Софийски университет „Св. Кл. Охридски”

Факултет по математика и информатика

Катедра „Софтуерни технологии”

Тодор Васков Цонков

ИЗСЛЕДВАНЕ НА МНЕНИЯ И ЧУВСТВА ОТ ТЕКСТОВЕ ОТ СОЦИАЛНИТЕ МРЕЖИ

Дисертационен труд *за присъждане на образователна и научна степен* ***„доктор”*** *в професионално направление 4.6 „Информатика и компютърни науки“. Научна специалност 01.01.12 „Информатика” (Изкуствен интелект)*

Научен ръководител:

проф. д-р Иван Койчев

СОФИЯ, 2016

Резюме

Настоящата разработка анализира актуалния проблем за извличане на мнения от текстове в социалните мрежи. В началото се прави обзор на най-популярните социални мрежи към момента на писане на дисертацията (2015 г.), поставят се основните проблеми в извличането на мнения от текст и възможните решения.

Основен принос на дисертацията е разглеждането на почти неизследвани дялове на извличането на мнение от текст - засичане на ирония, сарказъм и сатира в текст, за засичане на невярна информация и преувеличаване в текст (хипербола и литота). Приложените алгоритми са изследвани върху текстове от социалните мрежи на английски и български език и е представена тяхна паралелна версия, която може да работи с големи количества от данни. На базата на някои от съществуващите методи се изгражда система за автоматично извличане на мнения от основните социални мрежи (Twitter, Facebook, Google +, LinkedIn, Yahoo), в която по ключова дума се извличат мнения и се оценява автоматично мнението на потребителите относно думата (например продукт, личност или марка).

Освен разработката на приложение за извличане на мнения от социалните мрежи и система за автоматично оценяване на мненията, не на последно място от цитираните мнения може да бъде съставен корпус, който да се използва с изследователска цел.

Благодарности

Бих искал да изкажа специални благодарности на моя научен ръководител проф. д-р Иван Койчев, тъй като прие да съм негов докторант и по време на цялата ми докторантура участва активно в работата ми с ценни съвети, забележки и коментари по статиите ми.

Бих искал също да изкажа благодарност на членовете на катедра „Софтуерни технологии“, които през изминалите четири години ми помогнаха със съвети и напътствия.

Съдържание

1. Увод 8
1.1 Мотивация 8
1.2 Цели на дисертацията 13
1.3 Структура 14
2 Обзор на задачите и подходите за извличане на мнения от текст 15
2.1 Обработка на естествен език 16
2.2 Извличане на мнения от текст в социалните мрежи 18
2.3 Изследване на синтактичната и семантична структура на изречението 19
2.4 Обратно значение при обработката на естествен език 21
2.5 Технически предизвикателства в извличането на мнения от текст 22
2.5.1 Класификация на ориентацията на мнения 23
2.5.2 Интеграция на различните задачи в извличането на мнения от текст 23
2.6 Определяне на анафори в мнения 24
2.7 Разпознаване на наименуван обект 25
2.8 Инструменти за машинно самообучение 28
2.8.1 Система за класификация WEKA 28
2.8.2 Apache Gate 29
2.8.3 Големи данни 30
2.8.4 Наивен Бейсов Класификатор 31
2.8.5 K-Nearest Neighbours 31
2.8.6 Многослойна невронна мрежа - BACKPROPAGATION 31
3 Обобщен алгоритъм за автоматично засичане на текстове с двойно значение 33
3.1 Увод 33
3.2 Анализ на проблемната област 34
3.3 Подходи в областта на засичане на двойно значение от текст. 36
3.4 Използване на евристични правила 38
3.5 Проектиране на специални характеристики 39
3.6 Експерименти за проверка на предсказващата точност 41
3.7 Откриване на ирония в текстове на български език 45

3.7.1 Социалните мрежи и ирония в текстове на български език 46
3.7.2 Описание на алгоритъма за откриване на ирония на български 47
3.7.3 Експерименти за проверка на предсказващата точност 49
3.7.4 Анотация на ирония на други езици 50
3.8 Изводи 51
4 Алгоритми за определяне на невярна информация и преувеличение в текст 53
4.1 Алгоритми за определяне на невярна информация в текст 53
4.1.1 Представяне на проблемната област 53
4.1.2 Подобни разработки 55
4.1.3 Алгоритъм за проверяване верността на мнения 55
4.1.4 Експерименти за проверка на предсказващата точност 57
4.1.5 Извод 57
4.2 Автоматично определяне на надценяване и подценяване в текстове 58
4.2.1 Описание на главата 58
4.2.2 Представяне на предметната област 58
4.2.3 Подобни разработки 60
4.2.4 Алгоритъм за разпознаване на литота 61
4.2.5 Експериментална проверка на предсказващата точност 62
4.2.6 Заключение 63
4.3 Засичане на лъжи в текстови съобщения 63
4.3.1 Сложен отговор 64
4.3.2 Смяна на темата 64
4.3.3 Търсене на подробности 65
4.3.4 Неясни и уклончиви отговори 65
4.3.5 Смяна на граматичните времена 65
4.3.6 Необичаен граматически и стилово правилен отговор 66
5 Класификация на мнения и анализ на слухове от текстове на български език. Алгоритъм за подобрение на списъка от думи и изрази, носещи определено отношение в текст 67
5.1 Увод 67
5.1.1 Извличане на мнения от текст 68
5.1.2 Анализ на множеството от данни 69
5.1.3 Извличане на характеристики 69
5.1.4 Класификацията на ревютата 70

5.2 Подход, базиран на правила, за автоматично засичане на слухове 70
5.3 Експерименти за проверка на предсказващата точност 72
5.3.1 Получени изводи 73
5.4 Алгоритъм за подобрение на списъка от думи и изрази, носещи определено отношение 74
5.4.1 Въведение 74
5.4.2 Подобни разработки 75
5.4.3 Алгоритъм за подобряване на лексикона от думи 76
5.4.4 Експерименти за проверка на предсказващата точност 78
5.4.5 Заключение и обобщение 78
6 Изграждане на система за извличане на потребителски текстове от социалните мрежи и автоматично извличане мнения от текст 80
6.1 Изисквания 80
6.1.1 Дизайн на приложението 80
6.1.2 Изисквания към архитектурата 80
6.1.3 Технически изисквания към имплементацията на приложението 81
6.2 Подобни разработки 81
6.3 Имплементация на модул за извличане на мнения от социалните мрежи 83
6.4 Имплементация на модул за оценяване на мненията на потребителите 85
6.5 Създаване на потребителски интерфейс и статистики 87
7 Заключение 93
7.1 Обобщение 93
7.2 Приноси 93
7.2.1 Научни приноси 94
7.2.2 Научно-приложни приноси 94
7.2.3 Приложни приноси 94
7.3 Публикации 95
7.4 Доклади 95
7.5 Бъдеща работа 95
Използвана Литература 97
Декларация за оригиналност 102

Списък на Фигурите

Фигура 1 Най-популярните социални мрежи (vincos.it) 8
Фигура 2 Синтактично и семантично представяне на изречение 20
Фигура 3 Изход на системата Wikification 26
Фигура 4 Работен плот на Weka 29
Фигура 5 Схема на работа на Apache Hadoop и Yarn 30
Фигура 6 Условна независимост 31
Фигура 7 K-Means класификатор 31
Фигура 8 Многослойна невронна мрежа - BACKPROPAGATION 32
Фигура 10 Паралелен алгоритъм за откриване на ирония 49
Фигура 11 Примерен TUT файл 51
Фигура 12 Примерни мнения 55
Фигура 13 Алгоритъм за подобряване на лексикона от думи 77
Фигура 14 Примерна заявка от клиента 83
Фигура 15 Начална страница на софтуерния прототип 89
Фигура 16 Получени резултати 90
Фигура 17 Статистика и разпределение на резултатите 91
Фигура 18 Статистика на мненията 91

Списък на таблиците

Таблица 1 Статистики на мненията в социалните мрежи (forbes.com) 9
Таблица 2 Оценка на степента на чувство в текст 13
Таблица 3 Превод на най-често използваните термини **Error! Bookmark not defined.**
Таблица 4. Примери какво всъщност се има предвид в британския английски (Duncan Green 2002) 22
Таблица 5 Статистика за направените тестове 25
Таблица 6 Правила, зависими от езика, за определяне на двойно значение в текст 39
Таблица 7 Използване на правила, независими от езика, за откриване на двойно значение в текстове 39
Таблица 8 Описание на характеристиките на мненията 41
Таблица 9 Точност на правилата за определяне на двойно значение 43
Таблица 10 Анализ на резултатите на точността на характеристиките 43
Таблица 11 Точност на характеристиките 44
Таблица 12 Точност на класификаторите 44
Таблица 13 Точност на думите и на комбинациите от думи 45
Таблица 14 Описание на правилата за разпознаване на ирония 47
Таблица 15 Получени резултати при експериментална проверка за определяне на ирония в текст 49
Таблица 16 Описание на правилата 56
Таблица 17 Постигнати резултати 57
Таблица 18 Примери за литота 60
Таблица 19 Описание на правилата, класифицирани като надценяване и подценяване в текст 62
Таблица 21 Постигнати резултати 63
Таблица 22 Резултати за тестовото множество от данни 70
Таблица 23 Описание на правилата за засичане на слухове 72
Таблица 24 Сравнение на алгоритмите за класификация 73
Таблица 25 Получени резултати при подхода с правила 73

1. Увод

1.1 Мотивация

С развитието на социалните мрежи в Интернет потребителите генерират милиарди съобщения всекидневно, като броят им се увеличава постоянно. Потребители от цял свят и на всички езици използват социалните мрежи, като някои от най-популярните в световен мащаб са: Facebook, Twitter, Google +, YouTube Yahoo, Blogger, Digger. В редица страни обаче има локални социални мрежи като например VContacte в Русия, Orcut в Южна Америка, Mxit в Африка, QZone в Китай. По-долу може да се видят най-популярните социални мрежи във всяка държава в света:

Фигура 1 Най-популярните социални мрежи (vincos.it)

Дефиницията на социална мрежа в Интернет е: платформа или уеб сайт,

който се фокусира върху изграждането или отразяването на социалните взаимоотношения между хората. В зависимост от типа те биват:

1. За споделяне на контакти
2. За споделяне на предпочитания (новини, музика, видео, снимки, хипервръзки към интересни статии и др.)
3. За споделяне на авторско съдържание (картинки, музика, видео, статии и др.)
4. За споделяне на лична информация.

Съгласно статистики по времето на писането на текущата глава (10.2014) само за най-популярните - Facebook и Twitter могат да бъдат направени следните наблюдения:

социална мрежа	брой уникални посещения дневно	брой коментари дневно	средно прекарано време	приходи от реклама
Facebook	829 милиона	2.32 милиарда	20 минути	10 bln US $
Twitter	271 милиона	600 милиона	12 минути	1.7 bln US $

Таблица 1 Статистики на мненията в социалните мрежи (forbes.com)

На базата на данните от Таблица 1 може да се направи автоматичен анализ на текстовата информация, която се генерира от потребителите и да се извлече полезната информация, която те генерират.

Текстовата информация може да бъде класифицирана в две големи групи: факти и мнения.

Фактите са обективни изрази за лица, събития и техните свойства.

Мненията обикновено са субективни изрази, които описват настроения, оценки или чувства към лица, събития и техните свойства. Понятието мнение е твърде общо. В настоящата дисертация ще се съсредоточим върху мнения от социалните мрежи, чрез които изследваме положителното или отрицателното настроение на хората спрямо даден обект.

Характерни особености на мненията от социални мрежи са:

1. Относително кратки (В Twitter до 140 символа).
2. Огромен брой всекидневно.
3. Разпределени относително равномерно във времето.
4. На различни езици и от всевъзможни местоположения.

Всекидневно огромен брой мнения на различни езици и от различни местоположения се публикуват в интернет и в социалните мрежи. Извличането на информацията от тях е много важно за бизнеса. Очевидно е, че това не може да стане ефективно от хора, което прави задачата за автоматизирането на този процес изключително важна. Съществуващите подходи все още не са много ефективни особено когато става дума за ирония и сарказъм, което показва необходимостта от разработването на нови методи за решаването на тази задача.

За да разрешим проблемите в извличането на мнения от текст, трябва да дадем подходяща дефиниция на мнение. За нуждите на дисертацията ще използваме следното определение на мнение - това е израз, който се състои от два ключови компонента - тема и чувство на мнението.

Да разгледаме примерното мнение: “Обичам тази компания!”.

„Компанията“ е темата на мнението, а чувството, което е изразено, е “обичам”, което е положително.

В настоящата разработка разглеждаме проблемите за извличане на мнения от текст (opinion mining) и извличане на чувства от текст (sentiment analysis).

В научната литература по този въпрос двата термина обикновено се използват взаимозаменяемо. Съгласно (Peter Turney 2002) извличането на мнения от текст се стреми да определи отношението на автора на мнение към дадена тема или общото отношение или емоция, изразена в мнението.

Поради нарастващата нужда от използване на извличане на мнения от текст за практическите приложения, налице е значително нарастване на изследванията в тези две области в академичните среди, както и в приложенията им в индустрията. Към настоящия момент има поне двадесет компании, които предлагат услуги за извличане на мнения.

България е една от страните, в които се разработват продукти, автоматично оценяващи извличането на мнения от текст, и редица изследователи имат постижения в областта.

Анализът на мненията на потребители за специфични теми може да бъде водещ индикатор за предсказване на резултатите от избори, маркетингови кампании, пускане на продукт на пазара.

Причината за големия интерес на компаниите към извличането на мнения от текст е, че то може да им помогне да намерят отговори на въпроси като:

- Какво мислят потребителите за моята марка?
- Как потребителите приемат маркетинговите кампании относно продукт?
- Има ли начин да предвидим дали определена кампания (примерно реклама) ще е успешна или не, и така да преценим дали да инвестираме в маркетинг?

- Защо потребителите не купуват нашия продукт?

Докато някои от инструментите за автоматично отчитане на мнението на потребителите разчитат на оценка на чувството в думите на базата на тяхната полярност (т.е. позицията им в скалата между позитивното и отрицателното), точното отчитане разчита не само на ключови думи. но и на семантичното значение на думите. Семантиката е наука за отношението между думи, фрази, знаци и това, което всъщност те означават. Заради нея не можем веднага да предположим, че изречение, което съдържа положителна дума, непременно носи позитивно значение или изречение с отрицателна дума е отрицателно. Оказва се, че други думи могат да изместят чувството в думата например:

- **Изместващи чувството в текст**:

  Намирам това приложение по-малко полезно от предишните.

- **Свързващи думи**:

  Това приложение е всичко друго, но не и полезно.

- **Модални глаголи**:

  Предполага се, че това приложение би трябвало да е полезно.

- **Фразеологизми**:

  На седмото небе съм.

Думите “седмото” и “небе” сами по себе си не носят чувство в текста, но фразата “седмото небе” носи позитивно значение. Следователно, не можем да открием точно чувство в текст на базата на анализ дума по дума.

В извличането на мнения от текст може да имаме различни нива на анализа. Съгласно (Peter Turney 2002) се обособяват следните три нива:

- Документно ниво - анализиращо цялостното чувство, изразено в текста. Този подход работи добре, ако предполагаме, че целият документ е на една тема
- Ниво на изреченията - изследва чувството, разгледано във всичките изречения.
- Аспектно ниво - този анализ разглежда всяко мнение, което е в съдържанието на текста. Възможно е да има едновременно позитивни, отрицателни и неутрални аспекти в едно мнение, което затруднява значително неговия анализ.

Близко свързани до чувството в текст са два фактора, които оформят чувството в текста - субективност и емоция. Субективността изразява лични мнения, чувства и вярвания в текста, за разлика от обективното изречение, което изразява фактическа информация.

Пример за обективно изречение е: “Христо Стоичков е носител на златната топка за 1994.”, а пример за субективно: “Смятам, че най-добрият български футболист е Христо Стоичков.”

Субективните изречения не е задължително да изразяват чувство в текста - “Мисля, че Христо Стоичков е бил футболист на Харманли, но не съм сигурен” е пример за такова изречение.

Обективното изречение може да използва чувства:

“Христо Стоичков вкарва често голове за Барселона.“.

Въпреки че това е факт, пак се показва позитивното отношение на автора на изречението към играча.

Въпреки че понякога емоция и чувство са използвани взаимозаменяемо, те не са, особено в извличането на мнения от текст. Чувството е повлияно от емоция, субективно мнение или мисъл. Силата на чувството обикновено се свързва с интензивността на емоцията, почувствана от носителя на мнението. Важно е да се прави разлика между рационалната и емоционалната оценка:

- Рационална оценка: оценката е базирана на факти и логически връзки - “Христо Стоичков е вкарал много голове за Барселона.”
- Емоционална оценка: базирана на емоционални импулси и твърдения, които няма как да бъдат проверени -

“Христо Стоичков е истинска голова машина, невероятен голаджия, без него Барселона нямаше да има тези успехи.”

След изследването на тази разлика можем да направим система, която да оценява чувството в текста на базата на следните цифрови оценки на чувства:

Оценка	**Чувство**
+2	Емоционално позитивно
+1	рационално позитивно
0	неутрално
-1	рационално отрицателно

-2	емоционално отрицателно

Таблица 2 Оценка на степента на чувство в текст

Важно е да се уточни, че чувствата не са само позитивни или само отрицателни. Затова неутралното чувство също е толкова важно, тъй като, в много случаи то показва липсата на всякакво чувство, а не на наличието на обратното**.** Същевременно, особено, когато изследваме съдържанието на документ, можем да имаме противоречиви мнения или множество от мнение с различни чувства. Например:

“Харесва ми играта на футболиста в нападение, но не се старае достатъчно в защита”.

Това изречение в общия случай не е изцяло позитивно, не е и изцяло отрицателно, но не може да бъда класифицирано и като неутрално заради изразените чувства в текста. Ако разбием изречението на мнения, може да видим позитивно чувство в първата част (“Харесва ми играта на футболиста в нападение”) и отрицателно чувство във втората част (“но не се старае достатъчно в защита”).

Когато имаме същия брой положителни и отрицателни мнения в текст, имаме двойствено чувство в текста. Затова понякога се въвеждат чувствата позитивно-неутрално и отрицателно-неутрално. Това е полезно, когато трябва да се покаже различна степен на чувството, главно по сила. Примерно - “Стоичков не е лош играч.” и “Стоичков е най-великият футболист на планетата!” И двете са позитивни, но в различна степен. За съжаление, когато трябва да разглеждаме повече чувства, колкото повече са различните степени, толкова повече се увеличава вероятността за грешка. Това е един от проблемите в извличането на чувства от текст. По-детайлно те ще бъдат разгледани във втората част от дисертацията.

1.2 Цели на дисертацията

Целите на настоящата дисертация са:

- Обзор на съществуващите методи в областта на извличането на мнения и на изградените приложения до момента. Фокусиране върху проблемите в извличането на истинското мнение от текстове с обратно и двойно значение в текстове. Предлагане на възможни подобрения.
- Изследване на особеностите на мненията от текстове от социалните мрежи и предлагане на алгоритъм за подобрението на съществуващи речници за определяне на отношение към темата за конкретна предметна област.
- Разработване на нови начини за засичане на невярна информация, сарказъм, ирония, хипербола, литота и преувеличения в текстове.

Проверка на предсказващата точност на разработените алгоритми с мнения от социалните мрежи.

- Изграждане на приложение за извличането на мнения от текст, което да помогне на потребителите за определяне на чувствата в текста.

Стремежът е алгоритмите, които ще бъдат разработени в дисертацията, да са, доколкото е възможно, езикови и тематично независими. За подобряване на точността на предсказване е необходимо да се знаят предварително езикът и темата.

1.3 Структура

Дисертацията е с обем от 102 страници и е разделена на 7 глави. Приложени са 19 таблици и 26 илюстрации (фигури). Цитирани са 85 източника.

Дисертацията е организирана по следния начин:

- В глава 2. се разглеждат проблемите в извличането на текст. Описват се различните решения до момента и се обосновава защо е избрана предметната област за автоматично засичане на ирония, сарказъм, хипербола и неверни твърдения в текст.
- В глава 3. се описани разработените обобщени методи за автоматично откриване на текстове от социалните мрежи с двузначно значение. Планират се експерименти за оценка на емпирични правила, за избор на разделящи характеристики и за оценка на алгоритми за учене на класификатори. Разгледан е и случаят за откриване на ирония, сарказъм и сатира върху текстове на български език.
- В глава 4. се разглеждат разработените методи за определяне на невярна информация и преувеличение в текст. Представят се проведени експерименти с данни от социалните мрежи, които показват по-добри резултати от съществуващите методи.
- В глава 5. е представен подход за класификация на мнения от текстове на български език, основан на правила за автоматично засичане на слухове от текстове от социалните мрежи на български език и алгоритъм за подобряване на списъка от думи на от предметна област.
- В глава 6. е представена система за извличане на потребителските текстове от социални мрежи и са извлечени различни статистики, за да се покаже ползата за потребителя на приложението. Предлага се търсене по ключова дума и се определя чувството в текста по скалата (-5;+5).
- Глава 7. е заключение, обобщава постигнатите резултати и предлага препоръки за бъдещи изследвания по темата. Разглежда се библиография и е представено заявление за оригиналност.

2 Обзор на задачите и подходите за извличане на мнения от текст

Мотивацията за разглеждането на проблемите в извличане на мнения от текст е да се демонстрира защо задачата е толкова трудна, какви са актуалните тенденции към момента на писане на дисертацията и защо е избрана точно предметната област, разгледана във следващите глави.

Съгласно (Richard Pang 2010) има четири фактора, които ни спират в момента да разчитаме сляпо на автоматичните методи за извличане на чувства от текст:

- Контекст-позитивното и отрицателното чувство могат да има противоположна анотация в зависимост от контекста. Например: “Интернет доставчикът ми е най-добрият, щом става въпрос да ми вземе парите.”
- Двусмислие на изреченията: Изречения с положителна или отрицателна дума, които не е задължително да изразяват някакво чувство.

  Например: “Може ли да препоръчаш някоя добра книга по темата?”. Това изречение не показва никакво чувство, въпреки че използва положителната дума “добра”.

  Друг пример е: “Пътуваме много време и не се вижда края.”. Отново няма отрицателни думи, въпреки че отношението е явно отрицателно.

- Сарказъм, ирония, хипербола, лъжливи изказвания - разгледани са по-подробно в следващите глави. Общото е, че дума, която носи позитивно значение, може всъщност да носи отрицателно и обратното, ако се използва някоя от изброените по-горе литературни форми. Примерно: “Много съм щастлив, че противниците ни ни победиха.”
- Използван език - думите могат да променят значението си и чувството в тях в зависимост от използвания език. Това се вижда в жаргоните, диалектите и вариациите на езика.

  Например: “Направихме луд купон снощи” и “Ти си напълно луд!”, думата “луд” има позитивно и негативно отношение в зависимост от жаргона и контекста.

Да разгледаме проблемите при автоматичното извличане на мнения от текст.

2.1 Обработка на естествен език

Една от най-големите трудности в извличането на мнения от текст е при обработката на естествен език. Да разгледаме следното мнение:

Регистрирах се за пробна версия на продукта. Засега го обожавам. Отчетите са невероятни. Блогът им е информативен. Обаче, колегите ми мислят, че продуктът е много скъп.

