

Софийски университет "Св. Климент Охридски"

Факултет по математика и информатика

Катедра "Компютърна информатика"

ДИПЛОМНА РАБОТА

Система за анализ на настроението в текстови онлайн
документи

Елица Иванова Павлова

Магистърска програма „Изкуствен интелект“, фак. номер 23928

Ръководител:

проф. д-р Мария Нишева

катедра "Компютърна информатика", ФМИ, СУ „Св. Климент Охридски“

София,
юни 2014

Съдържание:

1. Увод	3
1.1. Преглед на областта	3
1.2. Понятието анализ на настроението в текст	4
1.3. Задачи в областта	4
1.4. Съвременни системи и решения в областта	6
1.5. Оценка на точността на анализа на настроението в текст	10
1.6. Обобщение на съдържанието на дипломната работа	11
2. Система за анализ на настроението в текст	12
2.1. Решаван проблем и задачи	12
2.2. Преглед на избрания подход	13
2.3. Поставени цели	14
3. Анализ на настроението в текст чрез класификация	15
3.1. Формално представяне на данните	17
3.2. Използвани данни	18
3.3. Мерки за оценка	19
3.4. Изследване на избраните алгоритми за класификация	21
3.4.1. Наивен Бейсов класификатор	21
3.4.2 SVM	24
3.4.3. К най-близки съседа	29
4. Изследвани методи за подобряване на резултатите от класификацията	33
4.1. Оценяване и избор на атрибути	33
4.1.1. Филтриране на стоп думи	33
4.1.2. Избор на информативни атрибути	35
4.2. N-грами	40
4.3. Отрицателни семантични конструкции	42
5. Софтуерна реализация	44
5.1. Архитектура на приложението	44
5.2. Функционалност	46
5.3. Потребителски интерфейс	47
6. Заключение	56
7. Литература	57

1. Увод

1.1. Преглед на областта

Емоционалната оценка на хората за заобикалящия ги свят е важна в различните дейности на живота и в голяма степен определя човешкото поведение. Тя има основно значение за вземането на решения. Отношението на една личност към обект, концепция или събитие може да повлияе и на околните. То може да се предаде чрез жест, реч или чрез текст. Текстът носи отношението на пишещия към темата, а и в не малка степен може да повлияе на мнението на четящия.

Днес обменът на текстова информация между хората в уеб пространството е много улеснен. Навлизането на блокове, социални мрежи, форуми, различни онлайн системи за споделяне на мнения и коментари или оценяване на конкретни продукти създава пространство с огромен по мащабите си обмен на информация, наситена с емоционални оценки. Много често човек съзнателно търси личното отношение на другите в мрежата, когато прави избор - чете препоръки, интересува се от класирания на продукти или услуги.

Това ново пространство на информация представлява интерес за изследване не само от научна гледна точка, но и от практическа. С навлизането на споделянето на мнения в интернет начинът, по който маркетингът работи, също се променя. Както хората, така и организациите започват да вземат решения на база информацията и споделяните предпочитания в интернет пространството. Вече не е необходимото правенето на специални проучвания или анкети, за да се разбере мнението на хората. Интернет пространството и в голяма степен социалните мрежи предоставят голямо количество информация, от която бизнесът би могъл да се възползва анализирайки я. Данните в интернет пространството са много и голяма част от тях носят отношението и настроението на потребителите. Но поради големия обем е неминуема и нуждата анализът да бъде автоматизиран. Част от този анализ неизменно е и анализът на настроението в текст. Така тази предизвикателна задача от гледна точка на научните изследвания се превръща и в практически значима. За компаниите е от основно значение да знаят мнението на клиентите си относно продуктите или услугите, които те предлагат. Появяват се нови софтуерни решения и компании зад тях, чиято единствена задача е автоматичният анализ на настроението в уеб пространството и откриване на тенденции, които биха могли да послужат за маркетинга на различни продукти. Големите корпорации разработват и разполагат и със собствени решения в областта на анализа на мнението в текст. Такива са Microsoft, Google, Hewlett-Packard, SAP.

Маркетингът на продукти и услуги не е единствената сфера, пряко повлияна от възходът на социалните мрежи и интензивното споделяне на информация, носеща емоционално отношение или оценка. Днес сме свидетели на много течения в социално-политическата сфера, които се развиват най-вече в интернет пространството. Подобни тенденции водят до значими промени в живота на обществото – създаването на организации, оформянето на политическия облик на страни по света. Така анализът на настроението в интернет пространството има значение и намира своите приложения в политика, социални събития и организации, държавни институции, маркетинг на продукти и услуги.

1.2. Понятието анализ на настроението в текст

Анализът на настроението (*sentiment analysis*) или анализът на мнението (*opinion mining*) е научна област, която се занимава с извличане на знание за субективното отношение на човек, т.е. мнение, емоционална нагласа, отношение и оценка към определена тема, събитие, продукт, услуга, организация или друго. Изследвания в тази област усилено се правят от началото на века ([1], [3], [7]). Една от причините за това е именно натрупването в интернет пространството на необходимите данни за подобен тип проучвания. До този момент е липсвал необходимият обем от текстови документи. Другата движеща сила се явяват нуждите на развиващия се бизнес, който намира в анализа на настроението много различни приложения. Тласкана от нуждите на бизнеса, тази научна област започва да стъпва все по-уверено и в сфери като социология, политология и икономика.

Анализът на настроението в текст се занимава с извличане на субективната информация от текст чрез използване на способите на обработката на естествен език, анализа на текст, компютърната лингвистика, статистика и различни аспекти на изкуствения интелект. Основната задача, която се поставя при този процес е определяне на полярността на настроението – дали то е положително или отрицателно. Анализът цели да определи отношението на създателя на текста към темата на текста. Настроението в един текст може да се дължи на емоционална обвързаност на пишещия към темата или на оценка на пишещия към темата. Възможно е и в текста да бъдат съзнателно внесена емоция, която да цели да повлияе на четящия. Настроението може също така да бъде определено и като неутрално, ако няма никакви белези на субективна оценка към темата от страна на създателя на текста.

По своята същност поставеният проблем е от областта на обработката на естествен език. Анализът на настроението обаче не се нуждае от пълно семантично разбиране на разглеждания текст, а само на разбиране за отношението или емоцията в текста и обекта, към когото са изразени. Така на анализа на настроението може да се гледа като на ограничен проблем от областта на обработката на естествен език.

1.3. Задачи в областта на анализа на настроението

Разглеждат се няколко основни направления от гледна точка на задачи и проблеми, които могат да бъдат поставени в тази област. Един от важните аспекти е целта, която се поставя пред анализа – в каква форма е необходимо да бъде представен резултатът и как да бъде формулиран. Целта произтича от практическите проблеми, които се решават чрез изследването на настроението. Разглеждат се няколко различни групи от категории или йерархии от категории, които да бъдат цел на анализа. Основна цел на анализа на настроението е да класифицира текст като положителен, неутрален или отрицателен. Това е най-основният проблем, който може да се решава в тази област. Може да се постави и за задача да се определи не само полярността на текст, но да бъде класифицирано настроението в текст в повече категории - тъжно, развълнувано, гневно, щастливо и т.н. Това е доста по-усложнен вариант на класическия анализ на настроението. Друга посока, към която може да се

поеме, е настроението да бъде класифицирано в предварително зададени подредени категории в някаква скала – например петобална скала за оценка [11]. В подобни скали често се дава оценка в много сайтове за продукти или услуги, за оценяване на филми или музика. Друга от възможните насоки, към която може да се подходи, е определянето на текст в категориите субективен или обективен. Всъщност тук не се определя какво е настроението, а по-скоро дали то присъства и дали има субективизъм в текста. Тази задача е дори по-трудна от определяне на полярността на настроението, защото обективността или субективността на едни и същи думи обикновено зависи от контекста на тяхното използване, а много често анализът на настроението се прави главно на база на полярността на използвани отделни думи в текста без оглед на техния контекст или подредба. Съществува и понятието аспекти базиран анализ на настроението. Той има за цел не просто да класифицира настроението в целия текст, а да определи субективната оценка или мнение в текста относно предварително зададени термини или ключови думи. Разглежда се отношението спрямо някой от аспектите на темата на текста, а не просто отношението в текста като цяло. Би могло текстът да издава положителна емоция, но оценката спрямо конкретно понятие в него да бъде отрицателна. Всички изброени задачи, които могат да се поставят пред анализа на настроението, се решават чрез различни методи.

Друг аспект на категоризация в областта е според различните нива, на които може да бъде правен анализа. Бинг Лиу [1] определя три изследвани до момента нива:

- Ниво документ : цели се да бъде определено отношението в текста като цяло. Най-често задачата е да се определи в коя от двете категории, положителен или отрицателен, е текстът. Проблемът, който се решава тук е класификационен. Това е и най-често поставяният проблем от практическа гледна точка - вземайки някакво мнение, новина, блог, рецензия или друго системата за автоматичен анализ на настроението в текст да определи положителен или отрицателен е той.
- Ниво изречение : анализът се извършва на ниво изречение и категориите, в които се прави класификацията, най-често са положителна, отрицателна и неутрална. В неутралната категория попадат изречения, които не съдържат изразени чувства или отношение, а по-скоро носят фактологична информация. Тук идва на преден план и проблемът за класификация в категориите субективност и обективност. В категория субективни попадат не само изречения, в които е изразено някакво настроение, но и такива, в които се долавя мнение или отношение на автора. Както бе споменато, това е също интересна и трудна задача често зависеща от контекста, в който е употребено изречението. Анализът на ниво изречение може да се възползва от лексикалните особености и граматическата свързаност на думите в изречението и така да даде повече информация за контекста на отделните думи в текста.
- Ниво дума : анализът на емоционалния заряд на дума е сравнително по-лека задача от описаните до момента. Съставят се лексикони с думи и съответстващото им настроение или отношение с цел тези структури да се използват в семантичния анализ на текстове. Този подход има и своите очевидни недостатъци. Не винаги една силно емоционална дума може да определи настроението на цяло изречение и дори наличието на настроение.

Проблем е и различният контекст, в който се употребяват думите, наличието на ирония и др.

- Аспектно ниво : това ниво Бинг Лиу [1] отделя не на база на разбиване на текст на отделни части. В изброените до момента различни подходи към анализа на настроението в текст не се дава отговор какво всъщност отделният човек харесва или не. Дава се отговор единствено на въпроса какво е отношението в текста погледнат като цяло. На практика по-често ни интересува мнението спрямо конкретно понятие и с това именно се занимава аспектното ниво на семантичния анализ. При анализа на мнението се отделят две основни понятия – настроение (положително или отрицателно) и обект на изразеното мнение. Отношение без намерен обект има малка полза. Едно изречение би могло да носи отрицателно настроение като цяло и въпреки това мнението към конкретен обект в него да бъде положително. Обектите на мнението се изразяват чрез същности и техните аспекти. Така анализът на това ниво е съсредоточен върху анализ на настроението спрямо същности и техните аспекти. Счита се, че това ниво има най-предизвикателните и трудни задачи в сравнение с анализа на ниво изречение или на ниво документ.

Като област с нарастващ научен и комерсиален интерес днес са разработени много различни методи и подходи към проблема за анализ на настроението на текст.

1.4. Съвременни системи и решения в областта

Автоматизираният анализ на настроението се базира на методи на машинното самообучение, техники за обработката на естествен език и статистика. Прилага се на практика върху големи обеми от данни – потребителски рецензии за продукти или филми, мнения в социалните мрежи, новини, коментари, блогове. Основният обект на анализа на мнението често са именно текстове, които могат да бъдат намерени онлайн.

Разгледан е автоматизираният анализ на настроението, защото обемите от данни, с които трябва да работят подобни системи, предопределят необходимостта от машинни ресурси за тази задача. Различните платформи за извличане на настроението от текст работят на база на различни научни методи и технологии. Част от тях са напълно автоматизирани, а други разчитат отчасти и на човешка намеса – началните данни, с които се обучава система са класифицирани от хора, или се отчита и анализира грешката в системата чрез съпоставяне с резултати дадени при анализ от хора. Основната част от работата обаче е автоматизирана. Съществуват и компании, които все още използват служители за следенето и извличането на мнението от текст, но те са все по-малко. Макар да се твърди, че точността на анализа от експерти е по-добра, някои съвременни решения предлагат сравними по-точност резултати. А количеството информация, което могат да следят и обработват не може да бъде сравнено с капацитета на анализа правен от хора. Дневно в съвременните социални мрежи се създава и запазва значителен обем от текстове. Например социалната мрежа Twitter има около 200 милиона активни потребители, които създават повече от 400 милиона кратки текста на ден. Голямата част от информацията в мрежите е публична и дава възможност на анализатори и компании да се възползват от нея.

Правени са проучвания и съпоставки на съвременните системи за анализ на настроението в текст [3]. Обусловени са два най-често използвани подхода за решаване на задачата – чрез използването на техники на машинното самообучение или чрез базирани на речници методи. Най-изследваните подходи в областта на машинното самообучение са използването на алгоритми от тип обучение с учител. Класификаторите изискват данни, които са предварително определени в категориите, в които се прави класификацията. Най-често това са категориите положителен или отрицателен текст. От особена важност е да бъдат подбрани подходящи данни за областта, в която се работи. Един от недостатъците на този подход е именно необходимостта от класифицирани предварително данни, което в някои области може да бъде скъпо или дори невъзможно. Предимство е, че един класификатор може чрез данните да бъде лесно адаптиран към конкретна област или контекст. Подходът с използването на техники на машинното самообучение ще бъде разгледан в настоящата работа. Другият разпространен метод е чрез използването на лексикално-базирани техники, които използват предварително дефинирани лексикони или речници от думи, в които всяка дума е асоциирана с определено настроение. Той е силно зависим от контекста на проблема, който се решава. Трудността тук е в създаването на лексикон, който да бъде използваем и точен в различни контексти. Представен е само общ обзор на подходите в съществуващите системи за анализ на настроението. Причината за това е, че малко се знае и е документирано за тях, тъй като те основно са комерсиални приложения и платформи. Трудно е и да се направи съпоставка и оценка, защото различните методи и приложения дават различни резултати в зависимост от конкретната задача, данните, върху които работят, контекста.

Описани са няколко популярни платформи и решения в областта на анализа на настроението в текст. Това са съществуващи в момента решения. Част от тях са комерсиални, други са отворени за употреба и разработвани предимно с научна цел.

Един от най-опростените подходи, използван в самото начало на опитите за анализ на настроението в текстове от социалните мрежи, е отчитане на наличието на емотикони в онлайн текстовете. Прави се предварителен речник на емотиконите и с всяка от тях се асоциира настроение. Този подход дава доста добра точност на практика, но има много малък обхват. Той може да работи само върху тестове, в които има емотикони – по-малко от 10% от текстовете, които се създават в социалните мрежи днес. Често той се съчетава с други методи на машинното самообучение.

LIWC (Linguistic Inquiry and Word Count, <http://www.liwc.net/>) е инструмент за анализ, който извлича емоцията от текст, както и структурните компоненти на текста. Той се базира на речник от думи и съответстващите им категории. Освен като положителен и отрицателен текст може да бъде определен и в повече и по-специфични емоционални категории. LIWC е комерсиален софтуер. Той предоставя възможност на своите потребители и да допълват използвания речник.

SentiStrength (<http://sentistrength.wlv.ac.uk/>) е софтуер, който може да бъде използван с комерсиални цели, но е свободно достъпен. Той комбинира даващите най-добри резултати техники на машинното самообучение. Авторите изследват множество класификатори спадащи към обучение с учител и обучение без учител – метод на опорните вектори (SVM), J18 класификационно дърво, Наивен Бейсов класификатор, многослоен перцептрон, AdaBoost, регресия чрез метод на опорните вектори и др. Те стъпват на базата на речника от думи в LIWC като го допълват. За тестване на

SentiStrength се използват данни от шест различни източника – коментари в YouTube, форума на BBC, форума на Runners World, Digg, Twitter, MySpace.

SentiWordNet (<http://sentiwordnet.isti.cnr.it/>) е софтуерно приложение за анализ на мнението в текст, което се базира на английски лексикален речник наречен WordNet. Този речник групира съществителните имена, прилагателните имена, глаголите и останалите граматични класове в множества от синоними. SentiWordNet асоциира по три числови резултата с всяко такова множество от синоними за да реши задачата за определяне на текста в категориите положителен, отрицателен и неутрален. Това асоцииране става на база на обучен класификатор. Трите числови резултата съответстват на трите категории и са между 0 и 1 като се сумират до 1. На всяка дума се дава тройка коефициенти за резултат, като категорията на думата се определя от най-голямото число в тройката.

SenticNet (<http://sentic.net/>) е метод за извличане на мнението и настроението от текст, който използва алгоритми и техники от областта на Изкуствения интелект и Семантичния уеб. SenticNet разглежда текста като го разбива на семантично значими единици, за разлика от SentiWordNet, който разчита на синтактични единици. Използвани са техники от обработката на естествен език за да се даде полярност на настроението на над 14 000 от най-употребяваните концепции в английския език. Анализът на настроението SenticNet първо прави като открива в текста концепциите, които разпознава, а след това оценява настроението на всяка една от тях чрез число между -1 и 1. Настроението в текста като цяло е обобщение на настроение в откритите в текста концепции. SenticNet е използван и тестван като инструмент за оценяване на полярността на мнението за здравната система в Обединеното кралство. Тестван е и върху блокове от LiveJournal.

SASA (SailAil Sentiment Analyzer, <https://code.google.com/p/sasa-tool/>) е платформа използваща методи от машинното самообучение. Тя има и свободна за използване версия. Тестван е върху 17 000 категоризирани предварително съобщения в Twitter за изборите в САЩ през 2012. Свободната за използване версия на платформата е тествана чрез Amazon Mechanical Turk (AMT) върху съобщения от Twitter.

Happiness Index [4] е решение, което използва популярната Affective Norms for English Words (ANEW). ANEW е колекция от 1 034 често използвани думи в английския език асоциирани с оценки за различни техни измерения. Happiness Index измерва щастието в текст по скала от 1 до 9 като се базира на речника от думи в ANEW. Системата е била използвана върху изречения от блокове, текстове на песни и заглавия на песни. Интересни са и получените резултати. Според тях нивото на щастие в текстовете на песни в периода от 1961 до 2007 е намаляло, а обратната тенденция се наблюдава в блоковете.

