилагането на подхода на Turney и Littman (17) не само върху групи от позитивно и негативно ориентирани думи, а върху семантично противоположни термини, определящи емоционални оси. Прилагайки изчислението (16) върху думите от Таблица 9 Референтни прилагателни имена, формиращи емоционални оси, 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 следната таблица:

	bing count	хубав-грозен	умен-глупав	активен-мързелив	верен-подъл	богат-беден	добър-лош	полезен-вреден	щастлив-нещастен
Нов	2100000	1.61	7.56	9.39	7.38	8.76	2.34	1.01	11.06
Жив	224000	8.38	6.57	6.41	6.39	-0.21	9.06	7.44	3.99
страхотен	224000	9.97	6.57	-0.25	6.39	-0.21	2.53	7.44	6.49
честен	50200	6.39	9.36	-0.25	6.39	6.45	10.65	-0.21	-0.17
Млад	283000	8.7	1.49	7.98	-0.27	2.24	4.07	7.44	6.49
истински	607000	8.97	-0.09	7.4	7.38	-0.21	3.52	6.45	4.34
мощен	139000	6.39	6.57	-0.25	6.39	8.76	9.28	-0.21	-0.17
Бърз	440000	7.97	8.15	0.74	-0.27	-0.21	10.48	8.44	-0.17
малък	693000	8.97	-1.08	7.98	6.39	-7.86	10.56	8.44	-0.17
делови	78900	8.38	-0.09	-0.25	6.39	6.45	8.48	-0.21	-0.17
Прав	217000	7.97	-0.09	8.72	6.39	-6.87	0.51	-0.21	9.48
успешен	151000	9.19	-0.09	-0.25	-0.27	6.45	2.28	-0.21	8.48

…

Стар	404000	0.47	-8.32	7.4	-0.27	0.37	2.14	6.45	-6.83
разумен	39700	-6.92	-0.09	-0.25	-0.27	-0.21	8.06	-0.21	-0.17
грешен	94500	-0.26	-0.09	-0.25	0.09	-0.17	0.03	-0.27	0.31
ужасен	99600	-0.26	-0.09	-0.25	-0.27	-0.21	0.32	-0.21	-0.17

…

напрегнат	40400	-0.26	-0.09	-0.25	-0.27	-0.21	-1.16	-0.21	-0.17

противен	136000	-8.5	-0.09	-0.25	-0.27	-0.21	6.49	-0.21	-0.17
незаконен	50200	-0.26	-0.09	-0.25	-0.27	-8.45	6.49	-0.21	-0.17
неприятен	43600	-0.26	-0.09	-7.9	-0.27	-0.21	6.49	-6.87	-0.17
мазен	39000	-7.91	-1.08	-0.25	-0.27	-0.21	-0.17	-0.21	-0.17
неправилен	40600	-6.92	-0.09	-0.25	-0.27	-0.21	-6.83	-0.21	-0.17
мръсен	42500	-6.92	-0.09	-0.25	-0.27	-0.21	-0.17	-0.21	-6.83
опасен	111000	-0.26	-0.09	-0.25	-0.27	-0.21	-7.82	-9.66	-0.17
ограничен	87500	-0.26	-7.74	-6.91	-0.27	-6.87	7.48	-6.87	-0.17
долен	102000	-6.92	-0.09	-0.25	-6.93	-0.21	-11.06	-0.21	-0.17

Таблица 11 Оценка на емоционално значими думи по семантични оси.

5.4. 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 Таблица 11 Оценка на емоционално значими думи по семантични оси. 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.

5.5. Обобщение

Настоящото изследване по информацията, достъпна на автора на дисертационния труд, 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.

6. Внедряване и практическо приложение

Практическото приложение на изложените методи е осъществено чрез построяването на онлайн инструмент за изследване на мнения. Инструментът е достъпен на адрес http://rankur.com и получи финансиране от фондът за развитие на стартиращи компании Launchub (http://launchub.com). Основните модули в инструмента са уеб робот, анализатор на уеб страници, анализатор на части на речта, модул за оценка на чувства, текстов индекс и релационна база от данни.

6.1. Изисквания

Както се вижда в края на тази глава, Ранкър е утвърден софтуерен продукт за анализ на онлайн репутация. В началото на разработката бяха поставени следните изисквания:

6.1.1. Изисквания към обработката на естествен език

- **Интегриране с популярните социални мрежи**: Общуването онлайн чрез популярните социални мрежи е основна активност на интернет потребителите през последните години. Съобщенията с публичен достъп са основен източник на свежи потребителски мнения в най-различни области.
- **Уеб робот, с подходяща политика за обхождане на уеб документи**: Друг източник на потребителски мнения са продуктовите страници, статии и политически сайтове, предлагащи възможност на потребителите да коментират основната тема на документа.
- **Анализатор на чувства от текст**: Откритите потребителски мнения трябва да бъдат оценени за предаваното в тях настроение с речник от контекстно независими думи.
- **Анализатор на демографски данни**: Една завършена система за обработка на мнения би трябвало да може да даде повече информация за авторите на мненията и да ги групира по географски, възрастов и други признаци. Етикетните последователни правила (Label Sequential Rules) предлагат добра възможност за откриване на твърдения,

свързани с възрастта и географското местоположение в описанията на публични профили в социалните мрежи.

6.1.2. Технически изисквания към реализацията

- **Текстово търсене**: Получените данни трябва да бъдат на разположение за допълнително и бързо търсене в свободен текст, като резултатът трябва да е независим от знаковия набор (codepage), използван от автора. Данните следва да бъдат съхранявани в структуриран вид като автор, информация за автора, дата и час, текст на мнението, оценка на настроението, части на речта, оценки за характеристики и др.
- **Преносимост**: Поради разнообразието от работни станции и мобилни устройства, би било подходящо проектът да е достъпен през уеб интерфейс, който не изисква предварителна инсталация на софтуер на клиентското устройство.
- **Скалируемост**: Изчисленията, извършвани от софтуерния пакет трябва да подлежат на паралелизация, която да осигури работоспособността на системата при големи натоварвания, както и защита от хардуерни сривове.

6.1.3. Изисквания към потребителския интерфейс

- **Скорост**: Във века на облачните технологии и методите за паралелни изчисления потребителите биха счели за неприемливо да изчакват повече от едноцифрен брой секунди за да получат нужната информация.
- **Удобен потребителски интерфейс**: Потребителският интерфейс следва да е базиран на съвременна среда (framework) за ефективна разработка, която да предлага леснота на работата със софтуера и лесен за поддържане програмен код.
- **Приспособимост на визуализацията**: Потребителите на уеб базирани приложения използват голямо разнообразие от устройства за достъп до глобалната мрежа. Следва софтуерът да е способен да оптимизира представянето на данните за размерите и параметрите на основните

класове устройства като настолни и портативни компютри, таблети и мобилни телефони.