Можем да разбием мнението на изречения:

(1) Регистрирах се за пробната версия на продукта.

(2) Засега го обожавам.

(3) Отчетите са невероятни.

(4) Блогът му е информативен.

(5) Обаче, колегите ми мислят, че продуктът е много скъп.

Въпреки че мнението като цяло е положително, има всякакви чувства в текста - положително, отрицателно и неопределено:

(1) е неутрално, тъй като няма никакво чувство в текста.

(2) изразява положително мнение за продукта като цяло.

(3) изразява положително отношение към отчетите на продукта.

(4) изразява положително отношение към блога на продукта.

(5) изразява отрицателно отношение към цената на продукта.

От този пример можем да достигнем до следните изводи:

Зад всяко съдържание има лице, което ние ще наричаме носител на мнението. Може да имаме и повече от едно лице, което е носител на мнението. В изреченията (1)-(4) носителят на мнението е авторът на мнението, а в (5) неговите колеги. За разлика от (1), (2)-(5) не могат да бъдат взети извън контекста, тъй като нямат много смисъл сами по себе си. Да вземем за пример (3) - можем да видим, че мнението е около един аспект от продукта - отчетите. Ако мнението е по-пълно, може да се разгледат и различни характеристики на този аспект (отчетите) - може да се разгледа и достоверността им, пълнотата им, дизайна, представяне и т.н. Следователно темите могат да бъдат разбити на различни аспекти.

Освен това мненията могат да се променят с времето в зависимост от различни фактори, които могат да бъдат субективни (промяна на настроението на потребителя), и обективни (промяна на цената на продукта). Затова е много важно да се включи като размерност и времето в изследването на чувства от текст. Сега имаме пет фактора, които да разглеждаме в извличането на чувства от текст:

- Темата - какви са основните идеи на дискусията;
- Аспектите (подтемите и атрибутите): какво относно тези теми е тема на изречението;
- Чувства - какво е чувството на съдържанието и на мненията, което се съдържа;

- Носител - кой е авторът на мнението. Има ли повече от един автор и ако се различават - по какво?
- Време - кога е публикувано съдържанието?

Изследването на тези 5 фактора прави извличането на чувства от текст изключително трудно.

В общия случай мненията могат да бъдат изразени върху всякаква цел - например продукт, услуга, личност, организация или събитие. Използваме този израз да определим целевата група, която е коментирана. Обектът може да има множество от компоненти (или части) и множество от атрибути, които ще дефинираме като черти на обекта. Например определена марка телефони е обект. Има компоненти (например: батерия и екран) и множество от атрибути (например: качество на звука и големина), които са свойства на обекта. Мнението може да бъде изразено за всяка черта на обекта, както и за самия обект (например: "Харесвам IPhone. Има страхотен дисплей."). Първото изречение изказва мнение за обекта, а второто - за свойство на обекта. Носител на мнението ще наричаме човека или организацията, която изразява мнението. В случай на мнения за продукти обикновено носителят на мнението е авторът на мнението. В новинарските статии например носителят на мнението често е някой друг и трябва да бъде разпознат. Например: "Президентът Обама каза, че ще направи здравна реформа."

Мнението за характеристика f или обект o е положително или отрицателно мнение или оценка на f или o от носител на мнение. Положителното и отрицателното са ориентация на мнението. На основата на тези понятия ще дефинираме модел на обект и модел на текст, в който се съдържа мнение. Това общо се определя като извличане на мнения от текст на основата на свойства.

Обект определяме като крайно множество от свойства, което включва и самия обект като специално свойство. Всяко свойство може да бъде описано с крайно множество от думи или фрази, които са синоними на свойството. Документ, който съдържа мнения, се дефинира като документ d, съдържащ мнения относно множество от обекти), като мненията са от множество от носители на мнения. Мненията за всеки обект са изразени като подмножество на характеристиките на подмножество F_j на черти. Мнението може да бъде един от двата типа:

- Директно мнение: Дефинираме като петорка (обект, характеристика на обекта, ориентация на мнението на обекта, носителя на обекта, време, когато е изразено мнението).

  Ориентацията на мнението за обекта може да бъде отрицателно, неутрално и положително.
- Сравнително мнение - изразява връзка или предпочитание на два или повече обекта на основата на някои от общите им черти.

Например: "Моят телефон е по-добър от твоя. Камерата е с по-високо качество и звукът се чува по-добре. "

В (Murthy Ganapathibhotla, Bing Liu 2011) се разглеждат по-детайлно

проблемите за извличане на мнение на текст от сравнителни мнения, като се разглеждат алгоритми за извличане на сравнителни думи и правила за определянето им. Сравнителните думи биват класифицирани на два основни типа (увеличаващи значението - например повече, по-дълго и намаляващи - по-малко, по-късо). Прави се подробна характеристика на сравнителните прилагателни в английския език и се прави втора класификация на контекстно зависими и контекстно независими.

(Hu and Liu 2009), (Liu et al 2005), (Popescu 2011) разглеждат извличане на мнения от текст от сравнителни изречения на основата на езикови правила.

2.2 Извличане на мнения от текст в социалните мрежи

Извличане на мнения от текстове от социалната мрежа Twitter се оказва по-труден проблем от конвенционалните проблеми при обработка на текстове като например разглеждането на документи. Това отчасти се дължи на малката дължина на туитовете, честото използване на неформални и неправилни думи, и бързата еволюция на езика в Twitter.

Голяма част от разработките в анализа на настроение в Twitter са базирани на избор на подходящи характеристики. (Go, A., Bhayani, R., Huang, L. 2009) разглеждат използване на различни характеристики на *n*-грами във връзка с POS тагове в обучението на смесени класификатори включително Naive Бейс (NB), Max Entropy (MaxEnt) и Support Vector Machines (SVMs). Те откриват, че MaxEnt класификатор, обучен от комбинация от униграми и биграми превъзхожда останалите модели обучени от комбинация от ПОС тагове и униграми с почти 3%.

(Barbosa, L., Feng, J.: 2010) твърдят, че използването на *n*-грами може да попречи на правилната оценка на мнението поради големия брой редки думи в Twitter. Вместо това те предлагат използването на микроблогинг функционалности като повторните туитове, хештагове, допълнения, пунктуация и емотикони. Те установяват, че използването на тези функции за обучаване на SVM-и подобрява точността на извличането на чувства в текстове с 2.2% в сравнение с SVM-и, обучени от само униграми.

Подобна констатация е достигната и от (Kouloumpis, E., Wilson, T., Moore, J.: 2011), които изследват характеристиките на съобщенията в Twitter, включително емотикони, съкращения и присъствието на усилващи думи и фрази като например фразеологизми и символни повторения за класификация на настроенията. Техните резултати показват, че най-добра точност се постига чрез използването на *n*-грами заедно с характеристиките на микроблогинг системи и функции на лексикона, като например думи с таг, който определя тяхната полярност.

(Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. 2011) също изследват POS (части на речта) характеристики както и особеностите на лексикона и микроблогинга. Освен просто комбиниране на различните характеристики, те също така проектират дърво на туитовете, където да комбинират много категории характеристики. Частично ядро за създаване на дървото се използва за пресмятане на близост между две дървета както е описано в (Moschitti, A. 2006). (Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R. 2011) откриват, че най-важните характеристики са тези, в които се комбинират полярността на думи заедно с техните ПОС тагове. Всички останали характеристики играят незначителна роля.

Освен това също така показват, че комбинирането на униграми с най-добрия набор от характеристики превъзхожда дървото на туитовете, което се получава в описанието по-горе и дава около 4% по-добра точност от униграмите на изходно ниво.

(Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J) изработват графика, която има някои от характеристиките на микроблогинга като маркери, специални символи и емотикони като добавят и потребители, туитове, текстообработващи униграми и биграми.

2.3 Изследване на синтактичната и семантична структура на изречението

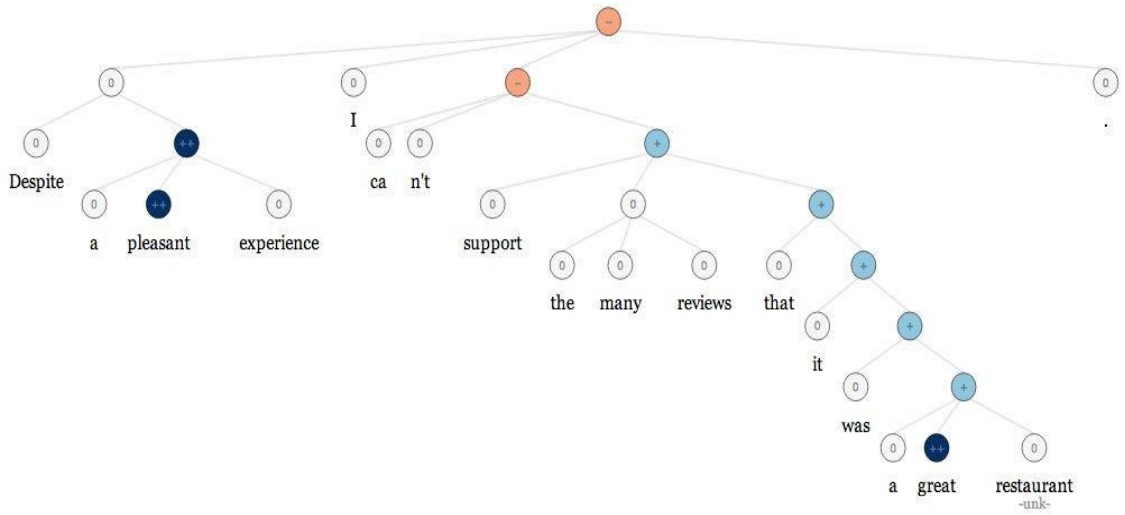
(R Socher, J Pennington, EH Huang, AY Ng, CD Manning 2011) разглеждат пример защо някои малки детайли в изречението променят смисъла. Начинът, по който изразяваме чувство в текст, е сложна смесица от лингвистични структури на нашите мнения и предполагаемото познание на хората, към които се обръщаме в това мнение. За разлика от идентифицирането на теми, извличането на няколко ключови фрази като “добър”, “лош”, “обичам/мразя” ще донесе до около 70% точност и затова имаме нужда от алгоритмите от машинно самообучение.

Да разгледаме класическия пример с изследването на мнения за ресторанти:

1) Това е страхотен ресторант.
2) Това би трябвало да е невероятен ресторант.
3) Ресторантът е невероятен с това, че прави всички бъдещи ястия много по-вкусни.
4) Въпреки приятното изживяване, не мога да подкрепя многото мнения, че е страхотен ресторант.

- Първото изречение е позитивно, но останалите са с отрицателно чувство, въпреки първата част на всяко от изреченията.
- Във второто изречение “би трябвало” показва желания изход, което имплицитно показва истинското чувство в текста.
- Третото изречение показва сарказъм, което много трудно може да бъде засечено от машина и дори човек лесно би могъл да го класифицира като позитивно, като го прочете по-бързо.

Четвъртото изречение е даже още по-сложно, защото започва с позитивно изречение и завършва пак с позитивно изречение, а цялостното отношение е отрицателно. Ключът към разбирането на по-сложни изрази е да се позволи на алгоритмите за извличане на мнения от текст да разбират на по-дълбоко ниво. В примера по-горе е възможно като се разбере синтактичната и семантичната структура на самото изречение. В (R Socher, J Pennington, EH Huang, AY Ng, CD Manning 2011) е илюстрирана фигура, която представя синтактично и семантично четвъртото изречение.



Фигура 2 Синтактично и семантично представяне на изречение

Този пример илюстрира начините, по които хората представят отрицателно чувство в текста. В някои случаи хората предпочитат да бъдат учтиви дори когато са недоволни (дори в Интернет) и в по-широк смисъл отрицателните мнения са по-креативно представени. Най-общо би могло да се каже, че позитивните чувства се представят по еднакъв начин, а отрицателните - по уникален. Това прави извличането на чувството в текст особено трудно за машинното самообучение, тъй като то се учи от примери. Колкото повече потенциални примери има, толкова повече се увеличава вероятността за поява на двусмислено изречение и това води до по-ниска точност. Абстракции като показаната на фигура 2, водят към правилния път за разделяне на изречение, но все още не може да се засекат точно сарказъм,

ирония и други форми на изразяване, където буквалното значение на текста е обратното на това, което всъщност се има предвид.

2.4 Обратно значение при обработката на естествен език

Социално определените отрицания стават особено сложни, когато попадаме на различни езици, тъй като във всеки език се проявява този проблем по различен начин. Например в английския можем да наречем нещо, че е малко, като по този начин да подсилим, че то е голямо. Подобни примери се наблюдават и в българския език. Дори в разновидностите на определен език може да има културни разлики, например в британския английски.

WHAT THE BRITISH C SAY	**WHAT THE BRITISH MEAN**	**WHAT FOREIGNERS UNDERSTAND**
I hear what you say	I disagree and do not want to discuss it further	He accepts my point of view
With the greatest respect	You are an idiot	He is listening to me
That's not bad	That's good	That's poor
That is a very brave proposal	You are insane	He thinks I have courage
Quite good	A bit disappointing	Quite good
I would suggest	Do it or be prepared to justify yourself	Think about the idea, but do what you like
Oh, incidentally/ by the way	The primary purpose of our discussion is	That is not very important
I was a bit disappointed that	I am annoyed that	It doesn't really matter
Very interesting	That is clearly nonsense	They are impressed
I'll bear it in mind	I've forgotten it already	They will probably do it
I'm sure it's my fault	It's your fault	Why do they think it was their fault?
You must come for dinner	It's not an invitation, I'm just being polite	I will get an invitation soon
I almost agree	I don't agree at all	He's not far from agreement
I only have a few minor comments	Please rewrite completely	He has found a few typos
Could we consider some other options	I don't like your idea	They have not yet decided

Таблица 3. Примери какво всъщност се има предвид в Британския английски (Duncan Green 2002)

За първи път тази таблица е представена от (Duncan Green 2002). Въпреки разликите на английския език в САЩ и Англия и другите държави, е относително лесно една машина да научи, че “леко разочарован” всъщност означава “раздразнен”, след като са били представени няколко примера.

Проблемът е да се разбере кога се използва “британски английски”, “американски английски” или друга разновидност. Това е най-важната следваща стъпка в извличането на мнения от текст - автоматично да се знае какъв вид анализ да се приложи на базата на жанр, език и източник на мнението.

Знаейки повече за социалната, културна среда и контекст на мнението, може да се направят по-точни предположения какво точно иска да каже авторът на мнението и по-прецизно да се изчисли предположението за чувство в текст в специфичните видове комуникации.

2.5 Технически предизвикателства в извличането на мнения от текст

Съгласно (Bing Liu 2010) целта на извличането на мнения от текст ни дава добри идеи относно основните задачи и техническите предизвикателства. Никой от проблемите, описани по-горе, не е разрешен. Да разгледаме по-сложен пример, за да илюстрираме горното твърдение:

(1) Вчера си купих Нокия, а приятелката ми Самсунг.

(2) Чухме се, когато се прибрах вкъщи.

(3) Звукът на моят телефон не беше много ясен.

(4) Камерата беше добра.

(5) Приятелката ми ми каза, че ме е чувала добре.

(6) Исках телефон с добро качество на звука.

(7) Затова не бях доволен и върнах телефона обратно.

Обектите, които могат да бъдат открити в мнението, са Нокия и Самсунг. Този проблем е важен, тъй като, ако не знаем обекта на мнението, то не е много полезно. Проблемът е подобен на класическия проблем за разпознаване на наименуван обект. Има разлика, която се изразява в това, че потребителят иска

да намери мнение за обекти, които се състезават помежду си (например продукти). Затова една система за извличане на мнения от текст трябва да раздели съществените обекти от несъществените.

Това е все още нерешен проблем в извличането на мнения от текст, въпреки че има много изследвания. За извличането на свойства и групирането им по синоними имаме висок процент на извлечени фрази, но ниска точност при извличането им. Освен това е много трудно да се извлекат глаголи, които са свойства. Трябва да се ползва списък от синонимен речник, където има еднакви свойства – например звук и аудио. Горният проблем е разгледан в детайли в (Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena 2012).

2.5.1 Класификация на ориентацията на мнения

Тази задача определя дали има чувства в мнение от изречението и ако има, дали е позитивно или отрицателно. Съществуващите методи са базирани на обучение със или без учител. Един от ключовите проблеми е да се определят думи (например добър, лош, страхотен, невероятен), които са ключови за извличането на мнение от текст. Проблемът е, че думите могат да носят съвсем различно чувство в зависимост от контекста, в който са използвани. Има разлика между значенията на думите в различните предметни области. Дори в същата предметна област една и съща дума може да има различно чувство в текста.

Например: “Батерията има дълъг живот”, думата “дълъг” е с позитивно чувство. В друг пример “Камерата се зарежда много дълго време” същата дума носи отрицателно чувство. Има още много проблеми, които трябва да бъдат решени.

2.5.2 Интеграция на различните задачи в извличането на мнения от текст

Интеграцията на различните задачи също е сложна, защото, ако разгледаме петте компонента, описани в глава първа, петорката (тема, аспекти, чувства, носител, време) трябва да бъде запълнена. В доста случаи изречението не споменава експлицитно информацията и тя трябва да бъде разбрана имплицитно например от местоименията, контекста или езикови детайли в текста.

За да се справим с тези проблеми, трябва да приложим техники от обработката на естествени езици в контекста на извличането на мнения от текст, например обработка на текст, двусмислие при използване на думи и определяне на връзките между думите. Например в изречението от фигура 2 е нетривиално за машина да открие какво се има предвид под “моят телефон” и под “нейния телефон”. Изречение (4) не споменава никакъв телефон и затова е

трудно да се определи за кого е камерата. Това е класически проблем в обработката на естествени езици, за който все още няма открито решение.

2.6 Определяне на анафори в мнения

Съгласно тълковния речник анафора (от гръцки ἀναφορά повторение) е стилистична фигура, повторение на звукове, думи или синтактични конструкции в началото на два или повече стиха, строфи или изречения.

Анафората се използва, за да усили значението на конкретната дума или фраза, да се придаде повече емоционалност и благозвучие на лириката. Анафората често се използва и в реторичната реч, както и във фолклорното песенно творчество. В задачите за извличането на мнение от текст може да се анализира съдържанието на повече от едно изречение, когато извличаме мнение.

Например: “Да го кажем директно, не ми е изобщо ясно как този телефон е най-продаваният. Отвратителен е!”.

Ако някой иска да извлече мнението във второто изречение, трябва да се използва алгоритъм за извличане на анафоричното мнение в текста. Извличането на подобни анафорични мнения е разгледано в много статии като отворен проблем - например в (Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006), които предоставят безплатен корпус от коментари за филми (Minqing Hu and Bing Liu. 2004), (Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003). Този проблем не е частен случай, тъй като (Kessler and Nicolov 2009) показват в техните данни, че 14% от целите на мнението са съществителни. В разработката си те разглеждат колекция от мнения за продукти, в които анафорични цели на мнения са анотирани. Тяхната база от данни не е публична.

Въпреки това задачата за определянето на анафора в текст не е решена и оценена до този момент. В (Niklas Jakob, Iryna Gurevych 2012) се разглежда подход за определяне на анафори в текст и се определя дали алгоритмите за определяне на анафора в текст могат да бъдат интегрирани успешно в алгоритъм за извличане на мнения от текст и дали по този начин може да бъде направено подобрение в алгоритмите по извличане на мнения от текст. (Zhuang et al. (2006)) използват подмножество от данни, които те публикуват (1829 документа), по-точно от тях използват 1100 документа, но не споменават кои документи използват от това подмножество. В техните експерименти те използват пълната база данни, която е достъпна, и показват, че приблизително 9.5% от целите на мнението са съществителни. Да разгледаме следната таблица, която описва по-подробно тестовете.

Тип данни	**брой**
Документи	1829
Изречения	24918
Токени	273715
цели + двойки в мненията	5298
цели, които са съществителни	504
съществителни	> 11000

Таблица 4 Статистика за направените тестове

Има три основни видове анафора:

- Съществителна анафора - това е най-често срещаният вид, където предметът на действието е изразен от местоимение. Например: “Джон намери хобито на своя живот”. Тук “своя” се отнася към Джон.
- Определено съществително: В този случай имаме определение от вида на “членувано съществително име”. Ако продължим примера от (1) “Футболът не успя да го заинтригува за дълго.” в случая “Футболът” се отнася за “хобито” от предишното изречение.
- Квалифициращо съществително: Анафората в случая е квалифициращо съществително като например в продължението на примера от (2) “Той започна ново хоби - баскетбола”. Да разгледаме следните ограничения:

    1. Анафората и този, който я ползва, трябва да набележат определени ограничения в атрибутите, за да се създаде сходство. Например пол (мъж, жена), число (единствено, множествено) и спазване на семантичните правила (например диск се копира, компютър се изключва, а не обратното).
    2. Тегловите предпочитания се използват, за да се сложи вероятност, с чиято помощ да определим избора на състезаващите се референции. Те включват близост (на фразата, която изразява анафората в текста), центриране (което определя центъра на обекта, който е под внимание) и семантичната и синтактична роля.

2.7 Разпознаване на наименуван обект

Разпознаването на наименуван обект е задача от извличането на информация от текст, която търси да намери и класифицира елементи в текст от предварително дефинирани категории като имена на личности, организации, места, изразяване на време, количества, парични стойности, проценти и т.н. Повечето проучвания на темата са структурирани като блок от текстове, подобен на този:

Петър си купи 300 акции на Хюлет-Пакард през 2007.

В този пример името на човека се състои от едно съществително, името на компанията и израз за време, който може да бъде засечен и класифициран.

В проблема за разпознаване на наименуван обект, думата наименуван ограничава задачата до тези обекти, за които има едно или много твърди обозначения, както е определено от (Kripke 2009) Например, автомобилната компания е създадена от Henry Ford през 1903 г. е по-нататък Ford или Ford Motor Company. Твърдите обозначения включват собствени имена, както и някои естествени термини като биологични видове и вещества.

Относно пълното разпознаване на име - имената често се разбиват, концептуално и по възможност и в съчетание, като два различни проблема: откриване на имена и класификация на имената от вида на субект който се отнасят (например лице, организация, населено място и други). Първата фаза е обикновено опростена относно проблема на сегментацията: имената са определени да бъдат разделяни на съседни участъци - например "Банк ъф Америка" е име, пренебрегвайки факта, че вътре в това име, подниза "Америка", сам по себе си, също е име. Този сегментационен проблем е подобен на задачата за клъстеризиране.

Въпреки големия брой набор от данни, проблемът за разпознаване на наименуван обект далеч не е разрешен. Основните усилия са насочени към намаляване на работата по анотацията чрез използване на самообучение с учител, добри резултати са постигнати в различни предметни области и при скалиране до обекти с различна големина. В последните години много от проектите са се обърнали към отворен код и мнения от различни потребители, което е обещаващо решение за получаване на висококачествени оценки от хора за самообучението.