PANAS-t [5] е психометрична скала предложена за откриване на настроенията в социалната мрежа Twitter. Методът се базира на Positive Affect Negative Affect Scale (PANAS), който е добре познат в психологията. PANAS-t използва голямо множество от думи определени в единадесет категории за настроение - веселост, увереност, спокойствие, изненада, страх, тъга, вина, враждебност, срамежливост, умора, и внимание. Платформата е създадена да следи измененията на настроението през времето, т.е. да проследява времеви тенденции.

Разгледани са още платформите Sysomos и Radian6 като пример за съвременни решения в областта категоризацията на текст и анализа на настроението в текст, които са изключително печеливши и известни в своята област. Те нямат научна цел, а единствено комерсиална такава.

Sysomos (<http://www.sysomos.com/>) е една от най-известните платени платформи, която се използва реално от маркетинга на редица известни компании. Ebay официално потвърди използването на Sysomos. Малко се знае за начина, по който е изградено това решение тъй като е изцяло комерсиално и насочено към бизнеса, маркетинга на различни марки, агенции и др, но то предлага редица впечатляващи функционалности. Компанията използва данни от Twitter, Facebook, YouTube, блогове и форуми за да създаде реална във времето картина на отношението в интернет пространството към определени продукти, марки, хора. Предлага статистики, анализи и решения с висока бизнес полза. Два са основните продукти, които компанията предлага. Media Analysis Platform (MAP) анализира данни от социалните мрежи. Това е основният продукт, който е сертифициран от социалната мрежа Twitter. Sysomos Heartbeat предлага следене в реално време на тенденциите в социалните мрежи.

Radian6 (<http://www.salesforcemarketingcloud.com/>) предлага функционалност подобна на Sysomos. В областта на мониторинга на социалните мрежи вече съществуват доста решения. Radian6 е едно от разпространените такива, част от платената платформа Salesforce Marketing Cloud. Тази платформа насочва своите услуги изцяло към маркетинга на марки, компании, индустрии. Radian6 е тази част от нея, която се занимава със следене на тенденциите в социалните мрежи – Twitter, Facebook, блогове и форуми. Разбира се, подобно на Sysomos, Salesforce Marketing Cloud предлага много по-богата функционалност освен анализ на настроението в текст, но то също е част от тези решения.

Внимание заслужават и платформите, които предоставят програмен интерфейс за анализ на настроението в текст. Има комерсиални или свободни за използване решения, които дават възможност за интегриране на потенциала на анализа на настроението в текст в други софтуерни проекти. Такива са

- Semantria (<https://semantria.com/>) – модерно и бързо развиващо се решение в областта на обработката на естествен език, базирано на Lexalytics' Saliency ядрото;
- AlchemyAPI (<http://www.alchemyapi.com/>) – може би най-популярната в света библиотека и платформа за анализ на естествен език и анализ на настроението в текст, която предоставя възможност за работа с най-разпространените езици за програмиране;
- Chatterbox (<http://chatterbox.co/>) – използва алгоритми от машинното самообучение за анализ в областта на социалните мрежи;
- Viralheat (<https://www.viralheat.com/>) – платформа за наблюдение на тенденциите в социалните мрежи и семантичен анализ;
- Bitext (<http://www.bitext.com/>) - семантична технология, предоставяща и платформа за анализ на настроението, която заявява резултати с най-добрата точност сравнение с всички останали известни до момента решения на пазара.

Всички изброени предоставят т.нар. REST API, който предоставя възможност на практически всеки език и платформа за програмиране да се възползва от

възможностите, които те предоставят в областта на семантичния анализ и анализа на настроението в текст. Огромните бази от данни и речници, с които работят тези платформи прави невъзможно тяхното използване от други софтуерни приложения без наличието на REST API.

1.5. Оценка на точността при анализ на настроението в текст

Една автоматизирана система за анализ на настроението в текст не би могла да бъде по-точна от анализ правен от хора. Хората могат да отчитат много по-голям диапазон на различни настроения и мнения. Лесно е за човек да улови нюансите на контекста или да разпознае ирония в писмен текст. Но хората не винаги се съгласяват един с друг относно това какво настроение носи текст. Без предварителни указания или контекст хората също не се справят без грешки. Оценката на настроението в текст, когато тя се дава от група от хора, съпада само в 80% от случаите [2]. Оценката за една система за анализ на настроението се базира на това до колко тя се доближава до оценката на човек. Тогава макар и малък, 80% точност е отличен процент за една автоматизирана система. Тук трябва да се вземе предвид както факта, че средно човешката оценка така или иначе не би съвпаднала в по висок процент, така и че системите за анализ на настроението се използват в области и за задачи, които често не изискват по-висока точност. Много от съвременните платформи, които бяха описани, постигат точност близка до 80% и дори по-добра.

Правени са научни изследвания за процента съвпадение на човешката оценка относно настроението в текст [6]. Учените Уилсън, Уилба и Хофман от университета в Питсбърг установяват 82% на съгласие между двама души при категоризирането на субективни твърдения от избрано тестово множество от документи. На таблица 1.1 са представени резултатите от изследването.

	Неутрална	Позитивна	Негативна	И двете	Общо
Неутрална	123	14	24	0	161
Позитивна	16	73	5	2	96
Негативна	14	2	167	1	184
И двете	0	3	0	3	6
Общо	153	92	196	6	447

Табл. 1.1. Резултати от изследването на Уилсън, Уилба и Хофман за съгласието между двама души относно настроението в текст

Предварително зададените категории, в които трябва да се оценят фрази, са неутрална, позитивна, негативна, и двете. Документите, върху които е правен анализът, са политически статии и новини. Самите твърдения, които са използвани за определяне на настроението в тях, са под формата на фрази и кратки изречения.

Майк Маршал от Lexalytics, компания предлагаща софтуер за анализ на текст, също провежда подобно изследване. То се отнася до определяне на настроението на ниво текст и показва точност на човешката оценка от 81,5%. Близко 81% от положителните документи са вярно определени като такива и 82% са вярно

определените от отрицателните. Така хората, които анализират и анотират текст, са склонни да се съгласяват един с друг в близо 80% от случаите. Тогава не бихме могли да очакваме по-висок процент на съгласие на човек с получените от автоматизирана система резултати в областта на определянето на настроението в текст.

Често точността на автоматизираните системи е под 80% и все пак те са предпочитани пред използването на хора за същата задача, защото притежават редица преимущества – цена, бързодействие, възможност за обработка на много данни за обозримо време, достъпност, възможност за работа на различни езици.

1.6. Обобщение на съдържанието на дипломната работа

Настоящото изложение представя обобщение на проучванията в областта на анализа на настроението в текст и преглед на реализираното софтуерно решение. Глава втора обобщава целите и задачите на дипломната работа. Дава се кратък преглед на избрания подход за решаване на проблема за анализ на настроението в текстови онлайн документи. Глава трета описва формалното представяне на проблема. Разглеждат се използваните данни и тяхното представяне, дава се аргументация за направения избор на данни. Представят се мерките за точност, които оценяват разгледаните алгоритми. Дава се обобщение на избраните алгоритми и се представят статистики за тяхното поведение при реално решаване на поставения проблем. Глава четвърта описва различните подходи, които са проучени с цел подобряване на класификационното поведение на избраните алгоритми. Отново са изведени статистики с получените резултати. Глава пета описва изготвеното софтуерно приложение, неговата архитектура и функционалност. Последната шеста глава е кратко обобщение на постигнатите резултати и възможните подобрения на реализираното решение.

2. Система за анализ на настроението в текст

2.1. Решаван проблем и задачи

Развитието на интернет пространството и улесненият достъп до него правят използването му всекидневие за много хора по света. Една от най-често извършваните дейности е търсенето използвайки различни онлайн платформи за търсене (Google, Bing, Yahoo, Ask) и посещаването на намерените сайтове. Възможно е освен от фактологична информация потребителят на интернет пространството да се интересува и какъв е емоционалният заряд на намерените от него резултати. Именно тази задача е водеща в настоящата работа и тя бива решавана от разработената система за анализ на настроението в онлайн текстови документи.

Един от най-разпространените шаблони за използване на интернет пространството е търсене по ключови думи на интересуващото потребителя и разглеждане на получените резултати. Това означава проследяване на интернет връзки (линкове) и разглеждане на сайтове, в които потребителят отново търси ключовите думи, описващи неговите интереси. Това е добре позната рутина на работа за всички потребители на интернет. Настоящото изложение обръща фокуса на вниманието към анализа и извличането на настроението в онлайн текстовете, които потребителят преглежда. Така бива разглеждано още едно измерение на потребителските интереси – какво е настроението спрямо тях в намерените онлайн текстове.

Задачите, които решава настоящата работа, са:

- Определяне на настроението на описанието за всеки един от релевантните резултати от търсенето на потребител използвайки определена онлайн платформа за търсене в интернет пространството;
- Намиране на пасажите от онлайн текстови документи, които съдържат ключовите думи на търсенето на потребител, и определяне на настроението в тях;
- Предоставяне на възможност за търсенето в един сайт по определени ключови думи (търсене в един и същи http домейн) и отново извличане на настроението на пасажите, съдържащите търсените ключови думи;
- Предоставяне на възможност за специфично търсене на новини относно интересите на потребителя и анализ на настроението в тях;
- Категоризиране на настроението в текст като положително, неутрално или отрицателно;
- Изготвяне на отчет на намерените резултати и систематизирането им според настроението в тях;
- Предоставяне на възможност за проследяване на времеви тенденции и изменения на настроението при търсене на едни и същи ключови думи и извеждане на помощни статистики;
- Запазване на резултатите от анализа на настроението на онлайн текстовете от търсенията, важни за потребителя;

Цели се обогатяване на възможностите за използване на интернет пространството чрез предоставяне на система, която да извлича и анализира

настроението в онлайн текстовете, които интересуват потребителя. Освен фактологичната информация ще бъде предоставен и поглед върху емоционалния заряд на темите, събитията, личностите и въобще обектите, които се търсят онлайн. Анализът на настроението на текстовете отнасящи се до определени теми може да бъде полезен за проследяване на тенденции, откриване на предпочитания в общественото мнение или промени в него, откриване на неочаквани нагласи. В зависимост от сферата на интереси подобен анализ би могъл да допринесе за политически и социологични изследвания, маркетинг на продукти, услуги или марки.

2.2. Преглед на избрания подход

Задачата, която се решава в настоящата работа, е извличане на настроението от онлайн текстове, намерени за определена тема, описана чрез ключови думи, и изготвянето на отчет от получените резултати. Първата стъпка е извличане на онлайн текстова информация. Втората съществена част е анализът на настроението в текст.

Разработен е уеб сайт, който дава възможност на потребителя да търси по ключови думи в онлайн платформа за търсене, в определена уеб страница или в сайт. Резултатите от търсенето в платформата за търсене са кратки описания и връзки към уеб страници. Резултатите от търсенето в сайт са откъси текст, които съдържат търсените ключови думи. И в двата случая от всеки от резултатите се извлича настроението в три категории - положително, неутрално и отрицателно. Разглеждат се само резултати на английски език, което е едно от съществените ограничения на системата. Уеб сайтът предоставя отчет за извършеното търсене и го запомня с неговите параметри – дали е търсене чрез платформа или в уеб сайт и при какви настройки от страна на потребителя. Предоставя се възможност за запазване на отчета в pdf файл и визуализирането му при необходимост. Търсенията на всеки потребител се пазят в история на търсенето, от където да могат да бъдат разгледани детайли за всяко едно от тях. Дава се възможност за визуализиране на статистики и проследяване на изменения в откритите настроения спрямо едни и същи ключови думи на търсене. Така могат да бъдат забелязани интересни тенденции. За осъществяване на изброената функционалност са разработени няколко софтуерни модула и са използвани и външни библиотеки. Осъществяването на търсенето по ключови думи в различните му варианти, използването на платформата за онлайн търсене, извличането на текст от HTML документ, извеждането на статистики и тенденции са проблеми, чиято софтуерна реализация ще бъде описана в глава 5 на настоящото изложение.

Съществена част от разработената система е анализът на настроението в текст. То се прави от отделен независим софтуерен модул. Подходът към този проблем може да бъде различен, както вече бе описано в уводната част. Поставената задача е текстът да бъде определян в три категории – положителен, неутрален и отрицателен. Категоризацията на текст се извършва на два етапа. На първия той е определян като субективен или обективен. Категорията субективен означава, че текстът носи някакъв емоционален заряд и в него може да бъде открито настроение и отношение на създателя на текста към темата. Обективен е текст, който носи фактологична информация, и това е всъщност категорията неутрален. Ако един текст бъде определен като неутрален, неговата категоризация приключва. Ако той е определен като

субективен, то той се подлага на нова категоризация, която да даде дали той е позитивен или негативен.

Според поставения проблем и произтичащите задачи е избран анализ на настроението на ниво изречение или пасаж като се разглежда цялостното настроение, което носи откъсът от текст. Подходът за категоризацията на текст в някой от описаните класове е чрез използването на методи на машинното самообучение от тип учене с учител (*supervised learning*). Задачата за определяне на текст в категориите обективен и субективен, положителен и отрицателен се свежда до задачата за класификация. Обучени са два класификатора със съответните данни и те извършват класификацията на тестове в описаните категории. Първият класификатор казва дали текстът е неутрален и ако не е, то вторият класификатор го определя като положителен или отрицателен. Това е най-общото описание на избрания подход за решаване на задачата за анализ на настроението в откъси от онлайн документи.

2.3. Поставени цели

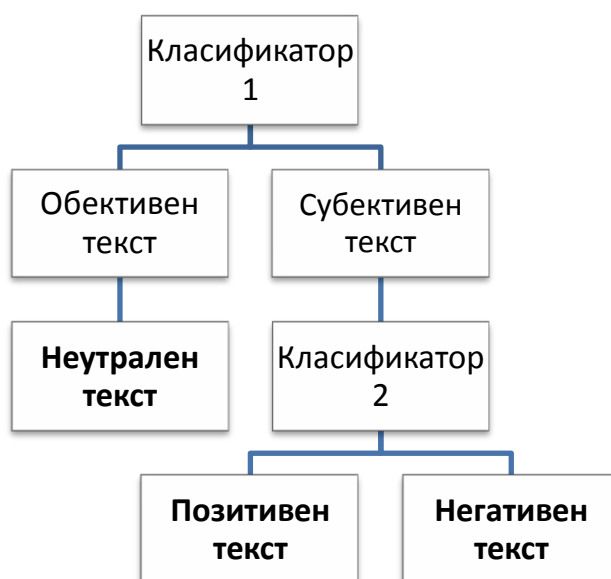
Основните цели на настоящата дипломна работа могат да бъдат формулирани както следва:

- Изследване на методи за обработка на естествен език с цел получаване на текстовата информация във вид удобен за автоматизирана обработка;
- Анализ, тестване и избор на класификационен алгоритъм за изграждане на класификатори, които да разпознават съответно обективност или субективност на текст и положително или отрицателно настроение в текст;
- Изследване на различни техники на машинното обучение и статистиката за подобряване точността на класификаторите;
- Създаване на онлайн система за анализ на настроението в текстове, получени като резултат от търсене на потребител по ключови думи в определена уеб страница или в онлайн платформа за търсене.

3. Анализ на настроението в текст чрез класификация

Използването на класификация за определяне на настроението в текст е само един от възможните подходи за решаване на тази задача. Анализът на мнението в текст е обширно изследвана тема, с множество публикации в последните години. Описани са различни начини за решаването на проблема и различни постановки на задачата.

Задачата, която е разгледана в настоящата глава, е анализ на настроението в кратък текст или откъс от текст. То трябва да бъде определено в категориите положително, неутрално или отрицателно. Според избрания подход се използват два обучени класификатора, които работят един след друг. Първият определя дали разглежданият текст е обективен или субективен. Ако той е определен като обективен, му се дава категорията неутрален и категоризацията приключва. При субективен текст, той се подава на втория класификатор, който дава дали настроението или изразеното мнение в текста е положително или отрицателно. Етапите на работа са представени на фигура 3.1. Тази йерархична схема на класификация на два етапа е избрана пред варианта за класификация в три класа. Една от причините за избора е, че се разчита точността на класификацията да бъде по-висока при класификатори обучени да разпознават по-специфични понятия. Подходът за определяне на субективността на текст преди да се определи полярността на настроението е тема на проучвания с доказано добри резултати ([6],[9]) и се използва при анализа на настроението в текст в някои библиотеки за работа с естествен език [8]. Панг и Ли [9] показват, че класификацията на полярността на настроението значително се подобрява, ако тя работи само върху субективни текстове, а обективните са отсети предварително. Двамата учени разработват модел с определяне на субективността на ниво изречение и определяне на полярността на ниво текст. Тъй като текущата работа има за задача анализ на кратки текстове и крайна класификация в три класа, то както определянето на субективността, така и на полярността, се прави на едно и също ниво – кратък текст или изречение.



Фиг.3.1. Етапи на класификация в разработената система

Извършва се анализ на настроението в текста като цяло. Не се разглежда определен обект, спрямо който да се търси отношението в текста, т.е. аспектично ниво на анализ на настроението в текст. Затова при решаването на задачата се правят няколко предположения [1]. Първото е, че текстът се отнася до единствена тема. В конкретния случай, който разглежда настоящата работа, текстовете се извличат на база на срещането в тях на търсени ключови думи. Ако потребителят търси с опцията за използване на онлайн машина за търсене, се разчита и на нейните функционалности за даване на релевантни резултати по темата на търсене. Така предварителната обработка на текстовете в тази ситуация гарантира до голяма степен, че текстът се отнася до единствена тема или засяга темата на търсене от потребителя. Второто предположение, което се прави при анализ на настроението в текст като цяло, е че има единствено изразено мнение и позиция и то носи единствен тип емоция. Това не може да бъде сигурно от дефиницията на задачата, която решава текущата система, но краткостта на текстовете предполага, че те няма да съдържат многоплатови мнения за темата на текста. Разбира се, при нарушаване на горните предположения избраният подход се оказва по-малко ефективен и точен. Не би могло да се очакват добри резултати, ако се разглежда настроението в текст като цяло, а той съдържа няколко мнения с различен емоционален заряд отнасящи се към различни обекти.