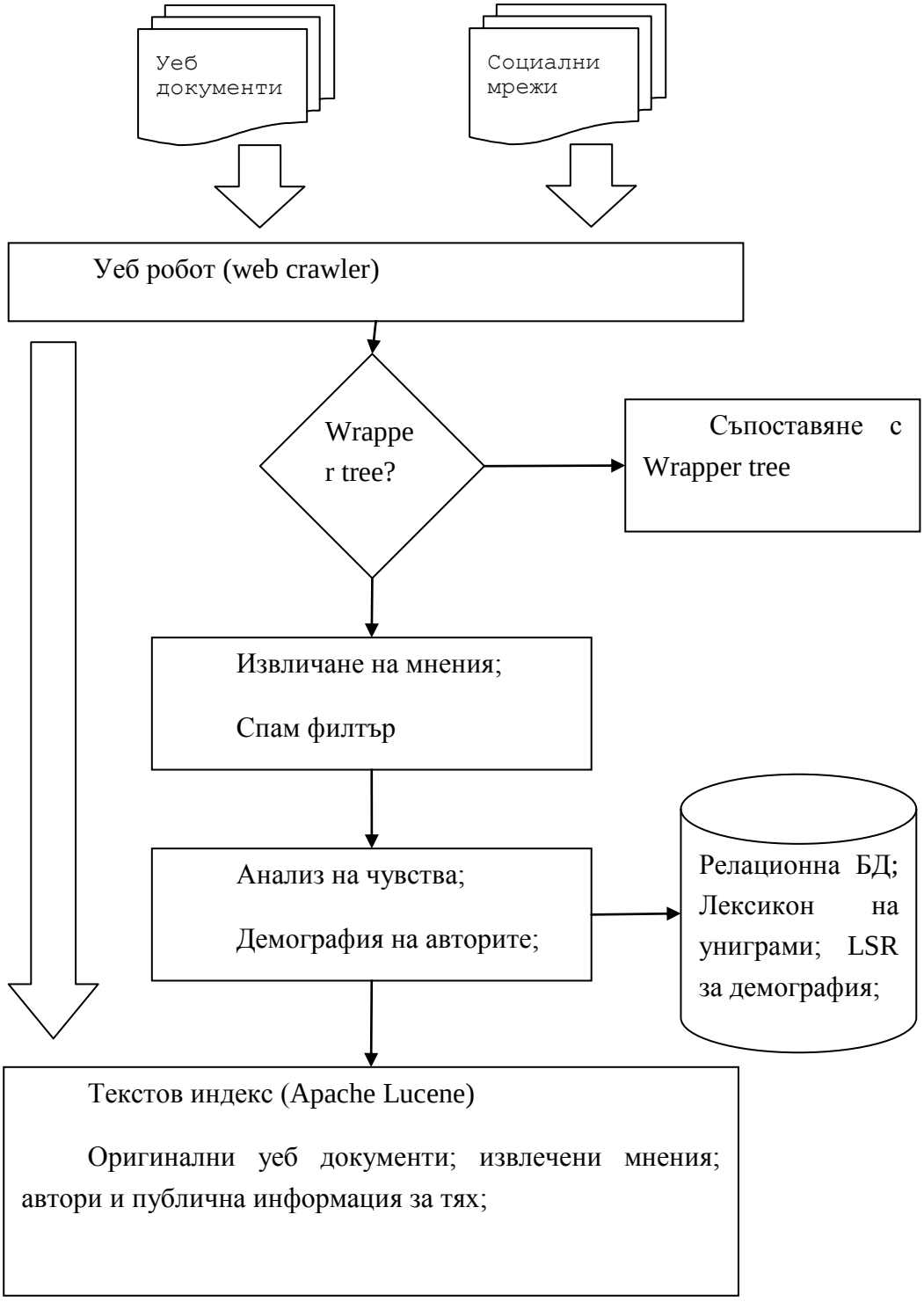
6.1.4. Изисквания към технологичния процес

- **Съвременно управление на програмния код**: Нужно е да бъде внедрена съвременна система за управление на програмния код, която да осигури качество и ефективност на разработките.
- **Управление на нови изисквания и поддръжка**: Всеки софтуерен проект има нужда от система за управление на неговия жизнен цикъл както по време на разработката, така и след реализацията на първата завършена версия.

6.2. Реализация на проекта

Във финалния вариант софтуерният пакет притежава програмен код от над 4 000 000 знака и е реализиран на 10 програмни езика. Това са процедурните: Java, PHP, javascript, PL/SQL, Unix shell scripts и експресивните: SQL, jQuery, Xpath, XSL и HTML5. Хардуерната реализация включва взаимозаменяемост на модулите, защита от външната среда като сривове и пикове в напрежението, географска разпределеност на данните с цел защита от тяхната загуба.

Функционирането на системата е представено на следващата диаграма. Макар и схемата да е в силно съкратен вид, тя дава представа за решението на поставените задачи.

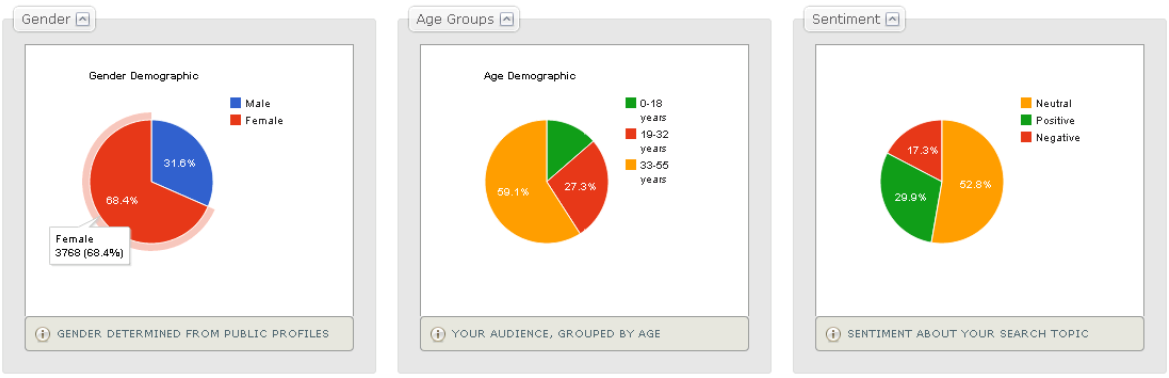


Фигура 22 Основни модули на класификатор на мнения в реално време

Уеб роботът и текстовият индекс използват методологията за разпределени изчисления Map-Reduce, докато програмната логика като съпоставяне на йерархични дървета (Wrapper Trees), оценка на чувства от текст, извличане на характеристики от мненията и описанието на техните автори е реализирано по начин, който позволява изпълнението на изчисленията в клъстер и паралелно изчисление с неограничен брой процеси. Крайната реализация на проекта е преносима и инсталируема на платформите за облачни (паралелни) изчисления на популярните хостинг компании като Amazon Web Service, RackSpace, SoftLayer и др.