Наскоро появила се задача е да се определят и свържат "важни изрази" в текст. Wikipedia може да се разглежда като пример за изключително успешно приложение на горепосочените алгоритми . В примера, който следва, може да се види примерен изход на системата Wikification, която разрешава горепосочения проблем:

<ENTITY url="http://en.wikipedia.org/wiki/Michael_I._Jordan"> Michael Jordan </ENTITY> is a professor at <ENTITY url="http://en.wikipedia.org/wiki/University_of_California,_Berkeley "> Berkeley </ENTITY>

Фигура 3 Изход на системата Wikification

(Lev Ratinov Dan Roth 2012) предлагат подход за разпознаване на наименувани обекти. Да разгледаме следния пример за анотация по-долу:

SOCCER - [PER BLINKER] BAN LIFTED . [LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped aworldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].

Системите за разпознаване на наименуван обект трябва да постигат високи резултати в различни предметни области, тъй като се очаква да може да се прилагат върху многостранно множество от документи - исторически текстове, новинарски статии, патентно право, уеб страници и т.н., затова в изследването са разгледани три корпуса от данни CoNLL03, MUC7 и множество от уеб страници.

В експериментите са разгледани способностите на алгоритмите да се приспособят към нови предметни области. Проверява се предсказващата точност върху базата от данни CoNLL03, като алгоритмите се обучават върху нея и след това се проверява върху други бази, без да се обучава допълнително. Разликите в схемите на анотацията в различни множества от данни създават различни предизвикателства при оценяването.

Да разгледаме базите от данни и методите за оценяване. Корпусът CoNLL03 е подмножество на новинарския корпус Reuters 1996, анотиран с 4 типа: PER, ORG, LOC, MISC. Важно е да се отбележи, че тренировъчното и реалното множество са новини от август 1996, докато тестовото множество съдържа новини от декември 1996. Наименуваните елементи в тренировъчното множество са значително по-различни от онези, които са в реалното множество. В резултат на това тренировъчното множество е значително по-трудно от реалното. След оценката се получава резултат по фрази F1. Множеството от данни MUC7 е подмножество на North American News Text Corpora, което е анотирано от широко множество от типове данни като хора, местоположения, организации, времеви обекти, валутни единици и т.н. Тъй като няма директен мапинг между обектите от MUC7 и CoNLL03 се измерва представянето само типовете LOC, PER и ORG. При оценяването възниква следният проблем - наблюдава се несъответствие между анотациите в MUC7 и CoNLL03.

Например, тъй като множеството MUC07 не съдържа MISC, в изречението "balloon, called the Virgin Global Challenger", фразата "Virgin Global Challenger" би трябвало да бъде класифицирана като MISC съгласно CoNLL03. Обаче в MUC7 се класифицирана като "balloon, called the [ORG Virgin] Global Challenger". Тези и други анотационни несъответствия принуждават да се намалят изискванията за намиране на точна граница на фразите и да се измерва представянето, използвайки ниво на токена F1.

Относно уеб страниците се използва ръчно анотирана колекция от 20 уеб страници, които включват лични, академични и страници на конференции в областта на компютърните науки. Множеството от данни съдържа 783 елемента (96-loc, 223-org, 276-per, 188-misc). Проблемите при оценяването са - наименуваните обекти в уеб страниците са силно многозначни и много различни от наименуваните обекти в тренировъчните данни.

Например в изречението “Hear, O Israel, the Lord our God, the Lord is one.” не може да се определи дали “O Israel” трябва да се класифицира като ORG, LOC или PER. Също така не може да се определи дали “God” and “Lord” е ORG или PER. Тези проблеми все още не са разрешени.

2.8 Инструменти за машинно самообучение

2.8.1 Система за класификация WEKA

Weka е софтуерен продукт, който съдържа колекция от инструменти за визуализация и алгоритми за анализ на данни и модел за предсказване, заедно с графичен потребителски интерфейс за лесен достъп до тези инструменти.

Оригиналът на първата версия на Weka се използва за моделиране на алгоритми, прилагани в други езици за програмиране, като например Makefile-базирана машина за учене чрез експерименти.

Предимствата на Weka включват:

1. Безплатна е под GNU General Public License.
2. Преносимост, тъй като тя е написана на езика за програмиране Java и по този начин работи на почти всяка съвременна компютърна платформа.
3. Изчерпателна колекция от методи за предварителна обработка на данни и методи за моделиране.
4. Лесен за употреба графичен потребителски интерфейс.

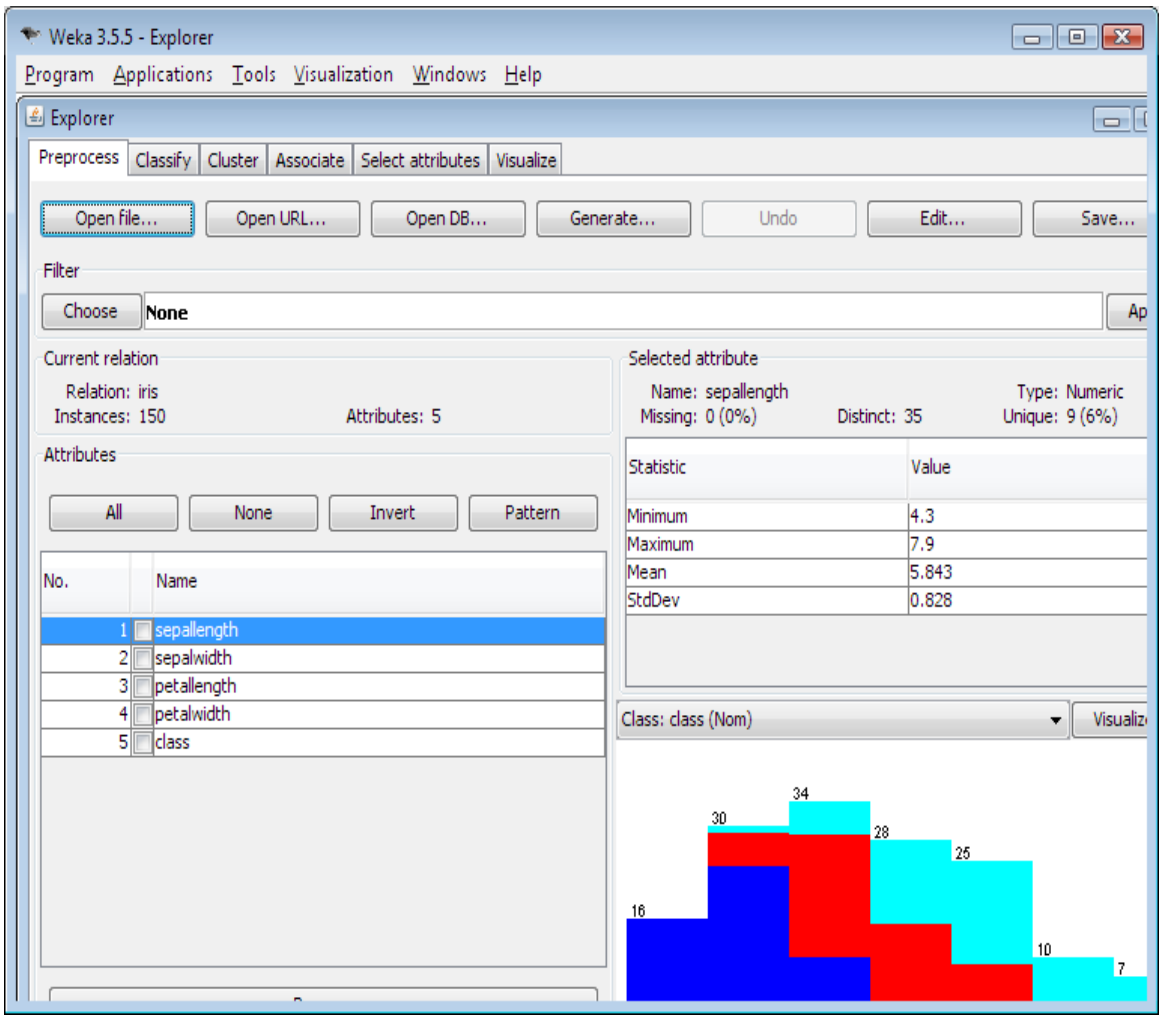
Weka поддържа няколко стандартни задачи за извличане на данни, по-конкретно предварителна обработка на данни, групиране, класификация, регресия, визуализация, както и избор на характеристики. Всички характеристики във WEKA се описват от фиксиран брой атрибути, които могат да бъдат числови или номинални. Weka осигурява достъп до бази от данни чрез SQL Database Connectivity на Java и може да обработва резултата, върнат със заявка от базата данни.

Тя не може да работи в състояние на мултирелационно извличане на данни, но разполага с отделен софтуер за конвертиране на колекции от свързани таблици в базата данни в единна таблица, която е подходяща за

обработка, използвайки системата.

Друга важна област, която в момента не се покрива от включените във Weka алгоритми, е моделирането на последователност от действия.

Изображението по-долу показва основния работен плот на приложението.



Фигура 4 Работен плот на Weka

2.8.2 Apache Gate

Gate е софтуерно приложение за решаване на проблеми в обработката на текст. За мнения на български се използва следната функционалност:

- Reset Document - Позволява документът да бъде възстановен до първоначалното му състояние, като се премахнат всички масиви от анотация и тяхното съдържание
- Tokeniser - разделя текста на много прости символи. BulStem - български Stemmer, който намалява получените думи до тяхната коренова форма.

- Groovy Script - самостоятелно работещ скрипт за bigram TF-IDF извличане на характеристики.

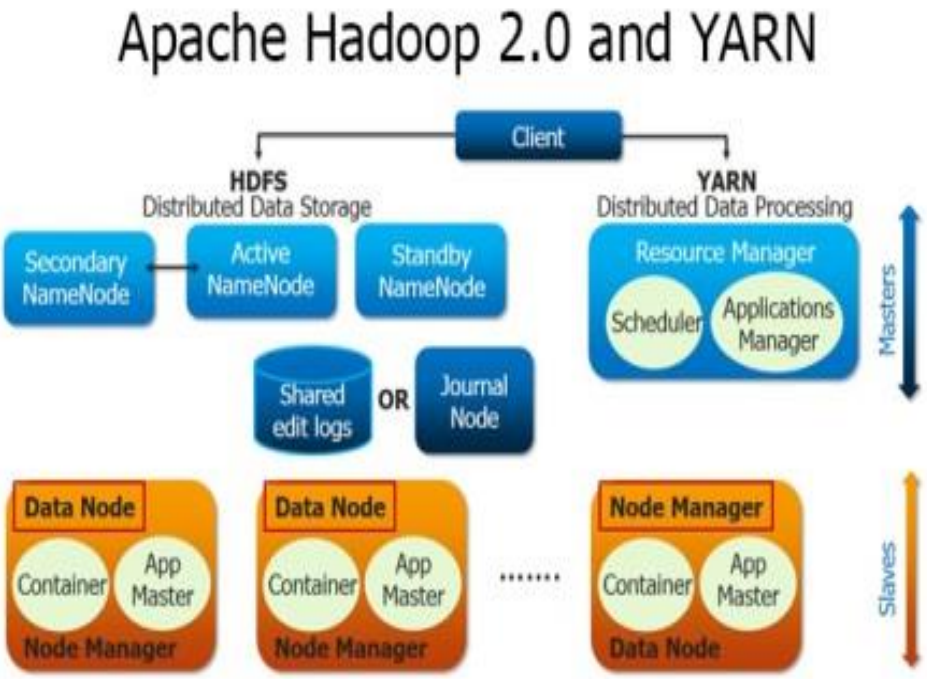
2.8.3 Големи данни

Големите данни (Big Data) предполагат голямо количество данни, които не могат да бъдат обработвани от една машина и нейните ресурси. Не само не може да бъдат съхранявани физически, но няма процесорно време за изпълнение до получаване на учебния процес и прогнози.

Apache Hadoop придобива популярност и се оказва приложимо в области като NLP. Hadoop разпределената файлова система (hdfs) и управлението на клъстера го правят да постига отлични резултати и да участва във важни сфери на бизнеса в световен мащаб.

Освен това, в експерименталните резултати, разгледани в петата глава от дисертацията, е използвана модерна Apache Spark архитектура, която работи на hdfs и използва yarn. Spark поддържа три езика - Scala, Python и Java. За език на изпълнение на заявките е избран Python. Spark изпълнява програмите до 100x по-бързо от Hadoop MapReduce, когато се изпълнява в оперативната памет, или 10 пъти по-бързо, когато се изпълнява на диска.

Примерна фигура, която илюстрира работата на Apache Hadoop, следва по-долу:



Фигура 5 Схема на работа на Apache Hadoop и Yarn

2.8.4 Наивен Бейсов Класификатор

Наивният Бейсов Класификатор разчита на предположението, че всичките атрибути са независими относно класа.

Независимост: Предполагаме, че позицията на думите няма значение.

Условна независимост: Предполагаме, че вероятностите на характеристиките *P(xi|cj)* са независими относно класа с

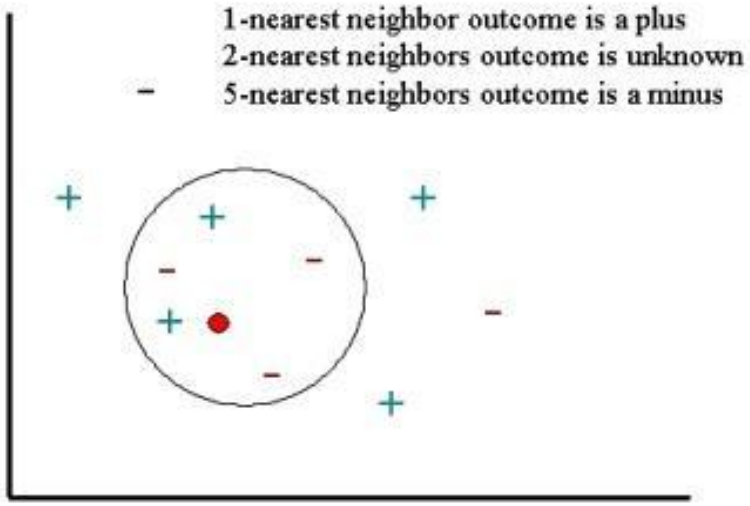
$$P(x_1,x_2,\ldots,x_n) = P(x1|c).P(x_2|c)\ldots.P(x_n|c)$$

Фигура 6 Условна независимост

За да класифицираме нови примери, избираме хипотезата, която е най-вероятна (постериорно).

2.8.5 K-Nearest Neighbours

Входните данни се класифицират по това кой от класовете съдържа най-много от съседите им. Ако k е 1, то примерната инстанция се определя като класа на най-близкия съсед. Примерна схема на работата на алгоритъма може да бъде видяна във фигурата по-долу:

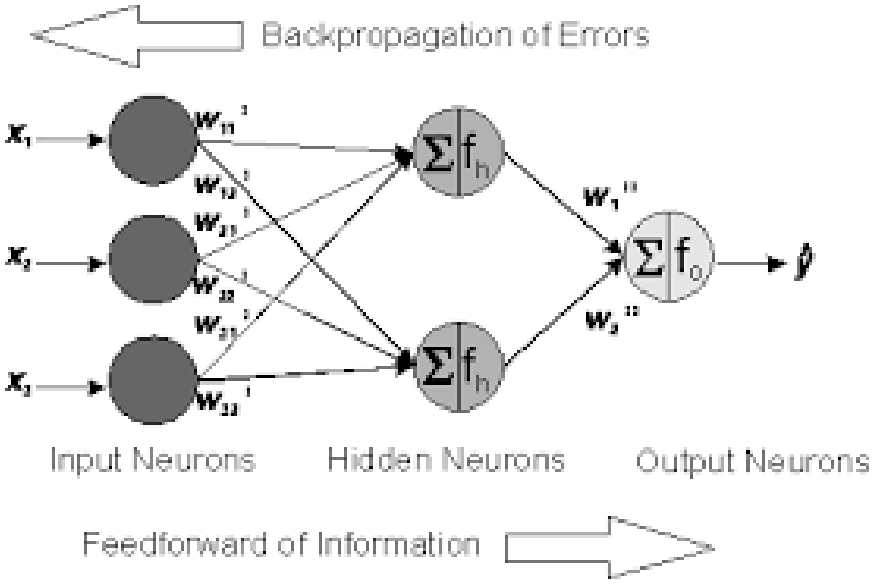


Фигура 7 K-Means Класификатор

2.8.6 Многослойна невронна мрежа - BACKPROPAGATION

BACKPROPAGATION реализира многослойна изкуствена невронна мрежа. Използва се оптимизационен метод - градиалното спускане, за да се

минимизира загубата на точност. Теглата се променят с итеративна процедура, състояща се от 500 епохи.



Фигура 8 Многослойна невронна мрежа - BACKPROPAGATION

3 Обобщен алгоритъм за автоматично засичане на текстове с двойно значение

3.1 Увод

Автоматичното откриване на двойно значение представлява голямо предизвикателство в областта на извличането на мнения от текст, като стандартните алгоритми не допринасят за очакваните резултати. В тази дисертация се представя един подход за автоматично откриване на двойно значение. Извличането на мненията от текстове от социалните мрежи става много широко разпространено в анализа на чувства от текст. Повечето от тези мрежи предоставят публични API-та, които позволяват потоци от мнения по тема да бъдат обработвани в реално време и непрекъснато анализирани за извличане на общественото мнение по определена тема (Albert Bifet and Eibe Frank 2005).

За целите на дисертацията ще дефинираме двойно значение като множеството от следните термини – ирония, сарказъм и сатира.

В частта от тази дисертация няма да правим разлика между ирония, сарказъм и сатира. Причината за това е, че обикновено потребителите в социалните мрежи не правят ясна разлика между трите вида двойно значение. Провеждаме експеримент, за да проверим и потвърдим тази хипотеза. В настоящата глава ще предложим нов метод за определяне на двойно значение в текстове.

Следва кратко обобщено описание на метода:

Предлагаме девет правила, създадени от автора на дисертацията.

Определяме шест характеристики и оценяване точността на предсказването на всяка от тях.

Освен това сравняваме достоверността на три различни класификатора - Naive Bayes, K-Nearest Neighbours и Support Vector Machine. Също така изучаваме точността на предсказване на всички думи и биграми. Проверяваме предсказващата точност на алгоритмите, описани по-горе, върху мнения от социалните мрежи Facebook, Twitter и Google+. Тези мнения са били извлечени чрез HTTP заявки, използващи един от хештаговете #sarcasm, #irony или #satire, като подбираме 3000 мнения за всяка от заявките.

Разработките, представени в настоящата глава от дисертацията са:

- Създаване на евристични правила за двойно значение за откриване и предоставяне на образци за лесна модификация, лесна проверка и лесна замяна на правилата.
- Създавана на корпус от мненията с двойно значение.
- Предложение на набор от характеристики, за предсказване на двойно значение. Провеждане на експеримент върху корпус от мнения от социалните мрежи за оценка на точността на предсказване на тези

характеристики.

- Провеждане на експерименти с подходи на машинното самообучение за автоматично определяне на двойно значение в текстовете от социалните мрежи.
- Осигуряване на експериментални доказателства, че потребителите в социалните мрежи не правят разлика между ирония, сарказъм и сатира.
- Експериментална проверка на точността на думи и биграми за определяне на двойно значение.
- Създаване и проверка на предсказващата точност на алгоритъм за ирония върху български текстове.

3.2 Анализ на проблемната област

Ако се прочетат мнения на клиенти на Amazon, може да бъде трудно да се определи кой всъщност е харесал продукта и кой е искал просто да се забавлява като пише коментар.

Що се отнася до Facebook, дори и да се прилага настоящият стандарт за анализ на настроения към Facebook харесванията, може да се пропусне нюанс какво точно е искал да каже потребителят.

(Matthew Russell 2013) коментира:

"Що се отнася до сарказъм, като всеки друг вид обработка на естествен език (NLP), контекстът има значение. Анализът на данните за естествен език е, по мое мнение, проблемът за следващите 2-3 десетилетия.“

Това е един невероятно труден въпрос и сарказмът и другите видове ироничен език по своята същност са проблемни за откриване от машини, когато се погледнат изолирано. Тогава е наложително да има достатъчно сложен и строг подход, в който съответният контекст да може да бъде взет под внимание. "

В статия на Personal Democracy в блога TechPresident е описано сътрудничеството между Facebook и организацията Politico, които се опитват да измерят репутациите на кандидатите в първичните избори за президент на САЩ и там се прави следната констатация относно Фейсбук и ирония:

Facebook констатират, че инструментът, който използват - лингвистични заявки и брой думи, е добре валидиран софтуерен инструмент, използван често в социалните и психологически изследвания за идентифициране на положителни и отрицателни емоции в текст.

Други изследователи са силно скептични за вземане на окончателни твърдения за това, което може да се появи в изследването на мнения от текст.

Марк Смит, директор на фондацията Social Media Research, казва: "Аз споделям скептицизма относно анализа на настроения в текст. Иронията е доста труден проблем." (Marc A. Smith 2011).

В някои скорошни мнения за ползване на продуктовата страница на Amazon, се виждат доказателства, че често се влага ирония и е възможно нарочно да се пишат положителни мнения с цел създаване на обратното мнение за продукта или антиреклама или просто потребителите да пишат мнения за забавление.

Не е трудно да се предположи, че много хора биха експлоатирали тази слабост в рекламния модел на Фейсбук и да автоматизират редица реклами или антиреклами. Трудно да се смята, че много рекламодатели ще искат да инвестират в толкова рисково начинание. Този пример води до следния по-общ извод. Емоционалната сложност на социалния Facebook граф прави ефективното търсене на клиенти от маркетинг специалисти изключително трудно.

Не само, че някои съобщения са двусмислени, но и както се оказва, приятелите във Facebook не са еднакви върхове в графа от приятелства.

Затова е трудно да се прецени какво тегло да се даде за всяка препоръка или мнение. Би могло мнението да е иронично или не, може “приятелят” във Фейсбук всъщност изобщо да не се познава с лицето. Рекламирането във Фейсбук може да е ефективно за стартъпи в областта на уеба, чиято цел е да популяризират сайта си и да добавят потребители, по възможност платени. Но за големите компании, които се опитват да накарат потребителите си да извършат определени действия, като например да купят кола от някое шоу, тези объркващи послания може да създадат твърде много двусмислие, което да не генерира ясен изход.

Ако прочетете мнения на клиенти на Amazon, може да бъде трудно да се определи кой всъщност е харесал продукта и кой е искал просто да се забавлява в писане на коментара.

Дори и да се прилага настоящият стандарт за анализ на настроения към езика на Facebook харесванията, може да се пропусне нюанс. Езикът на горното мнение е безспорно положителен, но освен това е и също ироничен.

Тогава е наложително да има достатъчно сложен и строг подход, в който съответният контекст да може да бъде вземан под внимание. "

3.3 Подходи в областта на засичане на двойно значение от текст.

Въпреки че автоматичното засичане на двойното значение има голям брой научни изследвания в областта на психологията на задачата за автоматично откриване в социалните мрежи е получила значително внимание през последните години. Задачата е много подобна на традиционната задача за анализ на настроения в ОЕЕ. За разлика от нея Хасан и колектив (Hassan et al, 2009) използват контролиран Марков модел, части на речта, както и зависимостта модели за идентифициране на полярност в теми, публикувани във форума за дискусии Usenet (Hassan et al, 2009). Барбери (Barbieri 2012) изследва определения на ирония и сарказъм.