Проблемът за определянето на настроението в текст може да бъде разглеждан като специфична форма на класификацията на текст. Най-разпространените алгоритми за класификация на текст са от типа обучение с учител. Те могат да бъдат приложени и за по-специфичната задача за класификация на настроението в текст. Изследваните алгоритми в настоящото изложение също са от типа обучение с учител.

Използването на обучение с учител за анализ и извличане на настроението от текст е добре изследвана област. Съществуват много научни разработки, които са посветени именно на този подход за решаването на проблема. Изследвани са различни алгоритми. Първата разработка, която използва техниките на машинното самообучение с учител, е на Pang, Lee и Vaithyanathan [7]. Изследваните от тях алгоритми са Наивен Бейсов класификатор и метод на опорните вектори. В последващите множество проучвания биват изследвани много други алгоритми и методи за подобряване на класификационната им точност. В голяма част от разработките в областта използваните алгоритми обаче отново са Наивен Бейсов класификатор и метод на опорните вектори (SVM).

От гледна точка на класификацията проблемът за определяне на настроението в текст има следната постановка. Създава се модел (класификатор), който при подаване на текст, определя към кой клас принадлежи. В зависимост кой от двата класификатора се разглежда, класовете при решаваният тук проблем са обективен и субективен текст или положително и отрицателно настроение в текста. Текстовете представляват примери, описвани с атрибути или още наричани признаци. Примери, които са предварително класифицирани, се използват за обучението на модела – т.нар. обучаващо множество. След като той е обучен може да определя класа на непознат до момента пример. В тази постановка на задачата очевидно атрибутите, с които биват описвани примерите, са от голяма важност за работата и точността на класификатора. Тъй като примерите са откъсци от текстове в уеб страници, то е важно да бъдат избрани подходящи атрибути, с които те да бъдат представяни формално. В проблемната

област категоризация на текст са възприети и изследвани няколко схеми за представяне на текст във вид на вектор от атрибути.

3.1. Формално представяне на данните

В класификацията на текст основно затруднение е работата с неструктурирани данни – именно свободен писмен текст. За нуждите на класификационните алгоритми е необходимо текстът да бъде представен в структуриран вид. Той трябва да бъде представен чрез множество от атрибути и избраното представяне има ключово значение за работата на класификационните алгоритми. Подборът на подходящи признаци, с които да бъде описан даден пример, се нарича извличане на атрибути (feature extraction). Начините, по които може да бъде направено извличането на атрибути при класификацията на текст, са няколко:

- Атрибутите, с които се описва всеки текст, да бъдат думите в текста, а стойностите на атрибутите да бъдат броя срещания на съответната дума в текста. Освен думите като признаци могат да бъдат разглеждани и т.нар. n -грами. Те са последователности от n думи в реда, в който се срещат в текста. Всяка n -грама се разглежда също като отделен атрибут. Една от възможностите за стойности на атрибутите- думи е мярката *tf-idf* позната от областта извличане на информация. Тази мярка дава относителна стойност за това колко е важна думата в текст на база на статистики. Показано е обаче, че *tf-idf* е мярка подходяща за големи текстове, каквито не биват разглеждани в настоящата работа. Друга възможност е атрибутите да бъдат бинарни и техните стойности да показват отсъствие или наличие на дума.
- При подбора на атрибути може да се вземе предвид каква част на речта са съответните думи. Логично би било при определяне на настроението в текст прилагателните имена да имат по-голяма тежест.
- Има думи във всеки език, най-често прилагателни имена и наречия, които семантично носят позитивен или негативен смисъл. Ако има предварително изработен списък от такива думи и фрази, те могат да бъдат използвани при подбора на атрибутите, с които се описва текст.
- Думи или конструкции в даден език имат свойството да сменят настроението в определена фраза. Такива са отрицанията. Добре е подобни конструкции да се взимат предвид при избора на атрибути.
- Също проучван подход към избора на атрибути при класификацията на текст е да се вземат предвид синтактичните зависимости на думите.

Макар и най-опростеният, изборът на атрибути думите в текста и техни стойности честота на срещане на тези думи, се оказва ефективен и често използван подход в класификацията на текст както и в класификацията на настроението в текст. Текущата разработка използва именно този подход. Поради краткостта на текстовете, с които работи разработената система, е направен избор атрибутите да бъдат бинарни. Също така броят срещания да дума в текст има основно значение при определяне на темата на текста, но не би имала главно значение за определяне на текст като обективен или субективен или като положителен или отрицателен. Разглеждат се като атрибути униграми, т.е. отделните думи. Стойностите на атрибутите са 0 или 1, т.е. наличие или отсъствие на дума. Проучена и тествана е възможността за използването

на биграми. Също е разгледана възможността за отчитане на отрицания и отрицателни конструкции в текста. Тези проучвания, подобренията и значимостта им, са описани подробно в следващата глава на изложението.

Така първата стъпка на работа е получаването на речник или още вектор от думи с техните честоти на срещане. Този подход често се среща под термина *“bag of words”*, който може да бъде преведен като множество или торба от думи. Смисълът е, че при този метод се губи информацията за последователността на думите в текста и така се губи и важна семантична информация за текста.

Нека разгледаме следния откъс от текст:

„The weather outside was nice and sunny. There were fluffy clouds in the sunny sky.”

Първата стъпка на работа е отделянето на думите в текста. Това е процесът на токенизация. Използван е доста лесен модел – думите се разделят по интервали, а препинателните знаци са игнорирани. Не биват разделяни думи като „don’t”, „won’t”, защото те имат значение за внесеното в текста настроение и трябва да бъдат разгледани като самостоятелни атрибути. Всички думи се преработват така че да се изписват с малка буква.

Втората стъпка е текстът да се представи във вид на речник или вектор от думи от типа:

{ “the” : 1, “weather” : 1, “outside” : 1, “was” : 1, “nice” : 1, “and” : 1, “sunny” : 1, “there” : 1, “were” : 1, “fluffy” : 1, “clouds” : 1, “in” : 1, “sky” : 1, }

Останалите елементи във вектора са думите, които присъстват в обучаващото множество текстове, но не и в конкретния пример. Те имат стойности 0 във вектора.

Първоначалните разглеждания са направени с този опростен модел за извличане на атрибути от текст.

3.2. Използвани данни

Избраният подход е обучение с учител. Затова от изключителна важност е подборът на данни, с които да бъдат обучени двата класификатора.

Поставената задача изисква класификаторите да работят с откъси от текст от уеб сайтове или с описанието на резултатите, които връща избраната платформа за онлайн търсене. Това означава, че в по-голямата част от случаите ще се работи със съобщителни изречения и езикът в тях ще бъде сравнително формален. Това е езикът, който може да бъде срещнат в голямата част от сайтовете в интернет пространството днес макар и това предположение да е не винаги вярно. Тогава трябва и класификаторите да бъдат обучени използвайки именно подобни данни. Поставената задача не ни позволява да ограничим текстовете от обучаващото множество в някоя конкретна сфера или по други критерии.

Формалността на езика е едно от важните изисквания. Класификаторите трябва да работят и с новини тъй като това е една от функционалностите на разработената система. Езикът в новинарските сайтове има дори по-голяма строгост на изказа от средната в интернет пространството. Тогава въпреки многообразието от обучаващи множества с текстове от социалните мрежи, те не биха били подходящи. Съобщение в социалната мрежа Twitter изглежда по начин сходен на показания на фигура 3.2.



Фиг.3.2.Съобщение в Twitter

Съобщението съдържа много линкове и т.нар. тагове за търсене (# tags). Изказването е на разговорен английски език. Обучавайки класификатор с подобни текстове, а в последствие използвайки го за друг тип текстове, би дало не добри класификационни резултати.

Избраното множество за обучение на класификатора, който разпознава обективност в текст, е свободно за използване. То съдържа текстове със сравнително формален изказ и предимно съобщителни изречения. Текстовете представящи класа обективен текст описват части от сюжети на филми. Текстовете представящи класа за субективен текст са рецензии на филми, в които е вложено отношение на пишещия към темата. В двата класа има по 5000 примера за обучение. Същото множество е използвано и в разработката на Панг и Лии [9]. То е избрано и заради размера на примерите, които по дължина на текста отговарят на очакваната дължина на онлайн откъсите, с които ще работи текущата система. Избраните данни съдържат текстове отнасящи се до филми и най-често срещаните думи са от типа „actor“, „actress“, „movie“, „plot“. Те се срещат във всички текстове, т.е. са атрибути във всички примери и не би трябвало да влияят на класификацията за обективността на текст. Очаква се тази специфика на обучаващото множество да не повлияе силно на класификацията на новите примери, върху които реално ще се работи (т.е. откъси от онлайн текстове) макар те да не са от същата проблемна област.

Данните, с които е обучен класификаторът на полярността на настроението, са от две множества. Те също се разпространяват свободно. И двете съдържат текстове, чиято тема са филми. Въпреки спецификата на областта и за тези две множества ще бъде направено изложеното горе предположение, че не се намалява драстично точността на класификатора обучен с тях. Множествата съдържат текстове със съобщителни изречения. В тези текстове са наблюдава по-голяма не формалност на изказа по причина, че те всички имат емоционална натовареност – положителна или отрицателна. Едното множество съдържа тестове с дължина в рамките на едно или две изречения, като позитивните текстове са 5331 и негативните са 5331. Другото множество съдържа сравнително по-дълги текстове. Броят примери в класа на негативните и позитивните е по 1000. И двете множества са за първи път използвани в публикациите на Панг и Лии ([7], [11]). Много последващи проучвания използват именно тези корпуси документи.

3.3. Мерки за оценка

При задачата класификация е важно предварително да бъдат определени критериите и мерките, с които ще бъдат оценявани използваните алгоритми.

Класификаторите са обучени с 3/4 от данните и тествани с 1/4 от тях. Данните за тестване не са използвани в процеса на обучение на класификаторите, за да се даде по-

реална оценка на точността им. Данните, с които се тества обучен класификатор, са т.нар. тестово множество. Не е разгледана по-сложна схема за тестване като кросвалидация или други, защото данните за обучение на класификаторите са сравнително достатъчно.

Основната мярка за оценяване на класификаторите е тяхната точност (*accuracy*). Точността е отношението на правилно класифицираните примери спрямо всички примери, които са подадени за класификация от тестовото множество.

За да има по-добър поглед върху работата на класификационните алгоритми не е достатъчно да бъде отчитана само тяхната точност. Тъй като се разглеждат класификатори, които оценяват пример в два възможни класа, то подходящи мерки за тяхната работа са прецизност (*precision*) и възвращаемост (*recall*). Те се изчисляват за всеки от двата класа на съответния класификатор. Мерките *precision* и *recall* произлизат от областта извличане на информация. В рамките на машинното самообучение се използват при бинарна класификация.

Ако модел класифицира в два класа - позитивен и негативен, то резултатите могат да се представят по схемата показана на таблица 3.1.

	Примери от клас „+”	Примери от клас „-”
Класифицирани като „+”	<i>Tp</i>	<i>Fp</i>
Класифицирани като „-”	<i>Fn</i>	<i>tn</i>

Табл.3.1. Разпределението на примери при класификация в два класа

tp е броят примери класифицирани правилно в клас „+”

fp е броят примери погрешно класифицирани в клас „+”

fn е примери погрешно класифицирани в клас „-”

tn е примери правилно класифицирани в клас „-”

Тогава мерките са:

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

Могат да бъдат изведени аналогичните отношения, ако за водещ клас е избран другия. Мерките са различни в зависимост от това кой клас е избран за водещ. В случая това е класа „+”.

Мярката *precision* измерва до колко е точен класификаторът, т.е. колко от примерите, които е определил като принадлежащи от даден клас, са действително от него. Мярката *recall* измерва пълнотата или чувствителността на класификатора, т.е. колко от всички примери от даден клас, той е успял да определи като принадлежащи на този клас. Очевидно според дефиницията на мерките трудно се постигат добри стойности и за двете. Най-лесният начин да се повиши резултатът от едната мярка е да се намали резултатът на другата. Затова изборът за целеви стойности тук е по-труден и не се очаква постигане на изключително добри стойности и за двете.

Ще бъдат разгледани следните критерии - точността за класификатора на субективността на текст и *precision* и *recall* за класовете обективен и субективен текст. Ще бъдат разгледани точността за класификатора на полярността на настроението в текст и *precision* и *recall* за класовете позитивен и негативен текст.

Съществува мярка, наречена F-measure, която комбинира *precision* и *recall* и е тяхното претеглено хармонично средно :

$$F\ measure = 2 \frac{Precision * Recall}{Precision + Recall}$$

Тази мярка обаче не дава допълнително знание за постигнатите резултати. *Precision* и *recall* са мерки, които предоставят по-задълбочен поглед върху представянето на модела при класификация на примерите в двата класа. Изчисляването на *precision* и *recall* и за двата класа дава информация за точността на класификатора за всеки клас поотделно. F мярката може да е подвеждащата тъй като нейната стойност е различна в зависимост от това кой от двата класа е избран за „позитивен“. При проблема, който се разглежда в текущата работа, и в двата етапа на класификация класовете имат една и съща тежест и не би могло да се избере т.нар. „позитивен“. Според поставените условия на текущата задача F мярката не би дала повече информация от мярката точност.

Проучени и сравнени са различни мерки за представянето на класификаторите [10]. Тези, които ще бъдат разгледани в настоящата работа, са точност и *precision*, *recall* за всеки един от класовете.

3.4. Изследване на избраните алгоритми за класификация

Изследваните алгоритми за класификация в настоящата работа са Наивен Бейсов класификатор, метод на опорните вектори (SVM), K най-близки съседа. Причината за избора на първите два алгоритъма е използването им в много проучвания в областта на определяне на настроението в текст ([7], [9], [11], [12]) и добрите резултати, които те показват. Това са алгоритми, които често се използват за по-общата задача за класификация на текст ([13], [14]). K най-близки съседа е алгоритъм от типа самообучение чрез запомняне. Настоящата работа изследва представянето на този алгоритъм върху задачата за определяне на настроението в текст. K най-близки съседа е също широко проучен алгоритъм даващ добри резултати в категоризацията на текст ([13],[14]).

3.4.1 Наивен Бейсов класификатор

Един от класическите алгоритми в машинното самообучение е Наивният Бейсов класификатор. Би могъл да се разглежда като общ модел на работа, който обединява няколко алгоритъма от типа обучение с учител.

Наивният Бейсов класификатор се базира на теоремата на Бейс като приема „наивното“ предположение за условна независимост между всяка двойка атрибути описващи примерите. Макар опростяването, което прави този подход, да е твърде голямо, той е известен като добре работещ алгоритъм включително и в областите класификация на текст и филтриране на спам. Той е един от сравнително бързо

работещите алгоритми за машинно самообучение и е използван в много реални приложения. Не се нуждае от много данни за обучение. При категоризацията на текст има едно много голямо преимущество. В тази проблемна област едно от основните затруднения е т.нар. „проклятие на размерността“, т.е. твърде голямото количество на атрибутите за описание на един пример. Правейки предположение за условна независимост между всяка двойка атрибути, Наивният Бейсов класификатор се справя ефективно с „проклятието на размерността“.

Нека y е даден клас, а векторът от атрибути на пример е (x_1, x_2, \dots, x_n) . Теоремата на Бейс твърди, че:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}$$

Използвайки предположението на Наивния Бейсов класификатор:

$$P(x_i|y, x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y), \quad \forall i = \{1, \dots, n\}$$

Тогава твърдението на Теоремата на Бейс се опростява до:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, x_2, \dots, x_n)}$$

Също така при даден вход знаменателят $P(x_1, x_2, \dots, x_n)$ е константа. Така Наивният Бейсов класификатор назначава клас на пример по следната формула:

$$\hat{y} = \arg \max_y P(y) \prod_{i=1}^n P(x_i|y)$$

Различните алгоритми от типа Наивен Бейсов класификатор правят различни предположения за разпределението $P(x_i|y)$. Два от тях са изследвани като посочени за много подходящи при класификация на текст – Мултиноминален и Бернулиев [15].

Мултиноминалният Наивен Бейсов класификатор прави предположението, че $P(x_i|y)$ е разпределено мултиноминално. Това разпределение е параметризирано по вектори от параметри $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$ за всеки клас y , където n е броят на атрибутите, с които се описва пример. В случай на класификация на текст n е размерът на речника, θ_{yi} е вероятността $P(x_i|y)$ за атрибут i в пример от класа y . Параметрите θ_{yi} се изчисляват на база на честота на срещане:

$$\theta_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

N_{yi} броят пъти, в които атрибут i се среща в примери от класа y в обучаващото множество T .

$$N_y = \sum_{i=1}^{|T|} N_{yi}$$

$\alpha \geq 0$ е параметър, който се използва, за да няма нула за вероятност при атрибут, който не се среща в обучаващото множество. Ще бъде използвана стойност на този параметър $\alpha = 1$, която постановка се нарича заглаждане на Лаплас.

Бернулиевият Наивен Бейсов класификатор предполага, че всеки атрибут е бинарен, т.е. има две възможни стойности. Тогава

$$P(x_i|y) = P(i|y)x_i (1 - P(i|y))(1 - x_i)$$

За разлика от предходния модел, в този случай липсата на някой атрибут в нов пример от обучаващото множество ще бъде наказана, а в горния случай просто ще бъде игнорирана.

Бернулиевият Наивен Бейсов класификатор се препоръчва при работа с по-кратки текстове. Представени са резултати при използването и на двата модела.

В задачата, която изследва текущата работа, примерите са вектори, получени от предварителната обработка на текстовете. Атрибутите на примерите са думите, а техните стойности са 0 или 1 в зависимост от наличието на думата. Обучени са два класификатора – единият разпознаващ субективен или обективен е текстът, а другият определящ дали настроението в текст е позитивно или негативно. Обучени са и по двата описани алгоритъма от типа Наивен Бейсов класификатор. Получените експериментални резултати за класификация в класовете обективен и субективен текст са представени в таблица 3.2.