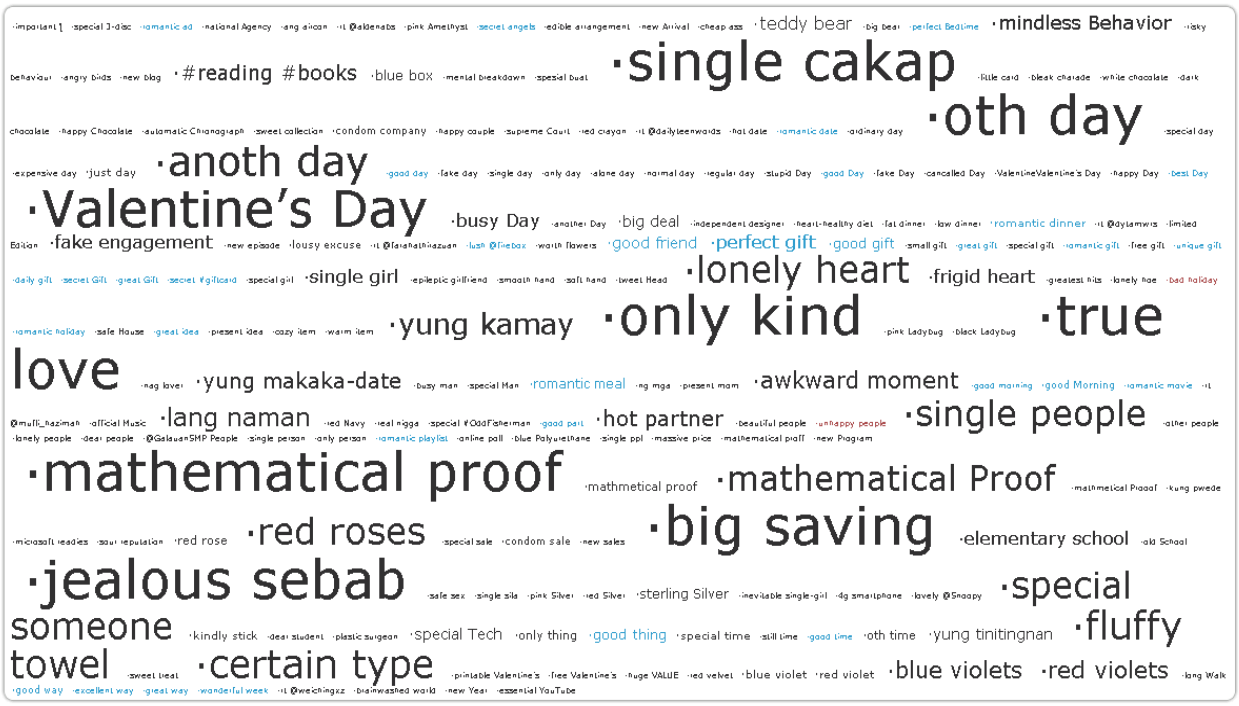
Крайната конфигурация позволява непрекъсната работа на системата дори при пикови натоварвания, например след неочаквана публикация в популярен уеб портал като SocialMediaToday или PuroMarketing. Следват графични примери на реализацията на проекта.

Следващата фигура представя демографските параметри на хората дискутиращи празника „Свети Валентин" непосредствено преди и по време на самия празник. Забелязва се силно дамска аудитория и вероятна изненада – превес на хората на средна възраст сред авторите на публични съобщения за празника. Наличието на отрицателни мнения показва противоречивите чувства, които темата предизвиква сред онлайн потребителите.



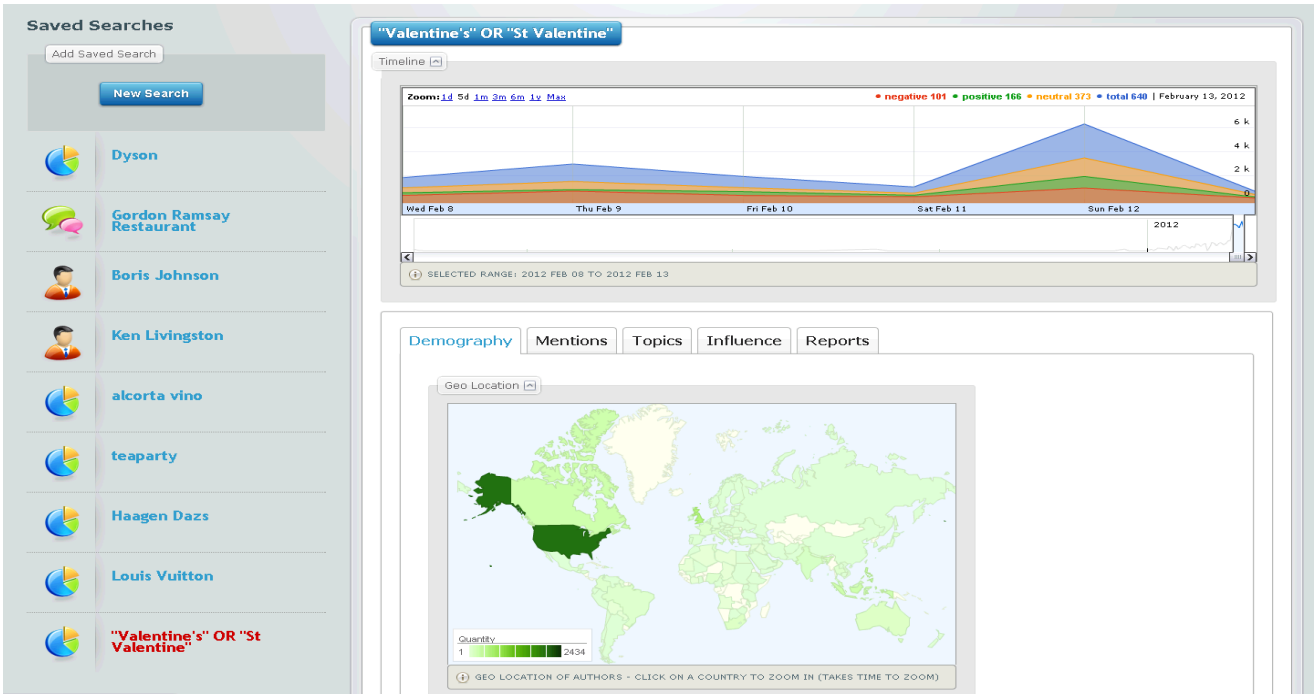
Фигура 23 Демографски анализ, извършен върху публични изказвания за празника „Св. Валентин"