Вербалната ирония е била дефинирана по различни начини през годините и все още не съществува единно определение. Стандартното определение за ирония е: "да се казва точно обратното на това, което искаш да кажеш" (Quintilien and Harold Edgeworth Butler. 1953), т.е. когато разликата между буквални и истински значения е много ясна. Те вярват, че иронията е риторична фигура, която нарушава максимата за качество: "Не казвайте това, което вярвате, че е невярно".

Има и няколко изчислителни модела, които откриват сарказъм на Twitter и Amazon (Davidov, D.; Tsur, O.; and Rappoport, A. 2010), (Roberto Gonzalez-Ibanez, Smaranda Muresan, and Nina Wacholder 2010), (Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013), но дори и да може да се спори, че сарказъм и ирония са едни и същи езикови явления, иронията е по-подобна на подигравката.

Други определят резултати за настроения в новини, истории и публикации в блогове на базата на алгоритмично генерирани лексикони на положителни и отрицателни думи. (Godbole et al 2007) показват експериментално, че въпреки честите случаи на нередовни модели на реч в туитове, Twitter може да се окаже полезен корпус за анализ на настроения. Разнообразието на Twitter потребители прави този корпус особено ценен, например за проследяване на заблуждаващи политически съобщения.

Заедно с многото предимства на Twitter като корпус за анализ на настроения, той има уникално предизвикателство, тъй като постовете са по-малко от 140 символа. Това означава, че те често могат да съдържат необичайни граматика и неконвенционални думи и символи.

Въпреки че са изпробвани различни подходи за формулиране на ирония като изчислителна единица, има малко резултати на автоматичното откриване на ирония в текст. Подобен подход като представения в тази дисертация е

направен от (Paula Carvahlo, Luis Sarmento, Mario J. Silva, Eugenio De Oliveira), които откриват ирония чрез осем шаблона, специфични за португалския език. (Aline A. Vanin, Larissa A. de Freitas, Renata Viera, Marco N. Bochernitsan**)** също така представят подход, в който се използват петнадесет шаблона и се класифицират в осем категории, както следва: (C1) списъци, (C2) точните изрази, (C3) част на речта, (C4) част от реч + точно изразяване, (C5) част на речта + списъци, (C6) част на речта + назовани лица, (C7) демонстративни местоимения + назовани лица, (C8) символи и проверка на предсказващата точност върху 2780 туитове (55663 думи) в бразилски португалски. Резултатите, които получили, показват 92,45% успеваемост върху списък от емоции и 71,77% успеваемост върху препинателни знаци (? , ! , . , и т.н.).

Да разгледаме други подходи.

Друг подобен подход е бил изследван от (Andrea Gianti, Christina Bosco, Viviana Patti, Andrea Bolioli, Luigi Di Caro) в статията им, като те добавят ново измерение от настроения, което отговаря специално за иронията. При този подход задачата се извежда до задача за класификация и се прилагат алгоритми от машинното самообучение.

(Tsur et all, 2010) предлагат подход, който упражнява смесено машинно самообучение за автоматично признаване на сарказъм в Amazon коментари, използвайки някои функции, които са специфични за Amazon. Техният метод използва два модула: за събиране чрез обучение на саркастични модели и класификатор. След това този метод се прилага към туитове от (Davidov, D.; Tsur, O.; and Rappoport, A. 2010), които постигат още по-добри резултати.

(Roberto Gonzalez-Ibanez, Smaranda Muresan, Nina Wacholder. 2011) изработват корпус от саркастични туитове и го използват за сравняване на съдебни решения, направени от хора и алгоритми за машинно самообучение, заключавайки, че никой от тях не се представя добре.

Reyes и др. (2013 г.), определят комплексен модел за идентифициране на сарказъм, който отива далеч зад повърхността на текста и се вземат предвид функции на четири нива:

1. автор на мнението
2. степен на неочакваност,
3. стил на изказа
4. емоционални сценарии.

Те показват, че тези характеристики не помагат на идентификацията, ако са разгледани в изолация от контекста. Въпреки това, те могат да бъдат използвани за точно предсказване, ако са коректни.

Barbieri и Saggion (2014 г.), фокусират техния подход за употребата на лексикални и семантични характеристики, като например честотата на думите в различни референтни корпуси, дължината на думите, както и броя на свързаните синонимни в WordNet (Miller и Fellbaum, 1998).

(Buschmeier, H., & Kopp, S. 2014) оценяват въздействието на функции, използвани в предишни изследвания и осигуряват важна основа за откриване ирония на английски език. Twitter потребителите често използват повтарящи се гласни (например "looooove") или капитализация (например "LOVE"), за да се подчертаят някои настроения или емоции.

Нормализирането се състои от премахване на повтарящи се гласни (например от "looooove" до "love"), а капитализацията (например от "LOVE" на "love"). Нормализираните форми могат да помогнат за подобряване на точността на анализа.

3.4 Използване на евристични правила

Използваме девет евристични правила, за да открием дали мнението е наистина сатирично, иронично или саркастично. Проверяваме дали конкретни предположения за него могат да бъдат успешни при определяне дали то наистина може да се определи като ирония, сатира или сарказъм. Таблицата по-долу показва подробно обяснение на тези правила. Използването на двойно значение (съдържащо един от хештаговете #irony, #sarcasm и #satire) се среща рядко при други мнения и когато нямаме някакви специални познания по представената тема, задачата става невъзможна да се разреши. Деветте правила са избрани от група от 20 правила, основани на най-добрата точност по отношение на тест набор от 200 произволни мнения. Използваме четири основни критерии, за да се опише правило. Описание на правилата следва в таблицата по-долу:

Описание	**Шаблон**
1. Съдържа поне три позитивни или поне три отрицателни думи (само позитивни ИЛИ само отрицателни, а не комбинация от двете)	[+дума] И [+дума] И [+дума] ИЛИ [-дума] И[-дума] И [-дума]
2. Съдържа положителна и отрицателна дума	[+дума] И [-дума]
3. Автоматично засичане на двойно значение в текст чрез използване на думи с противоположно значение от списък.	[дума1] and [дума2]
4. Съдържа положителна дума и отрицателна емоция.	[+дума] and [-емоция]
5 Разликата между често срещаните и рядко срещаните думи от списък	Чести думи - редки думи

Таблица 5 Правила, зависими от езика, за определяне на двойно значение в текст

Описание	**Шаблон**
6. Съдържа поне два от следните символа: [!], [?] или[...]	([!] or[?] or [...]) AND ([!] or [?] or [...])
7. Използване на три последователни прилагателни	прилагателно и прилагателно и прилагателно
8. Използване на дума в главни букви	[ДУМА]
9. Съдържа “” или ‘’	[‘’] OR [“”]

Таблица 6 Използване на правила, независими от езика за откриване на двойно значение в текстове

3.5 Проектиране на специални характеристики

Използваме пет допълнителни характеристики, насочени към подобряване на точността на подхода, основан на правила. (Francesco Barbieri and Horacio Saggion, 2014) също използват характеристики за извличане на чувство от текст. Използват се части от техните характеристики, заедно с авторски изследвания по темата.

Следва едно по-обстойно описание и обяснение защо тези характеристики са били избрани:

1. Разликата между думи, съдържащи положителни настроения и отрицателни настроения е показателно за двойно значение (Veale, T.,

and Hao, Y. 2010b) и от изследванията, направени от автора на тази дисертация. Изследваме броя на положителните и отрицателните думи. Те са избрани от списъците http://www.enchantedlearning.com/wordlist/opposites.shtml и са специфични за езика, на който се прави изследването.
Формално описана формулата е следната:

$$feature1 = count_{in_{opinion(positive)}} - count_{in_{opinion(negative)}}$$

2. Броят на препинателни знаци, емотикони и връзки е важен, защото както е показано в (Vanin,A.; de Freitas L.; Viera R.; Bochernistan M.2014), средният брой емотикони в иронични туитове е по-висок, отколкото в такива, които не са иронични. В допълнение, повече от един препинателен знак може да означава двойно значение, което е и целта на тази глава. Мнения с хиперлинкове обикновено изразяват мнението на автора за хиперлинка, което често може да бъде иронично, сатирично или саркастично. От 100 случайно избрани мнения за социалните мрежи, които са иронични, сатирични или саркастични, 65 съдържат хиперлинкове. Това може да се дължи на естеството на становищата на социалните мрежи, където се споделят много хиперлинкове.
Формулата, която го описва, е следната:

$$feature2 = number(links) + number(emoticons) + number(punctuation\ signs)$$

3. Броят прилагателни обикновено показва специална връзка към темата и по-често, отколкото не, изразява мнение, а не просто посочва факти. Дължината на мнението е важно, защото, както е показано в (Davidov, D.; Tsur, O.; and Rappoport, A. 2010), колкото по-дълго е мнението, толкова по-вероятно е то да е иронично или саркастично.
Използваме следната формула сумата от дължината на мнението + брой прилагателни умножено по 10.
Формално описана:

$$feature3 = length(opinion) + (number\ of\ adjectives) * 10$$

4. Определяме функцията интензитет, която се използва в (Dipankar D. and Sivaji B. 2012). Дефинираме интензитет като: интензитета на наречия и прилагателни и се изчислява резултатът на интензитетът като сбор от интензитетът както на наречията, така и на прилагателни.

$$intensity = \{-5, \dots, +5\}\ from\ a\ list\ of\ words$$
$$feature4 = \sum intensity(adjectives) + \sum intensity(adverbs)$$

5. Разликата между общи и редки думи измерва колко уникално мнението е в сравнение с другите. Дефинираме общи думи като един от най-често срещаните 2000 думи, които вземаме от:

http://www.talkenglish.com/Vocabulary/Top-2000-Vocabulary.aspx. Другите думи се считат редки. Разликата се изчислява като разлика между общите думи и редките думите на разделени на всички думи в изречението.

$feature5 = abs(\sum common\ words - \sum rare\ words)$

Номер	Описание на характеристиката
1.	Брой думи, съдържащи положително и отрицателно чувство в текста.
2.	Брой емоции, пунктуационни знаци и линкове
3.	Брой на прилагателните и дължина на мнението
4.	Интензитет на прилагателни и наречия
5.	Честота (разлика между общи и редки думи).

Таблица 7 Описание на характеристиките на мненията

Предсказващата точност на характеристиките е описана в частта експерименти по-долу.

3.6 Експерименти за проверка на предсказващата точност

В настоящата част от главата правим проверка на предсказващата точност за двойно значение върху 3000 становища от Twitter и Facebook, съдържащи един от хештаговете #sarcasm, #satire или #irony. 80% от мненията са от Twitter и 20% от Facebook.

Всички мнения са на английски език. Ако едно мнение изпълнява повече от едно правило, то ще бъде преброено във всички от правилата, които изпълнява. Филтрираме всички изречения, които не са на английски език. Също така разглеждаме всяко мнение, което съвпада с тези тагове, дали наистина е двойно значение.

Проверяваме предсказващата точност на всяко мнение върху всяко от тези двойни значения - ирония, сарказъм и сатира като всяко има по 1000 мнения. За експеримента, описан в тази глава от дисертацията, използваме социалните мрежи Фейсбук и Туитър и използваме софтуерно приложение, разработено на езика Java за извличане на постовете (чрез HTTP заявки и отговори). Всеки пост съвпада само с един от хештаговете "#sarcasm", #irony или #satire и ръчно се проверява от автора на главата.

За целите на настоящата разработка използваме мнения, които включват само един от хештаговете #irony, #sarcasm или #satire.

Примерни мнения, извлечени от изследванията:

1. Really delighted that I took the fantasy league captaincy off Diego Costa today #sarcasm

2. I know you "protestors" won't believe this, but the cops apparently just shot a white guy. #sarcasm

3. Because it is the gun that kills, not the humans holding it. #sarcasm

4. I'm SO glad racism is dead in this country. Whew! *wiping brow*#SARCASM

5. BLOODY IMMIGRANTS DRAINING THE STATE..... #irony

6. Black Donor's Sperm Mistakenly Sent To Neo-Nazi Couple"

Well. That's some serious #Irony. #KarmaIsWonderful

7. This just behooves me to share. The irony that the conference is happening to fight islamophobia while islamophobes rally outside. Note the facility is used for "several cultural and religious-based groups and events" according to the Garland ISD spokesman. #misinformation #iamamuslim #irony #equality?

8. The Borowitz Report: Queen Elizabeth II took to the airwaves to inform the people of Scotland that she "graciously and wholeheartedly" accepted their apology. #satire

https://www.facebook.com/hashtag/satire?source=feed_text&story_id=10152371028798869

Дизайнът на експеримента може да се опише по следния начин формално: За всеки пост от социалните мрежи направете следното:

1. Създаване на нишка, която обработва всяко правило, както е описано по-горе:
2. За всяко правило, описано като регулярен израз от схемата по-горе
3. Проверете дали мнението изпълнява правилото.
4. Ако поста изпълнява регулярния израз на текущото правило
5. Добавяне на мнение към списъка на становищата на сегашното правило
   За всеки списък на становища направете:
6. Ако правилото е правилно класифицирано увеличи брояча на правилно класифицирани правила .

7. За всяко правило да се изчисли степента на успех като: Правилно класифициран мнения, изпълняващи това правило, разделено на всички мнения, класифицирани като изпълняващи това правило.

Правило #	Ирония	Сарказъм	Сатира
1	78%	73%	76%
2	71%	69%	75%
3	80%	77%	81%
4	64%	66%	65%
5	84%	80%	83%
6	63%	67%	65%
7	58%	59%	62%
8	59%	55%	55%
9	61%	60%	58%
1-9	77.25%	75.75%	77.5%

Таблица 8 Точност на правилата за определяне на двойно значение

Резултатите по-горе показват защо задачата за автоматично откриване на двойно значение в текстовете е толкова трудна. Без да се знаят много специфични детайли, обикновено е невъзможно да се създадат точни правила, когато няма конкретна информация за темата и авторите на мненията.

От резултатите в последния ред на горната таблица можем да заключим, че няма разлика между ирония, сатира и сарказъм в текстове от социалните мрежи, тъй като различията в точността са малки.
В долната таблица показваме точността по отношение на правилно прогнозиране на двойно значение във всяка характеристика, описана по-горе.

Характеристика	Точност
1	74%
2	70%
3	63%
4	59%
5	67%

Таблица 9 Анализ на резултатите на точността на характеристиките

Точността при определяне на двойно значение на всяко от правилата по-горе показва, че чрез използването на думи самостоятелно е трудно да се постигне висока точност.

Сравняваме с подхода, базиран на правила и алгоритъма на използване на характеристики с други подходи. Обучаваме и проверяваме предсказващата точност на три различни класификатори - Наивен Бейсов класификатор, К-най-близки съседи и Support Vector Machine.

Избираме 4000 мнения от Facebook и Twitter социални мрежи и да проверим каква е прецизността на всеки класификатор, както е описано по-долу.

От тези мнения 2,000 съдържат тагове #irony , #satire или #irony, а други 2000 са избрани на случаен принцип от социалните мрежи.
В долната таблица изследваме предсказващата точност на характеристиките. Изследваме всички комбинации от две характеристики, тъй като, ако използваме три или повече характеристики, имаме твърде малко мнения, които да съответстват на тях, за да имаме смислена статистика.

Характеристики	Точност
1-2.	82%
1-4	77%
2-4	73%
1-5	77%
1-3	76%

Таблица 10 Точност на характеристиките

Точността е от гледна точка на мнения, които могат да бъдат класифицирани като иронично, саркастично или сатирично. Използваме пет характеристики, които са описани в частта на характеристиките като измерение и по този начин се създават пет тримерни пространства.

Разстоянието, което се използва, е нормализирано манхатъново разстояние, както и класификатори, като се използват реализираните в продукта с отворен код WEKA.

Резултатите са представени в Таблицата по-долу.

Класификатор	Точност
Naive Bayes	79%
K-NN	78%
Support Vector Machine	81%

Таблица 11 Точност на класификаторите

Освен тези характеристики, които са проверени върху всички думи и двойки от думи, за да се определи каква е точността на всяка, избираме думите и двойките думи и проверяваме каква е тяхната точност върху двойно

значение, като са включени всички думи и двойки от думи, които са били използвани в тестовото множество от мнения.

За целите на това изследване няма да се прави разлика между ирония, сарказъм и сатира. Проверено е върху двойно значение и резултатите могат да се видят в долната таблица и е измервана точността върху два класа мнения - двойно значение и пряк смисъл (не двойно значение). Проверяваме предсказващата точност върху всичките думи, които могат да бъдат открити най-малко пет пъти в мненията.

Това се прави, за да има достатъчен брой повторения на всяка дума/фраза, за да има смислена статистика.

Дума или двойка думи	Точност
She/He/It	52%
He/she/it has	57%
It is	64%
Finest	66%
Wish	60%
Smart	77%
Congratulations	80%

Таблица 12 Точност на думите и на комбинациите от думи

Резултатите от таблицата показват, че използването само на термини не е добър начин за постигане на висока точност при предсказване на двойно значение.

Въпреки това, думи като "поздравления" и "интелигентни" имат много добра предсказваща точност за двойно значение.

Тези резултати могат допълнително да бъдат използвани за оценяване на теглови характеристики.

3.7 Откриване на ирония в текстове на български език

Настоящата част от главата представя нов метод за откриване на ирония от социалните мрежи Facebook и Twitter в текстове на български език. Основната идея на алгоритъма е да прилага емпирични правила, специфични за българския език, които да откриват ирония в текста. Предложени са пет такива правила. Представени са и резултати от проведени експерименти с реални данни, които показват, че представеният алгоритъм постига относително висока точност на разпознаване на иронията. Резултатите показват, че висока точност може да се постигне, когато имаме специфични познания за приложната област и знаем предварително в какви случаи може да

се очаква ирония.

3.7.1 Социалните мрежи и ирония в текстове на български език

С продължаващото нарастване на социалните мрежи потребителите генерират милиарди коментари всеки ден и необходимостта от анализ на данните става все по-наложителна. Извличането на мнение от текст се опитва да реши задачата за класифициране на полярността на даден текст в документ и да определи дали той има положително, отрицателно или неутрално значение. Откриването на иронични изявления представлява голямо предизвикателство в областта на извличането на мнение от текст. Стандартните алгоритми за анализ на мнения не се справят добре с този случай.

В тази част от дисертацията ние представяме алгоритъм за откриване на ирония, който постига добри резултати при текстове на български език. За да се определи дали едно мнение е иронично, трябва да има ясна дефиниция за ирония. За целите ние определяме ирония като двойно изказване, в което говорещият предполага точно обратното на това, което всъщност се казва или изразява настроения в пряка опозиция на това, което всъщност се предполага. Няма да правим никаква разлика между ирония, сарказъм и сатира за целите на тази част от дисертацията, както показахме по-горе в главата. Интересен факт за иронията е, че потребителят обикновено иска иронията да бъде открита и разделя читателите на текста на два вида – такива, които разбират иронията, и такива, които не я разбират.

Ще използваме социалните мрежи Facebook и Twitter и Java приложение за извличане на мненията (чрез HTTP заявки и отговори). Всяко мнение е с хещтаг "#ДАНСWithMe " и се проверява за всяко от правилата и в случай, че има съвпадение постът се разглежда от автора на дисертацията и оценява дали наистина има ирония. Резултатите могат да се видят в част експерименти по-долу. Примерни мнения, извлечени от изследването:

1. “Подай #оставка, спаси домат! #ДАНСWithMe”.

2. „Номинирам този клип в категорията: „Жалка трагикомедия” – това Оскар, „Златен глобус”, „Златни малинки” и други плодови отличия – в кърпа са вързани: #ignorevolen #ДАНСWithMe

3. Фрустрирам Ви приятели #ДАНСWithMe.

4. К`во правиш, бе ?

Гледам Али Макбийл.

А, че що я гледаш?

Уча се от нея. Днес ме назначиха за зам.-министър на правосъдието. #ДАНСwithMe

5. С риска да ми се разсърдят няколко човека, бих казал, че е чудесен ден за барбекю :D Peace and #ДАНСwithme;(

6. Държавата - губещ хотелиер из "любимите вестници" #ДАНСwithmehttps://www.facebook.com/hashtag/дансwithme

Източник: OFFNews

3.7.2 Описание на алгоритъма за откриване на ирония на български

Използваме пет различни правила, създадени от нас, като се проверява дали всяко мнение ги изпълнява. В случай че дадено правило се изпълнява, проверяваме дали мнението наистина е иронично. В долната таблица може да се намери по-подробно обяснение на правилата. В конкретната тема има повече иронични мнения, отколкото обикновено, заради преобладаването на становищата (върху правителството).

Правила за разпознаване на ирония

Описание на правилото	Регулярен израз
1. Съдържа един от символите: !,? or …	! или ? или …
2.Съдържа емотикона::), ;) ;(	:) or ;) or ;(
3.Съдържа положителна и отрицателна дума (и #ДАНСwithme (хештаг	Намира се в списък от думи с позитивно и отрицателно значение в Българския
4. Съдържа кавички	[“”]
5.Съдържа две последователни позитивни думи	Положителна И положителна

Таблица 13 Описание на правилата за разпознаване на ирония

Формалното описание на експеримента за откриване на ирония:

За всеки пост от социалните мрежи, направете следното:

За всяко правило описано като регулярен израз от схемата по-горе:

1. Проверете дали постът съответства на правилото.

2. Ако постът съответства на правилото увеличи брояча на сегашното правило

3. Добави мнението към списъка от мнения на сегашното правило

За всеки списък от мнения

за всеки списък на становища направете:

ако правилото може да бъде класифицирано като иронично, увеличете настоящия брояч на точни правила.

За всяко правило да се изчисли точността като:

иронични правила, класифицирани като съвпадение това правило, разделени на всички правила, класифицирани като съвпадение това правило.

Прилагаме и паралелна версия на експеримента.

Паралелна версия на експеримента за откриване на ирония в текстова от социалните мрежи:

За всеки пост от социалните мрежи, направете следното:

Създай нова нишка, която го обработва. В тази нишка прави:

За всяко правило описано като регулярен израз от схемата по-горе:

1. Проверете дали поста съответства на правилото.

2. Ако поста съответства на правилото, увеличи брояча на сегашното правило

3. Добави мнението към списъка от мнения на сегашното правило

За всеки списък от мнения

за всеки списък на становища направете:

ако правилото може да бъде класифицирано като иронично, увеличете настоящият брояч на точни правила.

За всяко правило да се изчисли точността като:

Иронични мнения, класифицирани като изпълняващи условията на даденото правило, разделени върху броят на всички мнения, които изпълняват условието на даденото правило.

Фигура 9 Паралелен Алгоритъм за откриване на ирония.

3.7.3 Експерименти за проверка на предсказващата точност

Проверяваме предсказващата точност върху 2000 поста от Twitter и Facebook, съдържащи хещтаг # ДАНСwithme. 80% от постовете са от Twitter и 20% от Facebook. Всички постове са на български език. Ако едно мнение изпълнява повече от едно правило, то ще се брои към всички правила, които го изпълняват.

Филтрираме всички изречения, които не са на български и които не са написани на кирилица. Примери за мнения, които съдържат ирония са:

"Да подкрепим родното производство на ОРЕХОТРОШАЧКИ" #ДАНСwithme 38 http://t.co/L74aeVVzxn

Това мнение съдържа "", но истинската причина за ирония, е в думата "ОРЕХОТРОШАЧКИ", която се използва като препратка към фамилното име на премиера.