Класификация: обективен / субективен текст		
	Мултиноминален Наивен Бейсов класификатор	Бернулиев Наивен Бейсов класификатор
Точност	0.9236	0.9188
Precision (обективен текст)	0.9322	0.9388
Recall (обективен текст)	0.9136	0.8960
Precision (субективен текст)	0.9153	0.9005
Recall (субективен текст)	0.9336	0.9416

Табл. 3.2. Резултати на изследваните алгоритми от тип Наивен Бейсов за класификация в класовете обективен и субективен текст

Макар да не са направени никакви подобрения в избора на атрибути и работата на класификаторите (описано в следващата глава на изложението) представянето им е изключително добро за класификация на неутралността на текст. Въпреки очакваните не добри резултати поради по-голямата сложност на проблема за определяне на субективността на текст, точността, *precision* и *recall* са добри. Малко по-добро представяне има Мултиноминалният Наивен Бейсов класификатор.

За всички експериментални резултати е необходимо да се направи следното уточнение. Те биват изведени на базата на множество текстове, чиято тематика е свързана с филми. При реални условия се предполага не толкова добро представяне на класификаторите. Въпреки това наличието на думи от тази тематика не би трябвало да повлияе нито на съществени признаци определящи обективност на текст, нито на тези за полярност на настроението в текст.

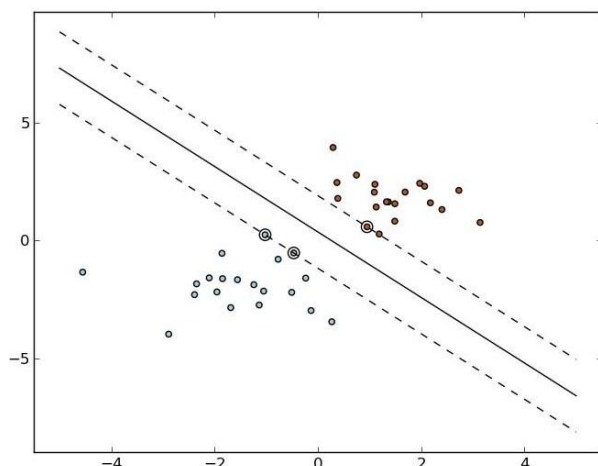
Класификация: позитивен / негативен текст		
	Мултиноминален Наивен Бейсов класификатор	Бернулиев Наивен Бейсов класификатор
Точност	0.7293	0.7407
Precision (позитивен текст)	0.8667	0.8685
Recall (позитивен текст)	0.5420	0.5673
Precision (негативен текст)	0.6668	0.6787
Recall (негативен текст)	0.9166	0.9141

Табл. 3.3. Резултати на изследваните алгоритми от тип Наивен Бейсов за класификация в класовете позитивен и негативен текст

Не толкова добри резултати се наблюдават при класификацията на полярността на текст. Те са изложени в таблица 3.3. И двата алгоритъма имат много нисък *recall* за класа позитивен текст. Това означава, че Наивният Бейсов класификатор има тенденция да отгатва като позитивни малка част от наистина позитивните класове. Затова се наблюдава и много малка прецизност за класа негативен текст. Малка част от определените като негативни текстове са били наистина такива. Резултатите показват, че при алгоритмите от тип Наивен Бейсов класификатор има тенденция позитивни текстове да бъдат определяни като негативни. Опит да се подобрят параметрите на получените резултати ще бъде описан в следващата глава.

3.4.2. Метод на опорните вектори (SVM)

Методът на опорните вектори (SVMs, *Support Vector Machines*)[16] може да се използва за решаване на задачите класификация и регресия. Той спада към методите в машинното самообучение от тип учене с учител. Основната формулировка на SVM работи с два класа. Методът представя обучаващите примери като точки в многомерно пространство. Примерите са проектирани в това многомерно пространство по такъв начин, че примерите от различни класове да са възможно най-добре разделени помежду си. Трябва те да бъдат линейно разделими от хиперравнина. Тази хиперравнина трябва да се избере по такъв начин, че да се намира възможно най-далеч от примерите и на двата класа. Класификацията на нов пример става като той се проектира в същото пространство и се определя класа му според това от коя страна на хиперравнината се намира. Примерна схема на работа на алгоритъма за двумерно пространство е представена на фигура 3.2. Методът на опорните вектори може да работи не само за линейна класификация, но са необходими модификации на първоначалната му формулировка.



Фиг.3.2. Примерно представяне на работата на метода на опорните вектори в двумерно пространство за класификация в два класа

Нека обучаващото множество е $\{x_1, x_2, \dots, x_n\}$, $x_i \in R^d$ и всеки от примерите от обучаващото множество има клас $y_i \in \{-1, 1\}$, $\forall i \in \{1, 2, \dots, n\}$.

Всички хиперравнини в пространството R^d са параметризирани от вектор \vec{w} (вектор ортогонален на хиперравнината) и от константа b .

$$wx + b = 0$$

Намирането на най-добрата хиперравнина се състои в решаването на следния оптимизационен проблем:

$$\text{Min} \left(\frac{1}{2} \|\vec{w}\|^2 \right)$$

с удовлетворяване на следните ограничения:

$$y_i(w^T x_i + b) \geq 1, \forall i$$

Това е първичната формулировка на метода на опорните вектори. Тя е проблем за удовлетворяване на ограниченията с d на брой променливи $\vec{w} = (w_1, w_2, \dots, w_d)$ – толкова колкото са атрибутите на примерите. За решаването му се използва метод на множителите на Лагранж.

Този оптимизационен проблем може да се представи в т.нар. „дуална форма“, която е оптимизационен проблем, но на n променливи – колкото е броят на примерите в обучаващото множество. Когато обучаващите данни се описват с много атрибути, както при класификацията на текст, това е изключително полезно. Проблемът се свежда до :

$$\text{Max} \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j \right)$$

с удовлетворяване на ограничението:

$$\alpha_i \geq 0, \sum_{i=1}^n \alpha_i y_i = 0$$

Където векторът w е дефиниран по следният начин:

$$\vec{w} = \sum_{i=1}^n \alpha_i y_i \vec{x}_i$$

Функцията, по която се класифицира нов пример, е:

$$f(\vec{x}) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i \vec{x}_i \cdot \vec{x} + b\right)$$

Разбира се, всичко до този момент е валидно ако данните са линейно разделими. Но има и начин да се работи с метод на опорните вектори и когато данните не са линейно разделими. Идеята е да се премине в пространство с по-висока размерност, където данните ще бъдат линейно разделими. Преминването от едното пространство в друго става чрез ядрото $\Phi: R^n \rightarrow H$.

В новото пространство имаме:

$$f(x) = \text{sign}(\vec{w} \cdot \Phi(\vec{x}) + b)$$

$$\vec{w} = \sum_{i=1}^n \alpha_i y_i \Phi(\vec{x}_i)$$

$$f(x) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i \Phi(\vec{x}_i) \cdot \Phi(\vec{x}) + b\right)$$

$$f(x) = \text{sign}\left(\sum_{i=1}^n \alpha_i y_i K(\vec{x}_i, \vec{x}) + b\right)$$

Тогава не е необходимо да се знае Φ , а да се дефинира функция

$$K(\cdot, \cdot): R^n \times R^n \rightarrow R$$

Не всяка функция може да бъде използвана. Трябва да отговаря на т.нар. условие на Мърсър [16]. Ядрото може да се представи като

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$$

Често използвани ядра са

- Линейно ядро: $K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j$
- Гаусово ядро (RBF, radian basis function): $K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \|\vec{x}_i - \vec{x}_j\|^2)$
- Експоненциално ядро: $K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \|\vec{x}_i - \vec{x}_j\|)$
- Полиномиално ядро: $K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q$
- Сигмоидално ядро: $K(\vec{x}_i, \vec{x}_j) = \tanh(k \vec{x}_i \cdot \vec{x}_j - \delta)$

Методът на опорните вектори е един от най-изследваните алгоритми за класификация на текст. Описани са следните негови преимущества в тази проблемна област [13]:

- Справя се добре при работа в пространства с голяма размерност, защото не зависи пряко от броя на атрибутите, с които се описват примерите.
- Векторите, с които се описват примерите при класификация на текст, имат в голямата си част нули за компоненти. При такъв тип данни *SVMs* се представят добре.
- Много от проблемите в областта категоризация на текст имат линейно разделими данни. Основната идея на метода на опорните вектори е да прави линейна класификация.

Направени са експерименти с линейния вариант на *SVM*, т.е. предположено е, че примерите са линейно разделими за класификацията на обективност на текст и полярност на настроението в текст.

Класификация: обективен / субективен текст	
Линеен метода на опорните вектори	
Точност	0.8864
Precision (обективен текст)	0.8791
Recall (обективен текст)	0.8960
Precision (субективен текст)	0.8939
Recall (субективен текст)	0.8768

Табл. 3.4. Резултати на линейния метод на опорните вектори за класификация в класовете обективен и субективен текст

Наблюдаваните резултати за линеен *SVM*, представени на таблица 3.4, са добри при класификацията на неутралността на текст, но класификаторите от типа Наивен Бейсов показват по-добри статистики. Получената добра точност обаче означава, че вероятно проблемът за определяне на неутралността на текст е линейно разделим или близък до линейно разделим, както се твърди и за други проблеми от областта класификация на текст.

Класификация: позитивен / негативен текст	
Линеен метода на опорните вектори	
Точност	0.7416
Precision (позитивен текст)	0.8525
Recall (позитивен текст)	0.5843
Precision (негативен текст)	0.6838
Recall (негативен текст)	0.8989

Табл. 3.5. Резултати на линейния метод на опорните вектори за класификация в класовете позитивен и негативен текст

Класификацията на полярността на текст показва доста сходни резултати с тези на двата алгоритъма от тип Наивен Бейсов класификатор. Те са дадени в таблица 3.5. Наблюдава се същата тенденция на погрешно определяне на позитивни текстове като негативни, за което говори доста ниският процент на *recall* за класа позитивен текст и ниският *precision* на класа негативен текст.

Изследван е метод на опорните вектори с полиномиално ядро. Изведени са резултати за степен на полинома 2, 3, 4.

Класификация: обективен / субективен текст			
Метод на опорните вектори с полиномиално ядро			
	степен 2	степен 3	степен 4
Точност	0.6468	0.5172	0.5
Precision (обективен текст)	0.9359	0.9987	0.9987
Recall (обективен текст)	0.3152	0.0344	0.0021
Precision (субективен текст)	0.5883	0.5088	0.5
Recall (субективен текст)	0.9784	0.9888	0.9888

Табл. 3.6. Резултати на метода на опорните вектори с полиномиално ядро със степен 2, 3 или 4 за класификация в класовете обективен и субективен текст

Показаните резултати на таблица 3.6. при използването на метода на опорните вектори с полиномиално ядро са значително по-ниски от линейния SVM. Изключително ниските стойности за *recall* на обективния клас и *precision* на субективния при степен на ядрото 2 говори за тенденцията моделът да класифицира обективни/неутрални текстове като субективни. Влошаването на статистиките при увеличаване на степента

на полинома потвърждава, че примерите са линейно разделими или близки до линейно разделими.

Класификация: позитивен / негативен текст			
Метод на опорните вектори с полиномиално ядро			
	степен 2	степен 3	степен 4
Точност	0.6380	0.6459	0.6488
Precision (позитивен текст)	0.7834	0.7992	0.7590
Recall (позитивен текст)	0.3816	0.3898	0.4359
Precision (негативен текст)	0.5912	0.5965	0.6043
Recall (негативен текст)	0.8945	0.9020	0.8617

Табл. 3.7. Резултати на метода на опорните вектори с полиномиално ядро със степен 2,3 или 4 за класификация в класовете позитивен и негативен текст

Отново при класификацията на полярността на настроението в текст линейният SVM се справя по-добре от SVM с полиномиално ядро. На таблица 3.7. се вижда, че много малко от позитивните текстове са разпознати като такива ($Recall \approx 40\%$) и при трите експеримента с различни степени на полинома. Наблюдаваната тенденция е позитивните текстове да бъдат разпознавани като негативни.

3.4.3. К най-близки съседа

К най-близки съседа е алгоритъм използван в различни задачи от областта категоризация на текст с добри резултати. Спада към алгоритмите използващи самообучение чрез запомняне (*instance-based learning, lazy learning*). Те се различават от описаните вече два класификатора по начина на изграждане на модела в етапа на обучение. Те не извеждат никакви правила или обобщаваща информация при обучението, а просто запомнят обучаващите примери. Разчитат на т.нар екстенционално описание да понятията, т.е. описание чрез представителни примери. Етапът на класификация също е различен. Определянето на класа на нов пример се осъществява на база на сходството между примера и един или няколко от запомнените обучаващи примери. Затова тези методи още се наричат базирани на сходство. Ключовата разлика при методите за самообучение чрез запомняне е, че при тях от обучаващите данни не се изгражда един модел за класификация, а при всеки нов пример се правят локални апроксимации. Това прави тези алгоритми подходящи при решаването на сложни проблеми неподлежащи на обобщение с единен модел.

Характерно при методите за самообучение чрез запомняне е, че при тях се изчислява сходство между примерите. Съществуват няколко подхода за намиране на сходство – геометричен, теоретико-множествен, статистически. Геометричният подход разглежда примерите като точки в n -мерно пространство и изчислява разстоянието

между тях. Най-често използваната мярка за изчисляване на разстоянието е Евклидовата. Теоретико-множественият подход разглежда сходството между обекти на база броя общи признаци между примерите, броя признаци уникални за единия и броя признаци уникални за втория.

К най-близки съседа е алгоритъм с лесна постановка на действие. Числото k е броят на най-близките примери, участващи в определянето на решение за класификацията на нов пример. Примерът се класифицира в съответствие с класа, който най-често се среща сред най-близките му k съседа. Ако повече от един клас се среща най-често сред най-близките k съседа, примерът обикновено се класифицира в съответствие с класа на най-близкия си съсед сред конкуриращите се класове.

Нека C е множеството възможни класове на класификация, D е множеството от обучаващи примери, f е функцията, която дава класа на пример, т.е. целевата функция, d е мярката за разстояние между примери.

$$D = \{ \langle x_i, f(x_i) \rangle \}$$

Нека се класифицира нов пример e и изчислените му най-близки съседи са $\{x_{i_1}, x_{i_2}, \dots, x_{i_k}\}$.

Тогава функцията, която използва К най-близки съседа за класификация е:

$$\hat{f}(e) = \arg \max_{c \in C} \sum_{s=1}^k \lambda(c, f(x_{i_s}))$$

$$\lambda(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

Една известна модификация на К най-близки съседа е да се претегля гласът на най-близките k съседа при класификация в зависимост от разстоянието им до новия пример, който се класифицира. Колкото по-близък е съсед, толкова по-тежък е неговият глас. Например :

$$\hat{f}(e) = \arg \max_{c \in C} \sum_{s=1}^k w_s \lambda(c, f(x_{i_s}))$$

$$w_s = \frac{1}{d(e, x_{i_s})^2}$$

Разбира се може да бъде избрана и друга мярка за претегляне на най-близките съседи. Описаната модификация се нарича алгоритъм за k най-близки съседи, претеглени по разстояние. Използването на претеглени гласове позволява да се неутрализира влиянието на изолирани обучаващи примери, в които предварително определеният клас е сбъркан, т.нар. зашумени данни. Обучаващите данни, които

използва текущата разработка не са зашумени и не се използва тази модификация на алгоритъма.

Изключително важно за работата на алгоритъма е изборът на параметъра k . Параметърът силно зависи от типа на данните. Изведени се резултатите от няколко от експериментите с различни стойности на k . Отново алгоритъмът е приложен и за двете задачи – анализ на неутралността на текст и анализ на настроението в текст.

Класификация: обективен / субективен текст					
K най-близки съседа					
	k = 10	k = 30	k = 50	k = 100	k = 200
Точност	0.7532	0.7368	0.7156	0.7360	0.7196
Precision (обективен текст)	0.7949	0.8246	0.8560	0.8554	0.8626
Recall (обективен текст)	0.6824	0.6016	0.5184	0.5680	0.5224
Precision (субективен текст)	0.7218	0.6864	0.6546	0.6766	0.6575
Recall (субективен текст)	0.8240	0.8720	0.9128	0.9040	0.9168

Табл. 3.8. Резултати на алгоритъма K най-близки съседа с различни стойности на параметъра K при класификация в класовете обективен и субективен текст

Таблица 3.8 показва, че алгоритъмът дава значително по-слаби резултати от Наивния Бейсов метод за същата задача. Точността е по-ниска и в сравнение с тази на линейния метод на опорните вектори. Тенденцията, която се наблюдава във всички опити, е обективните текстове да бъдат определяни като субективни – *recall* за класа обективен текст е ниска, както и *precision* за класа субективен текст. Интересно наблюдение е, че според показаните резултати точността на класификатора не се влияе силно от избора на броя на съседите, т.е. параметъра k . Вижда се, че малко по-добър резултат се наблюдава при малка стойност на k .

Класификация: позитивен / негативен текст					
K най-близки съседа					
	k = 10	k = 30	k = 50	k = 100	k = 200
Точност	0.5843	0.6074	0.6257	0.6323	0.6490
Precision (позитивен текст)	0.5945	0.5898	0.6296	0.6318	0.6858
Recall (позитивен текст)	0.5306	0.7056	0.6109	0.6342	0.5502
Precision (негативен текст)	0.5762	0.6336	0.6220	0.6328	0.6245
Recall (негативен текст)	0.6380	0.5092	0.6406	0.6304	0.7479

Табл. 3.9. Резултати на алгоритъма K най-близки съседа с различни стойности на параметъра K при класификация в класовете позитивен и негативен тест

K най-близки съседа за класификация на полярността на текст има по-ниска точност в сравнение с алгоритмите от тип Наивен Бейсов класификатор и линейния метод на опорните вектори. Резултатите от направените експерименти са представени в таблица 3.9. При силно увеличаване на броя на съседите, които се отчитат при вземане на решение за класификацията, слабо нараства точността на алгоритъма. Именно поради твърде малкото нарастване на точността в сравнение с увеличаването на k , не е оптимална стратегия алгоритъмът да бъде подобряван с нарастване на параметъра k .

В следващата глава на настоящото изложение са изследвани начини за подобряване представянето на класификаторите. На база на получените експериментални резултати за продължаване изследването на тяхното поведение са избрани следните алгоритми:

- за класификация неутралността на текст – Мултиноминален Наивен Бейсов класификатор, линейен метод на опорните вектори, K най-близки съседа за $k=100$;
- за класификация на полярността на текст - Бернулиев Наивен Бейсов класификатор, линейен метод на оперните вектори, K най-близки съседа за $k=100$ (по-голяма стойност на k прави алгоритъма доста неефективен по отношение на времето за класификация).