Приложението на анализа на части на речта в комбинация с етикетните последователни правила (LSR) показано на следната фигура, където се виждат двойките прилагателни и съществителни имена.



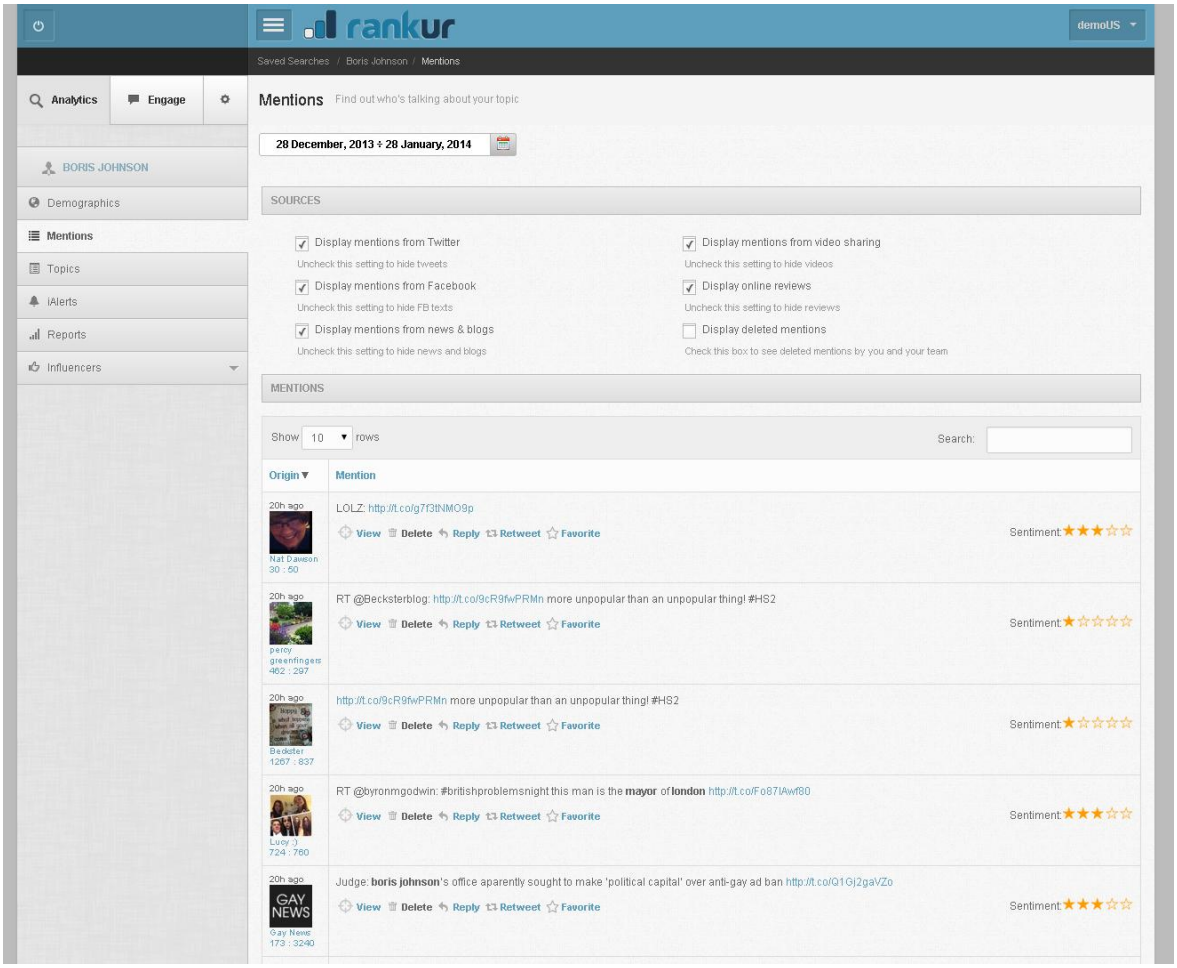
Фигура 24 Извличане на фрази чрез етикетни последователни правила (LSR) с оценка на обем и вложени емоции

Количественият анализ, както и сравнителното разпределение на авторите по географски признак са видни на следващата фигура:



Фигура 25 Количествен анализ на използваните чувства и географско разположение на източниците

През 2013г. потребителският интерфейс на продукта бе преработен, като бяха добавени модули за директен разговор с клиенти със средствата на социалните мрежи и планиране на публикации в най-ефективните периоди на денонощието. Пример на визуализацията на съобщенията е даден на следващата фигура:

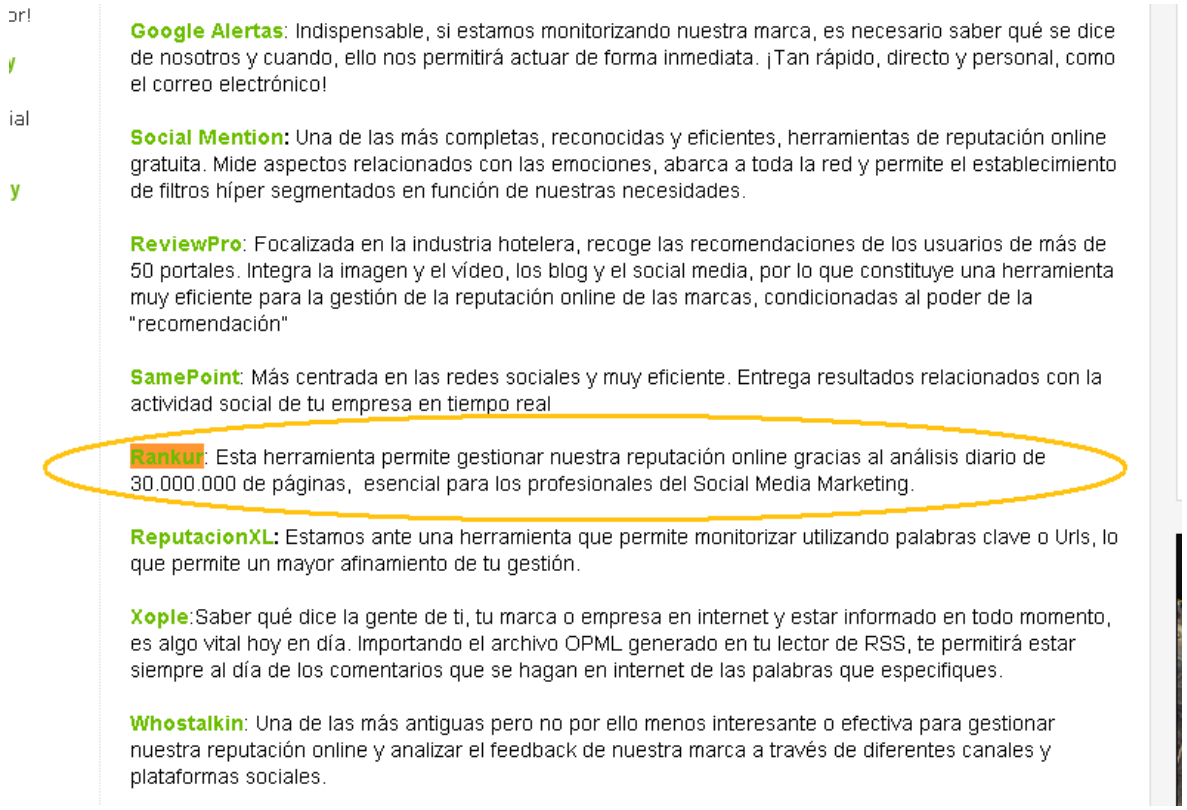


Фигура 26 Автоматично класифицирани мнения

6.3. Реакции и оценки на софтуерния пакет в публичното пространство

Инструментът е класиран от публиката като един от най-добрите в областта на управлението на онлайн репутация.

В началото на 2012г. популярният испански сайт puromarketing.com включи Rankur в препоръчителен списък на инструменти за онлайн репутация. Публикацията е видима на следния адрес: (http://www.puromarketing.com/42/12049/recursos-herramientas-gratuitas-para-seguimiento-nuestra-marca-gestion.html) и казва, че инструментът е от основна необходимост на специалистите по маркетинг в социални медии:



Google Alertas: Indispensable, si estamos monitorizando nuestra marca, es necesario saber qué se dice de nosotros y cuando, ello nos permitirá actuar de forma inmediata. ¡Tan rápido, directo y personal, como el correo electrónico!

Social Mention: Una de las más completas, reconocidas y eficientes, herramientas de reputación online gratuita. Mide aspectos relacionados con las emociones, abarca a toda la red y permite el establecimiento de filtros híper segmentados en función de nuestras necesidades.

ReviewPro: Focalizada en la industria hotelera, recoge las recomendaciones de los usuarios de más de 50 portales. Integra la imagen y el vídeo, los blog y el social media, por lo que constituye una herramienta muy eficiente para la gestión de la reputación online de las marcas, condicionadas al poder de la "recomendación"

SamePoint: Más centrada en las redes sociales y muy eficiente. Entrega resultados relacionados con la actividad social de tu empresa en tiempo real

Rankur: Esta herramienta permite gestionar nuestra reputación online gracias al análisis diario de 30.000.000 de páginas, esencial para los profesionales del Social Media Marketing.

ReputacionXL: Estamos ante una herramienta que permite monitorizar utilizando palabras clave o Urls, lo que permite un mayor afinamiento de tu gestión.

Xople:Saber qué dice la gente de ti, tu marca o empresa en internet y estar informado en todo momento, es algo vital hoy en día. Importando el archivo OPML generado en tu lector de RSS, te permitirá estar siempre al día de los comentarios que se hagan en internet de las palabras que especifiques.

Whostalkin: Una de las más antiguas pero no por ello menos interesante o efectiva para gestionar nuestra reputación online y analizar el feedback de nuestra marca a través de diferentes canales y plataformas sociales.

Фигура 27 Препоръка на Rankur.com в публичното пространство

Подобни публикации има във френския специализиран сайт outilsveille.com на следния адрес: http://outilsveille.com/2013/02/3-outils-pour-gerer-votre-reputation-en-ligne/ където Rankur е поставен под номер едно пред други две приложения. Специална публикация присъства и в белгийския икономически всекидневник L'echo: http://blogs.lecho.be/tzine/2013/03/petite-veille-grands-effets.html.

В Съединените Американски Щати, които представляват най-големия комерсиален пазар на автоматична обработка на мнения, Ранкър е класиран сред топ 7 и топ 10 инструменти на следните адреси:

- http://www.mckremie.com/blog/2012/05/7-of-the-best-online-reputation-management-tools/
- http://socialmediatoday.com/amanda-disilvestro/1227146/top-10-seo-reputation-management-tools-online

Една от последните публикации за софтуерното решение е на адрес https://www.openforum.com/articles/online-reputation-management-tools/, където

софтуерният пакет е споменат сред продукти, реализирани с многомилионни инвестиции на инвестиционни фондове за развитие на съвременни технологични компании.

6.4. Обобщение

Представеният софтуерен продукт е разработван от автора от края на 2010г., паралелно с настоящия дисертационен труд. По лично мнение на автора, реализацията на проекта отговаря на поставените функционални и технологични изисквания, имплементирайки една успешна система за управление и анализ на онлайн репутация.

Разработката на системата спазва съвременен работен процес и използва системи за управление на програмния код, проектните изисквания и заявките за поддръжка на софтуера. Посочените в т. 6.3 оценки показват, че става въпрос за една успешна система за управление и анализ на онлайн репутация, която е спечелила международно признание в държавите с развита онлайн комуникация.

7. Заключение

Широкото разпространение на публикуваните мнения в интернет като начин на споделяне на гледни точки, опит с даден продукт и коментари по неговите характеристики създава уникална възможност за оценка и обобщаване на чувствата на потребителите на продукти и услуги и подпомага дейности като бизнес анализ и маркетинг. Въпреки, че някои аспекти на анализа на емоции и чувства са добре изследвани, областта е нова и проблемът с точната класификация на изразените чувства от текст остава нерешен и предлага множество предизвикателства. В тази работа е направен опит да бъдат намерено решение на задачите, формулирани в т. 1.2. при извличането на мнения и анализа на чувства от текст, разгледани в следващата точка.