Правило:	точност
1	71%
2	68%
3	83%
4	77%
5	79%

Таблица 14 Получени резултати при експериментална проверка за определяне на ирония в текст

Някои иронични мнения, които не успяхме да засечем, са:

1. МВР ме брои за 0,1 човек #дансwithme http://t.co/dDbHk6Wmkxhttp://t.co/dDbHk6Wmkx

2. @Ivan_Petrov:RT @MGeorgieva_Elle: И сме терористи, и ни плащат. #ДАНСwithMe http://t.co/CX11V7mNkI

3. @vzlatanov: Благодаря на обединената независима журналистика, че ясно разобличи наетия протест на #ДАНСwithMe.

Тези примери показват защо задачата за определяне на иронията е толкова трудна - без да се знаят много специфични детайли понякога е практически невъзможно да се определи това, което всъщност има предвид потребителят.

3.7.4 Анотация на ирония на други езици

Проектът за развитие на Senti-Tut включва анотацията на езиковите данни по отношение на две разграничени нива. Докато първото включва морфологични и синтактични тагове, както обикновено например в синтактични банки, а второто се отнася за концепции, характерни за извличането на чувства от текст.

Морфологичната и синтактичната анотация се прави на базата на формат, разработен от проекта Turin University Treebank (за кратко TUT) (Bosco et al., 2000).

Този ресурс е публично достъпен и е разработен от групата по обработка на естествен език от Университета в Торино - както е описано на http://www.di.unito.it/˜tutreeb, което включва 102150 анотирани токена (около 3500 изречения) и се използва като тестова база за няколко кампании за автоматична обработка на италиански език (http://www.evalita.it/, (Bosco and Mazzei, 2012b) и (Bosco and Mazzei, 2012a)). Използва се този формат поради две основни причини: качеството на TUT формата за италианския език и наличието на множество приложения, разработени в TUT проекта като например Turin University Linguistic Environment (TULE, http://www.tule.di.unito.it/, (Lesmo, 2007) and (Lesmo, 2009)), който включва токенизация, морфологичен и синтактичен анализ.

Да разгледаме примерен туит от корпуса, представен в TUT формата:

TWSPINO-216 ‘La spazzatura di Napoli si sta decomponendo. Concorrera al ` Nobel per la chimica.’ (Боклукът в Неапол започва да изгнива. Ще може да кандидатства за Нобелова награда по химия.).

Би могло да се забележи, че TUT форматът описва много детайлни морфологични тагове, което се използва за описание на езици с голямо разнообразие от думи и голямо множество от граматически релации. За всяка дума и лема морфологичната категория и свързаните характеристиките са анотирани заедно. На фигурата по-долу следва примерно използване на формата:

```
*************** FRASE ALB-3 ***************
1 Slitta (SLITTARE VERB MAIN IND PRES INTRANS 3 SING) [0;TOP-VERB]
2 a (A PREP MONO) [1;PREP-RMOD-LOC+TO]
3 Tirana (TIRANA NOUN PROPER F #CITY) [2;PREP-ARG]
4 la (IL ART DEF F SING) [1;VERB-SUBJ]
5 decisione (DECISIONE NOUN COMMON F SING DECIDERE INTRANS) [4;DET+DEF-ARG]
6 sullo (SU PREP MONO) [5;PREP-RMOD-THEME]
6.1 sullo (IL ART DEF M SING) [6;PREP-ARG]
7 stato (STATO NOUN COMMON M SING) [6.1;DET+DEF-ARG]
8 di (DI PREP MONO) [7;PREP-RMOD]
9 emergenza (EMERGENZA NOUN COMMON F SING) [8;PREP-ARG]
10 . (#\. PUNCT) [1;END]
```

Фигура 10 Примерен TUT файл

3.8 Изводи

В частта от тази глава се описват и предлагат три различни начина за определяне на автоматично двойно значение в текстове на английски език.

Първият начин включва предложение на девет евристични правила за откриване на двойно значение в текстове.

Правилата са проверени от автора на дисертацията. Те показват уникални подробности за откриване на двойно значение въз основа на ръчно преглеждане на примерни мнения от социалните мрежи и са избрани от група от двадесет правила на база на най-голяма точност върху предсказване на мненията.

Добавянето на характеристики и на трите класификатора, описани по-горе, показват подобрение в точността в сравнение с базата на избора на правила. Класификаторите използват характеристики, описани в частта от главата - “оценка на характеристиките”. Те са създадени и експериментално проверени от автора на дисертацията въз основа на предишни изследвания по тази тема и авторско изследване на това, което трябва да се добави по спецификата на темата.

С цел да се провери дали резултатите са правилни, трябва да се филтрират становища, написани на други езици, и да се включват само мнения на английски.

Бъдещата работа може да включва усъвършенстване на правилата, за да се включат повече случаи на английски език, които съдържат твърдения с двойно значение.

Алгоритъмът, описан в частта от главата за ирония на български език, изисква да се знае предварително темата на становищата, за да бъде успешен. Все пак всички правила, описани по-горе, показват точност поне 68% в тестовете, което може да се счита за успешен резултат.

Най-успешното правило е: [положителна дума] + #ДАНСwithme) hashtag.

Обикновено становищата имат отрицателно чувство, тъй като те изразяват мнението на протестиращите).

С цел да се подобри предсказващата точност трябва да се филтрират мненията на различни езици освен български. Бъдещата работа може да включва подобряване на правилата, за да се включат повече случаи в българския език, които съдържат иронични изявления. Пример на правило, което може да бъде експериментално проверено, е - да е писано само с главни букви. Също така може да се добавят повече социални мрежи и повече съдии за определяне на полярността на мнението.

Идеите в тази част от дисертацията може да се развиват на други езици, което лесно може да бъде направено Можем да се добавят още правила и класификатори. Друга идея за подобрение е да се провери предсказващата точност върху други езици, както може да има някои езикови специфични правила и възможности.

4 Алгоритми за определяне на невярна информация и преувеличение в текст

4.1 Алгоритми за определяне на невярна информация в текст

Целта на настоящата глава е да се изследва точността на алгоритъм за автоматично засичане на неверни твърдения от социалните мрежи Facebook и Twitter в английски текстове. Основната идея на алгоритъма е да се изберат различни правила, специфични в английския език и да проверим дали всяко мнение ги изпълнява. Ще покажем, че предположенията ни показват добри резултати в определянето на верни твърдения. Предметната област е: извличане на мнения от текст, засичане на лъжливи текстове, анализ на текстове от уеб страници, социални мрежи.

4.1.1 Представяне на проблемната област

С текущото разрастване на социалните мрежи потребителите генерират милиарди коментари всекидневно и нуждата за анализ на данните става все повече и повече необходима и спешна. Извличането на мненията от текст се опитва да разреши проблема за определянето на полярността на определен текст в документа и да се определи дали има положително, отрицателно или неутрално значение. Засичането на заблуждаващи твърдения представлява голямо предизвикателство за областта на извличането на чувства от текст, тъй като стандартните алгоритми не получават очакваните резултати. В настоящата глава разглеждаме алгоритъм за автоматично определяне дали едно мнение е вярно или не, който показва високи резултати върху текстове относно предметната област футбол. За да определим дали едно мнение е заблуждаващо, трябва да имаме ясна дефиниция на заблуждаване. За целите на настоящата разработка дефинираме заблуждаване като „действие или твърдение, караще хората да повярват в информация, която не е вярна“.

Няма да правим разлика между заблуждаване, лъжа, дезинформация, подвеждане и подобни термини с цел опростяване на предметната област. Два интригуващи факта показват защо изучаването на заблуждаващ език е много важна част от естествените езици:

- Средностатистическият човек лъже 3 пъти по време на 10 минутен разговор.
- 65% от хората смятат, че е ОК да излъжеш, за да не нараниш някого.

За целите на текущия експеримент използваме Facebook и Twitter и

разработваме приложение, което да извлича публични мнения от социалните мрежи (чрез HTTP заявки и отговори). Всяко мнение съвпада с тага #rumours и се проверява от правилата и в случай, че отговаря на някои от тях, се проверява от авторът на дисертацията дали действително мнението е вярно или не. Резултатите са изложени в частта, описваща експериментите.

Следват примерни постове от мненията, които сме извлекли:

1. Juan Mata to United #fingercrossed #rumours #mata — feelingexcited.

2. #ELLIEGOULDING DISMISSES SECRET #WEDDING #RUMOURShttps://www.facebook.com/hashtag/rumours

https://www.facebook.com/hashtag/rumours

http://www.celebspy.co.uk/ellie-goulding-dismisses-secret-wedding-rumours-1318516_37540http://www.celebspy.co.uk/ellie-goulding-dismisses-secret-wedding-rumours-1318516_37540

http://www.celebspy.co.uk/ellie-goulding-dismisses-secret-wedding-rumours-1318516_37540

© pinterest - CelebSpy — with Ellie Goulding and 3 others.

When the lesbian teachers would go make out #**rumors**https://twitter.com/search?q=%23rumors&src=hash

.

3. #NBA #**rumors** page: Rajon Rondo doesn't like the Heat. Or anybody he plays. Read http://www.insidehoops.com/nba_rumors.shtml …http://t.co/KzFJWOcJad

http://t.co/KzFJWOcJad

4. . Sources Claim That Justin Bieber Is Using Drugs With 'Alarming Regularity' http://pitnb.com/LxIosS #**Rumors** #Updateshttps://twitter.com/search?q=%23Updates&src=hash

https://twitter.com/search?q=%23Updates&src=hash

Art Holladay @artholladay1hhttps://twitter.com/artholladay/status/426607233041182720

5. Interesting, if true. #**rumors** RT @verge: From earlier: Apple reportedly planning two larger iPhones for 2014 http://flip.it/0MMDc http://t.co/8auiJhZV5q

Фигура 11 Примерни мнения

Ще се концентрираме единствено върху твърдения, които лесно могат да бъдат проверени, тъй като понякога е много трудно да се провери истинността на мнения - например:

“Ислямска държава има 100 000 бойци, които чакат възможност за атаката.”

4.1.2 Подобни разработки

Въпреки че засичането на слухове е изследвано доста сериозно в областта на психологията, задачата за автоматично засичане в социалните мрежи нашумява значително през последните години. Задачата, която трябва да се разреши, е много подобна до традиционното извличане на мнения от текст. (Hassan et al. 2010) използват самообучение с учител, базирано на модели на Марков, части на речта и шаблони, базирани на взаимовръзки, за да определят полярност в изразеното отношение в мнения от дискусии в Usenet. Други са създали оценки за новини, истории и постове в блогове, базирани на алгоритмично генерирани лексикони на базата на позитивни и отрицателни думи (Godbole et al 2007), (Tsonkov & Koychev 2013). Извличането на данни от социалните мрежи става все по-популярно, тъй като публичното API предлага голям брой постове, които постоянно се обновяват по дадена тематика. (Bifet and Frank, 2010), (Pak и Paroubek, 2011) доказват експериментално че, въпреки честите срещания на неправилно формулирани мнения синтактично и граматично, Twitter може да предостави полезен корпус от мнения за извличане на чувства от текст. Разнообразието от потребители в Twitter прави корпуса много ценен. (Rakiewicz et al 2014) също използват Twitter, за да следят политически мемове, които заблуждават потребителите и ги карат да взимат определени решения по дадени въпроси.

Заедно с многото предимства на Twitter, има и недостатъци - например всичките текстове са не повече от 140 символа, освен това се наблюдават правописни и граматически грешки, символи, жаргонен език и т.н. В публикация използваме правила за засичане на ирония от текстове от социалните мрежи, което вдъхновява настоящата глава от дисертацията.

4.1.3 Алгоритъм за проверяване верността на мнения

Използваме шест евристични правила да засечем дали мнението наистина е вярно или не. Всяко от мненията може да изпълнява повече от едно правило. С използването на правилата искаме да проверим дали специфични предположения относно структурата и съдържанието на мнението могат да определят дали е вярно или не.

Таблицата по-долу показва подробно обяснение на тези правила. Специфичното в предметната област е, че има повече неверни твърдения от обикновено заради това, че мненията съдържат хештага #rumour, който предполага наличието на неверни твърдения.

Описание на правилото	**Шаблон**	**Вярно?**
1. Има дължина повече от 80 символа	Дължина на мнението <80	да
2. Съдържа удивителен, въпросителен знак или ...	[!] or [?] or [...]	не
3. Съдържа I, he, she, it	I or he or she or it	не
4.Съобщения в отрицателен тон	tone < 0	не
5. Съдържа положителна дума и отрицателна емоция	(+дума) и (-емоция)	да
6. Съдържа линк	[URL]	да

Таблица 15 Описание на правилата

Експерименталната част би могла да бъде описана по следния формален начин:

За всяко мнение от социалните мрежи правете следното:

За всяко мнение, описано чрез шаблона от таблицата по-горе

Проверете дали мнението изпълнява правилото

Ако мнението изпълнява правилото, увеличете брояча на текущото правило

Добавете мнението към списъка с мнения на текущото правило

За всеки списък от мнения за всяко правило прави:

Ако правилото е класифицирано правилно, увеличи броячът за правилно класифицирани мнения на правилото.

За всяко правило изчисли успеха на правилото като:

коректно класифицираните мнения, които изпълняват правилото разделено на всичките мнения, които изпълняват правилото.

4.1.4 Експерименти за проверка на предсказващата точност

Проверяваме предсказващата точност върху 1000 мнения от социалните мрежи Туитър и Фейсбук, които съдържат хештага #rumor както следва:

80% от мненията са от Туитър и 20% от Фейсбук. Всичките мнения са на английски. Ако едно мнение изпълнява повече от едно правило, то ще бъде броено към всяко от правилата, които спазва. Филтрираме изречения, които не са на английски. Примери за мнения, които съдържат горния хештаг са:

Andre Gomes to Liverpool!! #Rumours #LFChttps://www.facebook.com/hashtag/lfc

Arsenal are set to complete the signing of Julian Draxler this week. [Mail] #rumorhttps://www.facebook.com/hashtag/rumor

Описание на правилото	Предсказваща точност
1. Дължина поне 80 символа	52%
2. Съдържа въпросителен, удивителен или многоточие.	68%
3. Съдържа I, he, she, it	54%
4. Съобщения в отрицателен тон	56%
5. Съдържа положителна дума и отрицателна емоция.	84%
6. Съдържа линк	79%

Таблица 16 Постигнати резултати

Резултатите от таблицата по-горе показват защо определянето на вярност на твърдение в текст е трудна задача, ако не се знаят някои много специфични детайли относно потребителя и контекста, и на практика е невъзможно да се определи какво всъщност има предвид потребителят.

4.1.5 Извод

Настоящият алгоритъм показва начин за определяне на верността на твърденията в текстове на английски език. За да бъде успешен един алгоритъм, трябва да се знаят предварително авторите или тона на мненията, тъй като ще подобри цялостната точност. И все пак, всичките описани правила по-горе показват, че точността е поне 52% във всеки от тестовете, което може да се счита за успешен резултат.

С цел да се направят правилни тестове трябва да се филтрират мнения, написани на други езици и да се включват само мнения в английски. Бъдещата работа може да включва усъвършенстване на правилата, за да се включат

повече случаи на английски език, които съдържат заблуждаващи твърдения.

Можем да добавим повече правила като думи, които съдържат само главни букви. Друга идея за подобрение, е да се провери експерименталната точност върху други езици, като може да има някои специфични езикови правила. Също така могат да се добавят повече социални мрежи и повече съдии да определят полярността на мнението.

4.2 Автоматично определяне на надценяване и подценяване в текстове

4.2.1 Описание на главата

Целта на настоящата глава е да се изследва точността на алгоритъм за автоматично определяне на мнения, които съдържат надценяване и подценяване в себе си върху текстове на български език от социалните мрежи Facebook и Twitter. Основната идея на алгоритъма е да избере от различни правила, специфични за българския език и да провери дали всяко мнение ги изпълнява. Показваме експериментално, че висока точност може да бъде постигната, когато имаме специфично познание относно предметната област и какви отговори от потребителите можем да очакваме. Предоставяме стандартна както и паралелна версия на описаните алгоритми.

4.2.2 Представяне на предметната област

Един от проблемите при извличането на мнения от текст е, че в много постове потребителите използват непряка реч или двойно значение, което прави задачата много по-трудна. Откриването на мнения, които използват надценяване или подценяване, представлява голямо предизвикателство в областта на извличането на мнения от текст, тъй като стандартните алгоритми не получават очакваните резултати.

В тази глава представяме алгоритъм за откриване на надценяване и подценяване, който постига добра точност на текстове от социалните мрежи. Предлагаме и паралелна версия на алгоритъма, която може да се използва с големи данни.

Като **преувеличение** определяме един от следните термини:

- хипербола
- литота
- мейоза.

С цел да се определи дали мнението е хиперболизирано, трябва да имаме

ясна дефиниция на хипербола. За целите на тази глава определяме хипербола като вид стилистичен троп или фигура, която силно преувеличава герои, предмети и действия, така че авторите подчертават основните им характеристики и изразяват своето емоционално отношение към тях.

Хиперболата представя съзнателно преувеличаване на характеристиките на лице, предмет или действие. В реториката, някои противоположности на хиперболата са мейоза, литота, lackluster, prosaic, dull и bathos. За целите на тази дисертация те се считат за подценяване.

Пример за използване на хипербола:

"The bag weighed a ton." Хиперболата подчертава, че чантата е много тежка, въпреки че вероятно не тежи тон.

Други примери (запазен е оригиналния текст на английски):

1. I have not seen you for one hundred years!!!
2. When smartphone no internet . Im gonna die ! #**Hiperbola**https://twitter.com/search?q=%23Hiperbola&src=hash
3. Yeahhh ! we're leading ! 2-1 ! more gol please from Pastore - Ibra #AllezParis #PSGCHE #**HiperBola**https://twitter.com/search?q=%23HiperBola&src=hash
4. I walked a thousand miles but still can't get back home. Hope I'll make it on time!!!

Подценяването е фигура на речта, използвана от писатели или публични личности, за да карат ситуация умишлено да изглежда по-маловажна, отколкото всъщност е.

Например, ако спечелите 10 милиона долара в томбола, когато се каже на репортер: "Радвам се", се използва подценяване.

По същия начин, ако отбор загуби своята среща 10-0 във футболен мач и капитанът на отбора казва след мача: "Ние не се справяме добре", това е подценяване, защото той се опитва да се намали интензитета на загубата.

Подценяването обикновено се използва в ироничен смисъл, когато изявлението в отговор е точно обратното на това, което се очаква, но по-малко интензивно и с ироничен тон.

Например, ваш приятел купува новото си палто с петна на много места, в отговор можете да използвате подценяване:

"Това не изглежда чак толкова зле".

Ето защо, подценяването (литотата) е противоположна на друга фигура на речта - хипербола или преувеличение.

В реториката литотата е фигура на речта, която се използва за подценяване и реторичен ефект главно чрез двойни отрицатели. Например, вместо да каже, че нещо е привлекателно (или дори много привлекателно), може само да се каже, че "не е непривлекателно".

Литотата е форма на подценяване, винаги умишлено и с акцентирано намерение. Въпреки това тълкуването на отрицанието може да зависи от контекста, в това число и културен контекст. В речта то може да зависи и от интонация, и от акцент; Например, фразата "не е лошо" може да се каже по такъв начин, че да изрази значение от "посредствен" до "отличен". Тя може да се използва, за да се смекчат суровите изрази, подобно на евфемизъм.

Използването на литотата е общо за български, английски, руски, немски, холандски, иврит, арамейски, гръцки и френски език. Тя е изразно средство на староанглийската поезия и на исландските саги и е средство за много по-естестически издържан стил.

Думата литота произлиза от гръцката дума Литос, която означава "обикновен, малко или оскъдно". Пример следва по-долу:

Мнение	**Истинско значение**
"Not bad."	Good
She is not as young as she was.	She's old
He was not unfamiliar with the works of Dickens.	He was acquainted with the works of Dickens

Таблица 17 Примери за литота

За експеримента използваме социалните мрежи Фейсбук и Туитър и използваме инструмент за извличане на постовете (чрез HTTP заявките и отговори).

Всяко мнение се извлича от един от таговете "#hyperbole" или "#litosis" и се проверява от всяко от правилата и в случай, че има съвпадение на мнението, се преглежда от автора на настоящата глава и се оценява дали мнението наистина използва хипербола или не. Резултатите могат да бъдат видени в частта от главата “Експерименти”, която е по-надолу в текста. По време на писането на тази глава от дисертацията липсва съществуваща база от данни за откриване на надценяване или подценяване в тексове.

4.2.3 Подобни разработки

Липсват много разработки по темата преди. (Huifeng Tang et all 2012)

изследват как се е използвала литотата в текстове от литературни произведения.

В статия на подобна тема, (Цонков и Койчев (2013)) също използват правила за откриване ирония в текстове от социалните мрежи. Показваме, че евристичните правила могат да покажат добри резултати при определяне ирония в текстове и настоящата глава е вдъхновена от горната разработка. В статия на Guardian се разглежда разликата между „нещо добро“ и „нещо не-лошо“. Очевидно е, че става въпрос за подценяване и не се казва в прав текст какво се има предвид. Пример за литота е речта на британския политик Ed Miliband, който казва:

„Проблемът е, че при Дейвид Камерън имаме правителство, което е далече от това да обърне нещата и ги прави много по-зле. Не е случайно. Това е защото са водени от грешни представи за това как се управлява държава през 21-ви век.“

Друг нашумял пример за използване на литота е от реч на британския премиер Дейвид Камерън: През ноември 2014, той се изказва по темата за имигрантите, която може да предизвика доста конфликти. Той се изказва изключително внимателно –

„Трябва да сме ясни. Не е грешка да се изказва притеснение от броя хора, които пристигат в Обединеното кралство.“

Той би могъл да бъде максимално директен и да се изкаже : „Правилно е да се изрази притеснение от броя хора, които пристигат в Обединеното кралство“, но той се изказва доста по-тактично, използвайки „Не е грешка“, вместо „правилно“.

С помощта на литотата политикът се измъква от проблема, който разделя на два лагера партията на премиера и държавата. Литотата смекчава един политически проблем и двата лагера по темата могат да извлекат позитиви за каузата си от речта на британския премиер.

4.2.4 Алгоритъм за разпознаване на литота

Използваме три евристични правила, за да открием дали мнението е наистина пресилено или не. Искаме да проверим дали конкретни предположения за мнението могат да бъдат успешни при определяне дали това мнение е преувеличено или не.

Таблица 19 показва подробно обяснение на тези правила. Специфично за темата е, че има повече мнения, които са наистина пресилени, отколкото обикновено и се дължи на естеството на становищата (съдържащи таг

#hyperbola или #litotis, при които предполагаме авторът използва преувеличение във всеки случай). Формално описание следва по-долу

Описание на правилото	Шаблон
1.Съдържа позитивно правило и отрицателна емоция	POSITIVE [WORD] and NEGATIVE [EMOTION]
2 Съдържа три последователни прилагателни	ПРИЛАГАТЕЛНО и ПРИЛАГАТЕЛНО и ПРИЛАГАТЕЛНО
3. Съдържа число по-голямо от 1000	Число > 1000.