4. Изследвани методи за подобряване на резултатите от класификацията

4.1. Оценяване и избор на атрибути

Ключова роля в подобряване на ефективността и точността на алгоритмите в машинното самообучение има правилният избор на атрибути. В изложението до момента представянето на текстовете във структуриран вид като вектори от атрибути се прави чрез често използвания в класификацията на текст подход наречен множество от думи („*bag of words*“). Избраното представяне е широко застъпено и много проучвания и разработки използват именно него. То обаче има своите недостатъци и може да бъде подобро.

4.1.1. Филтриране на стоп думи

Едно от подобренията, които могат да бъдат направени в представянето на текстовете като вектори от атрибути, е да не бъдат разглеждани думи като „a“, „the“, „and“ като отделни атрибути на примерите. Подобни думи не носят никаква информация нито за полярността на текст, нито за наличието на субективност в текст. Те могат да попречат на работата на класификаторите като ги направят времево по-малко ефективни и дори да намалят точността им чрез внасянето на ненужна за проблема информация.

В обработката на естествен език такива думи се наричат стоп думи. Те могат да бъдат различни в зависимост от конкретната задача и област. Списъкът от стоп думи, използван в текущата система, е стандартен списък от стоп думи в английския език. Не са добавени допълнителни стоп думи тъй като текстовете, които могат да бъдат разглеждани от потребител на системата и съответно подавани за класификация, може да са от различни области. Използваният списък от стоп думи съдържа 127 думи като „a“, „an“, „the“, „and“, „but“, „if“, „or“, „because“, „as“, „until“, „while“, „of“, „at“, „by“, „for“, „with“.

Описан в разработките недостатък на филтрирането на стоп думите съществува при имащи семантично значение словосъчетания. Тогава ще бъде изгубена информацията, която носят думите заедно. Поотделно те могат да има напълно различно семантично значение. Но класификаторите на разглежданото решение работят с признаци от по една дума така че всъщност не би трябвало филтрирането да стоп думи да има отрицателни ефекти. Използването на значими за класификацията словосъчетания ще бъде засегнато в текущата глава на изложението.

При филтрирането на стоп думите избраните класификатори показват резултатите, представени в таблици 4.1 и 4.2.

Класификация: обективен / субективен текст			
	Мултиноминален Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9096	0.8684	0.6684
Precision (обективен текст)	0.9252	0.8556	0.6598
Recall (обективен текст)	0.8912	0.8864	0.6952
Precision (субективен текст)	0.8951	0.8822	0.6779
Recall (субективен текст)	0.9280	0.8504	0.6416

Табл. 4.1. Резултати на изследваните алгоритми за класификация в класовете обективен и субективен текст при предварително филтриране на стоп думите от признаците, с които се описват текстовете

Класификация: позитивен / негативен текст			
	Бернулиев Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.7590	0.7539	0.6342
Precision (позитивен текст)	0.8516	0.8295	0.7627
Recall (позитивен текст)	0.6273	0.6393	0.3898
Precision (негативен текст)	0.7050	0.7065	0.5902
Recall (негативен текст)	0.8907	0.8686	0.8787

Табл. 4.2. Резултати на изследваните алгоритми за класификация в класовете позитивен и негативен текст при предварително филтриране на стоп думите от признаците, с които се описват текстовете

Получените експериментални резултати показват много малки подобрения в работата на класификаторите за полярност на настроението в текст. Противно на очакваното К най-близки съседа показва дори по-ниски стойности за точност след филтриране на стоп думите. Мултиноминалният Наивен Бейсов класификатор и линейният метод на опорните вектори за класификацията обективен/субективен текст също дават по-ниски стойности за точност, както и за всяка от останалите метрики. Вероятната причина за не толкова доброто представяне на класификаторите за неутралност на текст е във филтрирането на думи като „i“, „me“, „my“, „myself“, „we“, „our“, „ours“, „ourselves“ показващи наличие на отношение на пишещия, а също и „too“, „very“ показващи засилена емоция. Не е добра идеята филтрирането на отрицателни частици като „no“, „not“, които могат да имат ключово значение за полярността на текста. Оказва се, че простото филтриране на стоп думите не дава добри резултати и за двата типа класификация – неутралност на текст и полярност на настроението в текст. Необходим е по-задълбочен анализ за намирането на правилните атрибути за представяне на примерите в поставените класификационни задачи.

4.1.2. Избор на информативни атрибути

Известен проблем в машинното самообучение е т.нар. „проклятие на размерността“ („*curse of dimensionality*“). Изразът за първи път е използван през 1961г. от Белман като обяснение на факта, че много от алгоритмите, които работят добре с данни с малка размерност, драстично влошават своето поведение при работа с данни с голяма размерност. Проблемната област на текущата работа е класификация на текст – област с наистина голяма размерност на данните. За пример по-голямото от двете обучаващи множества текстове в текущата система съдържа 57 029 различни думи.

От гледната точка на алгоритмите за машинно самообучение проблемът на голямата размерност има няколко аспекта. Алгоритмите от тип обучение с учител изграждат модели за класификация на базата да обучаващите данни. При фиксиран брой на примерите в обучаващото множество, колкото по-висока е размерността, толкова по-малко представителна става извадката от обучаващи примери, защото тя покрива все по-малка област от пространството от всички възможни примери. Затова способността на алгоритмите да правят коректни обобщения извън обучаващите данни експоненциално се влошава.

Размерността на примерите е важна за алгоритмите от тип обучение чрез запомняне (какъвто е използваният K най-близки съседи). Те още се наричат базирани на сходство, защото при своята работа изчисляват сходството между примерите. Мерките за изчисляване на сходство вземат предвид всички атрибути на примерите. Ако примерите са с голяма размерност, а много малка част от атрибутите, т.е. измеренията, имат значение за класификацията, то при изчисляване на сходството, информацията от релевантните атрибути ще допринесе много малко в крайната стойност за сходството. Информацията от релевантните атрибути ще бъде заглушена от случайната информация идваща от останалите атрибути. На практика тогава алгоритъм базиран на сходство ще прави напълно случайни предсказания.

Пред алгоритмите от тип най-близък съсед има още едно затруднение при примери с голяма размерност. С нарастване на размерността най-близките съседи на един пример не се оказват разположени наблизо в геометричен смисъл [18]. При голяма размерност на пространството всички примери изглеждат еднакво сходни. Изборът на най-близкия съсед или съседи става почти случаен.

В повечето проблемни области обаче се оказва, че примерите не са разпределени равномерно в пространството от примери [18]. Те са концентрирани в групи от примери в пространства с по-малка размерност. По тази причина за построяването на ефективни класификатори се използват алгоритми, които намаляват размерността на пространството на примерите.

За да се превъзмогне чувствителността на разгледаните алгоритми към нерелевантните атрибути в описанието на примерите са приложени методи за намирането на важните за класификационната задача атрибути. Правени са изследвания на множество методи [17]. Те базово се разделят в два подхода – избиране на атрибути (*feature selection*) и претегляне на атрибути (*feature weighting*). Първата група методи избират определено подмножество от всички атрибути на примера, а останалите не се отчитат при етапа на обучение и класификация. Вторият подход използва всички атрибути на примера, но претегля различно приноса на всеки

от атрибутите за класификацията. Подходът на настоящата работа е избиране на подмножество от атрибути и работа само с тях тъй като проблемната област предполага един пример да се описва с хиляди атрибути (колкото е речника на обучаващото множество текстове) и претеглянето и използването на всички атрибути прави класификацията много неефективна от изчислителна гледна точка.

Методите за премахване на неинформативните атрибути са същите, каквито са тези за оценяване на атрибути, но просто теглата, които те дават на всеки атрибут са 0 или 1. Основната идея е за всеки атрибут да бъде изчислено тегло w_i , което да дава относителната важност на този атрибут. Според начина на действие съществуват две групи методи за даване на тегла – итеративни (*online*) и статистически (*batch*). Итеративните методи за оценяване на атрибути се базират на обратна връзка с използвания алгоритъм за класификация. Текущият подход изследва два от статистическите методи за избор на релевантни атрибути. Избраните мерки са едни от най-често използваните за проблемната област класификация на текст – χ^2 и мярка за взаимната информация.

В статистиката мярката за взаимна информация (*mutual information*) на две случайни променливи е мярка за тяхната взаимна зависимост – определя степента, с която се намалява неопределеността в стойността на едната променлива при наличието на някакви знания за стойността на другата променлива. Когато се използва за определяне на значимостта на атрибут при класификация тя се разглежда между стойностите на атрибута и класа на обучаващите примери.

Нека C е множеството от всички възможни класове на класификация, X е произволен пример от обучаващото множество, $P(c_X = c)$ е вероятността, че класът на произволния пример X е c , $P(x_i = v)$ е вероятността, че стойността на неговия i -ти атрибут е v , V_i е множеството от всички възможни стойности на i -тия атрибут. Тогава мярката за взаимна информация дава тегло на всеки атрибут по следната формула:

$$w_i = \sum_{v \in V} \sum_{c \in C} P(c_X = c \wedge x_i = v) * \log \frac{P(c_X = c \wedge x_i = v)}{P(c_X = c) * P(x_i = v)}$$

Намира се теглото w_i за всеки от един от атрибутите описващи примерите. Избират се най-добрите n атрибута подредени според теглото си и само те участват в описанието на примерите. Числото n силно зависи от проблема и данните.

Направени са тестове със стойности на n между 1 000 и 20 000. Посочените в таблици 4.3 и 4.4 резултати са най-добрите открити за всеки един от алгоритмите.

Класификация: обективен / субективен текст			
	Мултиноминален Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9196	0.8860	0.7124
Precision (обективен текст)	0.9268	0.8767	0.8098
Recall (обективен текст)	0.9112	0.8984	0.5552
Precision (субективен текст)	0.9127	0.8958	0.6616
Recall (субективен текст)	0.9280	0.8736	0.8696

Табл. 4.3. Резултати на изследваните алгоритми за класификация в класовете обективен и субективен текст при избор на най-релевантните атрибути чрез мярката за взаимна информация

Резултатите за класификацията на неутралността на текст са дадени в таблица 4.3. При алгоритмите Наивен Бейсов класификатор и метод на опорните вектори точността остава почти непроменена и дори е по-ниска от първоначално постигнатата. Отстраняването на нерелевантните атрибути би трябвало да даде съществени резултати при алгоритмите базирани на сходство, т.е. К най-близки съседа. Представянето и на този метод обаче е по-ниско от първоначалното и то с 4%.

Класификация: позитивен / негативен текст			
	Бернулиев Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.7334	0.7252	0.6273
Precision (позитивен текст)	0.8633	0.8373	0.7202
Recall (позитивен текст)	0.5546	0.5590	0.4163
Precision (негативен текст)	0.6719	0.6690	0.5895
Recall (негативен текст)	0.9122	0.8913	0.8383

Табл. 4.4. Резултати на изследваните алгоритми за класификация в класовете позитивен и негативен текст при избор на най-релевантните атрибути чрез мярката за взаимна информация

При класификацията на полярността на текст не се наблюдава почти никаква промяна в стойностите на мерките дадени от трите алгоритъма. Резултатите са представени в таблица 4.4. При К най-близки съседа има силно понижение в *recall* за класа позитивен текст и в *precision* за класа негативен текст, което показва склонност позитивните текстове да бъдат определяни като негативни.

Намаляването на атрибутите на база мярката за взаимна информация не подобрява поведението на изследваните алгоритми. Едно от възможните обяснения за получените експериментални резултата е, че използването на тази мярка предполага атрибутите да са независими. При решаване на задача с множество взаимодействащи си признаци (каквато се оказва задачата за класификация на текст) се получава деградация на класификационното поведение на използващите мярката алгоритми.

Изследвана е още една мярка за оценяване релевантността на атрибутите за описание на примерите. χ^2 се често използван статистически тест за независимост. Оценява дали наблюденията за стойностите на две случайни променливи са независими едно от друго, т.е. нулевата хипотеза е, че те са статистически независими. Нека всяко наблюдение е записано в таблица като всяко измерение на таблицата (*contingency table*) отговаря на стойностите на двете случайни променливи. Нека таблицата има r реда и c колони. Теоретичната честота на всяка клетка от таблицата имайки предвид нулевата хипотеза за независимост е:

$$E_{i,j} = \frac{(\sum_{n_c=1}^c O_{i,n_c}) \cdot (\sum_{n_r=1}^r O_{n_r,j})}{N}$$

$O_{k,l}$ е наблюдавана честота, N е сумата от всички клетки в таблицата. Тогава стойността на тестовата статистика е

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

При класификация тестът се използва да измерва отклонението от очакваното разпределение ако се приеме, че наличието на признак е независимо от стойността на класа. При разглеждания от текущата работа проблем класът може да се представи като една бинарна променлива, както и всеки атрибут описващ примерите. Когато имаме бинарни променливи, т.е. таблица 2x2, χ^2 има следната зависимост с квадрата на коефициента ϕ :

$$\phi^2 = \frac{\chi^2}{N}$$

ϕ измерва зависимостта на две случайни бинарни променливи. Това е всъщност коефициентът за корелация на Пиерсон за две случайни бинарни променливи. Коефициентът на корелация на Пиерсон е мярка за линейната корелация на две случайни променливи X и Y . Изчислява се по следната формула:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

В конкретната задача имаме таблица 2x2 и случайни променливи клас и наличие на дума. Таблица 4.5 показва разпределението на примерите спрямо стойностите на двете случайни променливи.

	w	$\sim w$	
Клас „+”	n_{11}	n_{10}	$n_{1\cdot}$
\sim Клас „+”	n_{01}	n_{00}	$n_{0\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 0}$	N

Табл. 4.5. Разпределението на примери спрямо класа и наличието на атрибута w

Изчислява се

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1.}n_{0.}n_{.0}n_{.1}}}$$

По посочената горе зависимост от ϕ се намира χ^2 .

В таблици 4.6 и 4.7 са дадени получените експериментални резултати при използване на χ^2 за оценяване на атрибути. Направени са тестове със стойности на броя атрибути (n) между 1 000 и 20 000. Посочените резултатите са най-добрите постигнати от съответните алгоритми.

Класификация: обективен / субективен текст			
	Мултиноминален Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9252	0.8940	0.7784
Precision (обективен текст)	0.9242	0.8768	0.7939
Recall (обективен текст)	0.9264	0.9168	0.7520
Precision (субективен текст)	0.9262	0.9128	0.7644
Recall (субективен текст)	0.9240	0.8712	0.8048

Табл. 4.6. Резултати на изследваните алгоритми за класификация в класовете обективен и субективен текст при избор на най-релевантните атрибути чрез мярката χ^2

Използването χ^2 мярката за претегляне на атрибутите при класификация в класовете обективен и субективен текст показва подобрение на поведението на всички алгоритми. Конкретните стойности за точност, *precision* и *recall* са дадени в таблица 4.6. Подобрението е доста малко за Наивния Бейсов класификатор и за методът на опорните вектори и по-значително при алгоритъма К най-близки съседа.

Класификация: позитивен / негативен текст			
	Бернулиев Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9072	0.8540	0.7640
Precision (позитивен текст)	0.8964	0.8583	0.8204
Recall (позитивен текст)	0.9208	0.8480	0.6760
Precision (негативен текст)	0.9185	0.8498	0.7245
Recall (негативен текст)	0.8936	0.8600	0.8520

Табл. 4.7. Резултати на изследваните алгоритми за класификация в класовете позитивен и негативен текст при избор на най-релевантните атрибути чрез мярката χ^2

При задачата за определяне полярността на текст класификационното поведение и на трите алгоритъма е значително подобро. Както се вижда от таблица 4.7, точността се повишава в рамките на 10-15%. Получените резултати са отлични за проблем от областта анализ на настроението в текст, където точност в рамките на 80% е приемлива. Разбира се, трябва да се има предвид, че тестовите данни са текстове отнасящи се до филми. При реално използване на класификаторите за извличане на настроението на текст от различни сфери се очаква по-малка точност.

Използването на χ^2 мярка за избирането на най-информативните атрибути повишава точността, *precision* и *recall* при всички направени тестове. Затова последващите разглеждания ще бъдат направени с използването на този подход. Като обобщение от направения експеримент с премахването на нерелевантните атрибути може да се твърди, че този подход значително подобрява класификационното поведение на използваните алгоритми.

4.2. N-грами

Използвайки представяне от тип *“bag of words”* се губи важна семантична и синтактична информация за текста. Подредбата на думите има значение за посланието, което носи един текст. Тогава тя има значение и за решаваните в настоящата работа два класификационни проблема.

Нека са дадени следните примерни словосъчетания - „not good”, „exclusively bad”. Когато един класификатор работи с текст съдържащ подобни фрази и той използва подхода всяка дума да бъде отделен атрибут е много вероятно да бъде сбъркан класа на разглеждания текст. Ако алгоритъмът отчете наличието на „good”, „exclusively”, то той би определил настроението като позитивно, което е погрешно. Причината за това е, че думите от тези фрази разгледани поотделно носят противоположно значение. Описаната ситуация е пример на едно от несъвършенствата на подхода текст да бъде представен като множество от думите си.

В обработката на естествен език се използва понятието *n*-грами. Това са последователности от *n* на брой думи в реда, в който се срещат в текста. Възможно е описаният проблем да бъде отчасти решен с използването и на *n*-грами като атрибути, с които да бъдат описани примерите. За кратки текстове, с каквито работи текущата система, е препоръчително да се работи с биграми, т.е. последователности от две думи. Тогава в атрибутите на горния пример ще бъдат включени и „not good”, „exclusively bad”.

Използването обаче на всички биграми би означавало, че атрибутите, с които се описват текстовете, доста ще се увеличат. Не всички от тези биграми носят и значима информация за класификацията. Както бе вече описано наличието на нерелевантни атрибути влошава класификационното поведение на алгоритмите. Затова към атрибутите описващи примерите ще бъдат включени само значимите биграми. Изборът на това кои биграми да бъдат включени ще бъде направен по същия начин, по който се избира кои униграми да бъдат включени – оценяване на атрибутите чрез χ^2 мярката.

Колко биграми да бъдат включени е определено експериментално. Нека *m* е този брой. Направени са тестове със стойности на *m* между 100 и 1000. Разликата в класификационната точност при различните тествани стойности на *m* се оказва малка.