7.1. Приноси на дисертационния труд

Приносите на настоящата дисертационна теза са следните:

7.1.1. Научни приноси

- Подобрена е изчислителната сложност на алгоритъм за съпоставяне на йерархични структури и извличане на информация от HTML документи.
- Разработен е метод за класификация на онлайн мнения.
- Разработен е метод за автоматична класификация на прилагателните имена в българския език по положителност и по емоционални оси.
- Направен е обзор на съвременните методи за извличане на информация от уеб и анализ на чувства от текст.

7.1.2. Научно-приложен принос

- Реализиран е цялостен проект за извличане и анализ на мнения от уеб и социални мрежи. Проектът е публично достъпен и има международна популярност. В проекта има и следните внедрявания:
- Внедрен е алгоритъм за извличане на информация от HTML документи с линейна изчислителна сложност, описан в глава 3.
- Внедрен е метод за анализ на продуктови мнения, описан в глава 4.

- Внедрен е метод за оценка на положителност на прилагателни имена, описан в глава 5.
- Използвани са съвременни методи за паралелна обработка на големи количества неструктурирана информация.

7.2. Бъдеща работа

Бъдещата работа по темата включва изследване на филтри за фалшиви мнения – изкуствено добавени, с цел прикриване и намаляване важността на истинските мнения или за злонамерено увреждане на репутацията на конкурентни продукти. Методите за това филтриране могат да се основават на емоционален профил. Използвайки резултатите от последната глава и класифицирайки мненията по емоционални оси, може да се постигне геометрична класификация на използваните емоции, която да е ключът към откриване на отклоняващи се от общия поток мнения.

Друга посока за развитие е предвиждане на полезността на дадено мнение. Ползата от такова изследване би подобрила представянето на информация на бъдещи потребители на стоки и услуги, предоставяйки по-ценните мнения по-напред в списъка със споделени оценки. За постигане на такава оценка могат да се използват методите на линейната регресия описани в (28).

8. Публикации и цитирания

8.1. Публикации

Kraychev, B. and Koychev, I. **Computationally Effective Algorithm for Information Extraction and Online Review Mining**. In Proc. of International Conference on Web Intelligence, Mining and Semantics June 13-15, 2012, Craiova, Romania, ACM

Ivan Koychev, Boris Kraychev. **Classification of Books over Semantic Axes**, Семинар "Семантични технологии в хуманитаристиката" Созопол, 7-8 юни 2012 г., ИИКТ-БАН. http://www.smart-book.org/publications/#bs

Ясен Кипров, Иван Койчев, Борис Крайчев, **Социален подход към семантичното анотиране на Е-книги**. Сборник доклади на Петата Национална конференция "Образованието в информационното общество", Пловдив, 31.05-01.06.2012 г., ARIO

Kraychev, B., Koychev, I.: **Classification of Online Reviews by Computational Semantic Lexicons** – In Proc. of International Conference S3T'11 (Track Intelligent Content and Semantic), Advances in Intelligent and Soft Computing,Volume 101, 2011, pp 9-16, Springer Berlin Heidelberg.

Kraychev, B: **E-learning and the Social Media - automatic opinion analysis and summarization**, In Proc. of E-Learning and Distance education conference 2011, pp. 128-131

8.2. Доклади

Б. Крайчев, **Методи за извличане на мнения от текст и тяхното приложение за следене на онлайн репутация.** Годишна научна сесия на ФМИ, 26 март 2011г.

8.3. Цитирания

Kraychev, B. and Koychev, I. **Computationally Effective Algorithm for Information Extraction and Online Review Mining**. In Proc. of International

Conference on Web Intelligence, Mining and Semantics June 13-15, 2012, Craiova, Romania, ACM:

- Н.А. Сарры, превод на руски език http://masters.donntu.edu.ua/2012/iii/sarry/library/trans.htm
- Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar: **Opinion Mining and Analysis: A survey**, IJNLC Vol. 2, No.3, June 2013
- Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, **An Approach for Online Analysis using Expectation Maximization** International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 4, June 2013
- Maxim Bakaev and Tatiana Avdeenko, **Data Extraction for Decision-Support Systems: Application in Labour Market Monitoring and Analysis**, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1, February 2014

8.4. Цитирания на софтуерния проект

- http://www.puromarketing.com/42/12049/recursos-herramientas-gratuitas-para-seguimiento-nuestra-marca-gestion.html
- http://outilsveille.com/2013/02/3-outils-pour-gerer-votre-reputation-en-ligne/
- http://blogs.lecho.be/tzine/2013/03/petite-veille-grands-effets.html.
- http://www.mckremie.com/blog/2012/05/7-of-the-best-online-reputation-management-tools/
- http://socialmediatoday.com/amanda-disilvestro/1227146/top-10-seo-reputation-management-tools-online
- https://www.openforum.com/articles/online-reputation-management-tools/

9. Библиография

1. **View London Ltd.** View London. *www.viewlondon.co.uk.* [Онлайн] 2010-2014 г.

2. **Kim S., Hovy E.** *Determining the sentiment of opinions.* н.м. : Proceedings of the international conference on on computational linguistics, 2004. COLING-2004.

3. **Hu M., Liu B.** *Mining and summarizing customer reviews.* н.м. : Proceeedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004. KDD-2004.

4. **Liu B.** *Sentiment Analysis and Opinion Mining.* н.м. : Morgan & Claypool, 2012.

5. **Myrick, R. D. & Erney, T. .** *aring and Sharing: Becoming a Peer Facilitator.* н.м. : Educational Media Corporation, 1984.

6. **Levin, Beth. .** *English verb classes and alternations. A preliminary investigation. .* Chicago, London : The University of Chicago Press, 1993.

7. **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.

8. **Pennebaker, J.W., Francis, M.E., Booth, R.J. .** *Linguistic Inquiry and Word Count (LIWC): LIWC2001.* Mahwah, NJ : Erlbaum Publishers., 2001.

9. **Whissell, C.M. .** *The dictionary of affect in language. Plutchik, R., Kellerman, H. (eds.): Emotion: Theory, Research, and Experience.* New York : Academic Press. pp. 113–131, 1989.

10. **George A. Miller.** *WordNet: A Lexical Database for English.* н.м. : Communications of the ACM Vol. 38, No. 11: 39-41, 1995.