Таблица 18 Описание на правилата, класифицирани като надценяване и подценяване в текст

Експериментът може да бъде описан формално по следния начин:

За всеки списък от мнения за всяко правило прави:

Ако правилото е класифицирано правилно, увеличи броячът за правилно класифицирани мнения на правилото.

За всяко правило изчисли успеха на правилото като:

коректно класифицираните мнения, които изпълняват правилото, разделено на всичките мнения, които изпълняват правилото.

Паралелна версия на алгоритъма е описана по-долу:

За всяко мнение от социалните мрежи:

Създай нишка за всяко от правилата, описано чрез шаблоните по-горе, и паралелно прави следното:

За всеки списък от мнения за всяко правило прави:

Ако правилото е класифицирано правилно, увеличи броячът за правилно класифицирани мнения на правилото.

За всяко правило изчисли успеха на правилото като:

коректно класифицираните мнения, които изпълняват правилото, разделено на всичките мнения, които изпълняват правилото.

4.2.5 Експериментална проверка на предсказващата точност

Проверяваме предсказващата точност върху 100 мнения от Facebook и Twitter. 50% съдържат хештага хипербола, 50% са произволни. Всичките

мнения са на английски език. Ако едно мнение изпълнява повече от едно правило, ще го броим във всичките правила, които изпълнява. Филтрираме ръчно всички мнения, които не са на английски и всички мнения, които съдържат хештага #hyperbola или #litotis и се има предвид термина от областта на геометрията хипербола.

Правило	Предсказваща точност
1.	80%
2.	80%
3.	60%

Таблица 19 Постигнати резултати

Резултатите от таблицата по-горе показват защо определянето на хипербола или литота на твърдение в текст е трудна задача. Ако не се знаят някои много специфични детайли относно потребителя и контекста, е на практика невъзможно да се определи какво всъщност има предвид потребителят.

4.2.6 Заключение

Настоящият алгоритъм показва начин за откриване на надценяване и подценяване в английски текстове. За да бъде успешен един алгоритъм, трябва да знаят предварително авторите или тона на мненията, тъй като това ще подобри цялостната точност. Много важно е да се знае контекста на мнението, като без него би било много трудно да се правят предположения.

С цел да се подобри предсказващата точност, трябва да се филтрират мнения, написани на други езици и да се включват само мнения на английски. Бъдещата работа може да включва усъвършенстване на правилата, за да се включат повече случаи на английски език, които съдържат усилване на твърденията в текстове (например хипербола) и разбирания. Можем да се добавят повече правила като думи, които съдържат само главни букви. Друга идея за подобрение е да се провери предсказващата точност върху други езици, като може да има някои езикови специфични правила. Също така могат да се добавят повече социални мрежи и повече съдии да определят полярността на мнението.

4.3 Засичане на лъжи в текстови съобщения

Да разгледаме следните характеристики, които помагат за автоматичното засичане на лъжи в текстови съобщения. Те биха могли да бъдат използвани за съставяне на евристични правила или характеристики с подхода, описан по-горе.

4.3.1 Сложен отговор

Ако е зададен доста прост въпрос, като например : "Какво направи снощи?" и събеседникът Ви изпраща обратно три параграфа от текст, то може да бъде, за да се компенсира фактът, че той не казва истината. Някои хора смятат, че по-подробният им отговор ги прави да са по-убедителни, но всъщност може да бъде точно обратното особено ако лицето обикновено не ви изпраща дълги и описателни текстове.

Ако лицето описва всеки малък детайл от неговата или нейната нощ, когато всичко, което искате, е прост отговор, тогава той може да се опитва да ви убеди, че каквото и да казва, е истина.

Ако лицето е наистина новак в лъжите, тогава може дори да се върне към ситуацията, за да се коригира по време на обмена на съобщения.

4.3.2 Смяна на темата

Ако човекът се опитва да смени темата, това е друг сигнал, че той или тя се опитва да излъже.

Това би било вярно и за ситуацията в лице; ако човекът, с когото се води разговорът, не казва истината, то е съвсем естествено, че той или тя не би искал да се занимава с темата. Ако лицето ви дава бърз отговор и след това пита един по-сложен въпрос, това може да бъде един чудесен начин да се създаде отклоняване от факта, че той или тя се опитва да излъже.

Може да се каже нещо подобно: "Аз бях навън късно с Джон. Ами ти? Как мина нощта? "

Още едно 100% сигурно поведение, което сигнализира лъжа, е, ако някой каже, че трябва да напусне разговора, след което рязко Ви дава отговор.

Ако човекът, с когото говорите, прави това, тогава най-вероятно не е майстор лъжец. Ако сте водили чудесен разговор и след това изведнъж го хванете в лъжа, а след това и следващото нещо, което той каже, е, че трябва да прекрати разговора, това е може би един не толкова фин начин, по който показва, че не иска да се справи с последиците от лъжата.

Особено подозрително е, ако лицето напусне без никакви обяснения и ако не мислите, че той е имал някакви планове през това време.

Ако събеседникът се опитва да не дава точен отговор, това е сигнал, че иска да не го хванете в лъжа. Ако човекът, с когото говорите, опитва да променя темата, след като е казал лъжа, тогава най-вероятно иска да излъже.

Ако събеседникът обикновено не Ви прави комплименти, а изведнъж направи вместо да отговори на въпрос или пък дава завързан отговор на един от вашите въпроси, е вероятно, че това е просто се опитва да прикрие лъжата.

4.3.3 Търсене на подробности

Потърсете много незначителна подробност. Ако лицето ви дава твърде много незначителни подробности и обикновено не говори толкова много върху дадена тематика, това би могло да бъде, защото се опитва да покрие следите си и неговата история да изглежда наистина правдоподобна.

Ако човекът Ви казва каква музика е свирела в клуба, когато всичко, което искате да знаете, е с кого е бил, тогава това може да е знак, че той лъже.

Ако Ви казват нещо като: "Аз бях навън късно в закусвалнята с Джим. Той говореше за успеваемостта на нападателите на противниковия отбор. Имахме най-добрите картофки със сирене, въпреки че... ", когато обикновено никога не Ви дава подробности като тази, то това може да е знак, че събеседникът Ви лъже.

4.3.4 Неясни и уклончиви отговори

Търсете дали лицето дава неясни или уклончиви отговори.

Друг начин да се хване лъжец е да се види, ако лицето откаже да формулира наистина своята история и се стреми да Ви дава неясни отговори, като така по-трудно да може да го хванете. Ето някои фрази, които може да се използват за засичане на лъжа в текстове:

"Трябва да е било около полунощ, когато ..."

"Може би това е така, защото ..."

"Аз може би се прибрах у дома около два."

"Аз не съм сигурен, че..."

"Изглеждаше, нещо като ..."

4.3.5 Смяна на граматичните времена

Смяната на граматичните времена или прескачането между минало и сегашно време е още един знак, че лицето лъже. Ако човек прави това, то означава, че може да лъже, защото е толкова зает да измисля историята в главата си, че той забравя, че това се е случило в минало време, а не в настоящето.

Ако човекът, с когото говорите нормално, изведнъж започва да сменя минало и сегашно време и ако подозирате, че може да лъже, това е сигурен белег за лъжа.

Внимавайте, когато някой говори така: "Снощи излязох с момчетата за няколко питиета. Но пих само няколко бири. След това аз се връщам у дома преди полунощ ... "

Хората може да са напрегнати, когато лъжат, защото, след като започнат измислянето на историята в главите си, може естествено да преминат към сегашно време.

4.3.6 Необичайно граматически и стилово правилен отговор

Пример за подозрителен отговор е, ако отговорът на събеседника Ви е необичайно полиран. Ако лицето, което обикновено не се опитва да има перфектно изработени и граматически правилни изречения, изведнъж построи отговор като по учебник, то това е знак, че той/тя се опитва твърде много да изглежда, че казва истината.

Разбира се, ако лицето, обикновено се говори по този начин, тогава това не означава нищо.

Ако събеседникът обикновено използва много съкращения, не винаги използва препинателни знаци или думи с главни букви или просто по принцип не пише правилно, тогава той/тя може да се опитва лъже, ако изведнъж ви изпраща четири перфектни граматически изречения.

5 Класификация на мнения и анализ на слухове от текстове на български език. Алгоритъм за подобрение на списъка от думи и изрази, носещи определено отношение в текст

Първата част от настоящата глава е направена съвместно с докторант Гергана Лазарова.

5.1 Увод

С развитието на уеб технологиите се появяват големите масиви от данни, които стават лесно достъпни. Новите алгоритми и технологии трябва да се справят с огромното количество от данни. Обработката на естествен език получава бурно развитие, ползвайки новите технологии за бързо извличане и обработка на данни. Актуално направление в тази област е класификацията на мнения. Настоящата част от дисертацията има за цел да приложи алгоритми за класификация от машинното обучение в тази област, използвайки филмови ревюта и мнения от социалните мрежи на български език.

Хората като интелигентни агенти, за да могат да чувстват емоции, изразяват чувства и комуникират помежду си и за да се разбират, трябва да разчитат на скрити модели на емоция. Можем да формулираме едно изречение като неутрално твърдение, но може и да вложим определени емоции, което го прави мнение.

Извличането на мнение от текстове е модерна дисциплина за получаване на субективна информация, разбирайки емоциите в текста. Използват се множество от алгоритми, за да се постигне целта. Извличането на мнения от текст е интердисциплинарен предмет, който комбинира алгоритми от изкуствения интелект, обработката на естествен текст, извличането на информация, статистика и т.н. и симулира интелигентно поведение.

Социалните мрежи (facebook, twitter, google+, linkedin) съдържат големи количества от данни - коментари, статии и т.н., които изразяват лично мнение. Какво мислят хората по дадена тема - положително или отрицателно е мнението им? Големите компании искат да разберат какво мислят потенциалните им клиенти за техните продукти и за продуктите на техните конкуренти. Политиците се интересуват относно това как ги възприемат хората, футболният отбор – дали е популярен сред феновете. Тези важни проблеми трябва да бъдат разрешени с наличните ресурси и за да се отговори на тези въпроси, трябва да обработваме “големи данни” и да се научим как да засичаме чувство в текстовете.

Има много разработки на английски език в областта на извличането на чувство от текст, това е област, която се е разраснала бързо, за да стане една от най-активните предмети за изследователска дейност в областта на обработката на естествен текст (NLP). Повечето хора в Интернет владеят английски език, изразяват мнения на английски и голяма част от съдържанието на световната мрежа е на английски.

(Bing Liu 2000) анализира видовете извличане на мнение от текст (на ниво изречение и текст) и показва широко разнообразие от практически приложения. (Kennedy and Inkpen) представят два метода за определяне на чувство в текст, изразено от ревюта на филми на английски. В българския език няма достатъчно класифицирани примери. Малкото данни, с които разполагаме за обучаващ процес, са едновременно непълни и редки. В настоящата глава ревютата на английски са преведени на български. Тези данни се използват в обучителния процес и няколко класификационни алгоритми се сравняват на базата на тестово множество, съдържащо реални ревюта на български филми.

Също разработваме модул за извличане на мнения от социалните мрежи на български и проверяваме точността на метод, базиран на правила, върху мненията, извлечени по-горе.

5.1.1 Извличане на мнения от текст

Извличането на мнения от текст определя отношението на автора на мнението към дадена тема. Той прогнозира полярността на документ в зависимост от контекста .

Има два основни подхода за извличане на мнение от текстове (Bo Pang; Lillian Lee and Shivakumar Vaithyanathan 2002)**.** . Първият се основава на лексикон от настроения (Blitzer J., Dredze M., Pereira F 2007) и разчита на полярността на самите думи на документите. Обикновено ние пресмятаме отрицателните и положителните думи и окончателното решение се прави въз основа на сумиране. Вторият подход използва алгоритъм за класификация, за да се направи разграничение между класовете. Обикновено се използват два класа – положителен и отрицателен клас, но понякога може да се използва и неутрален клас.

Съвременни подходи в анализа на настроения в текстове включват:

- Личностни черти - стабилни личностни нагласи и поведение, характерни тенденции (нервни, неспокойни, безразсъдни, мрачни, враждебни, ревниви);
- Емоции - ядосани, тъжни, радостни, страшни, срамежливи, горди, въодушевени;

- Положителни / отрицателни становища с вариации - прогнози за рейтинги (и положителни, и отрицателни) могат да бъдат използвани като обучение на К-класификация или като рейтинг на функция на регресия);
- Откриване на ирония в документи;
- Засичане на плагиатство, признаване на авторски права.

Трябва също да се отбележи, че мнението обикновено е субективно. Двама души може да не се споразумеят за настроенията на мнението. Освен това някои конкурси по извличане на мненията от текст (примерно "SemEval") премахват мнения, ако съдиите не се споразумеят по тях.

5.1.2 Анализ на множеството от данни

Трудността в извличането на мнения в текстове на български език се крие в липсата на данни за обучение. Класифицирани мнения се намират трудно. Един учител трябва ръчно да опише всички примери за обучение.

От друга страна, Интернет изобилства от класифицирани примери на английски - Amazon, TripAdvisor и т.н. Така че един английски източник на класифицираните случаи е лесно да се намери и да се използва. Следователно, тези средства могат да бъдат използвани за комплекта за обучение. Тези документи използват мнения от Amazon, които са преведени с Google Translate за български. Броят на мненията е 992.

Тестовият набор от филмови ревюта на български се извежда от www.cinexio.com. Той се състои от 38 анотирани примера. Всяко мнение има рейтинг между 1 и 5. Той съдържа указание за обратна информация за филмите. Тези мнения са много кратки и непълни.

5.1.3 Извличане на характеристики

Използва се bigram модел. Теглото на всяка функция следва TF-IDF стандарт. Разглеждаме честотата и обърната честота при срещането на термини за:

- Честота на срещането на термин - броят пъти, в които терминът *t* се среща в документ *d;*
- Обърната честота на срещането на термини - измерва дали терминът се среща често или рядко във всичките документи;

В долната фигура е изобразен форматът за пресмятане на термовете:

TF-IDF:

$$TF(t) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document}{Total\ number\ of\ terms\ in\ the\ document}$$

$$IDF(t) = \log(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ with\ term\ t\ in\ it})$$

Final Value = Tf(t) * IDF(t)

Формулата на честота на теглата е описана по-долу:

$W_{t,d}$ = 1 + $\log_{10} tf_{t.d.,}$ ако $tf_{t,d}$ >0

= 0, в другите случаи

Извличаме 17099 характеристики от 992 ревюта. По-подробна информация може да бъде видяна в таблицата по-долу.

База от данни	**брой характеристики**	**брой примери**
Amazon reviews –train set	17099	992
Cinexio.com– тестово множество	17099	38

Таблица 20 Резултати за тестовото множество от данни

5.1.4 Класификацията на ревютата

Класификацията е процесът на идентифициране към кое множество от категории принадлежи нов пример на базата на тренировъчно множество. Обикновено, примерът (ревю на филм) ще използва клас - положителен или отрицателен. В тази глава се използват 5 класа, които съответстват на броя звезди, които ревютата биха могли да имат. Всеки рейтинг се счита за отделен клас.

За класификационния процес използваме три различни класификатора, които са описани в глава 2 -

- Наивен Бейсов Класификатор (*Naïve Bayes Classifier (NB))*
- *kNearestNeighbours (kNN)*
- *BACKPROPAGATION*.

5.2 Подход, базиран на правила, за автоматично засичане на слухове

Въпреки че алгоритмите за автоматично засичане на слухове са

изследвани в областта на психологията, задачата да се проверява информацията в социалните мрежи, получава значително внимание през последните години.

Задачата е много подобна на традиционния анализ на настроения и чувства в текст. Според (Hassan et al, 2010.) използването на контролиран Марков модел, части на речта, както и зависимостта между модели, за да идентифицират нагласата на полярността в теми, публикувани в Usenet дискусионни мнения, което води до добра точност.

Други определят резултати за настроения в новини, истории и публикации в блогове на базата на алгоритмично генерирани лексикони на положителните и отрицателните думи (Godbole et al, 2007) , (Цонков и Койчев, 2013).

Извличане на данни от социалните мрежи се превръща в тенденция в анализа на настроения, тъй като публичното API осигурява голям брой мнения, които постоянно се подновяват. (Bifet и Франк, 2010) . (Pak and Paroubek 2010) показват експериментално, че въпреки честата поява на нередовни модели на речта в туитове, Twitter може да се използва като полезен корпус за анализ на настроения. Разнообразието на Twitter потребители прави този корпус особено ценен. (Rakiewicz et al 2007) също да използват Twitter за проследяване на подвеждащи политически мемове.

(Цонков и Койчев 2013) също използват правила за откриване ирония в текстове от социалните мрежи. Те показват, че евристичните правила могат да показват добри резултати при определяне ирония в текстове и вдъхновяват настоящата глава от дисертацията.

Използваме шест евристични правила, за да открием дали мнението е наистина достоверно или не. Правилата определят дали мнението не е или е достоверно.

Искаме да проверим дали конкретни предположения за мнение могат да бъдат успешни при определяне дали то е достоверно или не.

Таблицата по-долу показва подробно обяснение на тези правила. А специфично за темата е, че има повече неверни мнения, отколкото обикновено и това се дължи на естеството на становищата. Формално описание следва по-долу.

Описание на правилото	**Шаблон**	**Истина или лъжа**
1. Дължина повече от 80 символа	дължина (мнение) < 80	Истина

2. Съдържа удивителен въпросителен или многоточие	[!] или [?] или [...]	Лъжа
3. Съдържа лично местоимение	аз ИЛИ тя ИЛИ той ИЛИ то	Лъжа
4. Мнения в отрицателен тон (които съдържат поне една дума с отрицателно значение)	(-word) от списък	Истина
5. Съдържа положителна дума и отрицателна емоция	(+word) И (-emotion) от списък	Лъжа
6. Съдържа линк	[URL]	Истина

Таблица 21 Описание на правилата за засичане на слухове

Алгоритъмът може да бъде описан по следния формален начин:

За всяко мнение от социалните мрежи правете следното:

За всяко мнение, описано чрез шаблона от таблицата по-горе

Проверете дали мнението изпълнява правилото

Ако мнението изпълнява правилото, увеличете брояча на текущото правило

Добавете мнението към списъка с мнения на текущото правило

За всеки списък от мнения за всяко правило прави:

Ако правилото е класифицирано правилно, увеличи броячът за правилно класифицирани мнения на правилото.

За всяко правило изчисли успеха на правилото като:

коректно класифицираните мнения, които изпълняват правилото, разделено на всичките мнения, които изпълняват правилото.

5.3 Експерименти за проверка на предсказващата точност

След извличането на характеристики, трите алгоритъма се изпълняват върху тренировъчно множество. Класификационната точност на алгоритмите се сравнява на базата на тестово множество от реални български ревюта. Може да се види в таблицата по-долу, че BACKPROPAGATION се представя по-добре от другите класификатори, постигайки 42.11% класификационна точност (т.е.

пропорцията от коректно класифицирани примери в тестовото множество). Наивният Бейсов класификатор се оказва с най-слаба класификация - 36.84%, докато kNN постига 39.47%.

Класификационната точност е на базата на оценката на ревютата, която е от 2 до 6, т.е. очакването за всеки от класовете е 20%, а класификаторите я подобряват почти двойно.

Алгоритъм	**Точност**
kNN	39.47%
Naive Bayes	36.84%
BACKPROPAGATION	42.11%

Таблица 22 Сравнение на алгоритмите за класификация

Правило	Точност
1.	58%
2.	62%
3.	52%
4.	56%
5.	71%
6.	79%

Таблица 23 Получени резултати при подхода с правила

5.3.1 Получени изводи

Може да се заключи от резултатите в експериментите, че анализът на настроения на базата на буквалните преводи на английски филмови рецензии на български език, дава добри резултати. Оказва се, че многослойната невронната мрежа (BACKPROPAGATION), надминава другите стандартни алгоритми за класификация. Невронната мрежа е итеративна процедура, прилагана за големи данни и се оказва, че е по-лесно да се провери върху hdfs, yarn и spark. За бъдещата работа следва да бъдат проучени и изследвани повече области от данни.

Алгоритъмът, описан по-горе дава резултати също така за определяне на доверието в българските текстове. За да бъде успешен един алгоритъм, трябва да знае предварително авторите или тона на мненията, тъй като това ще подобри цялостната точност. И все пак, описаните правила по-горе показват, че точността е поне 52% във всеки от тестовете, което може да се счита за успешен резултат.

С цел да се провери дали резултатите са правилни, трябва да се филтрират мнения, написани на други езици и да се включват само такива на български.

Бъдещата работа може да включва усъвършенстване на правилата, за да се включат повече случаи на български, които съдържат заблуждаващи твърдения.

Можем да се добавим повече правила, като например думи, които съдържат само главни букви. Друга идея за подобрение е да се провери върху други езици, като може да има някои езикови специфични правила. Също така могат да се добавят повече социални мрежи и повече съдии да определят полярността на мнение.

5.4 Алгоритъм за подобрение на списъка от думи и изрази, носещи определено отношение

Целта на представената разработка е да предложи алгоритъм за подобрение на списъка от думи и изрази, носещи определено отношение (положително или отрицателно) към даден обект. Този списък се използва по-късно за автоматична класификация на нови мнения за обекти. Алгоритъмът използва мнения, класифицирани от човек, след това ги анализира, за да намери най-често срещаните думи и изрази, които не са в списъка от "спиращи думи". Такъв списък се изгражда за определена предметна област. Проведените експерименти с текстове от различни предметни области, показват че предложеният подход подобрява точността на класификацията на мнения.

5.4.1 Въведение

Задачата за определяне на чувства в текст е обект на сериозен изследователски интерес през последните години. Предпоставка за това са все по-голямото използване на социалните мрежи (Facebook, Twitter и др.), блогове и форуми. Това налага създаването и използването на методи, които да оценяват дали едно мнение е положително, отрицателно или неутрално. Тези методи имат различни приложения в маркетинг анализа, мнението на хората относно политиката и т.н. Например, ако искаме да закупим лаптоп, освен цена, характеристики, производители т.н., биха ни интересували мненията на други хора (най-вече такива, които считаме за експерти в областта), които биха ни дали съвети. За производителите на такива лаптопи мнения на клиентите са изключително полезни за уточняване на маркетинг стратегията и за улавяне на погрешни мнения - например "На този лаптоп не върви новият Уиндоус."

Въпреки че на пръв поглед задачата изглежда лесна, изречения от рода на: "Този филм би трябвало много да ми хареса, актьорите са невероятни,

сюжетът изглежда повече от обещаващ. Но накрая нещо не се получава.” са много трудни за класификация от машина, тъй като се съдържат предимно думи, носещи положително значение.

Една от основните стъпки при извличането на чувства от текст, е използването на ключови думи, които да носят позитивен, отрицателен или неутрален контекст. Например думи като “харесвам”, “обичам”, “мразя” или “не харесвам”. На базата на списък от тези думи може чрез различни алгоритми (например наивен бейсов класификатор) да определим чувството в текста.

Възможно е да се използват различни класификатори – например support vector machines, наивен бейсов и други.

Идеята на настоящата глава от дисертацията е да предложи алгоритъм за подобрението на този списък, тъй като освен думите, които носят положително или отрицателно отношение във всяка предметна област (като харесвам, мразя и т.н.), се срещат и други думи или поредица от две думи, които да носят определено отношение в зависимост от конкретната предметна област (например в областта на футбола - “отборът се защитаваше много” носи отрицателно значение, а никоя от думите сама по себе си не е с такова). Характерна особеност за социалните мрежи (например Twitter) е, че някои съкращения, например тагове, могат да носят сами по себе си значение за мнението в текста.