Класификация: обективен / субективен текст			
	Мултиноминален Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9260	0.9148	0.7228
Precision (обективен текст)	0.9339	0.9118	0.7562
Recall (обективен текст)	0.9168	0.9184	0.6576
Precision (субективен текст)	0.9183	0.9178	0.6970
Recall (субективен текст)	0.9352	0.9112	0.7880

Табл. 4.8. Резултати на изследваните алгоритми за класификация в класовете обективен и субективен текст при включване в признаците на най-добрите биграми и оценяване на атрибутите чрез χ^2 мярката

Експерименталните резултати за класификация в класовете обективен и субективен текст са представени в таблица 4.8. Включването на допълнителни признаци в класификационния проблем оказва негативно влияние върху алгоритъма К най-близки съседа. Той, като алгоритъм базиран на сходство, е и най-чувствителен към наличието на нерелевантни признаци. При другите два алгоритъма обаче се наблюдава повишаване на точността им, макар и слабо.

Класификация: позитивен / негативен текст			
	Бернулиев Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9152	0.8760	0.7400
Precision (позитивен текст)	0.8885	0.8917	0.7857
Recall (позитивен текст)	0.9496	0.8560	0.6600
Precision (негативен текст)	0.9459	0.8615	0.7069
Recall (негативен текст)	0.8808	0.8960	0.8200

Табл. 4.9. Резултати на изследваните алгоритми за класификация в класовете позитивен и негативен текст при включване в признаците на най-добрите биграми и оценяване на атрибутите чрез χ^2 мярката

При проблема за определяне на текст в класовете позитивен и негативен резултатите, представени в таблица 4.9, са много подобни на получените при определяне на субективността на текст. Наивният Бейсов подход и методът на опорните вектори повишават слабо своята точност, а К най-близки съседа показва понижение на точността и на всички останали мерки за оценка.

4.3. Отрицателни семантични конструкции

Още един от проблемите, който възниква при използване на представяне от тип „bag of words”, е загубата на информация за отрицателните конструкции в езика. Нека е дадено изречението „This afternoon didn’t go very well”. При представянето му като вектор от атрибути ще бъде отчетено наличието на отрицателната конструкция “didn’t”, но тъй като представянето разглежда всяка дума поотделно ще бъде отчетено и наличието на позитивната дума „well”. Това е доста подвеждаща информация особено за класификационната задача за определяне на полярността на настроението в текст.

Разгледан е прост алгоритъм, който се опитва да се справи с описания горе проблем. Неговата идея е при срещане на някаква отрицателна конструкция като „not”, „no”, „cannot”, „non”, „didn’t”, „won’t” всяка следваща дума в изречението да бъде добавена два пъти като атрибут описващ текста. Първият път се добавя думата, каквато е в изречението, и още веднъж се добавя атрибут във формата <„not_” дума>. Това означава, че по описания горе пример ще бъдат добавени атрибутите „well” и „not_well”. Алгоритъмът също отчита наличието на нова отрицателна конструкция в изречението и не добавя отрицателни форми за думите, които са след нея.

Представени са резултатите от използването на този подход като отново атрибутите са филтрирани спрямо оценката на χ^2 мярката.

Класификация: обективен / субективен текст			
	Мултиноминален Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9184	0.9124	0.7668
Precision (обективен текст)	0.9197	0.9101	0.7126
Recall (обективен текст)	0.9168	0.9152	0.8944
Precision (субективен текст)	0.9171	0.9147	0.8582
Recall (субективен текст)	0.9200	0.9096	0.6392

Табл. 4.10. Резултати на изследваните алгоритми за класификация в класовете обективен и субективен текст при добавяне на нови признаци от отрицателните конструкции в текста и оценяване на атрибутите чрез χ^2 мярката

Отчитането на отрицателни конструкции в изречението няма решаващо значение за проблема за неутралността на текст. Това показват и получените резултати от поведението на класификаторите в таблица 4.10. Използването на описания алгоритъм за добавяне на нови признаци дори слабо намалява точността на метода на опорните вектори и К най-близки съседа. Вероятната причина е, че добавянето на нови признаци по-скоро внася шум, отколкото релевантна информация за проблема.

Класификация: позитивен / негативен текст			
	Бернулиев Наивен Бейс	Линеен метод на опорните вектори	К най-близки съседа
Точност	0.9044	0.8420	0.7540
Precision (позитивен текст)	0.8934	0.8519	0.8191
Recall (позитивен текст)	0.9184	0.8280	0.6520
Precision (негативен текст)	0.9160	0.8327	0.7109
Recall (негативен текст)	0.8904	0.8560	0.8560

Табл. 4.11. Резултати на изследваните алгоритми за класификация в класовете позитивен и негативен текст при добавяне на нови признаци от отрицателните конструкции в текста и оценяване на атрибутите чрез χ^2 мярката

Подобно на резултатите за определяне неутралността на текст, определянето на полярността на текст също не се повлиява добре от реализирания алгоритъм. Резултатите на трите алгоритъма, дадени в таблица 4.11, са близки до вече получените при филтриране на релевантните атрибути и дори по-ниски в рамките на 1-2%. Макар очаквано класификаторите на полярността на текст да се повлияят положително от отчитане на отрицателните конструкции в езика, оказва се, че по-скоро предложения алгоритъм внася твърде много нерелевантни нови атрибути.

Може да бъде обобщено, че най-добри резултати и за двата класификационни проблема, които се разглеждат, дават алгоритмите от типа Наивен Бейсов класификатор и линеен метод на опорните вектори. Оптимални статистика показва Наивният Бейсов класификатор при включване на биграми в признаците и филтриране на релевантните атрибути на база оценката от χ^2 мярката. Това е класификаторът, заедно с описаните подходи за подобрене, който финално е използван в разработена софтуерна система за извличане на настроението в текстове намерени онлайн.

5. Софтуерна реализация

Реализирано е софтуерно приложение, което да осъществява анализ на настроението в текстови документи намерени онлайн. Отворената за потребители част на система представлява уеб сайт, който предоставя достъп до функционалността на системата.

5.1. Архитектура на приложението

Архитектурата на системата се състои от два независими модула. Първият от тях представлява библиотека, чиято основна функционалност е определяне на настроението в текст. Наречена е „*sentiment*”. Тази част на система предоставя единствено програмен интерфейс и е изградена независимо от останалите използвани модули. По тази причина може да бъде включвана и в други софтуерни приложения. Ограничението в използването ѝ обаче е, че е необходимо да се използва езика *Python*.

Разработената библиотека дава възможност за работа с двата описани вече в изложението корпуса документи. Единият от тях съдържа текстове класифицирани като обективни и субективни. Това множество служи за обучение и тестване на класификаторите на неутралността на текст. То съдържа по 5000 текста във всеки от двата класа. Вторият корпус документи съдържа текстове класифицирани като положителни и отрицателни. Във всеки от двата класа са налични по 6331 документа. Този корпус е съчетание от два други, описани вече в настоящото изложение. Той се използва като обучаващо и тестово множество за класификаторите на полярността на настроението в текст. Реализирана е логика за превръщането на всеки от текстовете във вектор. Всяка от неговите компоненти отговаря на отделна дума, а стойностите на наличието на думата в съответния пример. Тези вектори са необходими като вход за обучението на класификаторите.

Съставен е отделен подмодул на системата, който е отговорен за избора на атрибути, с които да бъдат описвани примерите. Преди текстовете да бъдат подадени на класификаторите се определя как да бъдат представени те във вектор от атрибути и кои техни думи или производни на думи да бъдат включени като атрибути в описанието им. Този подмодул има логика за определяне на най-информативните думи и биграми, за филтриране на стоп думи, за добавяне на специални форми на думи отговарящи на отрицателните конструкции в езика.

Модулът „*sentiment*” предоставя възможност да бъдат обучени и тествани описаните вече алгоритми с различни стойности на техните параметри и с различни подходи за избиране на атрибутите. Могат да бъдат избирани следните настройки:

- използвано обучаващо множество, като възможностите за избор са два;
- алгоритъм за реализация на класификатора;
- при алгоритъм K най-близки съседа, може да се избере параметър K;
- дали да се използва всяка дума от обучаващото множество като атрибут за описание на примерите;
- дали да бъдат филтрирани стоп думите;

- дали да се филтрират нерелевантните атрибути при описанието на примерите;
- при филтриране на нерелевантните атрибути може да се направи избор между две метрики за оценяване на атрибути – взаимна информация и χ^2 ;
- може да се избере какъв да е броят на най-информативните атрибути, които да участват в описанието на примерите;
- дали да бъдат използвани биграми като атрибути и колко от най-релевантните биграми да се включат като атрибути;
- дали да се използва описаната схема за обработка на отрицателните конструкции в езика и така да се включат допълнителни атрибути в описанието на примерите;

При тестването на алгоритмите на стандартния изход се изписват стойностите за точност на алгоритъма и *precision*, *recall* за всеки от двата класа, чиито имена също се изписват.

Имплементирана е функционалност за запазване в постоянната памет на компютъра на обучените класификатори. Те могат да бъдат възстановени без повторно обучение и използвани за класификация. Това е удобен подход, който дава възможност да не се прави обучение при всяко повторно пускане на системата.

Основната функция, която предоставя библиотеката „*sentiment*” приема като вход текст. Той не е необходимо предварително да бъде обработен или да отговаря на специални критерии освен да бъде на английски език. Функцията връща класа, в който е определен текстът. Намира се първо класа според класификатора за неутралност на текст. Ако изходът от класификацията е субективен текст, то той се подава на класификатора на полярността на настроението. Изходът на самата функция може да бъде „*obj*” за неутрален текст (обективен), „*pos*” за позитивен и „*neg*” за негативен.

Вторият основен модул на системата е наречен „*sentisite*”. Той представлява сайт, който дава потребителския интерфейс на разработеното решение.

В „*sentisite*” са реализирани важни за работата на системата задачи. Едната от тях е връзката с онлайн платформа за търсене и извличането от нея на резултатите от търсенето. Използвана е платформата *Bing*, която предоставя свободен достъп до ресурсите си при ограничение от 5 000 заявки на месец. Съставен е модул за работа с *Bing* и обработване на резултатите, които биват връщани при търсене.

Като отделен модул е обособена и функционалността за обработка на *html* файлове и извличане на текст от тях. Тъй като една от задачите на системата е извеждане на всички текстове, съдържащи ключовите думи на търсене, то е необходимо *html* файловете да бъдат обработени. Премахват се т.нар. тагове от тях, които са всъщност системната информация, която потребителите не виждат. След това се определят частите от текста, в които се съдържат ключовите думи на търсенето. Към тях е приложен е прост алгоритъм за извличане на смислени синтактични части, пасажии завършени изречения, които ще бъдат показани на потребителя.

В „*sentisite*” е реализирана и логика за обхождане на всички страници на един сайт чрез търсенето на връзки към тях в *html* файловете. Потребителят въвежда адреса на сайта, който иска да обходи, и от тази първоначална страница системата започва да търси всички достижими от нея страници и ги посещава. Това се изисква от задачата системата да предоставя търсене както в отделна страница, така и в сайт.

Съставен е и подмодул за създаване на файлове във формат *pdf*, в които да бъдат запазени резултатите от реализирано от потребителя търсене.

Двата основни модула на разработената система са реализирани на езика *Python* (версия 2.7). Той е базиран език, което го прави подходящ за математически изчисления. На този език съществуват свободни за използване библиотеки за работа с матрици, за често срещани изчислителни задачи, за обработка на естествен език, за проблеми от машинното самообучение. Разработената система се възползва основно от две външни библиотеки – *nltk* (<http://www.nltk.org/>) и *scikit-learn* (<http://scikit-learn.org/stable/index.html>). Част от функционалността за работа с корпусите от текстове е реализирана благодарение на библиотеката *nltk*. От библиотеката *scikit* е използвана имплементацията на алгоритмите за класификация.

Уеб сайтът е реализиран на най-използваната уеб платформа за езика *Python* – *Django* (версия 6). Използваната база от данни е *SQLite*.

5.2. Функционалност

Функционалността, която разработената система предоставя на потребител, е:

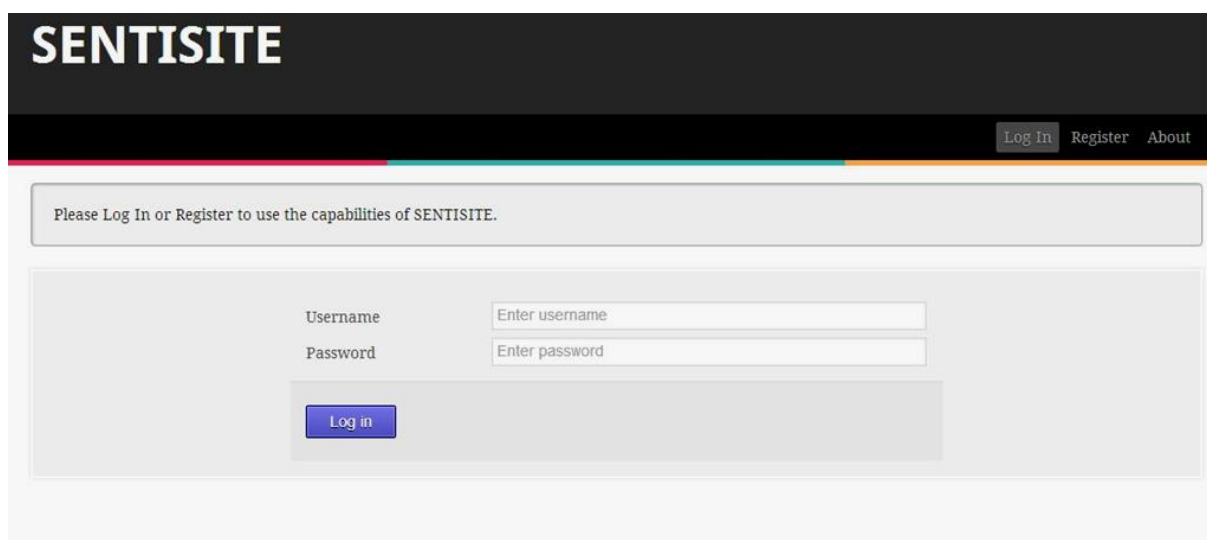
- Регистрация на потребител;
- Вписване на регистриран потребител;
- Търсене по ключови думи в онлайн платформата за търсене *Bing* и извеждане на отчет с намерените резултати подредени в категории позитивен, неутрален и негативен текст;
- Предоставяне на възможност за определяне на броя на връщаните резултати при търсене чрез *Bing*;
- Включване на новини при търсене в онлайн платформата за търсене *Bing*;
- Търсене по ключови думи в определен сайт или уеб страница и извеждане на отчет с намерените резултати (всички намерени откъси текст съдържащи ключовите думи) подредени в категории позитивен, неутрален и негативен текст;
- Предоставяне на възможност за запазване на отчета на всяко търсене във файл с формат *pdf* и преглеждането на файла при необходимост;
- Преглеждане на история с всички направени от регистрирания потребител търсения;
- Преглеждане на подробна информация за всяко направено от потребител търсене, включително и настройките, които потребителят е избрал при започването му;
- Извеждане на списък с търсени повече от веднъж ключови думи;
- За търсения на едни и същи ключови думи повече от веднъж изобразяване на графика с изменението във времето на процентите за позитивни, неутрални и негативни текстове, възможност за избор на времеви интервал за проследяване на графиката;

Функционалността на разработената система предоставя на потребителя ѝ възможност за проследяване на настроението в интернет пространството спрямо неговите интереси.

5.3. Потребителски интерфейс

Разработен е уеб сайт, който да предоставя описаната функционалност на потребителите на системата.

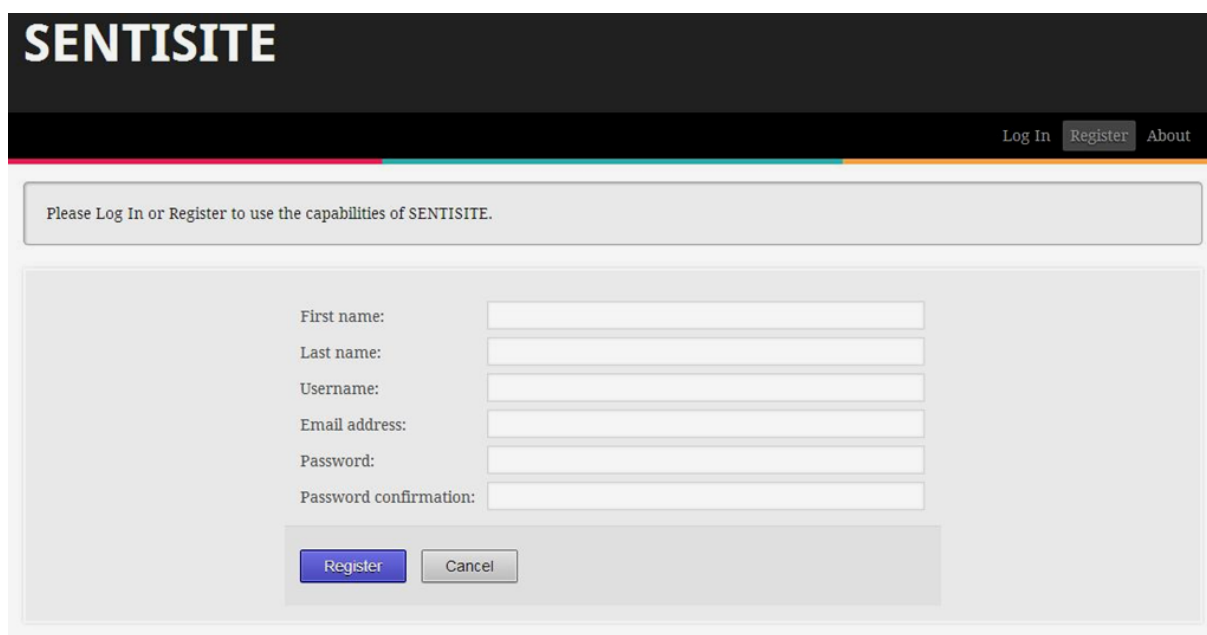
Нерегистрираните потребители на сайта могат да извършват следните действия – да се регистрират или да получат информация за платформата и нейните възможности чрез “About” страницата. При зареждане на сайта се показва форма за въвеждане на потребителско име и парола, представена на фигура 5.1.