11. Wordnet in RDFS and OWL. [Онлайн] w3c.org, 2004 r. http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html.

12. **Budanitsky, A., Hirst, G.** *Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures.* Pittsburgh : Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association for Computational Linguistics, 2001.

13. **Kamps, J., Marx, M., Mokken, R. Rijke, M. .** *Using WordNet to measure semantic orientation of adjectives.* . н.м. : Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, IV. pp. 1115–1118., 2004.

14. **Andreevskaia, A., Bergler, S. .** *Mining WordNet For a Fuzzy Sentiment: Sentiment Tag Extraction From WordNet Glosses.* . н.м. : Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics. pp. 209-216, 2006.

15. **Hassan A., Radev D.** *Identifying Text Polarity Using Random Walks.* 2010 : Proceedings of the Association for Computational Linguistics.

16. **Esuli, A., Sebastiani, F.** *entiWordNet: A Publicly Available Lexical Resource for Opinion Mining.* н.м. : Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation., 2006.

17. **Turney, P.D., and Littman, M.L.** *Measuring praise and criticism: Inference of semantic orientation from association.* н.м. : ACM Transactions on Information Systems (TOIS), 21 (4), 315-346, 2003.

18. **Liu, H., Lieberman, H., Selker, T. .** *A Model of Lexical affect Sensing Using Real-World Knowledge.* Miami, US : Proceedings of IUI-03, the 8th international conference on intelligent user interfaces,ACM Press. pp. 125–132, 2003.

19. **Liu B.** *Sentiment Analysis and Subjectivity.* н.м. : Handbook of Natural Language Processing, 2010.

20. **Yi, J., Nasukawa, T., Bunescu, R:, Niblack, W.** *Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques.* Melbourne, US : Proceedings of ICDM-03, the 3ird IEEE International Conference on Data Mining IEEE Computer Society. pp. 427– 434, 2003.

21. **Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.** *Psychological aspects of natural language use: Our words, our selves.* н.м. : Annual Review of Psychology, 54. pp. 547–577, 2003.

22. **Chung, C., Pennebaker, J.** *The psychological function of function words.* New York : Psychology Press, 2007.

23. **Brody S., Diakopoulos N.** *Cooooooooooooooollllllll!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs.* н.м. : Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2011), 2011.

24. **Mairesse F., Walker, M., Mehl, M., Moore, R.** *Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text.* н.м. : Journal of Artificial Intelligence Research (JAIR), 30. pp. 457–500, 2007.

25. **OpenTable International Limited, Inc.** toptable. *www.toptable.co.uk.* [Онлайн] http://www.toptable.co.uk/.

26. **Hatzivassiloglou, V., and McKeown, K. R.** *Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning.* н.м. : Proceedings of 31st Annual Meeting of the Association for Computational Linguistics, pp. 172—182, 1993.

27. **Pang B., Lee L., Vaithyanathan S.** *Thumbs up? Sentiment classification using machine learning techniques.* н.м. : Proceedings of the 2002 Conference on Empirical Methods of Natural Language Processing (EMNLP) , 2002.

28. **Liu, Yang.** *Review Mining from Online Media.* Saarbrücken : VDM Verlag Dr. Müller, 2010.

29. **Christopher M. Bishop.** *Pattern Recognition and Machine Learning.* н.м. : Springer, 2006.

30. **Godbole N., Srinivasaiah M., Skiena S.** *Large-scale Sentiment Analysis for News and Blogs.* н.м. : ICWSM, 2007.

31. **Liu B., Hu M., Cheng J.** *Opinion Observer: Analyzing and comparing opinions on the web.* н.м. : Proceedings of WWW, 2005.

32. **Pang B., Lee L.** *Opinion Mining and Sentiment Analysis.* н.м. : Now Publishers Inc., 2008.

33. **Aue A., Gamon M.** *Customizing sentiment classifiers to new domains: a case study.* н.м. : Proceedings of Recent Advances in Natural Language Processing (RANLP-2005), 2005.

34. **Blitzer J., Dredze M., Pereira F.** *Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification.* н.м. : Proceedings of Annual Meeting of the Association for Computational Linguistics, 2007. ACL-2007.

35. **Pan S.J., Xiaochuan N., Jian-Tao S., Qiang Y., Chen Zh.** *Cross-domain sentiment classification via spectral feature alignement.* н.м. : Preceedings of International Conference on World Wide Web, 2010. WWW-2010.

36. **He Y., Chenghua L., Harith A.** *Atomatically extracting polarity-bearing topics for cross-domain sentiment classification.* н.м. : Proceeedings of the Annual Meeting of the Association for Computational Linguistics, 2011. (ACL-2011).

37. **Gao Sh., Haizhou L.** *A cross-domain adaptation method for sentiment classification using probabilistic latent analysis.* н.м. : Proceedings of the ACM Conference on Information and knowledge management, 2011. CIKM-2011.

38. **Xia R., Zong C.** *A POS-based ensemble model for cross-domain sentiment classification.* н.м. : Proceedings of the 5th International Joint Conference on Natual Language Processing, 2011. IJCNLP-2011.

39. **Ding X., Liu B., Yu P.** *A holistic lexicon-based approach to opinion mining.* н.м. : Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008), 2008.

40. **Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.** *Lexicon-based methods for sentiment analysis.* н.м. : Conputational Linguistics, 2011.

41. **Joachims T.** *Making Large Scale SVM Learning Practical, Advances in Kernel Methods - Support Vector Learning.* н.м. : MIT Press, 1999.

42. **Shawe-Taylor J., Christianini N.** *Support Vector Machines.* н.м. : Cambridge University Press, 2000.

43. **Santorini B.** *Part-of-speech tagging quidelines for the Penn Treebank Project.* н.м. : University of Pennsylvania, 1990.

44. **Montague R.** *Formal Philosophy, selected papers of Richard Montague.* н.м. : Yale Univerisy Press, 1974.

45. **Dowty D., Wall R., Peters S.** *Introduction to Montague Semantics.* н.м. : Springer, 1981.

46. **Gamon M.** *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis.* н.м. : Proceedings of International Conference on Computational Linguistics, 2004.

47. **Pang B., Lee L.** *A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.* н.м. : Proceedings of Meeting of the Association for Computational Linguistics, 2004.

48. **Mullen T., Collier N.** *Sentiment analysis using support vector machines with diverse information sources.* н.м. : Proceedings of EMNLP-2004, 2004.