В настоящата глава от дисертацията са направени тестове в няколко предметни области – футбол и политика, като използваме социалните мрежи Facebook, Twitter, Google+, Blogspot, Digg, Yahoo News. Използвани са стоп думи от https://code.google.com/p/twitter-sentiment-analysis/source/browse/trunk/files/stopwords.txt?r=51.

Разгледано е определяне на резултати с помощта на класификатора SentiStrength (който може да се използва за научни цели) върху 1000 произволни мнения от социалните мрежи. Показва се, че точността на класификация се подобрява с няколко процента при ползването на подобрения речник с ключови думи.

Класификацията на мненията на позитивни/отрицателни/неутрални е направена от автора на дисертацията с помощта на програмни средства за верификация на получените резултати (SentiStrength) и верификация на резултатите от автора на дисертацията, т.е. всяко мнение, класифицирано като положително и отрицателно, е проверено от автора.

5.4.2 Подобни разработки

През последните години се наблюдава засилен интерес в областта на

извличането на чувства от текст. Ранната работа в тази насока е на (Turney и Pang 2000), които прилагат различни методи за намиране на полярността на ревюта на продукти и на ревюта на филми. Това се прилага на ниво документ.

Документите могат да бъдат класифицирани и на степени на отношение - например в скалата от (-5; +5) изречението:

“Много много харесвам този продукт - най-добрият, който някога съм ползвал!”, може да се оцени като +5, а мнение от рода на “Това не е лошо” да се оцени като +1.

Изречение от рода на “Изключително много мразя този продукт!!!” може да бъде оценено като -5.

В настоящата разработка се спираме изцяло на класификация - положително/отрицателно/неутрално (т.е. дали е засечено отношение или не). Pang и Snyder разглеждат точно такъв тип класификация.

За верификация на получените резултати е използван софтуерният инструмент SentiStrength (http://sentistrength.wlv.ac.uk/), който има резултати, достигащи класификационните способности на човек. Доказано е, че човек има средно 79% успеваемост в класифицирането на мнения и дори програма да има 100% точност, двама произволни човека ще имат разлика в класификацията около 20% на произволно избрани мнения (както и на която и да е тема). Горепосоченото софтуерно приложение осигурява Java код, който е базиран на списъци от думи, носещи положително и отрицателно значение.

В началото се използват списъците с позитивни/отрицателни думи, които предлага SentiStrength, които се обогатяват с всяка предметна област. В настоящия алгоритъм не се засичат т.нар. emotion icons, което би могло да бъде подобрение, тъй като в доста мнения (като “I’m walking outside :))))))))))))))))”) чувство може да бъде засечено в текста. Редица софтуерни продукти с отворен код използват решения като машинно самообучение, статистка, и обработка на естествен език, за да автоматизират извличането на чувства от текст в уеб страници, онлайн новини, дискусионни групи в Интернет, онлайн ревюта, уеб блогове и социалните медии. Системи, базирани на знания, ползват публично достъпни ресурси като WordNet-Affect, SentiWordNet и SenticNet, за да извличат семантичната информация с методи на обработката на естествен език.

5.4.3 Алгоритъм за подобряване на лексикона от думи

Списък позитивни думи = {Предварително зададен списък от думи, които носят универсално положително значение}

Докато списъкът от думи не се променя, повтаряй:

Извлечи произволни 1000 положителни кратки изречения от социалните мрежи (примерно от определена област или по ключова дума). Класификацията се извършва от учител.

Намери най-често срещаните N думи в изреченията, класифицирани като позитивни, които не се съдържат в даден списък от стоп думи.

Обнови списъка с думи, като добавиш тези, които се срещат по-често от досегашните на базата на отношение - срещане/брой думи в изреченията.

Алгоритъмът трябва да бъде направен и за отрицателни думи по същия начин. Той е изцяло еквивалентен на гореописания. Следва илюстрация на алгоритъма и за думите с отрицателно отношение:

1. Добави най-често срещаните думи с положително отношение.

2. Добави най-често срещаните думи с отрицателно отношение.

3. Класифицирай отново с помощта на някой класификатор, ползващ списък от думи с положително и отрицателно значение, и сравни резултатите с вече получените.

Фигура 12 Алгоритъм за подобряване на лексикона думи

Предметната област на тестовете може да бъде зададена предварително. Алгоритъмът разчита на потребителя да бъде обучен от предварително зададена тема.

Да дадем следния пример – Извличаме по 1000 мнения от социалните мрежи, които съдържат дадена дума (“football” или “politics”), при което намираме следните подобрения в списъка - думи като “scored”, “goal” носят положително значение, докато думи като “defended” или дори биграма като “zonal marking” може да носят по-скоро отрицателно отношение. От всички извлечени мнения броим най-често срещаните думи и биграми, които не попадат в списъка от стоп думи и ги добавяме към списъка от думи. След това проверяваме дали има подобрение на класификацията отново с 1000 произволно извлечени мнения от предметната област.

Предметната област може да бъде лесно заменена, посочените примери са в области, които са с широк публичен интерес, всички мнения се извличат на база ключова дума. Може да се направи, така че да се търсят биграми или поредица от думи, но тогава те трябва да бъдат добавени в списъка от получените стоп думи, тъй като се срещат във всяко мнение и не носят значение.

5.4.4 Експерименти за проверка на предсказващата точност

Гореописаният алгоритъм е проверен експериментално в следните предметни области - футбол и политика. След изпълняването на алгоритъма и подобряване на списъка от думи, в областта на футбола се подобрява класификацията с наивен бейсов класификатор със средно 3,5%, а в областта на политиката – с около 5%. Посочената разлика се дължи на това, че в политиката повече думи носят отрицателно значение.

Използвани са социалните мрежи Facebook, Twitter, Google+, Digg, Bing Search, като се извличат най-актуалните мнения при всяка проверка. Почти всички от социалните мрежи предлагат възможност за извличане на публични мнения от тях, като лесно могат да се добавят социални мрежи, тъй като в повечето случаи извличаният формат е един и същ - JSON.

При всеки тест извличаме произволни 1000 мнения от социалните мрежи. Разпределението по социални мрежи е следното:

Приблизително 600 (т.е. 60%) от мненията са от Twitter. Приблизително 300 (т.е. 30%) са от Фейсбук. Останалите 10% са от другите мрежи, в зависимост от най-новите резултати. Всички мнения са сортирани по дата (т.е. най-нови) и са публични, т.е. имаме права да ги използваме. Използваме най-често Twitter и Facebook, тъй като има най-много мнения, публикувани там.

След анализ на футболни мнения откриваме, че най-често срещаните думи, които са специфични за областта и не са стоп думи: ball, cup, player, Match, Win, Lose, Play, Team, Goalkeeper, Striker, Goal, Kick, Pass, Tackle, Cross, Dribble, Shoot, Strike, Score, Foul, Defend, Attack, Referee, Penalty, Red.

От изброените по-горе думи положително значение в контекста на областта носят: cup, goal, strike, score, team, shoot.

Думите goalkeeper, lose, referee, red, penalty носят отрицателно значение.

След добавянето им в списък с думи с общо положително значение, при класификацията с помощта на софтуерният инструмент SentiStrength се постига подобрение от 4% в получените резултати - от 71% на 75% и от 66% на 71%.

В областта на политиката голяма част от мненията се класифицират като отрицателни (около 40%), като средното за другите области е около 10-15%. С добавянето на специфични термини се увеличава точността на предсказване с около 8% - например най-често срещаните думи и биграми са: politician, tax, elections, president, minister, new president, tax cut.

5.4.5 Заключение и обобщение

В настоящата част от главата се изследва метод за подобряване на резултатите при извличането на чувства от мнения, базиран на извличане на специфични думи от предметна област и добавянето им. С направените тестове се вижда, че се подобряват получените резултати. За бъдещо развитие ще бъдат разгледани други предметни области, ще се подобри избирането на

специфичните думи и ще се проверява с по-голямо множество от мнения.

6 Изграждане на система за извличане на потребителски текстове от социалните мрежи и автоматично извличане мнения от текст

6.1 Изисквания

6.1.1 Дизайн на приложението

Съществена част от работата по дисертацията е изграждането на практическо приложение, което да илюстрира идеите, разработени в публикациите. Идеята за разработка на подобен софтуерен продукт възниква още в самото начало на работата по дисертацията и практическата реализация е много полезна за разбиране на материята и опознаването на материала и какво търсят потребителите.

Голяма част от мненията, публикувани в Интернет, са в социалните мрежи. Затова е изключително важно да могат да бъдат обработвани в реално време, тъй като информацията и мненията се променят бързо.

Скоростта на изпълнение е от критично значение и поради тази причина се налага имплементация на сървър, който да приема паралелна обработка на информацията. Друга възможност е да се извличат мнения от различни уеб сайтове с уеб кроулер, но в социалните мрежи се публикуват достатъчен брой мнения за анализ на чувството в текст за целите на дисертацията.

Основният замисъл на приложението е да се анализира чувството в текста. По-долу е описан детайлно използваният алгоритъм, но той трябва да предоставя точна оценка на мнението в зависимост от езика. Разработените алгоритми са за английски език, но могат лесно да бъдат променени да ползват друг език, без загуба на точност.

Важна подробност за правилен анализ е местоположението на потребителя, тъй като, в зависимост от него, дори един и същи език може да има различно значение и смисъл (пример: британски и американски английски).

6.1.2 Изисквания към архитектурата

Архитектурата на приложението включва сървър, който обработва заявки от уеб клиент. Заявките се въвеждат от потребителя и клиентът ги праща до сървъра, който търси мнения в социалните мрежи, оценява чувството в текста и праща резултатите на клиента. Клиентът се грижи за показването на резултатите от сървъра и ги предоставя на потребителя в удобен за него вид.

Всяка социална мрежа трябва да може да се обработва отделно, затова използването на отделна нишка за всяка социална мрежа за паралелно извличане и обработване е добро разрешение на проблема и предлага решение как да се обработват мнения в реално време.

6.1.3 Технически изисквания към имплементацията на приложението

Данните трябва да могат да бъдат достъпни бързо и лесно, като резултатът трябва да е независим от устройството, което потребителят използва.

Получените данни трябва да бъдат на разположение за допълнително и бързо търсене в свободен текст, като резултатът трябва да е независим от encodinga на страницата и езикът, който се използва. Данните следва да бъдат съхранявани в структуриран вид като автор, информация за автора, дата и час, текст на мнението, оценка на настроението, език, социална мрежа, дължина, и т.н.

Поради наличието на различни клиенти - например от компютри и мобилни устройства, би било подходящо проектът да е достъпен през уеб интерфейс, който не изисква предварителна инсталация на софтуер, което би затруднило използването на проекта от краен потребител.

Изчисленията, които се извършват, трябва да може да бъдат лесно разпалелирани, тъй като при наличие на голям брой заявки се очаква забавяне на скоростта на приложението. Трябва да се създаде протокол, по който да комуникират клиентът и сървърът.

6.2 Подобни разработки

Да разгледаме няколко от най-популярните приложения, които измерват чувство в текстове от социалните мрежи -

Meltwater: Преценява тона на коментара като прокси за репутацията на марката и открива нови детайли, които помагат да се разбере целевата група.

Google Alerts: Опростен, но много полезен начин да се наблюдават заявките, които търсим. Използва се за засичане на съдържанието на мнения и се получават регулярни имейли с обобщение на последните релевантни резултати от Гугъл. Това е добро начин да се следят влиятелни личности, тенденции и конкуренти.

People Browser: Намира всички споменавания на марка, индустрия и

участници и анализира чувството в текста. Това приложение позволява да се сравнят обема на споменавания преди и след маркетингови кампании с цел да се изчисли успеваемостта им.

Google Analytics: Мощно приложение, което открива кои канали влияят на вашите купувачи и клиенти. Може да се създават ръчни отчети, анотации, за да се поддържат непроменени записи на вашите действия, свързани с маркетинг и уеб дизайн. също така и напреднали сегменти, за да се разбият данните за гостите на сайта и да се придобият ценни данни за техните действия и преживявания онлайн.

Hootsuite: Предлага се безплатно приложение, което позволява да се управлява и измерва поведение на потребителите в страниците ви в определени социални мрежи. Срещу 6$ месечно може да ползва платена версия, която дава подобрен анализ на мненията и поведението на потребителите.

Tweetstats: Предлага се безплатно приложение, което показва графиката на статусите в Twitter, приложението е лесно - просто се въвежда потребителят и паролата и приложението работи без повече действия от страна на потребителя.

Facebook Insights: Когато имате поне 30 харесвания на страницата си във Фейсбук, можете да измервате нейното представяне с горното приложение. Може да се види общият брой харесания, броят фенове, активните потребители всеки ден, новите харесвани и махане на харесвания и т.н. Източниците, които харесват страницата, демографските данни, броят гледания на страницата и уникалните потребители, които са я посетили, статистиките относно табовете, външните референти, използване в медиите и т.н.

Pagelever: Още едно приложение за измерване на активността на facebook страницата. Приложението дава възможност да се измери точно във всяка стъпка какво съдържание се използва и как се споделя във Фейсбук.

Social Mention - Еквивалентът на Google Alerts в социалните медии, това е полезно приложение, което позволява да се следят споменаванията на определени ключови думи и тагове във видео, блогове, микроблогове, събития, бележки, коментари, новини, въпроси и отговори, хеш тагове и дори аудио. Там също се отбелязва дали мненията са позитивни, отрицателни или неутрални.

Marketing Grader: Marketing Grader е приложение, което оценява цялостната разработка на маркетинга. Използва над 35 метрики, които да изчислят оценката на мнението като се гледа къде се публикува, колко често се ползват Twitter и Facebook и се превръщат мненията на потребителите в цялостна оценка.

6.3 Имплементация на модул за извличане на мнения от социалните мрежи

Модулът за извличане на мненията на потребителите от социалните мрежи е реализиран на езика за програмиране Java. Извличат се мнения от социалните мрежи и сайтове Facebook, Twitter, Blogspot, Digg, Google+, NyTimes, Yahoo answers. За всяка социална мрежа се пуска отделна нишка, която да извлича мненията. В началото клиентът се връзва на порт 12111 към сървъра като за всеки клиент се създава отделна нишка.

Клиентът праща заявка на сървъра какъв лимит от мнения да търси за всяка социална мрежа, а сървърът праща или мненията, или код 204 при грешка.

Примерна заявка от клиента следва на фигурата по-долу. В началото се изпраща заявката, за която търсим данни, след това се праща лимита за всяка от социалните мрежи. Единственото задължително поле е заявката, а останалите полета не са задължителни. Стойността от всяка социална мрежа е 100, ако няма заявка от клиента, а ако има поле [all], тогава сумата от мненията от всички социални мрежи трябва не го надвишава (тогава се разпределят всички мнения пропорционално).

{query:'jobs', [all]:1000, [gplus_limit]:10, [nytimes_limit:100], [twitter_limit:100], [facebook_limit:100], [bing_limit:100], [blogs_limit:100]}");

Фигура 13 Примерна заявка от клиента

Всяко мнение се пази във наредена петорка (чувство, източник, локация, мнение, дата). Всички мнения се запазват или в база или в текущата памет. Базата периодично се изтрива, когато датата на мнението надвишава 3 месеца и се пазят само статистиките, които не съдържат мнения.

Статистиките съдържат следните данни: множество от всичките позитивни мнения, отрицателни мнения, неутрални мнения. Броячът на мненията по позитивни, отрицателни и неутрални. Чувството в текста може да бъде от интензивност от (-5; +5) и се пазят статистиките в колко мнения се среща всяко чувство. В следващата част от тази глава на дисертацията ще се уточни как се пресмята всяко чувство. Пресмятаме и сумата от постовете с линкове и постовете без линкове. Това се прави с цел анализ на мненията, които са за линкове, тъй като най-често мнението, което е коментирано, е относно съдържанието в пратената уеб страница.

Статистиките се изпращат на клиента в json формат - като json масив.

Полетата, които се пращат може да се видят в списъка по-долу:

positive_posts, neutral_posts, negative_posts

posts_per_time_positive,posts_per_time_neutral, posts_per_time_negative

positive_sentiments, negative_sentiments, posts_with_links

posts_without_link. num_positive. num_neutral, num_negative

Търсенето се осъществява по сходен принцип за всички социални мрежи. Ще разгледаме детайлно описание на извличането на мнения по мрежа:

- Фейсбук - изпраща се заявка към
  "https://graph.facebook.com/search?type=post&locale=en_US&limit=200&q=" + query; /*+ "&offset=" + page;*/

  Оттам се взимат данните и се обработват мненията като се попълват данните по зададената петорка по-горе. Ако мненията са повече от 200 - извличат се порции по 200.

- Туитър - използва се библиотеката Twitter4j. Характерна особеност е, че всяка заявка не може да извлече повече от 200 мнения и затова се налага както във Фейсбук да се извличат мнения на части от по 200 мнения.
- Гугъл+ - праща се HTTP заявка към публичните мнения като извличането на мненията не е толкова лесно колкото в останалите мрежи да се извлекат мненията от социалната мрежа като се налага да се извлекат резултатите, които са създадени в JSON формат.
- Bing - Извличат се мненията от публичната търсачка като

  http://api.bing.net/xml.aspx?Appid=58003B78464C927F8668E7060A755A426F106D8F&sources=news&web.count=40&web.offset=

  - като се взимат новините между <news:Snippet> </news:Snippet>, освен това извличаме стойността на тага <news:Date>
- Blogspot - Извличаме публични мнения от

  "https://ajax.googleapis.com/ajax/services/search/blogs?v=1.0&scoring=d&rsz=8&q=https://ajax.googleapis.com/ajax/services/search/blogs?v=1.0&scoring=d&rsz=8&q=

  - като се взимат обектите от тага <responseData> и след това

се извличат като <results>, освен това взимаме датата от тага <publishedDate>.

- Yahoo Answers - Извличаме отново публичните мнения от:

  http://answers.yahooapis.com/AnswersService/V1/questionSearch?appid=It1iOM34&results=30&query=

  - взимаме мненията от <Content> и </Content> и датата от <Date> и </Date>.
- Google Talking - Извличат се от
  - "http://api.whostalkin.com/search?key=d963ba347735ac1f5662a0269fa2beb4&q="
- Digg - Извличаме публичните мнения от
  - "http://services.digg.com/2.0/search.search?query=" + query + "&count=20";
  - Взимат се данните от <stories> тага и <description> и <submit_date> за довършване на наредената петорка:

(чувство, източник, локация, мнение, дата),

- NY Times - Извличаме от
  - "http://api.nytimes.com/svc/search/v1/article?"+ "&api-key=" + NEWSWIRE_API_KEY +"&query=" + query;
  - като NEWSWIRE_API_KEY се взима от сайта на NY Times, като изисква предварителна регистрация.
- Отново се извличат резултатите от results тага, както и от body и date таговете.

6.4 Имплементация на модул за оценяване на мненията на потребителите

Съществената част на разработката на приложението е модул за оценяване на мненията на потребителите.

Оценката на мнението на потребителите е в интервала (-5; +5) като -5 е максимално отрицателно, а +5 максимално позитивно. Системата използва речници от думи и специални символи, на базата на които се прилагат алгоритмите.

Прилагаме описание на файловете, които се използват:

- WordLookUpTable.txt е просто списък на думите, носещи емоция, като на всеки ред имаме думата, след това таб, а след това цяло число от 1 до 5, или -1 до -5. Този списък може да се редактира и

разшири. Трябва да се има предвид, че силните страни на +` и -1 нямат ефект върху програмата. Има някои, в списъка, просто за да покаже, че те думи са взети предвид, но не е използван. Всяка дума може да приключи с уайлд кард *, но той може да отиде само в края.

- EmoticonLookUpTable е подобна на таблицата думи, описана по-горе, но за списък с емотикони.
- EnglishWordList.txt е просто списък на английските думи - тя се използва за тази част от алгоритъма, който се опитва да коригира думи с нестандартни начини на изписване.
- NegatingWordList.txt обръща поляритета на следващите думи например - не е щастлив е отрицателно.
- BoosterWordList.txt увеличава интензивността настроения - например много щастлив е по-положителен от щастлив.
- SlangLookupTable.txt - замества общият жаргон с еквивалентни думи или изрази.
- IdiomLookupTable.txt- заменя силата на настроенията на отделните думи във фразата със силата на цялата фраза.

Обхванатото в тази дисертация е само малък процент от всичко свързано с анализа на мнения от текст - не е разгледано и не са правени опити за:

- Анализ на настроения по съдържание отвъд текст (как може да се определи настроение във Vine видео или Instagram снимка например?)
- Анализ на настроенията на социалните взаимодействия (може ли да се класифицира едно харесване и един коментар като две положителни взаимодействия?

Може ли да се класифицира дадено “unfollow” във Фейсбук като отрицателно взаимодействие, каква е ролята на пунктуацията в настроенията:

(например разликата м/у "Аз съм добре !!!" и "Добре съм….")?

Това води до два въпроса:

- Могат ли инструменти да анализират точно настроение?
- Какво е бъдещето на анализа на настроенията в текст?

Инструментите не могат да интерпретират фактори като сарказъм (където очевидното чувство се изследва от инструмента, а мълчаливото и внушено противоположно настроение не се засича), и социална семантика, където някои изрази могат да се интерпретират на същия език, в зависимост от това къде е било казано (например American English срещу British English).

Както описахме, въпреки че човешкото възприятие на настроенията е по-зряло и точно от анализа на софтуерен продукт, то все още не е 100% точно.

Нека да вземем този преглед на форум например:

"Аз заменям дружество А за дружество В, заради най-новата им изключителна оферта, която не мога да пропусна! Обслужването на клиенти на дружество А е много по-добро все пак, така че аз категорично ще се върна."

Докато един инструмент може да определи горното мнение като неутрално, защото на броя на положителни и отрицателни становища е равен, ние може и да не сме съгласни с решението. Този преглед може да се възприема по различен начин от фирма А и фирма Б, в зависимост от това кои аспекти търсят. Ако фирма В търси обратна информация за най-новите си оферти, това което се споменава определено е положително, въпреки че тя има отрицателно споменаване за обслужването на клиентите. Следователно, докато може да има универсален анализ на настроения и чувства, също така могат да се променят въз основа, на това кой прави анализа.

В крайна сметка, точността на инструмента е просто броя пъти, в които човешката преценка е съгласна с решението на инструмента. Точността на човешкото съгласуване е между 70% и 79%.

Вземайки под внимание това, можем спокойно да кажем, че една добра точност за инструменти за анализ на чувства е 70%. Проблемът е, че един инструмент, който е точен по-малко от 70% от времето, не е точен, и "перфектният" инструмент, който е точен 100% от времето ще изготви данни, които не сме съгласни приблизително 30% от времето.

6.5 Създаване на потребителски интерфейс и статистики

Да разгледаме в детайли потребителския интерфейс и статистиките на приложението. На фигура 14 по-долу се вижда началната страница на приложението, която се зарежда за потребителя.