The screenshot shows the login page of the SENTISITE application. At the top, there is a dark header with the 'SENTISITE' logo on the left and three navigation links ('Log In', 'Register', 'About') on the right. Below the header, a light gray box contains the text: 'Please Log In or Register to use the capabilities of SENTISITE.' The main content area features a login form with two input fields: 'Username' (placeholder: 'Enter username') and 'Password' (placeholder: 'Enter password'). Below these fields is a blue 'Log in' button.

Фиг.5.1. Начална страница на уеб сайта

Ако посетителят на сайта не е регистриран, то той може да направи това чрез избиране на “Register” от лентата с опции в горния край на прозореца. Бутонът отваря страницата за регистрация дадена на фигура 5.2.



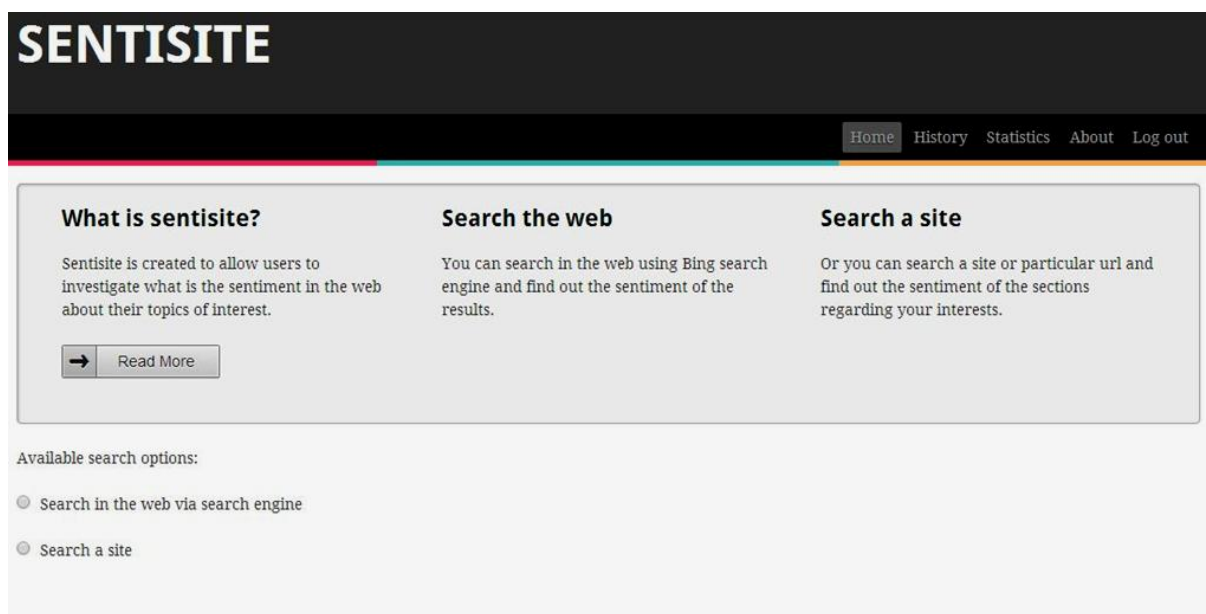
The screenshot shows the registration page of the SENTISITE application. It has the same header as the login page. Below the header, a light gray box contains the text: 'Please Log In or Register to use the capabilities of SENTISITE.' The main content area features a registration form with six input fields: 'First name:', 'Last name:', 'Username:', 'Email address:', 'Password:', and 'Password confirmation:'. At the bottom of the form are two buttons: a blue 'Register' button and a gray 'Cancel' button.

Фиг.5.2. Страница за регистрация на потребител

При регистрация е необходимо потребителят да въведе потребителско име, двете си имена, адрес на електронна поща и парола. При некоректно въвеждане на някое от полетата, потребителят е пренасочван към същата страница и бива посочено полето, което съдържа погрешно въведените данни. За да бъде завършена регистрацията на потребител на системата, трябва той да е въвел в регистрационната форма верен адрес на електронна поща и да е изписал правилно паролата си и в двете полета, където тя се изисква. При такива условия регистрацията успешно се реализира при натискане на бутона *“Register”*. Тя може да бъде прекратена от бутона *“Cancel”*.

Отново от меню лентата в горния десен края на прозореца може да бъде посетена страницата *“About”*, която съдържа подробна информация за разработената система, нейната функционалност и как тя може да бъде използвана.

При вписване на потребител, се показва основната страница на сайта. Тя е представена на фигура 5.3.



Фиг. 5.3. Основна страница на уеб сайта

Дадени са кратки обяснения за опциите на търсене, които са предоставени на потребителя. Едната от тях е той да осъществи стандартното търсене с онлайн платформа за търсене (*Bing*). При избиране на опцията *“Search the web via search engine”* се показва форма, в която се въвеждат ключовите думи на търсене, може да се избере броят на върнати резултати, дали да се генерира и запази pdf файл с резултатите, дали в търсенето да бъдат включени и новини. Фигура 5.4 показва потребителския интерфейс на формата за търсене в *Bing* и на полетата за избор на съответните параметри на търсене.

Available search options:

☒ Search in the web via search engine

☐ Search a site

Keywords:

Top results:

☐ Include News

☐ Generate PDF

Фиг.5.4. Форма за търсене чрез онлайн платформата Bing

При избиране на опцията *“Search a site”* се визуализира формата за търсене в конкретен сайт или уеб страница. Тя е представена на фигура 5.5. В нея трябва да се въведат ключовите думи за търсене и адрес. Дава се опцията или да бъде посетена и претърсена само страницата отговаряща на адреса, или да бъде осъществено търсене в целия сайт. Това означава, че ще бъдат претърсени и всички страници от същия домейн, които са достижими от посочения адрес. Техните адреси се намират от *html* кода на първоначалната страница. Предоставена е и опция за избор дали да се генерира и запази *pdf* файл с резултатите от осъщественото търсене.

☒ Search a site

Keywords:

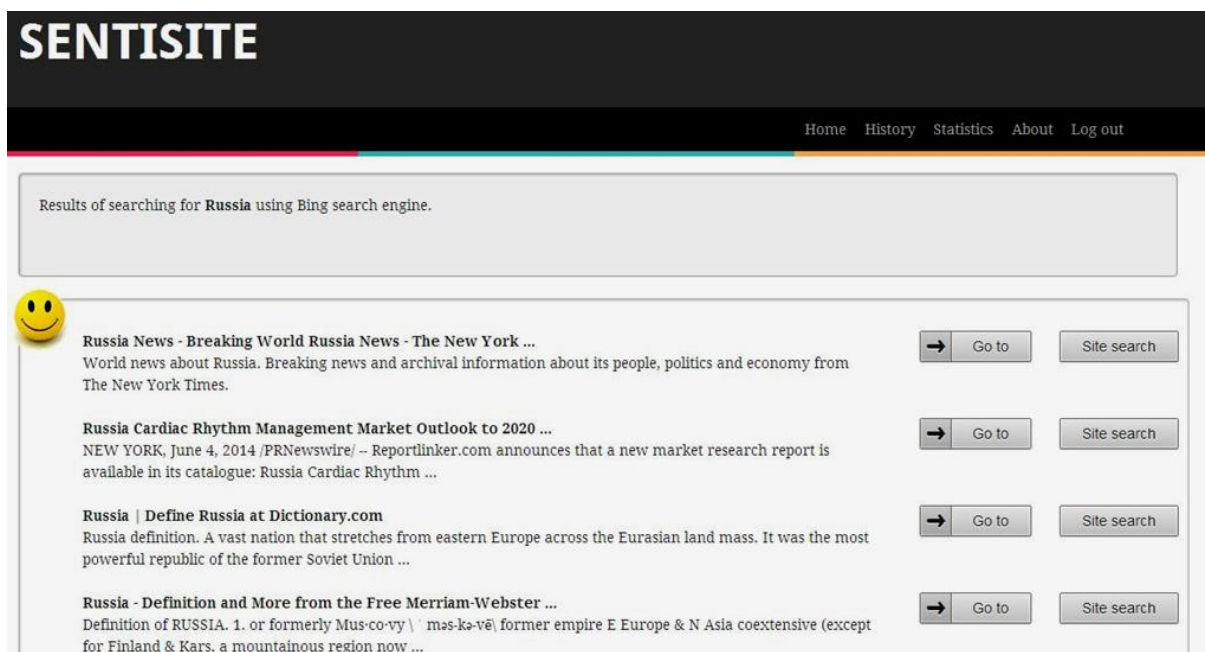
Site url:

☐ Search only in this url content

☐ Generate PDF

Фиг.5.5. Форма за търсене в сайт

След като системата е осъществила търсенето с настройките избрани от потребителя, тя форматира получените резултати и се извлича настроението в избраните откъси от онлайн текстове. Изготвя се отчет и потребителят е пренасочен към нова страница, в която са изведени резултатите. Първата секция в страницата показва намерените положителни текстове. До всеки от тях има бутон *„Go to”*, който отваря оригиналната страница, където е намерен текста.



Фиг.5.6. Секция с положителните резултати от търсене

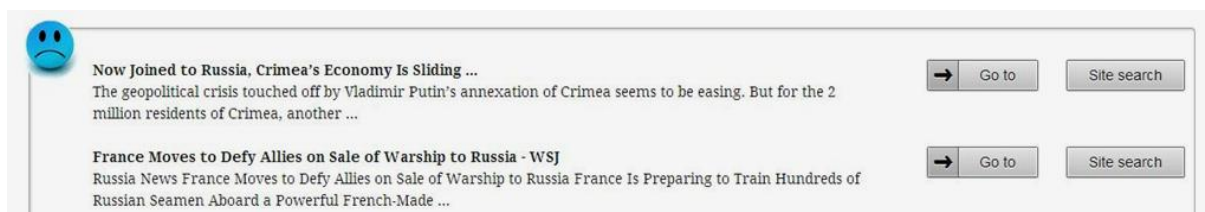
Фигура 5.6 показва резултатите при търсене чрез *Bing* на ключовата дума “Russia”. Когато се осъществява търсене в сайт, отчетът изглежда по напълно сходен начин на изображения. Липсва единствено бутона „Site search”. Той дава възможност да се осъществи допълнително търсене по ключовите думи, но този път в конкретния сайт. Цели се схема на използване на сайта подобна на тази, която потребителят би следвал при търсене в други онлайн платформи. Но “sentisite” предоставя в допълнение и знание за настроението в намерените онлайн текстове.

След позитивните резултати в отделен сегмент се дават и неутралните резултати от осъщественото търсене. Те имат същата структура, както позитивните. Тяхна примерна извадка за същото търсене е да дадена на фигура 5.7.



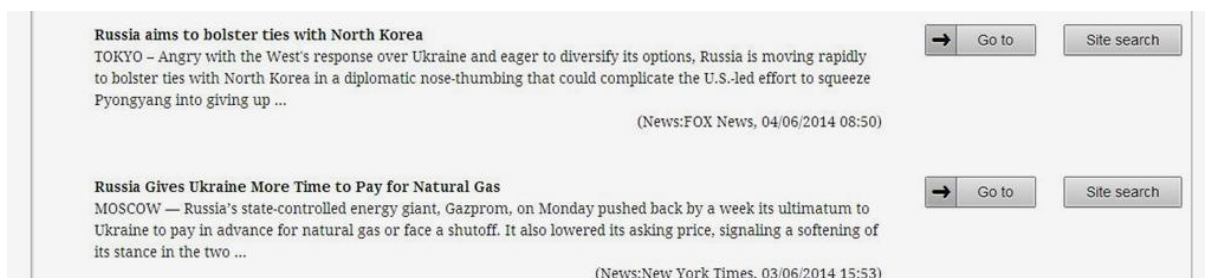
Фиг.5.7. Секция с неутралните резултати от търсене

На последно място са изведени негативните резултати от търсенето, представени на фигура 5.8.



Фиг.5.8. Секция с негативните резултати от търсене

При попълване на формата за търсене чрез *Bing* потребителят може да избере дали да бъдат включени и новини. Ако той е избрал тази опция, на края на всеки от трите секции на отчета в зависимост от настроението в тях, са изброени и намерените новини.



Фиг.5.9. Резултатите от тип новини при търсене чрез *Bing*

Фигура 5.9. представя сегмент с новини открити при търсене с ключовата дума *"Russia"*. Под всяка новина е посочена и допълнителна информация в скоби – източника ѝ и датата на нейното публикуване.

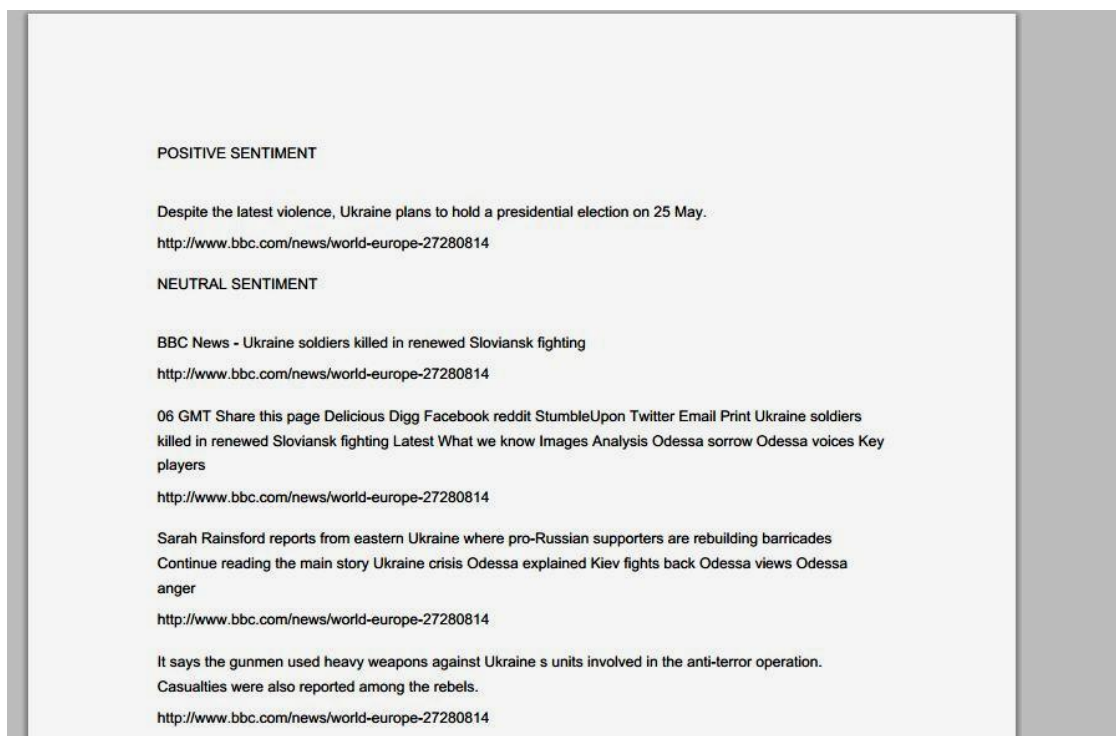
От менюто в горния десен ъгъл на прозореца може да бъде избрана опцията *"History"*. Тя пренасочва към страница, която дава информация за всички осъществени търсения на регистрирания потребител в *"sentisite"*. На фигура 5.10 е представен потребителският интерфейс на страницата с история на търсенията.

SENTISITE							
				Home	History	Statistics	About Log out
Search history							
Ukraine	May 1, 2014, 10:03 p.m.	site search	5% +	47% 0	21% -	PDF	More
Ukraine	May 5, 2014, 10:50 p.m.	site search	1% +	49% 0	15% -	PDF	More
Ukraine	May 6, 2014, 4:11 p.m.	site search	5% +	47% 0	21% -	PDF	More
Ukraine Russia	May 12, 2014, 1:03 a.m.	web search	4% +	80% 0	16% -		More
Russia	May 12, 2014, 11:58 p.m.	web search	31% +	46% 0	22% -	PDF	More
Russia	May 15, 2014, 12:32 a.m.	site search	7% +	46% 0	14% -		More

Фиг.5.10. Страница с информация за всички осъществени от потребител търсения

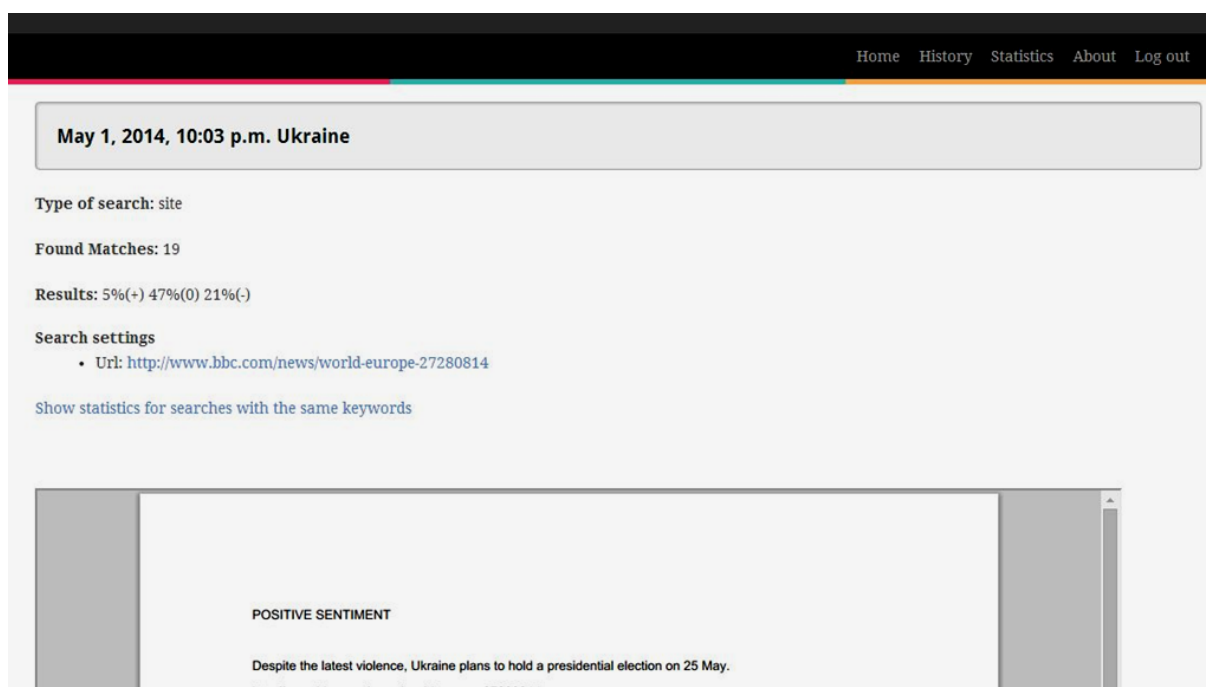
За всяко едно от търсенията са изброени ключовите му думи, кога е било направено, дали е било търсене в онлайн платформа или в сайт, процентни резултати за настроението в намерените текстове, бутон *"PDF"* и бутон *"More"*. Ако при търсенето

си потребителят е избрал да бъде генериран *pdf* файл с резултатите, то тогава той се запазва в системата и може да бъде достъпен чрез натискане на бутона “PDF”. Самият *pdf* файл съдържа единствено текстова информация с отчета за търсенето. В него са изредени три секции – позитивни, неутрални и негативни текстове. За всеки текст е записан и адресът на сайта, в който е намерен. Ако потребителят не избере генериране на отчет, то такъв няма да бъде запазен. Посочената фигура 5.11 е *pdf* отчет изготвен при търсене с ключовата дума “Ukraine”.