49. **Kennedy A., inkpen D.** *Sentiment classification of movie reviews using contextual valence shifters.* н.м. : Computational Intelligence, 2006.

50. **Ng V. Dasgupta S., Arifin S.M.** *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews.* н.м. : Proceedings of COLING/ACL 2006, 2006.

51. **Abbasi A., Hsinchun C., Salem A.** *Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums.* н.м. : ACM Tranzations on Information Systems, 2008.

52. **Li T., Zhang Y., Sindhvani V.** *A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge.* н.м. : Proceedings of Annual Meeting of the Associaation for Computational Linguistics (ACL-2009), 2009.

53. **Martineau J., Finin T.** *Delta TFIDF: An improved feature space for sentiment analysis.* н.м. : Proceedings of the Third International AAAI Conference on Weblogs and Social Media, 2009.

54. **Xia R., Zong C.** *Exploring the Use of Word Relation Features for Sentiment Classification.* н.м. : Proceedings of COLING 2010, 2010.

55. **Bespalov D., Bing B., Yanjun Q., Shokoufangeh A.** *Sentiment Classification Based on Supervised Latent n-gram analysis.* н.м. : Proceedings of the ACM conference on Information and knowledge management, 2011.

56. **Kamal A., Abulaish M., Anwar T.** *Mining Feature-Opinion Pairs and Their Reliability Scores from Web Opinion Sources.* н.м. : Second Conference on Web, Intelligence, Mining and Semantics, 2012.

57. **Zhai Y., Liu B.** *Web Data Extraction based on Partial Tree Alignment.* н.м. : roc. of the 14th Intl. World Wide Web Conference (WWW‘05), pp. 76–85, 2005.

58. **Ferrara E., Baumgartner R.** *Combinations of Intelligent Methods and Applications* . н.м. : Springer, 2011.

59. **Baxter I.D., Yahin A., Moura L., Sant‘Anna M., Bier L.** *Clone Detection Using Abstract Syntax Trees.* н.м. : Proc. of ICSM’98, 1998.

60. **Muslea I., Minton S., Knoblock C. A.** *A Hierarchical Approach to Wrapper Induction.* н.м. : Proc. of the Intl. Conf. on Autonomous Agents (AGENTS’99), pp. 190–197, 1999.

61. **Yang W.** *Identifying Syntactic Differences between Two Programs.* . н.м. : Software Practice Experiment, 21(7), pp.739–755, 1991.

62. **Zhai Y., Liu B.** *Extracting Web Data Using Instance-Based Learning.* н.м. : Proc. of 6th Intl. Conf. on Web Information Systems Engineering (WISE‘05), pp. 318–331, 2005.

63. **Freitag D. .** *Information Extraction From HTML: Application of a General Learning Approach.* . н.м. : Proc. of the 15th National Conference on Artificial Intelligence (AAAI-98), 1998.

64. **Crescenzi V., Mecca G. .** *Automatic Information Extraction from Large Websites.* New York : J. ACM, Vol. 51, Nr. 5 ACM (2004) p. 731—779., 2004.

65. **Satpal S., Bhadra S., Sundararajan S., Rastogi R., Sen P.** *Proceedings of the 20th international conference companion on World Wide Web.* New York : ACM, 2011.

66. **Tekli J., Chbeir R., Yetongnon K.** *An overview of XML similarity: Background, current trends and future directions.* н.м. : Computer science review, vol. 3, no. 3, pp.151-173, 2009.

67. **Ferrara E., Fiumara G., Baumgartner R.** *Web Data Extraction, Applications and Techniques: A Survey.* н.м. : Technical Report, 2010.

68. **Liu B.** *Web Data Mining.* Berlin : Springer-Verlag, 2007.

69. **Gusfield D.** *Algorithms for Strings, Trees and Sequences.* н.м. : Cambridge University Press, 1997.

70. **Peng F., McCallum A.** *Accurate Information Extraction from research papers using conditional random fields.* н.м. : HLT-NAACL04, p. 329-336, 2004.

71. **Jindal N., Liu B.** *Opinion spam and analysis.* н.м. : Proceeedings of the Conference on Web Search and Web Data Mining, 2008. WSDM-2008.

72. WordNet - A lexical database for English. [Онлайн] Princeton University. http://wordnet.princeton.edu/.

73. **Liu B., Hu M.,Cheng J.** *Opinion Observer: Analyzing and comparing opinions on the web.* н.м. : Proceedings of WWW, 2005.

74. Weka 3: Data Mining Software in Java. [Онлайн] 1997 r. http://www.cs.waikato.ac.nz/ml/weka/.

75. Morph Adorner Part Of Speech Tagger. [Онлайн] http://morphadorner.northwestern.edu/morphadorner/postagger/.

76. **Deese, J.** *The Associative Structure of some Common English Adjectives.* н.м. : Journal of Verbal Learning and Verbal Behavior 3(5), pp. 347-357, 1964.

77. **C., Levinson S.** *Pragmatics.* н.м. : Cambridge University Press, 1983.

78. **Berlin B. and Kay P.** *Basic color terms: their universality.* Berkeley : Oxford: University of California, 1969.

79. **Lehrer, A.** *Semantic Fields and Lexical Structure.* London : North Holland, 1974.

80. **Lasswell, H.D. and Namenwirth, J.Z.** *The Lasswell Value Dictionary.* 1969 : Yale University Press, New Haven.

81. **Wiebe, J.** *Learning subjective adjectives from corpora.* н.м. : Proceedings of AAAI/IAAI 2000, pages 735—740, 2000.

82. **Hindle, D.** *Noun classification from predicate argument structures.* н.м. : Proceedings of the 28th Annual Meeting of the ACL, pages 268-275. ACL., 1990.

83. **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.* н.м. : Computing Attitude and Affect in Text, Springer, 2006.

84. **Brave, S. and Nass, C.** *Emotion in human computer interaction/.* Mahwah, NJ : Lawrence Erlbaum Associates, Inc., 2002.

85. **Riloff, E, Patwardhan, S., Wiebe, J. .** *Feature Subsumption for Opinion Analysis.* Sydney : Proceedings of EMNLP-06. Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. pp. 440–448., 2006.

10. Декларация за оригиналност

Във връзка с провеждането на процедура за придобиване на образователна и научна степен „Доктор” в Софийски университет „Св. Климент Охридски” и защита на представения дисертационен труд „Извличане и анализ на мнения и чувства от текст от онлайн източници”, декларирам, че резултатите и приносите са оригинални и че не са заимствани от изследвания и публикации, в които нямам участие.