Всички добри интерфейси споделят осем качества или характеристики:

1. Яснота - Интерфейсът трябва да избягва двусмислие, като направи всичко ясно чрез езика си, последователността от действия, йерархията и метафорите за визуалните елементи.
2. Стегнатост. Лесно е да се направи един интерфейс чрез свръх-изясняване и етикетиране всичко, но това води до претоварване, когато има прекалено много неща на екрана по едно и също време. Ако твърде много неща са на екрана, намирането на това, което търсите, е трудно, и

така интерфейсът става досаден и труден за използване. Истинското предизвикателство при правенето на по-сериозен потребителски интерфейс е да се направи кратък и ясен в същото време.

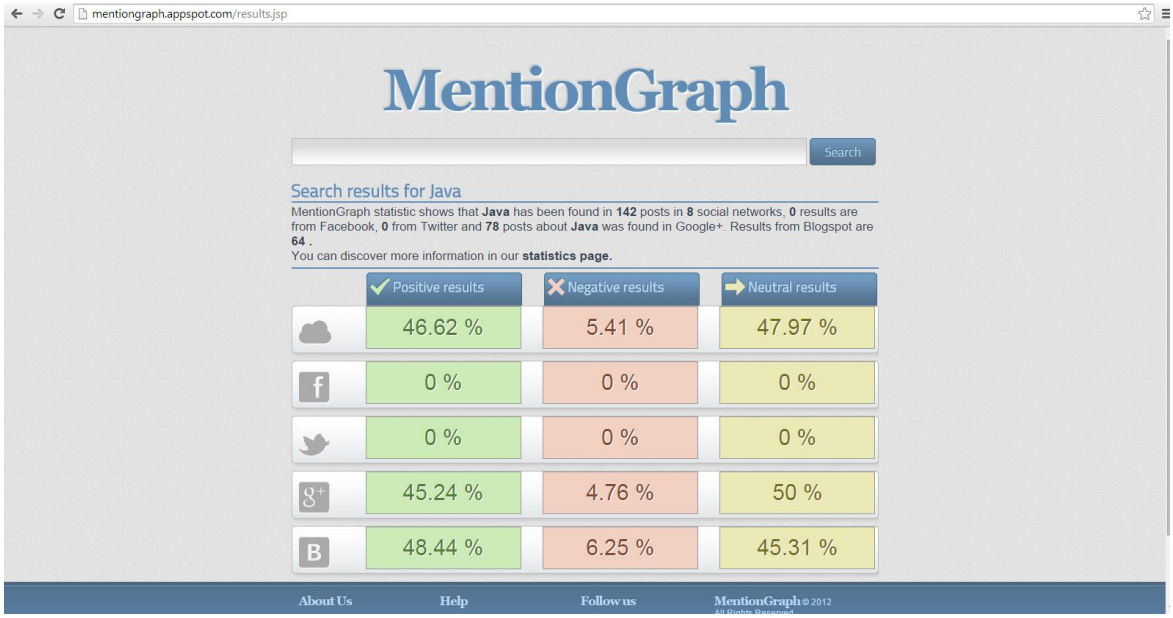
3. Познаваемост. Дори ако някой използва интерфейс за първи път, някои елементи може вече да са познати. Метафорите от реалния живот могат да се използват, за да придават значение.
4. Отзивчивост. Един добър интерфейс не трябва да изглежда муден и неясен. Това означава, че интерфейсът трябва да осигури ясни отговори на потребителя за случващото се и дали входът на потребителя е обработен успешно.
5. Съвместимост. Интерфейсът трябва да може да работи на различни платформи и да предоставя еднаква последователност от действия.
6. Ефективност. Времето е пари и един добър интерфейс трябва да направи потребителя по-продуктивен чрез комбинации и добър дизайн.
7. Отношение към потребителя - Един добър интерфейс не трябва да наказва потребителите за техните грешки, а вместо това трябва да се осигурят начини за отстраняването им.
8. Принципът на най-малкото учудване -Принципът на най-малко удивление и учудване (POLA) е общ принцип при проектирането на всички видове интерфейси. Той се основава на идеята, че човешките същества могат да отделят цялото си внимание върху едно нещо в даден момент, което води до заключението, че новостите трябва да бъдат сведени до минимум.

Началната страница на приложението може да бъде видяна фигурата по-долу.

Фигура 14 Начална страница на софтуерния прототип

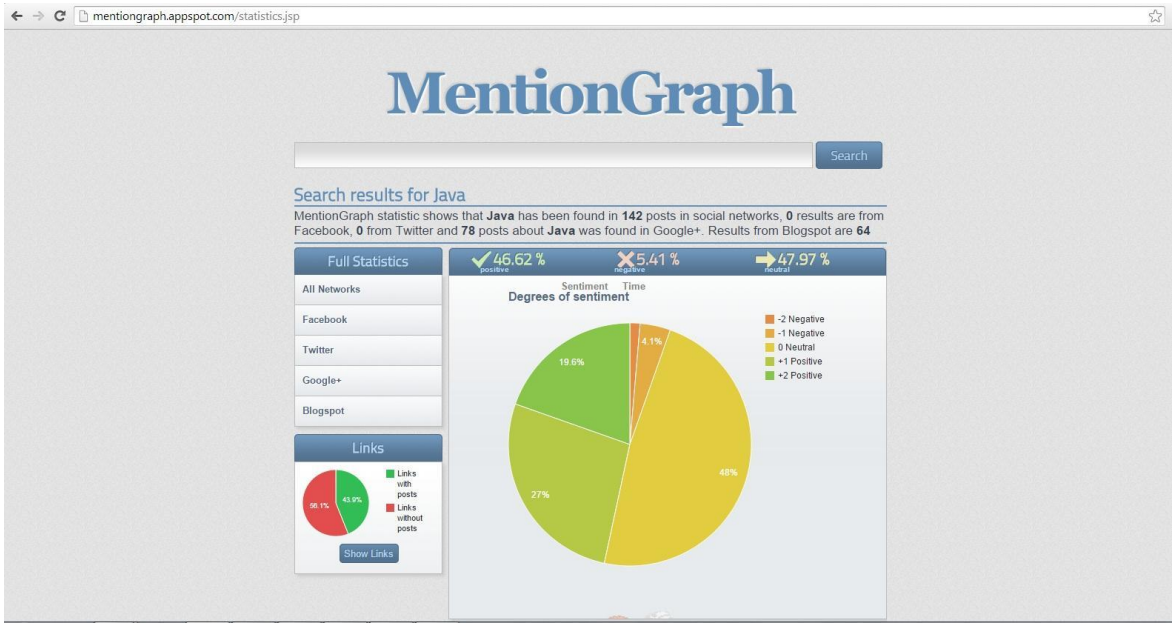
Началната страница е съставена от опростен дизайн, в който потребителят въвежда търсенето си. Може да се търсят всякакви думи и фрази. Езикът, на който работи, е английски. Понякога търсенето може да продължи до минута, тъй като ресурсите на сървъра, на който е качен софтуерният прототип, са ограничени. Предлагат се различни опции на потребителя за връзка с автора на приложението, което дава възможности за подобряване на качеството и точността на оценката на мненията на потребителите.

След търсене на определен низ - в случая java, се виждат резултатите от различните социални мрежи, разпределени по положителни, отрицателни и неутрални мнения, оформени в удобен дизайн за потребителя. Социалните мрежи, които са включени са: Фейсбук, Туитър, Гугъл+, Блогспот, Digg. Причините, поради които са избрани, са количеството на мненията и популярността на социалните мрежи, както и лесния достъп до мненията - публичните мнения в изброените сайтове могат да бъдат достъпни с публично API и заявки в json формат.



Фигура 15. Получени резултати

Следващата страница на приложението е разгледана на долната фигура и показва получените резултати при търсене. Пресметнати са всичките резултати сумарно, след което са диференцирани резултатите по социални мрежи. Освен това разглеждаме нивата на чувство в текста, разделени от -2 до +2, като -2 е най-отрицателното чувство, а +2 е най-позитивното. 0 е неутрално и множеството от мненията са именно неутрални. Предлага се разпределение на мненията по социални мрежи и е представено визуално оформление на получените резултати, както е показано на фигура 16. Предоставя се на потребителите възможност да получат отделно и линковете, които се съдържат в мненията. Това е полезно, тъй като повечето мнения на потребителите изразяват отношение относно линковете. Дизайнът на приложението е разработен с HTML5 и CSS. Графиките са създадени с езика за програмиране Javascript и се зареждат динамично.



Фигура 16. Статистика и разпределение на резултатите

Отделно се предоставят примерни постове от социалните мрежи с цел да се покаже, че мненията са реални и оценката им е правилна. На долната фигура са показани част от извлечените мнения, като се цитира и статистика, която включва разпределенията на чувството в текста, статистики на базата на социални мрежи, които използваме, както и брой мнения, които съдържат линкове, и такива, които не съдържат.

Фигура 17 Статистика на мненията

Използва се различна цветова гама, която да представи визуално разликите между чувствата в текста. Мненията съдържащи линкове са много важни за експертите по SEO и поради тази причина са отделени в отделна

статистика. Статистиките на базата на социални мрежи описват чувството в текста на базата на различните мрежи и могат да извлекат мненията и да ги предоставят на потребителя.

7 Заключение

7.1 Обобщение

В дисертацията е разгледана задачата за извличане на мнения от текст от социалните мрежи. Описани са проблемите в предметната област, техническите предизвикателства и е аргументирано защо е избрана насоката на разработките.

Представен е алгоритъм за подобряване на списъка от положителни и отрицателни думи в българския език и експериментално се доказва подобрението на предсказващата точност.

Предложени са алгоритми за автоматична класификация на мнения от текстове на български език и са анализирани слухове в твърдения.

Основна част от дисертацията са алгоритми за автоматично определяне на двойно значение в текстове от социалните мрежи. Като двойно значение определяме множеството от ирония, сарказъм или сатира. Освен правила, са използвани характеристики и е изградено пространство на характеристиките, чрез което върху тестово множество е предсказана точността на алгоритмите. Проверена е и предсказващата точност на три класификатора за предсказване на ирония.

Разгледани са и алгоритми за предсказване на лъжа, невярна информация, преувеличение и подценяване в текстове. Показано е, че на основата на определени предварителни знания относно събеседника е възможно да се постигне добра предвиждаща точност.

Описана е изградената система за оценяване на мнения относно дадена тема, която съдържа потребителски интерфейс за лесно и бързо използване на системата, и чрез нея експериментално се потвърждава предсказващата точност на използваните алгоритми.

Изпълнена е целта на дисертацията, а тя е разглеждането на двойно и обратно значение в текстове (ирония, сарказъм, сатира, хипербола, литота, лъжи и невярна информация).

7.2 Приноси

Основен принос на дисертацията е разглеждането на почти неизследвани дялове на извличането на мнение от текст - засичане на ирония, сарказъм и сатира в текст, за засичане на невярна информация и преувеличаване в текст (хипербола и литота). Приложените алгоритми са изследвани върху текстове от социалните мрежи на английски и български език и е представена тяхна паралелна версия, която може да работи с големи количества от данни. Изгражда се система за автоматично извличане на мнения от основните социални мрежи (Twitter, Facebook, Google +, LinkedIn, Yahoo), в която по ключова дума се извличат мнения и се оценява автоматично

мнението на потребителите относно думата (например продукт, личност или марка).

7.2.1 Научни приноси

1. Разработени са правила за откриване в текстове от социалните мрежи на:
    a. ирония,
    b. сарказъм,
    c. сатира,
    d. хипербола,
    e. литота
    f. слухове
2. Изграждат се характеристики за откриване на двойно значение.
3. Разработване на алгоритъм за подобряване на списъка от положителни и отрицателни термини за извличането на мнения от текст.

7.2.2 Научно-приложни приноси

1. Експериментална проверка на точността на предсказване на различни класификатори за определяне на двойно значение на мненията от социалните мрежи.
2. Разработване и експериментална проверка на точността на предсказване на алгоритъм за засичане на слухове от социалните мрежи.
3. Експериментална проверка на предсказващата точност на предложените характеристики (описани в горната част).

4. Създаване на следните корпуси за определяне на двойно значение:
    a. ирония,
    b. сарказъм,
    c. сатира,
    d. хипербола и литота

7.2.3 Приложни приноси

1. Изгражда се система за извличане на мнения от социалните мрежи, която и оценява мненията от текстовете.
2. Създава се система, която предлага оценки на мненията от -5 до +5, където -5 е максимално негативно чувство в текста, а +5 е максимално положително.

3. Създадени са шаблони за правилата за засичане на двойно значение в текстове, които лесно могат да бъдат променяни и използвани.

7.3 Публикации

Публикациите на автора на дисертацията, свързани с темата на дисертацията:

1) Todor V. Tsonkov, Ivan Koychev, Automatic Detection of Double Meaning in Texts from the Social Networks, Proc. of BCI-2015, Editors: Costin Badica et al, published in: CEUR WP, 2015
2) Gergana Lazarova, Todor Tsonkov - Класификация на мнения от текстове на български език., Journal of the Technical University – Sofia Plovdiv branch, Bulgaria, "Fundamental Sciences and Applications" Vol. 21, 2015
3) Todor V. Tsonkov, Ivan Koychev, Detecting Credibility in Rumours in Texts from the Social Networks, Proc. of 125 years University of Sofia Conference, издателство:Софийски Университет, 2014
4) Тодор Цонков, Иван Койчев - Разпознаване на ирония в текстове от социалните мрежи на български език - Сборник доклади на Осма Национална конференция „Образованието и изследванията в информационното общество", Пловдив, 2014
5) Тодор Цонков, Иван Койчев, Подобряване на списъка от положителни и отрицателни термини при извличането на мнения от текст, Сборник доклади на Шеста Национална конференция „Образованието и изследванията в информационното общество", Пловдив, 31.05.-01.05.2012 г, АРИО, 2013, стр.268-273.

7.4 Доклади

1) Todor Tsonkov, "Irony in the Social Networks", FET EYE Lab Surfing workshop, Солун, Гърция, Март 2014
2) Тодор Цонков – Извличане на мнения от текст – Първо Докторантско Училище, София 2013
3) Todor Tsonkov - "Double meaning in bulgarian texts" - Third MUMIA School, Crete, Greece, July 2014

7.5 Бъдеща работа

Базата от мнения за двойно значение – ирония, сарказъм, сатира, неверни твърдения, умаления и хиперболи би могла да бъде разширена.

Могат да се добавят различни алгоритми освен базирани на правила, класификатори и характеристики – например използване на графи, които да описват връзките в мненията и да се обхождат по различни начини в зависимост от конкретната задача. Биха могли да се добавят повече източници за извличане на мнения от потребителите, например ревюта на продукти, мнения от блогове и т.н. Други алгоритми може да използват подобрение на статистическия анализ на ирония, сарказъм и сатира и добавяне на по-подробен анализ на зависимостите между думите в мненията като например

линейна регресия. Също могат да бъдат добавени различни класификатори– например maxEntropy.

Перспективите за бъдещо развитие на софтуерния проект са да бъде популяризиран и разширен като проект с отворен код.

Използвана литература

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)

Albert Bifet and Eibe Frank, 2005- Sentiment Knowledge Discovery in Twitter Streaming Data, University of Waikato, Hamilton, New Zealand

Alec Go; Richa Bhayani; Lei Huang; Twitter Sentiment Classification using Distant Supervision; Technical report, Stanford University

Aline A. Vanin, Larissa A. de Freitas, Renata Viera, Marco N. Bochernitsan - Some Clues on Irony Detection in Tweets. 2012

Andrea Gianti, Christina Bosco, Viviana Patti, Andrea Bolioli, Luigi Di Caro- Annotating irony in a Novel Italian corpus for sentiment analysis

Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING. pp. 36–44 (2010)

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

Bing Liu - Sentiment Analysis, a multifaceted problem, IEEE Intelligent Systems, 2010

Bing Liu *(2001)* - Sentiment analysis and Subjectivity. Morgan &. Claypool

Bing, L. (2010) *Sentiment Anlaysis and Subjectivity.* Invited Chapter for the Handbook of Natural Language Processing, Second Edition.

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.

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.

Bo Pang; Lillian Lee (2005). "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales". "Proceedings of the Association for Computational Linguistics (ACL)". pp. 115–124**.**

Bo Pang; Lillian Lee and Shivakumar Vaithyanathan (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques". "Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)". pp. 79–86

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.

Buschmeier, H., & Kopp, S. (2014). A dynamic minimal model of the listener for feedback-based dialogue coordination. In *SemDial 2014: Proceedings of the 18th Workshop on the Semantics and Pragmatics of Dialogue*, pp. 17–25, Edinburgh, UK.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. WASSA 2013, page 29.

Christopher M. Bishop. Pattern Recognition and Machine Learning. : Springer, 2006.

Common Opposites - Antonyms Vocabulary Word List http://www.enchantedlearning.com/wordlist/opposites.shtmlhttp://www.enchantedlearning.com/wordlist/opposites.shtml

Cunningham, H., Maynard D, Bontcheva K. (2011) *Text Processing with GATE,* University of Sheffield Department of Computer Science

Davidov, D.; Tsur, O.; and Rappoport, A. 2010. Semisupervised recognition of sarcastic sentences in twitter and amazon. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning, 107–116. Association for Computational Linguistics.

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.

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.

Dipankar D. and Sivaji B. 2012, Identifying Emotional Expressions, Intensities and Sentence level Emotion Tags using a Supervised Framework*

Duncan Green - British English table, 2002, https://oxfamblogs.org/fp2p/what-brits-say-v-what-they-mean-handy-de-coding-device/

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In Proceedings of Language Resources and Evaluation Conference, pages 392–398.

Facebook statistics - https://blog.kissmetrics.com/facebook-statistics/https://blog.kissmetrics.com/facebook-statistics/

Francesco Barbieri and Horacio Saggion, 2014 1- Automatic Detection of Irony and Humour in Twitter

Fresh Minds - Turning conversations into insights: A comparison of Social Media Monitoring Tools; A white paper from FreshMinds Research 14th May 2010;FreshMinds 229-231 High Holborn London WC1V 7DA Tel: +44 20 7692 4300 Fax: +44 870 46 01596 www.freshminds.co.uk.

George A Miller. 1995. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41.

Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009)

Godbole et al – "Large-Scale Sentiment Analysis for News and Blogs" 2007

Hart L. 2013, The Linguistics of Sentiment Analysis, Portland State University, PDX Scholar 2013, http://pdxscholar.library.pdx.edu/honorstheses/20

Hassan Saif, Yulan He and Harith Alani – "Semantic Sentiment Analysis of Twitter"

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.

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).

Hinrich Schutze - Introduction to information retrieval

Holmes A., (2014) *Hadoop in practice*

Hu, M and Liu, B. (2004) *Mining and summarizing customer reviews*. KDD'04.

Jason Kessler and Nicolas Nicolov. 2009. Targeting sentiment expressions through supervised ranking of linguistic configurations. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, May

Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In Proceedings of the 3rd IEEE International Conference on Data Mining, pages 427–434, Melbourne, FL, USA, December.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. 2013. TwitIE: An Open-Source Information Extraction Pipeline for

Microblog Text. In Proceedings of Recent Advances in Natural Language Processing Conference.

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

Kennedy, A., Inkpen D., (2006) *Sentiment classification of movie reviews using contextual valence shifters*, Computational Intelligence, vol. 22, pp. 110–125.

Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011)

Kripke, A. Presupposition and Anaphora: Remarks on the Formulation of the Projection Problem, 2009

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th International Conference on World Wide Web, WWW '03, pages 519–528, New York, NY, USA. ACM.

Lev Ratinov Dan Roth 2012 - Design Challenges and Misconceptions in Named Entity Recognition

Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In Proceedings of the ACM 15th Conference on Information and Knowledge Management, pages 43–50, Arlington, VA, USA, November.

Li Zhuang, Feng Zhing, Xiao - Yang Zhu – Movie Review Mining and summarization, 2006

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

Martin J. R. and White P. R.R. 2005, The Language of Evaluation: Appraisal in English, Palgrave, London, UK

Matthew Russell - Mining the Social Web, 2nd Edition, 2013

McAuley, J., Leskovec J..(2013) *Hidden factors and hidden topics: understanding rating dimensions with review text*. RecSys.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 168–177, Seattle, WA, USA, August.

Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of the European Conference on Machine Learning. pp. 318–329 (2006)

Murthy Ganapathibhotla. Bing Liu - Mining Opinions in Comparative Sentences

Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena - Large-Scale Sentiment Analysis for News and Blogs

Niklas Jakob, Iryna Gurevych 2012 - Using Anaphora Resolution to Improve Opinion Target Identification in Movie Reviews

Pak and Paroubek – "Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 2010

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.

Paula Carvahlo, Luis Sarmento, Mario J. Silva, Eugenio De Oliveira - Clues for detecting irony in User-Generated Contents: Oh…!! It's "so easy" ;-). , 2009

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.

Pentreath, N., *(2015), Machine Learning with Spark, Amazon*

Peter Turney (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". "Proceedings of the Association for Computational Linguistics". pp. 417–424. arXiv:cs.LG/0212032.

Potts, C. 2011. Developing adjective scales from usersupplied textual metadata. NSF Workshop on Restructuring Adjectives in WordNet. Arlington,VA

Quintilien and Harold Edgeworth Butler. 1953. The Institutio Oratoria of Quintilian. With an English Translation by HE Butler. W. Heinemann.

R Socher, J Pennington, EH Huang, AY Ng, CD Manning - Semi-supervised recursive autoencoders for predicting sentiment distributions, Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011

R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," SIGKDD Explorations Newsletter, vol. 8, pp. 41–48, 2006.

Rachel Giora. 1995. On irony and negation. Discourse processes, 19(2):239–264.

Roberto Gonzalez-Ibanez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In ACL (Short Papers), pages 581–586. Citeseer.

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.

Shamma, D., Kennedy, L., Churchill, E.: Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of the first SIGMM workshop on Social media. pp. 3–10. ACM (2009)

Simon Colton - Inductive Logic Programming, Imperial college of London ?

Songbo Tan, Gaowei Wu, Huifeng Tang, Xueqi Cheng: A novel scheme for domain-transfer problem in the context of sentiment analysis, 2007

Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP pp. 53–63 (2011)

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2010). *Lexicon-based Methods for Sentiment Analysis.* Journal of Computational Linguists.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2011)

Tony Veale and Yanfen Hao - Detecting Ironic Intent in Creative Comparisons.

Top 2000 Vocabulary Words http://www.talkenglish.com/Vocabulary/Top-2000-Vocabulary.aspxhttp://www.talkenglish.com/Vocabulary/Top-2000-Vocabulary.aspx

Tsonkov and Koychev - Detecting Irony in texts from the social networks: the Bulgarian language case (2013).

Utsumi- A unified theory of irony and its computational formalization. In Proceedings of the 16th conference on Computational linguistics, pages 962–967, Morristown, NJ, USA, 1996. Association for Computational Linguistics

Vanin, A.; de Freitas L.; Viera R.; Bochernistan M. 2014 Some clues on Irony Detection of Tweets

Veale, T., and Hao, Y. 2010b. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. Minds and Machines 20(4):635–650

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

Yu, B.: An evaluation of text classification methods for literary study. Literary and Linguistic Computing 23(3), 327–343 (2008).

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 текстове от социалните мрежи”, декларирам, че резултатите и приносите са оригинални и че не са заимствани от изследвания и публикации, в които нямам участие.