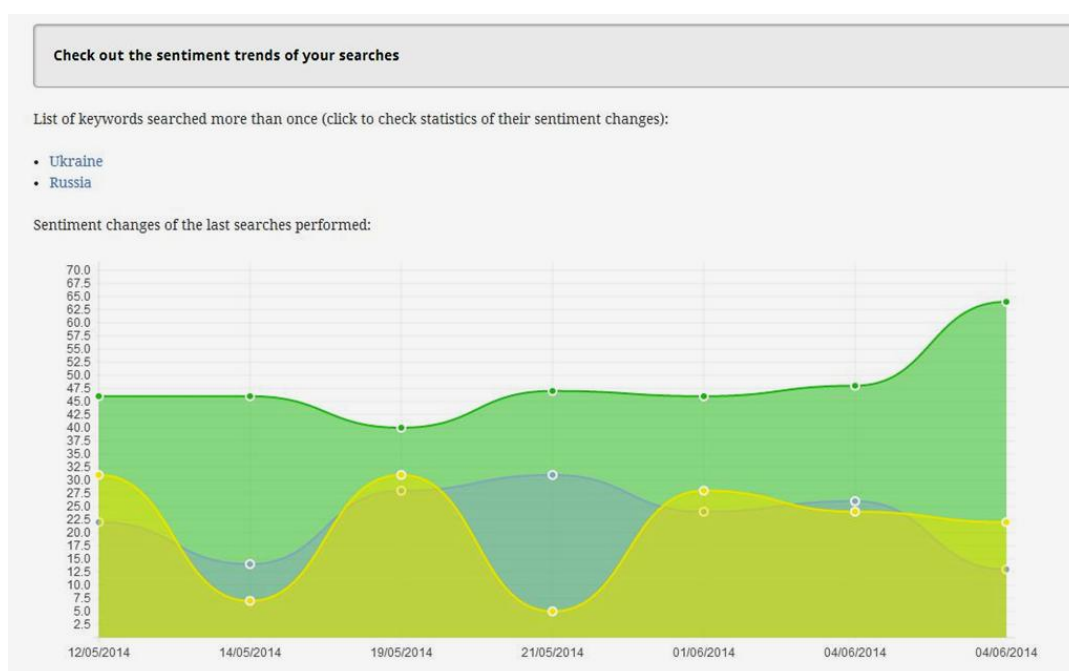
Фиг.5.11. Отчет на търсене в *pdf* формат

От страницата с историята на търсенията може да бъде избран и бутон “More”. Той отваря страница с детайлна информация за конкретно търсене. Фигура 5.12 показва страница с информация за търсене на думата “Ukraine” в сайт. Ако търсенето е в платформата Bing, на страницата за детайлна справка се посочва дали резултатите включват новини и какво е зададеното от потребителя ограничение за техния брой. Ако търсенето е сайт, се дава неговият адрес. И в двата случая се посочват резултатите от анализа на настроението в проценти. Изобразява се и прозорец с *pdf* отчета, ако е генериран такъв.



Фиг.5.12. Страница с подробна информация за конкретно търсене

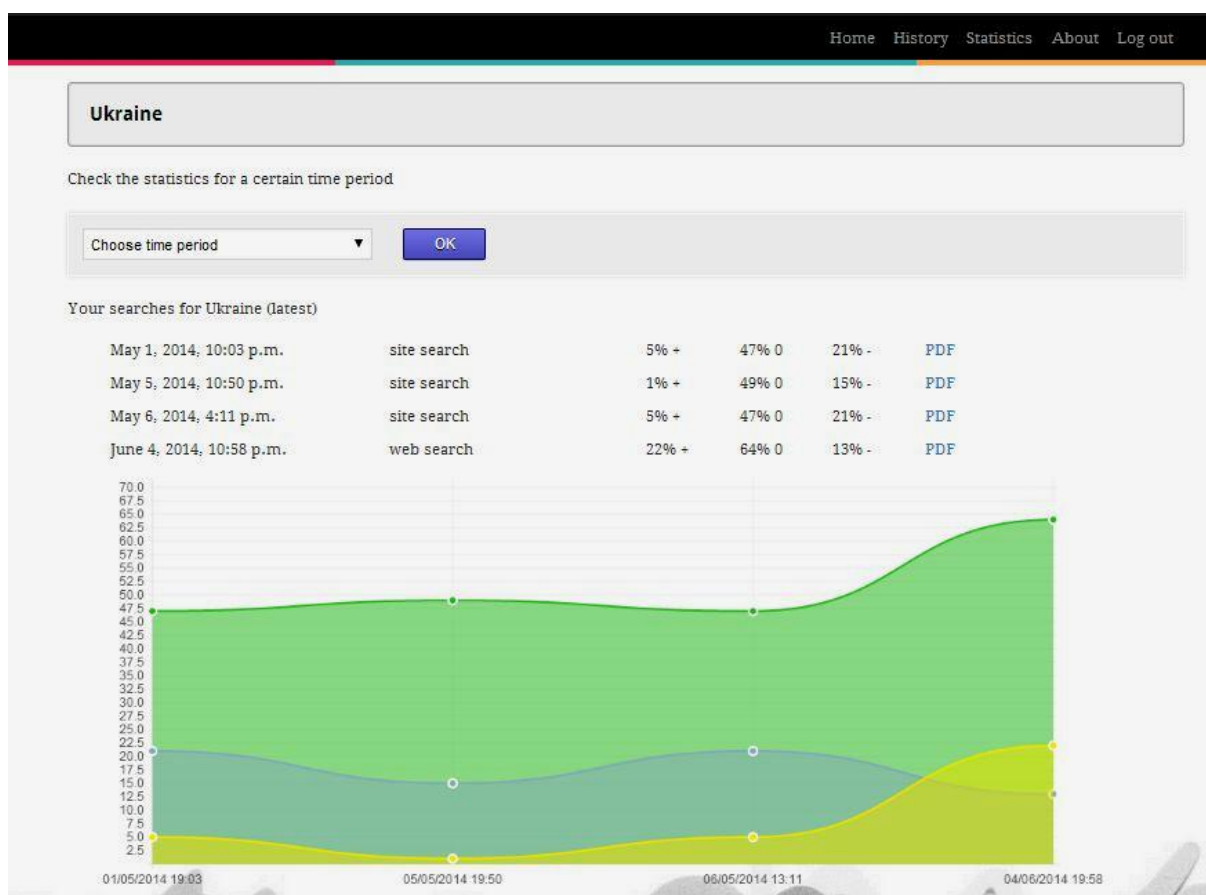
От менюто в горната лента може да бъде избрана и опцията “Statistics”. Тя пренасочва потребителя към нова страница. Примерна страница със статистики за търсенията на потребител на системата е дадена на фигура 5.13.



Фиг.5.13. Страница със статистики за търсенията на потребител

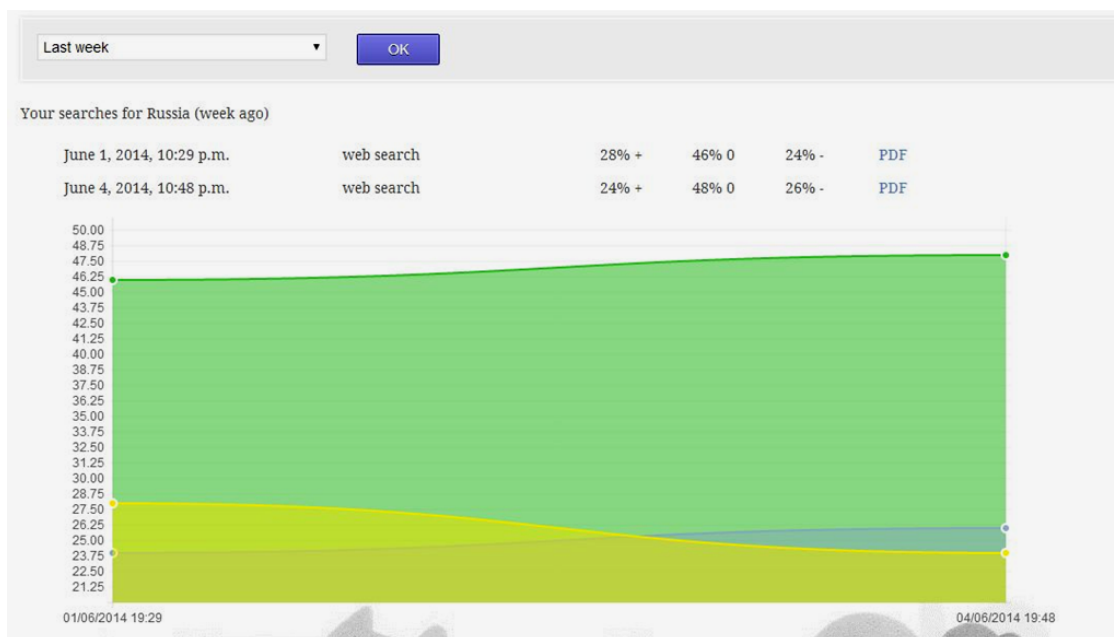
В края на страницата е изобразена диаграма с изменението на настроението в последните търсения на потребителя. Зелената графика показва изменението на неутралното настроение, жълтата на позитивното, синята на неутралното. По абсцисата са отбелязани датите на последните търсения, а по ордината процента на намерените

резултати в трите класа за настроение. В горния край на страницата са изброени ключовите думи, които потребителят е търсел повече от веднъж. Така той може да проследи времевите изменения на емоционалните нагласи в интернет пространството спрямо конкретна тема. При избиране на някои от тези ключови думи, той е пренасочен към нова страница. Това е страница за конкретната тенденция, т.е. за изменението на настроението спрямо избраните ключови думи. При пренасочване към нея се изобразяват последните търсения на потребителя за тези ключови думи. За всяко търсене е посочена дата, тип на търсенето, процентни резултати за настроението и връзка към *pdf* файла. Под списъка с търсенията има графика с изменението на настроението. До тази страница може да бъде достигнато и от страницата за подробна информация за конкретно търсене. На фигура 5.14 е представена една такава страница със статистики на търсения с ключовата дума “Ukraine”.



Фиг.5.14. Страница със статистики за конкретни ключови думи на търсене

За да се предостави по-голям контрол на потребителя в горната част на страницата се намира форма, в която той може да посочи времеви интервал, в който да бъдат показани резултатите за осъществени търсения – последния ден, седмица, месец или година. При избиране на период и натискане на бутона “OK”, страницата се презарежда и се показват резултатите само от избрания времеви интервал.



Фиг.5.15. Страница със статистики на търсения за последната седмица

Фигура 5.15 показва изглед към статистиките за търсения с ключовата "Ukraine" при избран времеви период – последната седмица.

В менюто в горния десен край има опция "About", която отваря страница с кратки разяснения за същността на "sentisite" и как може да бъде използван. До тази опция има и "Log out", която прекратява текущата сесия на потребител и го пренасочва към страницата за вход.

Разработеният уеб сайт предоставя удобен интерфейс на потребител на системата за осъществяване на различни видове търсения в уеб пространството, извличане на знание за настроението на получените резултати и проследяване на времеви изменения на настроението спрямо конкретна тема.

6. Заключение

Анализът на настроението в текст е област, към която се проявява все по широк интерес. С нарастване на значението на социалните мрежи и другите средства за споделяне на мнение в интернет пространството се натрупват значителни количества текстови данни носещи емоционална оценка. Анализът на подобна информация означава проследяване на нагласата към определен обект на интерес и това може да бъде полезно за маркетинга на продукти и услуги, бизнес решения, социално-политически анализи. Заради възможностите, които анализът на настроението предоставя в различни сфери на живота, се правят много научни изследвания в областта и също така се създават много комерсиални платформи за анализ на мнението.

Анализът на настроението в текст е предизвикателна задача от научна гледна точка. Трудностите произтичат от необходимостта да се работи с неструктурирани данни при обработката на естествен език. Формалното представяне на текст не съумява да го предаде в неговата пълнота. Пропускат се важни семантични аспекти като контекста, наличието на ирония, специфичните характеристики на писане на всеки създател на текст.

Текущата работа представя изследване на методи от машинното самообучение за осъществяване на анализа на настроението в текст. Решават се два класификационни проблема – определяне на неутралност на текст и определяне на полярността на настроението в текст. Проучени са три от известните алгоритми в областта класификация на текст – Наивен Бейсов класификатор, метод на опорните вектори, K най-близки съседа. Изследвани са методи за подобряване на класификационното поведение на използваните алгоритми. Постигнатите резултати върху тестовите множества се оказват добри, макар при използване на класификаторите в реална среда да се очаква точността на бъде по-ниска.

Бъдещо развитие на разработеното решение би могло да бъде насочено към подобряване на избора от признаци, с които формално се представя текст. Освен наличието или отсъствието на определени думи би могло да се отчита и някаква семантична, граматическа или синтактична информация, която думите от текста носят. Представянето на отделни думи като атрибути, какъвто е възприетият подход в настоящата работа, води до загуба на голяма част от семантична информация в текст. По-добри резултати също така могат да се очакват при използване на повече обучаващи текстове с теми от различни сфери. Това би подобрило представянето на класификаторите при реалното използване на система, в което се работи с текстове с много различни теми.

Разработеният уеб сайт предоставя възможност за използване на интернет пространството поставяйки фокус върху анализа на настроението в текстовете, намерени онлайн. Предоставя се удобен интерфейс за търсене по ключови думи на интересувашото потребителя и същевременно обобщаване на настроението на намерените текстове. Изграденото приложение дава възможност за придобиване на поглед върху мнението и настроението в интернет пространството по определена тема и проследяване на времевите тенденции за тези емоционални оценки.

7. Литература

- [1] Liu B., *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, May 2012.
- [2] Grimes S., *Expert Analysis: Is Sentiment Analysis an 80% Solution?*, InformationWeek, 2010
<http://www.informationweek.com/software/information-management/expert-analysis-is-sentiment-analysis-an-80--solution/d/d-id/1087919?> (посетен на 20.03.2014г., 12:30)
- [3] Goncalves P., Araujo M., Benevenuto F., Cha M., *Comparing and Combining Sentiment Analysis Methods*
<http://homepages.dcc.ufmg.br/~fabricio/download/cosn127-goncalves.pdf> (посетен на 20.05.2014г., 15:40)
- [4] Dodds P. S. and Danforth C. M., *Measuring the happiness of large-scale written expression: songs, blogs, and presidents*, 2009
- [5] Goncalves P., Benevenuto F., Cha M., *PANAS-t: A Psychometric Scale for Measuring Sentiments on Twitter*, 2013
- [6] Wilson T., Wiebe J., Hoffman P., *Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*, 2005
<http://people.cs.pitt.edu/~wiebe/pubs/papers/emnlp05polarity.pdf> (посетен на 21.05.2014г., 18:00)
- [7] Pang B., Lee L., Vaithyanathan S., *Thumbs up? Sentiment classification using machine learning techniques*, 2002
<http://www.cs.cornell.edu/home/llee/papers/sentiment.pdf> (посетен на 21.05.2014г., 19:10)
- [8] Bird S., Loper E. and Klein E. (2009), *Natural Language Processing with Python*, O'Reilly Media Inc.
- [9] Pang B., Lee L., *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*, 2004
<http://www.cs.cornell.edu/home/llee/papers/cutsent.pdf> (посетен на 24.05.2014г., 11:15)
- [10] Labatut V., Cherifi H., *Accuracy Measures for the Comparison of Classifiers*, 2012
http://hal.archives-ouvertes.fr/docs/00/61/13/19/PDF/17_ICIT11_VL.pdf (посетен на 25.05.2014г., 13:05)
- [11] Pang B., Lee L., *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*, 2005
<http://www.cs.cornell.edu/home/llee/papers/pang-lee-stars.pdf> (посетен на 01.06.2014г., 17:15)

[12] Manning C., Raghavan P. , Schütze H., *Introduction to Information Retrieval*, Cambridge University Press. 2008

[13] Joachims T., *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, 1997

http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf (посетен на 01.06.2014г., 20:00)

[14] Yang Y. and Liu X., *A Re-examination of Text Categorization Methods*, 1999

<http://www2.hawaii.edu/~chin/702/sigir99.pdf> (посетен на 03.06.2014г., 14:10)

[15] McCallum A., Nigam K., *A comparison of event models for Naive Bayes text classification*. 1998

[16] Cortes C., Vapnik V., *Support-vector networks*, 1995

[17] Forman G., *An Extensive Empirical Study of Feature Selection Metrics for Text Classification*, 2003

[18] Domingos P., *A Few Useful Things to Know about Machine Learning*, 2011

Използвани корпуси документи за обучаващи и тестови множества:

http://www.cs.cornell.edu/people/pabo/movie-review-data/rotten_imdb.tar.gz (посетен на 24.03.2014г., 17:00)

http://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz (посетен на 24.03.2014г., 17:30)

<http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.tar.gz> (посетен на 01.04.2014г., 10:40)